博士論文

Development of optimal decision-making system for road-reconstruction considering human mobility by applying reinforcement learning

(深層強化学習に基く人の移動性を考慮した道路ネットワーク復旧戦略の最適化)

朱　秀賢

Development of optimal decision-making system for road-reconstruction considering human mobility by applying reinforcement learning

Soohyun JOO

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Engineering

University of Tokyo

2021

Reading Committee:

Yoshihide Sekimoto, Chair

Koji Ikeuchi

Riki Honda

Makoto Chikaraishi

Yudai Honma

Takashi Fuse

Program Authorized to Offer Degree:

Civil Engineering

University of Tokyo

Abstract

Development of optimal decision-making system for road-reconstruction considering human mobility by applying reinforcement learning

Soohyun JOO

Yoshihide Sekimoto:
Professor, Dr. Engineering
Center for Spatial Information Science, University of Tokyo

Western Japan experienced record-breaking heavy rain from June 28 to July 8, 2018. Approximately 600 road sections were closed due to flooding in Hiroshima and Okayama Prefecture. The government develops road-recovery plans to return human mobility to a specific level as soon as possible, however, restoring neighborhood roads was delayed for a week after this flooding. This means that their own plan might not be effective to achieve their own objective. There three limitations government has: 1) lack of prior knowledge, 2) absence of evaluation indicators and 3) the difficulty of estimating human mobility.

With the demand for increased efficiency, we suggested the road reconstruction plan for rapid human mobility recovery with Deep RL. In addition, we utilized origin-destination pairs from mobile phone GPS data, and digital road map to estimate and evaluate human movement under recovery operation at each time step.

The agent in our model is one operation crew. Input layers and reward consist of the information related to each damage road's recovery, inter-road connectivity with the results of traffic allocation, the travel time. With single agent RL and multi agent RL, the agents could establish the optimal policy for at least 15 roads and up to 45 roads. Multi agent RL might consider a recovery plan for almost damage roads in Hiroshima Prefecture. The agent in our model could identify the recovery effect and the importance of each disrupted roads. It selected disrupted roads with high effect of human mobility recover preferentially after learning progress. Moreover, the operation crews in multi-agent systems could learn the concept of cooperation through information about road usage in O-D. In this study, approximately 1000 kinds of O-Ds' route choice models could identify the change of traffic volume with the sequence reconstruction operation process, and the visualization data would allow the government officials to response further to abnormal traffic phenomena.

The final human mobility recovery rate with their optimal policy is 25% better on average than the lowest recovery rate when working randomly. Furthermore, the system in this paper could solve the optimization problem for the number of cases in $6.13 * 10^{34}$ in less than three hours. With the comparison of previous studies, this model could examine the number of cases greater than 10^{7}times for the computation time similar to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The author very much appreciates the support by Professor Yoshihide Sekimoto, my academic advisor. Thank you for giving me the opportunity to be a part on sekimoto lab and do research activities. I could complete this paper for your general support. I would like to thank Dr. Ikeuchi, Dr. Honda, Dr. Chikaraishi, Dr. Honma and Dr. Fuse. They thankfully became the committee member and gave me many constructive feedbacks and advises in three presentations and personal interview. I did additional analysis and verification and their advises help the meaning of my thesis be enriched. I sincerely acknowledge the support of Assistant Professor Ogawa. He helped me a lot and a little during my Ph.D. program. With his support, I learned a lot about how to write a paper, the framework of a model, and the visualization method. He always helped make my presentation, manuscript easier for others to understand. Furthermore, I have received much support from Assistant Professor Kashiyama and Dr. Seto. Even if I asked a sudden question, they always answered kindly. Especially, Dr. Kashiyama help me improve the calculation speed of this model. It must have been possible to make more practical decision-making system with his technical support. Before entering the Ph.D., I had no knowledge of programming and reinforcement learning. Sheofeng Yang help me write this paper by checking errors in my programs and sharing the knowledge and paper on reinforcement learning. So, I wish to express thanks for his support.

Chapter 1. INTRODUCTION

1.1 BACKGROUND AND PURPOSE

For the past few decades, there has been an increase in the number of days with heavy rain (100mm / day). The number of occurrence of events with extreme precipitation (50mm / hour) has also been increased [1]. As a result, there have been occurred large-scale damage by serious slope failure and downpours. Ministry of Land,

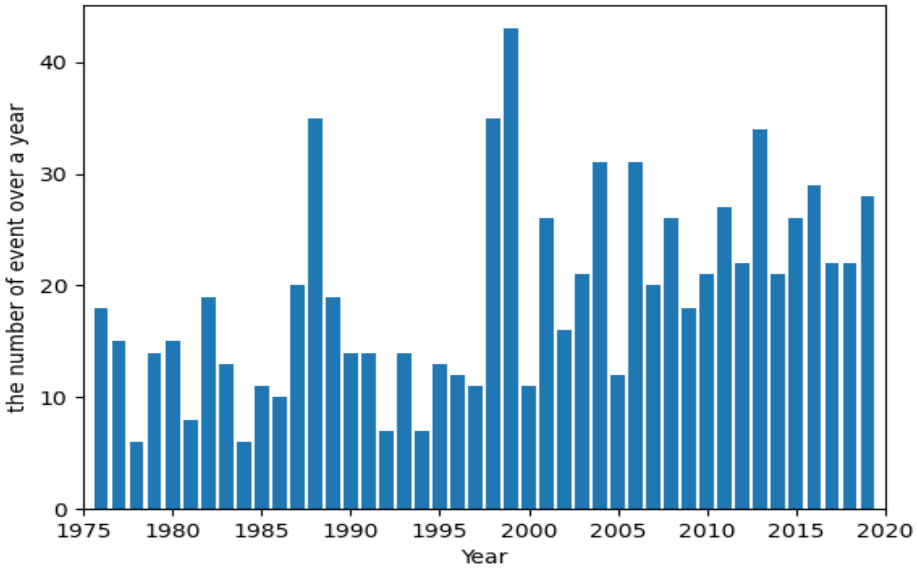


Figure 1. Annual total number of appearances with precipitation of 80mm [2]

Infrastructure, Transport and Tourism (MLIT) has reported the annual economic losses and damage occurred by natural disaster since 1961. We could identify that the trend in damage amount caused by flood has been on a constant rise every year. Especially, the cost of flood damage in 2019 totaled 2.15 trillion yen, the biggest of all time.

Approximate 520 billion yen (24.4%) is related to the damage of public civil infrastructure, such as river and roads [3].

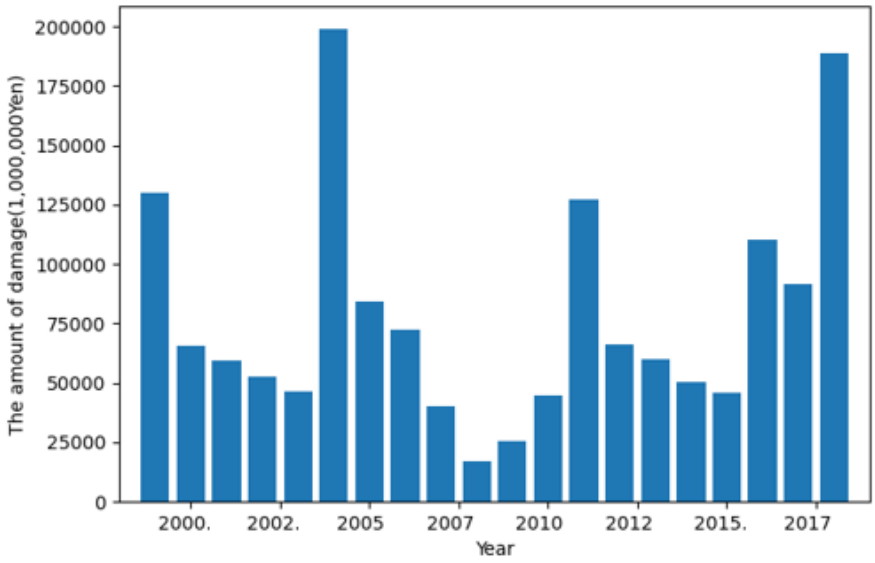


Figure 2. The amount of damage to road facilities caused by flooding [4]

With the increases of low-probability, high-impact multilocation hazards [5] like Western Japan flooding or Great East Japan Earthquake, it is growing more importance to set effective management plans. Disaster management plans would be designed before disaster to carry out post-disaster reconstruction operation rapidly [6]. There are numerous effect factors with the effectiveness of post-disaster reconstruction: 1) the available of resource [7], 2) economic and political actors [8], and 3) the influence and coordination of funding agencies [9]. In 1961, Japan government enacted the Disaster Countermeasure Basic Act, which define the institutional responsibility for disaster prevention and management. They often do the amendment based on limitations and deficiencies which they have learned since mid and large-scale disaster. Figure 3 describes the detail of each administration's responsibility. The municipal government

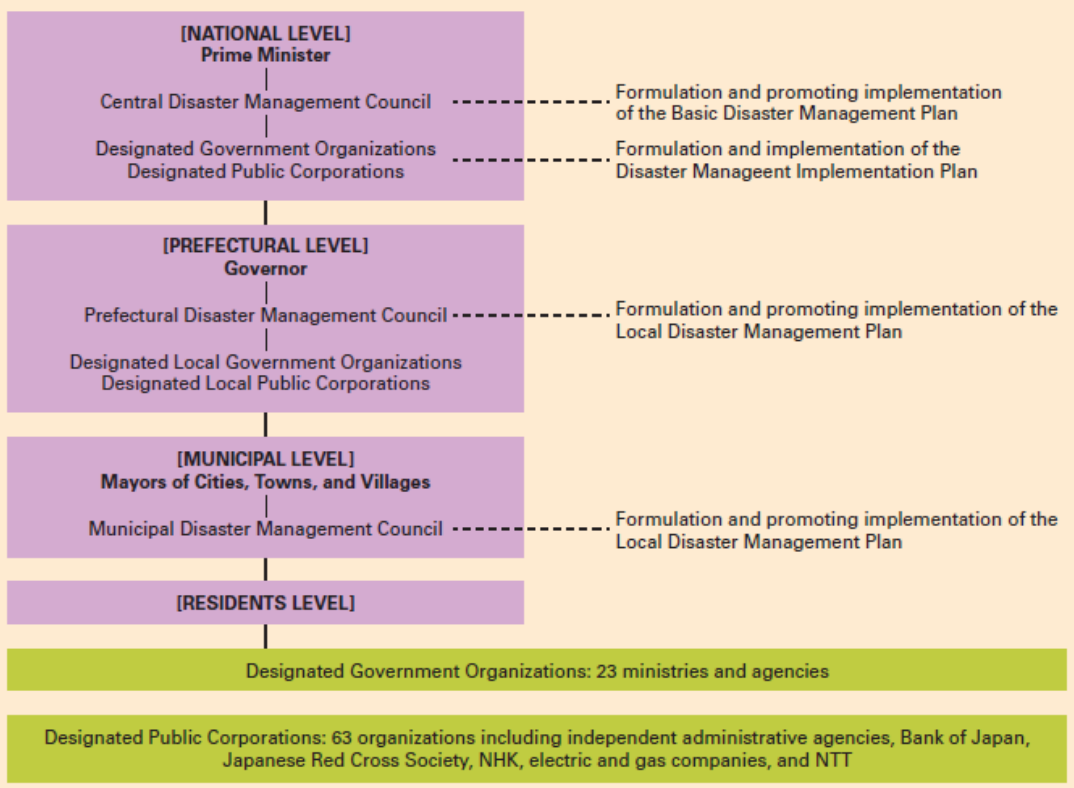


Figure 3. The detail of administrations' responsibility [5]

is the key player for local disaster management. If they do not fundamental function properly with the occurrence of disaster damage beyond their capability, the prefectural government takes the municipality's duty. Accordingly, we could identify that the majority effect factor in Japan's disaster management is political actors. Government's countermeasure is very important because it determines reconstruction's success and failure.

The damage of social infrastructure includes housing, water, electricity, gas, and transport network. The reconstruction of road infrastructure would not be provided attention and fund with the comparison of the necessities of life's damage [10]. Chang et al [11] mentioned that administrations have consider the rehabilitation of road network to be less than significant in recovery project performance. However, the road-network is one important part of the urban infrastructure system under both pre-disaster

and post-disaster situation. There are two reasons. Firstly, fundamental function is to secure basic mobility services and manage emergence situation. Secondly, it is a critical factor influencing the spatial variation of many other social and economic activities [12].

Road-network tend to be disrupted by several external shock. External shocks are as follows: 1) Outside force in daily life such as human errors, rush hour or technological breakdowns and 2) unexpected situation such as adverse weather change or natural disaster. It is designed to withstand a certain amount of external disruption. However, transportation network would often fail to withstand the impacts of natural disaster. It eventually ends up losing prevention and response capabilities. As road networks fail to do fundamental function, there are challenges not only in rescue and emergency activities, but also in activities for the restoration of other infrastructure.

Local government has made post-disaster management plan related to road network based on the past disaster situation. This plan includes the information of alternative routes, emergency routes and the priority order of reconstruction. However, it is hard to utilize management plan government made manually while implement [13]. This is because the afflicted area might be similar to the past disrupted area, but the damage patterns vary depending on the present disaster. In other words, management plans based on the past cases might not be appropriate for the current situation. Furthermore, there are some limitations which government need to overcome for effective measurements: the uncertainty of estimated damage, the missing road information, confusion of damage information transfer [14], and the occurrence of abnormal human mobility [15]. More specifically:

- Initial response is a critical for effective recovery operation, but immediate response phases are characterized by a variety of deficiencies.
- Administration finds it challenging to anticipate and respond to the possible change with disaster situation. They choose locally optimal strategies in present situation or depend on their past experiences.
- Skillful expert is required to estimate another complexion on human mobility caused by changed geo-physical condition and government's countermeasures.
- There is no quantitative method of evaluating the subsequent transition. In other words, they have no way of verifying interim result.
- Even if they have numerical evaluation method, confusion of information makes the assessment each workgroup made quite vary.

The purpose of this paper is to develop road reconstruction plan for rapid recovery of disrupted human mobility to original level. We utilize GPS data from smartphone and Digital road map to estimate the change of human mobility under reconstruction and evaluate the human mobility recovery rate with the comparison of original state. Furthermore, the method we select is Deep Reinforcement Learning. Reinforcement learning (RL) is one method of machine learning, which is known to show outstanding performance in a variety of fields in recent years. This method would be suitable for solving choice or control problems through agent imitating human intelligent. With no prior knowledge or basic knowledge, we could determine the model systems including changes and/or uncertainties by utilizing RL [16]. Furthermore, RL could overcome formidable given the scale, so it is possible to consider the exponential number of input

factors. Based on this advantage, we could make the agent in RL derive the optimal reconstruction plan using large amount of human mobility data.

1.2 LITERATURE REVIEW

1.2.1 *Identifying Effective Road-Reconstruction Strategies in Disaster*

There have seen many changes with technologies and environment improvement: rapid urbanization, the population concentration, the degradation of resilience, and the negative knock-on effect between regions. The risk of unpredictable and serious disaster damage has been increased, so the necessity of disaster risk management is growing. Aydin et al [12] proposed a methodology to evaluate road recovery strategies for restoring connectivity after blockage due to an earthquake. Yamada Y et al [17] examined the restoration order with constraints on available human resources and materials. Chang S.E. [18] utilized the concept of accessibility to evaluate and enhance the performance of urban transportation systems in the aftermath of disaster. Balal et al [19] proposed five concepts for measuring urban highway network resilience and recommended that other research should define resilience measures to meet project requirements. Masafumi H [20] considered the prospect of possible indirect road network paths as a standard for evacuation with the consideration of available personnel staff, machinery, and road crew cooperation for road recovery.

1.2.2 *Irregular Human Mobility in Disaster Situation*

Recently, GPS and call detail records (CDR) of mobile phones are being used for human mobility analysis [21]. The application of these data has been extended father. Some studies have used this to analyze the human mobility [22] after disaster.

When sudden disaster occur, irregularity of people's movement is increased. The analysis of mobile phone GPS/CDR data could improve the ability to understand the change of human mobility. Wako et al [15] found that people selected routes, transportation method, or their destination abnormally after Tohoku Earthquake in 2011. Lu X et al [23] said that population movements during disasters may be significantly predictable. These findings help relief organizations to efficiently reach people in need. Song X et al [24] developed the model of finding population mobility patterns after severe natural disaster. They confirmed that it is critical for planning disaster management and long-term reconstruction to understand and predict human movement. Yabe et al [25] proposed a framework to estimate the evacuation hotspots after Kumamoto Earthquake using location data collected from smartphones. They said that official could find where victims are effectively with this framework.

1.2.3 *Application of Reinforcement Learning in Disaster*

Reinforcement learning (RL), one of the model-free algorithms, generate insights and identify optimal answer through the interaction with the ever-changing environment. The agent in RL could deal with uncertainty that could be difficult for decision makers to fully consider. The use of RL in disaster management has recently attracted much attention because it has potential to replace human decision making or supplement expert judgement and traditional response method. Nguyen et al [16] scheduled the effective distribution of volunteers to rescue victims by proposing a heuristic multi-agent RL. Saravi et al [26] proposed an algorithm for collecting information with RL. They noted that the information helped to improve resilience, prevent damage, and save lives in case of flooding. Su et al [27] proposed a path

selection algorithm based on Q-learning to provide disaster response as quickly as possible. Yang S et al [28] suggested the optimal policy for recovering the inter-firm transaction network in the supply chain with multi-agent RL. Companies had better secure alternative business partners first. They believe that it is possible for them to recovery efficiently by utilizing this model.

1.3 THESIS ORGANIZATION

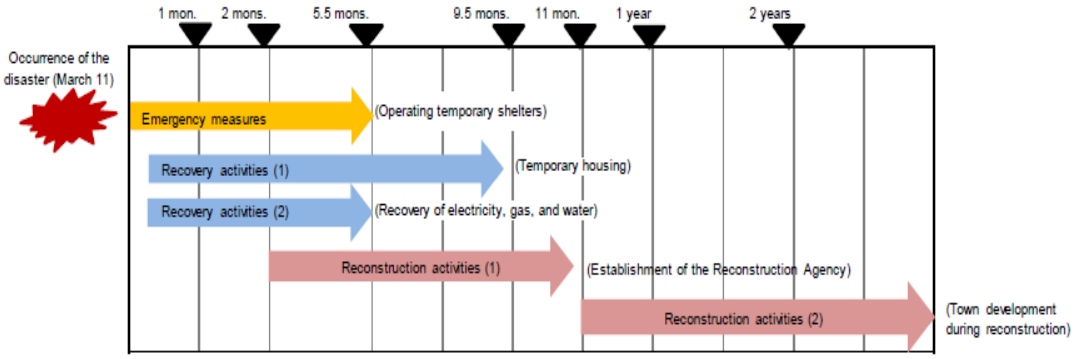
The remainder of this paper is organized as follows. In Chapter 2, we explain disaster management system. Section 3, we explain the suggest decision-making system. Section 4 gives some information of the Western Japan Flooding, as a case study and describes the digital road map and mobile phone GPS data of Hiroshima Prefecture. In Section 5, we describe the result of single agent RL system and multi agent RL system and explain the additional analysis of agent's learning result and the basic reward setting. Section 6, we summarize the conclusion, the limitation and future work.

Chapter 2. DISASTER MANAGEMENT SYSTEM

The overall process in disaster management consists of four phases: pre-disaster planning, rescue operations, recovery and reconstruction [5]. For setting effective disaster management system, we need a clear understanding of the characteristic and the objective of each stage.

2.1 OUTLINE OF DISASTER MANAGEMENT SYSTEM

When natural disaster occurred, many infrastructures sustain great damage at the same time. The source of disaster damages could be divided into three main categories: 1) Damges to housing, 2) damages to lifeline infrastructure (e.g., electricity, water supply etc.,), and 3) damages to road-network [29]. These damages have a fatal impact on local economic activities and livelihoods. So, appropriate restoration strategies are necessary to straighten out disrupted infrastructure facilities. As Figure 4 shown, the government generally proceeds with the reconstruction process as following a timeline from the emergency relief stage to the recovery state and then the reconstruction state [30]



Source: JICA Study Team

Figure 4. Processes up to Recovery and Reconstruction from the 3.11 Earthquake [30]

The restoration of infrastructure other than road-network focus on increasing the utility of afflicted population throughout the whole process. On the other hand, the purpose of road-network's reconstruction depends on recovery operation's stage. More specifically:

- In the emergency relief stage, administrations want to use roads for saving victims and transporting emergency goods. The recovery operation is generally to eliminate debris and secure at least one lane of disrupted roads for the passage of specific vehicles.
- From one week after disaster, government would allow general traffic to start use road-network. Roads with large traffic on normal days has higher priority than others for the increase of general users' utility.
- During reconstruction state, operation focus on restoring to its original form and recreation for better resilience than before.

Unlike reconstruction phase, emergency relief stage and recovery stage are related to improving road usage. Understanding the change of mobility pattern under dynamic road situation enhances prevention and response capability during disaster events [31].

2.2 EMERGENCY RELIEF STAGE

The role of transport-network in relief operation is to provide emergency support to the victims and the operation crew [29]. In detail, operation crew transport emergency goods and service from distribution centers to afflicted population as well as basic rescue operation (e.g., save lives and perform victims' search). The vehicles

with special authorization (e.g., ambulances, police vehicles etc.,) are only permitted to pass through. Because delayed delivery makes the possibility to save lives or relieve disturbed situation decreased. In addition, it is not only difficult to ensure safe passage to other vehicles but the permission to pass the general vehicles could also increase confusion in dealing with emergency situation.

Local officials (e.g., road managers, police, firefighters etc.,) usually have searched an emergency route that allows emergency vehicles to pass through. After disaster, road managers check the damage status of predetermined emergency routes and confirm the roads that be accessible. In addition, they identify isolated areas and required resources for restoration operation. The fundamental objective is to ensure that authorized personnel have access to isolated areas and other regions. Therefore, inaccessible detour route, expressway and truck road have higher priority for repair than others in this stage. This is because that the restoration of expressway ensures connectivity with other areas which did not be afflicted and makes it comfortable to get support from other regions and central governments.

It is necessary to identify which blocked roads have the impact on restoring the accessibility of the entire transport network. Although there are many model and measures of evaluating accessibility of each blocked road, the increase in travel time and travel distance might be the most common measures [32]. Wisetjindawat W et al [29] estimate travel time of the shortest path to evaluate response ability based on three disaster scenarios. Chang and Nojima [33] evaluated the accessibility with post-disaster shortest paths between all node pairs. Sohn [34] modified this method with the consideration of populations and traffic densities. Toshihiro A et al [35] decided the

recover order by observing the change of betweenness centrality of road network. The recovery order based on the evaluation of individual roads does not consider the transition probability. So, it is determined with Greedy algorithm. This method is to make the locally optimal choice at each stage with a reasonable amount of computational effort [36]. The recovery order could not always be guaranteed to be optimal.

2.3 RECOVERY OPERATION STAGE

Recovery stage involves mid and long-term measures to stabilize the community and restore normalcy after the disaster's immediate impact has passed [37]. Citizens gradually commence resuming normal activity and are allowed the use of road network. In other words, road users extend to ordinary people and road-network plays role of securing basic mobility services as well as managing contingency situation.

In this stage, government focus on how to increase the utility of users for the entire road-network as soon as possible. It would be best to evenly improve the performance of each part, but there is the limitation of resource (e.g., material, men powers). Their task is to determine which road is better to recover first for the rapid recovery of human mobility service. Government administrations utilize scoring method based on ten items which is included in four classification: 1) Risk or damage, 2) importance of road, 3) geological factors and 4) stability. Recovery operation is started from high-scored damage road. Furthermore, there are two additional measures for securing portability rapid: 1) Designation of alternative route and 2) early traffic opening with the utilization of present usable lanes.

As we mentioned in 2.2, previous researchers related to recovery operation also utilized these two methods: scoring method, traditional supervised learning. As the utility of road users become more important, some studies reproduce transport network and estimate human mobility under the change of road-network. Considering the probability of the change in the post-disaster situation is necessary to estimate human mobility under reconstruction. This means that a model with uncertainty and complexity needs a different methodology from the previous one. So, the optimization method is utilized to deal with solving optimization problem with increased complexity.

Osawa S et al [38] introduced potential accessibility indicators with free travel time and evaluate disrupted roads. They set Kumamoto earthquake as the case study and utilize road network with 3,142 links and 2,106 nodes. Sugimoto H et al [39] focused on the cooperation of operation teams and solved a restoration process in national highway network with Genetic algorithm. Chen X.Z et al [31] estimate the accessibility of different travel model in flooding scenarios and tried to consider the difference of people's travel behavior. David R et al [40] tried to minimize the total network travel time. Their model is designed on a realistic transport network with the consideration of two disaster scenarios. Sakamoto J et al [41] proposed an accessibility-based model for the priority order of road reconstruction. They set 75 O-Ds which are important to secure the connectivity and estimate accessibility using O-Ds travel time. With predetermined parameters, they consider the level of disaster, recovery capacity and the interruption of road usage. Hori et al [42] utilize multi-agent simulation for recovery process of lifeline. There are some challenges such as lack of generality and

covered areas. Bhatia et al [43] proposed hybrid method combining topology and optimization for recovery damaged network on real world transportation system.

2.4 CONCLUSION

Previous studies related to post-disaster response would be conducted with specific model setting. They utilized historical or random post-disaster cases to specify the parameters or the standard required for the model. In addition, the algorithm of vehicle routing problem (VRP) is often applied to evaluate disaster response plan and find the optimal strategies. However, it is challenging to apply current method to the actual disaster management. The reason is as follows:

- The parameters are estimated to fit the sample data and are fixed values. In other words, the solution in this model would not do anything in unexpected situation which is not included in sample datasets.
- With the curse of dimensionality, road network is man-made or covers with small-size region and movement subject is limited (e.g., operation crews or fewer ordinary people). The bias might arise from predicting the utility of road users.

We suggest how to combine mobile phone GPS data with Deep Q-Network (DQN) to effectively determine and develop the optimal recovery strategy with disaster scenarios. The expected improvements are as follows:

- Deep Q-network does not have specific model which includes the transition probability distribution and the reward function. The agent in DQN determine

its own model with accumulated knowledge on the environmental changes and effects of several actions.

- The agent could identify approximate function related to optimal policy repeated interaction with variable environment. This function could get approximate reward value with unpredictable situation.
- The agent could find the trends in influencing factors with neural network although there is a great deal of considerations. So, we could deal with more than a thousand user information and the network analysis result of one prefectural area without the curse of dimensional.
- The agent basically wants to choose the action which help the cumulative future reward maximized [44]. We could consider the optimal policy from a mid-to long-term perspective.
- We utilize digital road map and origin-destination pairs from mobile phone GPS data to estimate human mobility in the condition that closely resemble real situation.
- We evaluate current strategies periodically based on the change of traffic volume, the representative of huma mobility. In other words, we could confirm the effect of each reconstruction operation with the result of VRT.
- By using geographic information system (GIS) application, estimated traffic volume could be visualized at each time step. Officials could see at a single glance human mobility's change with reconstruction and consider additional measures.

Chapter 3. DECISION-MAKING SYSTEM

The proposed decision-making system is to identify the optimal strategies for improving road users' performance. DQN is one of the methods of Reinforcement learning, which refers to an algorithm that learns agents who make the best choice under a given condition through repetitive trial and error based on Deep artificial neural network. Section 3.1 describes DQN, one of RL methods, and we describe four components in decision-making system with DQN in Section 3.2. The framework of proposed model is presented in Section 3.3.

3.1 Deep Q-Learning Algorithm

3.1.1 *Outline of Reinforcement Learning*

The proposed optimal decision-making system utilize the estimation of traffic volume, the operation progress rate as input data of deep artificial neural networks. Reinforcement Learning (RL) is one method of three basic machine learning with

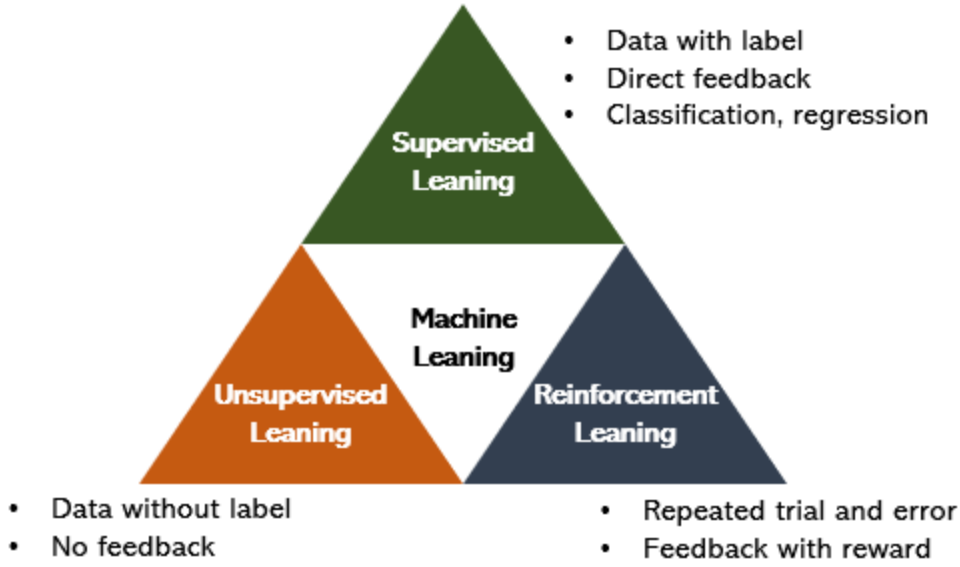


Figure 5. three basic machine learning [47]

supervised learning and unsupervised leaning. The agent concerned with which action should be taken in specific environment for the maximization of cumulative reward [45]. Supervised leaning finds the unknown function that connects known inputs to unknown outputs and is specialized in regression and classification problems [46]. Unsupervised leaning is one machine learning method that specialized in clustering or finding patterns in a data set with no pre-existing labels [47]. Figure 5 shows main features of three types of machine learning. Unlike supervised and unsupervised learning, RL aims to make intelligent agents study the optimal decision by observing the environment and undergoing repeated trial and error [45].

RL started from psychology studies of animal behavior. The term "reinforcement" derives from the animal behavior's experiment using the Skinner box. The animal which does not know the relationship between food and buttons learns this mechanism through repeated trial and error. Sutton R. S. et al [48] define reinforcement as the strengthening of a pattern of behavior as the result of receiving a stimulus. RL refers to the process of applying these psychology studies to mechanical leaning. Intelligent agents learn the policy by experiencing various process of obtaining reward with various action and state. They identify action value function which means policy that maximize the total sum of reward through several trials [45]. To solve some problem through RL, we need two types of processes: 1) Expressing problem mathematically and 2) making the agent learning optimal policy. This whole process is called RL algorithm.

RL began with the dynamic program proposed by Richard Bellman in the 1950s to solve the optimal control problem. Richard Bellman suggested how to express the

problem of choice in real situation mathematically and solve it through policy evaluation. However, there are major disadvantage of the initial RL: 1) Complexity of calculation and 2) uncertainty of proposed model. These limitations make it difficult to apply RL to solve problem that are comparable to real situation. Model-free RL has been developed by applying Temporary Difference method (TD) proposed by Watkins and Dayan in 1992 and Policy Gradient technique proposed by Ronald J. Williams. Accordingly, it has begun to be applied to real problem through various methods.

Artificial intelligence based on supervised learning system is made to replicate the decision of human experts, but we need reliable learning data for good learning result and there is the ceiling on the performance [49]. In the other hands, RL systems train them from their own experience, in principle allowing them to exceed human capabilities, and to operate in domains where human expertise is lacking [46]. Present studies, especially published by Google's DeepMind team, have proposed various RL algorithm with deep neural network and shown the ability to solve complex problem above an expert level. This method is being put to use in various fields.

RL algorithms are divided into two groups: 1) Model-based RL and 2) model-free RL. Thomas M et al [50] defined that model-based RL is learning of a global policy or value function based on known or given samples. The agents relatively easy accomplish their own task by exploits previous learned model. This method has some disadvantages: 1) Uncertainty of known model and estimation's result, 2) requirement of high volumes of prior data, and 3) complex data processing. Typical model-based RL include Monte-Carlos Tree Search (MCTS) applied to AlphaGo Zero [51]. On the other hand, model-free RL is a method of training action value function or policy

functions through repeated interaction with environment. Without known model of environment in advance, this method makes it easy to calculate and apply to a variety of environments because agents could identify the optimal policy function based on the accumulated samples through repeated trial and errors. However, the amount of learning data is more required than model-based RL, and it is difficult to deduce the relationship to how input data affects the result, which is the problem of black box.

3.1.2 *Learning Method of Deep Q-Learning Algorithms*

RL is modeled on Markov Decision Process (MDP). The components in RL could be expressed mathematically through MDP. MDP has proven to be applicable to problems with Markov Property, in which the current state is the complete determinant of the next state and the next state is independent of prior history [48]. Finite MDP could be expressed with state, action, reward. Agents in RL learn to select action having the maximum reward through interaction with environment. Specifically, the agents observe and analyze surrounding environment and do one action based on the present state identified from the observation. Whole series of this process is referred to as sequential decision making.

The passage of time in RL represents the information acquired in time unit and this information is defined as a state, s_t, at time t. When a set of whole states that could be acquired in given environments is called $\mathcal{S}$, the relationship between $\mathcal{S}$ and s_t is $s_t \in \mathcal{S}$. In the same way, the action which the agent take at time t and the set of all action are represented as a_t, $\mathcal{A}$ respectively. The relationship between these two factors is $a_t \in \mathcal{A}$. At each time step, agent choose one action, a_t, in the set of actions, $\mathcal{A}$, based on its own policy. Policies could be expressed in a particular function

form that assume input value and output value as s_t, a_t respectively. This value function is Policy function and represented as follows [48]:

$$a_t = \pi(s_t) \tag{1}$$

Environment transmits s_{t+1} and r_{t+1} at time $t+1$, which is the change of information based on at. Figure 6 describe the conceptual diagram of the repetitive interaction between the environment and the agent.

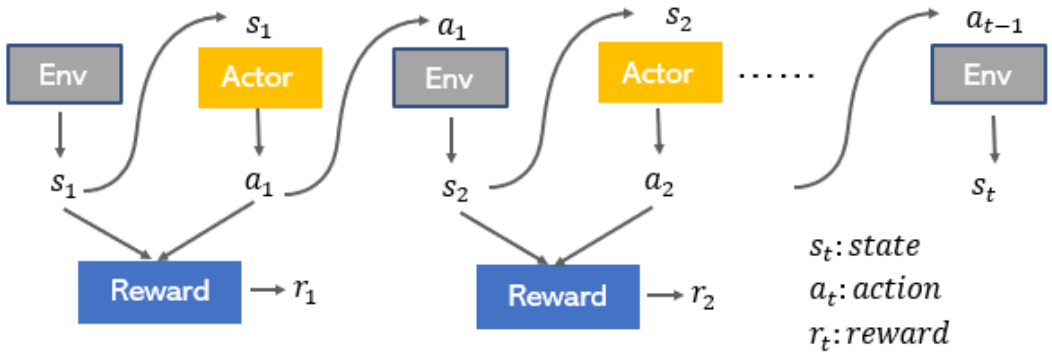


Figure 6. Diagram of the interaction between environment and agent

Problem-solving with RL is to find policy that make agent decide the optimal action for a certain situation by utilizing mathematically represented components of MDP. The optimal policy would instruct the agent to select the action that maximize the sum of reward under any circumstances. There are various methods of optimizing policy in RL and in this study, we utilize Deep Q Learning as the method of optimizing policy.

Deep Q-learning algorithm is techniques for determining optimal policy using action value function defined in the form of deep artificial neural networks. Almost all RL algorithms involve estimating value functions that is mainly divided into two types:

1) state value function and 2) action value function [48]. State value which is denoted $\mathcal{V}^{\pi}(s)$, is the expected return in the case that agent follows present policy π, in a state, s_t. If one system wants to consider the reward at each time differentially, the concept of the discount rate, γ, is applied to calculate the cumulative reward. The agent in the system would select action, at to maximize the expected discount return, denoted G_t [48]:

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad (2)$$

r_{t+1} is reward obtained by action, a_t at time t, and γ is a parameter, $0 \leq \gamma \leq 1$, called the discount rate. The discount rate, γ, is the determinant of the present value of future rewards. It is used to prevent the total sum of all expected rewards from increasing indefinitely by setting the range of 1 or less, so finite learning process is guaranteed. In addition, we could make the agent consider the sense of time (e.g., myopic, future-oriented).

Similarly, Sutton R.S. et al [48] define the value of taking action a_t in state s_t under a policy π, denoted $Q_{\pi}(s, a)$, as the expected return, G_t, starting from s_t, taking the action a_t, and thereafter following policy:

$$Q_{\pi}(s, a) = E_{\pi}[G_t | s_t = s, a_t = a] \qquad (3)$$

In other words, the expected value at time t, denoted is expressed with the maximum value of value function, $Q_\pi(s_{t+1}, a_{t+1})$, and the present reward, r_{t+1}. The equation is as follows:

$$Q_\pi(s_t, a_t) = r_{t+1} + \gamma \, max \, Q_\pi(s_{t+1}, a_{t+1}) \quad (4)$$

Assuming that there exists the optimal policy, π^*, that returns the maximum value of the expected reward in whole states, action value function under optimal policies, π^*,would always return the highest value. In other words, the policy that make the value of action value function the largest is the optimal policy, π^*[48].

$$Q_\pi^*(s_t, a_t) = r_{t+1} + \gamma max Q_\pi^*(s_{t+1}, a_{t+1}) \quad (5)$$

To make action value function, $Q_\pi(s_t, a_t)$, under any policy, π ,close to the optimal policy, π^*, it is necessary to minimize the difference between action value function, $Q_\pi(s_t, a_t)$, and the optimal action value function, $Q_\pi^*(s_t, a_t)$. However, we do not have the information about the optimal policy. So, we assumed the target Q-function and utilize it as the substitute of the optimal action value function. This target function is updated periodically to allow for the latest optimal Q-function. Concurrently, action value function learns how to be optimal under updated target Q-function. The target Q-function is expressed as $\hat{Q}$ and the loss function $\mathcal{L}$ is as follows [46]:

$$\mathcal{L} = r_{t+1} + \gamma max \hat{Q}_\pi(s_{t+1}, a_{t+1}) - Q_\pi(s_t, a_t) \quad (6)$$

Agent tries to find optimal policy by updating both of target Q-function and action value function (Q-function) through learning process. The target Q-function play the objective function of Q-function. The update frequency of target Q-function should be adjusted accordingly to ensure that Q-function exploit properly this function. If the cycle of target's update is too short, the objective function would be changed before Q-function have finished present learning, thus preventing proper learning.

The agents could not determine the value of actions that have not been experienced if they always choose their action under the action value function. So, they have to pick action at random, accumulate various experience and explore the optimal policy. This process is called Exploitation and exploration. Exploitation chooses the greedy action to get the most reward. On the other hand, exploration allows the agent to improve its current knowledge about each action. The experience gained through exploration might be better one than what has gained so far, or it might be worse. If the experience gained through exploration might be better one than what has gained so far, it might be useful in updating Q-function in better direction. Otherwise, the attempt to explore would be serve as the waste. Therefore, we need to make the agent prefer exploration until enough experience has been accumulated, and then change their preference from exploration to exploitation. A simple method for performing the balancing between exploration and exploitation is the method called ε-Greedy [52]. With ε -greedy, the agent selects at each time step a random action with a fixed probability, $0 \leq \varepsilon \leq 1$, instead of selecting greedily one of the learned optimal action with respect to the Q-function [52]. Google's DeepMind team introduced replay memory to address the problem of poor learning performance arising from the

correlation of samples of time-series data. Replay memory is a method of selecting randomly samples which would be utilized for updating action value function. Through this method, we could prevent the degradation of learning performance due to the strong correlation of adjacent time-series data.

3.2 Component of Decision-making System

RL solves the problem of sequential decision process which is needed to be seen in perspective. There are four primary components in RL: 1) the agent, 2) the action, 3) the reward, and 4) the environment. In sub-chapter, we illustrate the concept of each components and define each component with road reconstruction.

3.2.1 *Agent*

Agents are the representative of making decisions (e.g., individual, firm, machine or system) based on the feedback. They select action in specific state which presents virtual environment as a vector and gain different reward / punishment with various actions and states. Their fundamental role is to learn something from trial and errors and then make better decisions about the given situation. In other words, they utilize the knowledge through learning and adapt good action automatically although starting to execute with basic knowledge or without prior knowledge.

The municipal government is the key decision maker of road reconstruction operation. They deploy reconstruction teams immediately in the wake of flooding to restore damaged areas to normal state. The problem related to the reconstruction work includes deciding when, where and how many operation teams need to be dispatched. Accordingly, we set the agent in this model as the road reconstruction crew.

3.2.2 *Action*

The action means what agent do at each time step. This factor makes environment's change, reward. In other words, agent could interact its environment through action. RL agents typically have either a discrete or a continuous action space [48]. With a discrete action space, the agent decides which distinct action to perform from a finite action set. With a continuous action space, action is expressed as a single real-valued vector [53].

The road recovery operation consists of three processes: 1) Inspection, 2) planning and 3) implementation. We limited agent's operation to something performed during the implementation process. Disaster damage considered in this paper is mainly landslide by heavy rain. The agent is assumed to remove soil and rock from the landslides, replaces and compacts demolished road sections. This operation is known as "Excavation and Embankment". Figure 7 shows overall operation process.

Figure 7. Excavation and embankment operation process

Hiroshima Prefecture set the amount of work available per a day (8 hours) for one worker who operates the machine depending on the type of restoration work. Table 1 describe the operation hour it takes to complete $100m^2$. Based on these standard, we assumed that the daily workload of one worker would be $256m^2$. In summary,

reconstruction crew (agent) selects specific damage road, so the shaping of actions is discrete action space. The size of action space is the number of damage roads that are the target of the operation. The meaning of action is that the agent selects one damage road at each time step which is one day. And then it had to work within daily workload that had already been established.

Table 1. The operation hour of excavation and embankment [54]

(h /$100m^2$)

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

After disaster, disrupted roads would be closed, and road users are restricted from traveling. Government's objective is to recover restricted human mobility to normal state as soon as possible. In the term of the civil engineering project, that means to secure the available lanes and increase the number of passable traffics. Accordingly, we could think that at least the entiral of the damaged road was needed to secure all available lanes. We determined the maximum amount of work on each disrupted road as the area of each road.

Thrun and Schwartz [55] said that the agent who has large action spaces has the tendency to converge to a suboptimal policy. For example, if the number of actions is $\mathcal{N}$ and the number of trial steps is $\mathcal{M}$, the number of possible combinations is $\mathcal{N}^{\mathcal{M}}$.

That is, it is difficult to identify the optimal policy in finite number of simulations as the number of actions increases. In the actual restoration process, operation crew does not do any more operation if the maximum amount of work is done. However, a set of actions in basic RL remain unchanged, so agent choose afflicted road that no longer requires recovery operations. In other words, it only hinders the converge to optimal policy to allow agent to choose meaningless action.

T. Zahavy et al [56] propose the Action-Elimination Deep Q-Network (AE-DQN) that combines a Deep RL algorithm with an Action Elimination Network (AEN) that eliminates sub-optimal actions. Through the elimination signal, agent could know which actions not to take, thus mitigate converging to sub-optimal policy. We adopt this method and let the agent does not select damage road which the maximum workload is completed.

3.2.3 *Reward*

Reward shaping attempts to model the conduct of the learning agent by adding additional localized rewards that encourage a behavior consistent with some prior knowledge [57]. On each transition, the environment judges the experience and send a corresponding reward value to agent. Through reward, agent could identify whether its action at each time step is good or bad for achieving its goal. It is important that the reward function make use of prior knowledge adequately. Because reward value make agent perform good selection and accelerate the process of converging to optimal policy.

The reconstruction goal is to improve performance and recover human mobility to normal state as fast as possible. The government has not only this fundamental goal,

but also goals for the recovery period and construction cost. In this model, we have set the recovery period (the number of steps) and the amount of resource (the number of workers) in advance. The construction cost is limited to the travel time of the agent from pre-determined starting point to current operation place. We want that the agent learns the effect of restoration of each damage road on human mobility recovery and recovers first from the disrupted road with high effect value. In addition, it considers the moving cost and the relationship among target disrupted roads. There are three consideration for reward setting: 1) Accessibility, 2) the degree of human mobility recovery and 3) the connection between disrupted roads. The reason is as follows:

- Accessibility means the cost of the agent's operation. This factor allows the agent to have the tendency to move to other damage roads after finishing a certain amount of current operation.
- The change in human mobility recovery rate represents the impact of agent's action at each time step on human mobility recovery. The agent perceive action with high reward as good choice. By using the change in human mobility recover rate, the damage road with high impact would be chosen first.
- Road-network has the intimate connection. We provide the agent with knowledge of the connectivity of damage roads that people pass through. This value is calculate based on the traffic volume which passed through workplace of the previous step and of the current step at the same time.

The recovery operation involves multi-objective optimization (MOO). MOO is the process of simultaneously optimizing multiple objectives which can be complementary, conflicting as well as independent [58]. With reward shaping, Tim B

et al [59] suggested the combination of an extra reward ($\mathcal{F}$) and the basic reward ($\mathcal{R}$). This reward makes the agent drive the exploration behavior incorporating heuristic knowledge of the system designer [59]. The Q-learning update rule with extra reward system as follows [59]:

$$\begin{aligned}\hat{Q}(s_t, a_t) \leftarrow (1-\alpha_t)\hat{Q}(s_{t-1}, a_{t-1}) \\ + \alpha_t[\mathcal{R}(s_t, a_t, s_{t+1}) + \mathcal{F}(s_t, a_t, s_{t+1}) \\ + \gamma \max_{a_{t+1}} \hat{Q}(s_{t-1}, a_{t-1})]\end{aligned} \tag{7}$$

We define the fundamental reward ($\mathcal{R}$) utilizing the human mobility recovery rate. The extra reward ($\mathcal{F}$) includes the agent's travel time and the connectivity value between the agent's action at step $t-1$ and at step t. We would expect the agent to choose the action having better effect of human mobility recovery with the consideration of behavioral guidelines.

3.2.4 *Environment*

Environment is everything surrounding the agent. For example, the bicyclist (agent)'s environment includes the road, bicycle, driver's body. The environment response to agent's action and provide the latest situation. In addition, the environment provides special numerical values, rewards, to agents for their optimization problem which maximize the total sum of discount value. An environment is complete specification of agent's task [48]. In other words, we should provide as much as

environmental information as possible to ensure that the agent finds the optimal policy effectively.

The state refers to environmental information in a vector form, the input value for determining action value function. Therefore, the state should contain as much evidence as possible regarding the reward received according to the action of each time step. As we mention in 3.2.3, the reward consists of these three factors: 1) agent's travel time, 2) the change in human mobility recovery rate, and 3) the ratio of traffic volume passing through both previous damage road and current damage road to the total of traffic volume. Accordingly, the state space basically includes the operation progress rate and human mobility recover rate of each damage section, travel time (hour), previous action, and the average of human recover rate.

3.3 Framework of Proposed Model

We describe the framework of our model and the process of estimating human mobility. As Figure 8 shown, there are four processes which present agent's action and the accompanying changes in the environment.

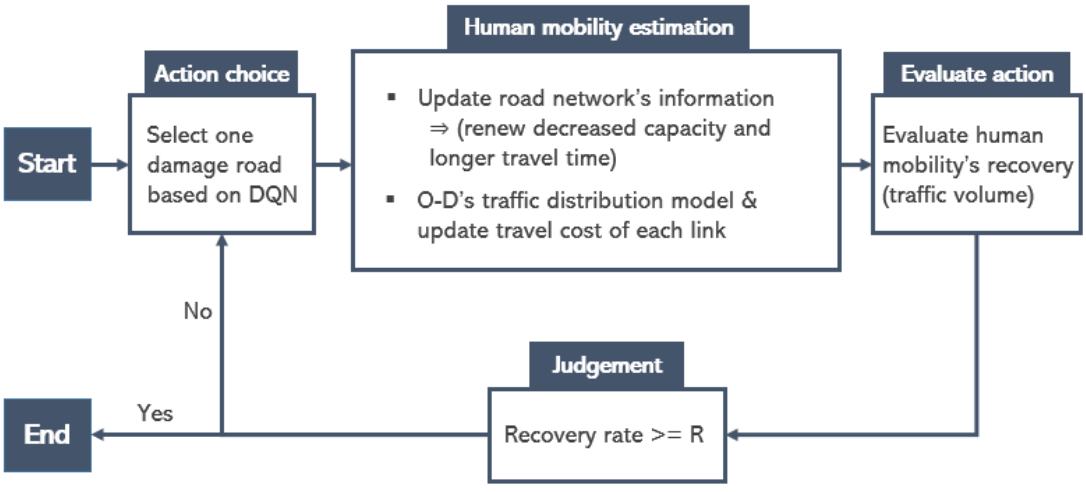


Figure 8. the framework of proposed model

Agent selects one damage road based on DQN at each time step. Road-network attribute and O-Ds' traffic volume are updated with accordance of all target roads' operation progress rate. Based on updated information, traffic flow on each link is estimated and human mobility recover rate is evaluated. At last, human mobility recovery rate is above target value, the whole simulation is over. Otherwise, agent do these processes again up to the maximum number of trials.

3.3.1 *Calculation of Operation Progress Rate*

The disrupted area and the degree of damage are different depending on the type of road and the surrounding environment (e.g., a riverside road or mountain area). In other words, the more disrupted or large-scale roads, the greater the workload required for restoration. Conversely, daily workload the agent perform at each time step is constant regardless of the damaged road sections' size. One disrupted road with small area might be accomplished sooner than with large area. So, we adjust the rate of increase in work progress rate using the relationship between the total workload and the cumulative workload of each disrupted road.

Productive efficiency might vary in each phases of construction project. Because balanced performance requires operation experience and repetitive practice, and it takes time to do this. S-curve would best represent a cumulative flow of material or money over a time period. The methodology could be utilized for estimating manpower utilization rate [60]. Accordingly, we would predict the progress rate with the agent's action by applying sigmoid function. Sigmoid function is one of S-shape curve which represents the cumulative progress rate [48]. Road's cumulative progress rate is calculated with Equation 8:

$$\mathcal{R}_{m,t} = \frac{1}{1 + e^{-0.8x_t}}$$
$$\because x_t = -7 + 13(\frac{\sum_{k=0}^{t} \mathcal{W}_k}{\mathcal{S}_m}) \quad (8)$$

$\mathcal{R}_{m,t}$ 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 road.

3.3.2 *Human mobility Estimation Process*

Traffic assignment estimates loads, user volumes on each segment of a transportation network [61]. These would be 24-hour traffic volumes, peak hour transit volumes, or yearly volume of freight flow [62]. Among the estimated user volumes, we estimate peak hour transit volumes to identify the effect of the given transportation system at each time step on traffic generations. The required data is the network topology and O-D matrix. In this paper, mobile phone GPS data and digital road map are utilized to estimate traffic volume similar to the real situation.

3.3.2.1 The change in Road Capacity with Reconstruction

Immediately after the flood, the government blocks the disrupted roads to identify the extend of the damage and relieve emergency situations. Initial basic capacity and traffic volume vary depending on the extent of the damage, but we have no information related to damage in current situation. We assumed that all demolished roads totally lost its function and vehicle access is not allowed without restoration work.

Road's capacity is the maximum flow obtainable on a given roadway using all available lanes. Reconstruction in this paper is to recover human mobility to normal

state. In other words, recovering traffic volume to normal state means to increase the number of usable lanes from zero to the original value. Based on the agent's workload in 3.2.1, we could think that the workload of the product of the length of road and the width of lane is needed to make one lane available. Therefore, it could be assumed that the degree of capacity recovery is same as the rate of operation progress of corresponding road. The assumptions of traffic capacity are as follows:

- **Assumption 1.** The initial value of road's capacity is zero.
- **Assumption 2.** The degree of capacity recovery is the same as the rate of operation progress of the corresponding road.

3.3.2.2 The change in Travel Generation with Reconstruction

MLIT and municipal government provide the information on the road reconstruction process to ensure the convenience for road users as much as possible.

OD	Presence of detour	D10's progress rate	D3's progress rate	D11's progress rate	Traffic volume (K)
A	Y	0.9	0.3	0.1	0.2*K
B	N	0.9	0.3	0.1	0.1*K

Figure 9. Example of traffic generation at each step

These information include the operation process and available alternative routes. In unusual road network, trip generation is determined based on the notification by road reconstruction managers. Specifically, if the road network is seriously damaged and there exist no detour route, O-Ds are unable to move and isolated. In addition, we could expect the trip generations to be increasing as the recovery rate of damage road they pass through is increasing. Accordingly, we make one assumption (Figure 9) with the trip generation under reconstruction:

- **Assumption 3.** The amount of each O-D's movement is influenced by the possibility of indirect routes, the minimum value of the cumulative progress of damage road that O-D pass through on normal days.

3.3.2.3 Traffic Allocation Assignment

Traffic Assignment (TA) model simulates how travel demand and transport networks interact in transportation system [63]. We could calculate the travel cost and traffic flows on each link with O-Ds' demand and route choice. There are two types of traffic allocation algorithms: 1) Static traffic assignment (STA) and 2) dynamic traffic assignment (DTA). The detail explanation is as follows:

- In STA, the performance of each link is not affected at all by other thing (e.g., traffic flow, congestion) and is fixed at a constant value. Route choice of O-Ds is determined with unchangeable variable and is easier to be expected.
- DTA models allow the changes of components (e.g., link costs, travel demands). It is possible to observe the change in traffic volume and trajectory

selection of O-Ds according to specific situation (e.g., evacuation, commute hours etc.,)

We introduce DTA models to estimate vehicular flow with changed transportation network under reconstruction. Three processes have to be used for traffic allocation with road network. The preceding two steps refer to find out the assignment route. The shortest path tree is utilized to find the shortest path from source node to all other nodes. This tree notified all road links that traffic with specific source node could pass through. Lastly, travel demands may be allocated to several path between source node and target node with the consideration of the road performance.

There are many factors representing road network performance (e.g., travel cost, capacity, accessibility etc.,). Among them, travel time is quite important thing to select trajectories. However, the road network during the reconstruction is in the unstable state, so O-Ds choose their own trajectories based on the capacity of all selectable paths. The assumption with the allocation of O-D's flow is as follows:

- **Assumption 4.** The amount of traffic allocated to one of trajectories that O-D could pass depends on the minimum capacity of the link that constitutes this route.

The reason for this assumption is as follows. Disrupted road has decreased basic capacity. This attribute cause travel time to fluctuate easily even if traffic volume on this link is quite low. In other words, the use of disrupted road under restoration implies uncertainty in the surge of travel time, although it is difficult for road users to estimate the actual travel time. It is more likely to secure stable travel time by considering the saturation of use.

Meng Q et al [64] explained the basic notations, assumptions in DTA conditions. We would like to use their descriptions to explain the traffic allocation algorithm in this model. $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ refers to transportation network, where $\mathcal{N}, \mathcal{A}$ are sets of nodes and link, respectively. O-D pairs are described (r, s) with origin node (r) and destination node (s) . $\mathcal{R}$ and $\mathcal{S}$ mean the set of Origin, Destination respectively. Denoted by $\mathcal{K}_{rs}$ the set of paths connecting O-D (r, s) , by $q_{rs,t}$ travel demand of O-D (r, s) at each time step t.

With Assumption 4, denoted by $\mathcal{C}_{kt}^{rs}$ the minimum capacity, by $\mathcal{U}_{kt}^{rs}$ road usage rate on path $k \in \mathcal{K}_{rs}$ at each time step t. $\mathcal{F}_{kt}^{rs}$ means the traffic flow on path $k \in \mathcal{K}_{rs}$ at each time step t and is calculated with Equation 9:

$$\mathcal{F}_{kt}^{rs} = q_{rs,t} * \frac{\mathcal{C}_{kt}^{rs}}{\sum_{k \in \mathcal{K}_{rs}} \mathcal{C}_{kt}^{rs}}$$
$$\because \mathcal{U}_{kt}^{rs} = \frac{\mathcal{C}_{kt}^{rs}}{\sum_{k \in \mathcal{K}_{rs}} \mathcal{C}_{kt}^{rs}} \tag{9}$$

The traffic flow of each link $\mathcal{V}_{at}$ at each time step t would calculate with the fundamental flow equations [64]:

$$\mathcal{V}_{at} = \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{K}_{rs}} \mathcal{F}_{kt}^{rs} \delta_{ak}^{rs}, a \in \mathcal{A}$$
$$\sum_{k \in \mathcal{K}_{rs}} \mathcal{F}_{kt}^{rs} = q_{rst}, r \in \mathcal{R}, s \in \mathcal{S} \tag{10}$$

where $\delta_{a\hbar}^{rs} = 1$ if path $\hbar \in \mathcal{K}_{rs}$ between O-D pair (r, s) traverse link $a \in \mathcal{A}$, and 0 otherwise.

3.3.2.4 Evaluation of Human mobility recovery

We define the human mobility recovery rate $\mathcal{R}_t$ at each time step t as the total amount of loads with respect to the total average traffic volume on normal days. Equation 11 is as follows:

$$\mathcal{R}_t = \frac{\sum_{a \in \mathcal{A}} \mathcal{F}_{a,t}}{\sum_{a \in \mathcal{A}} \mathcal{F}_{a,n}} \tag{11}$$

where $\mathcal{F}_{a,n}$ is the estimated traffic flow on link a on normal days, $\mathcal{F}_{a,t}$ is the estimated traffic flow on link a at each time step t .

We found all possible road links that O-Ds travel and predicted the traffic volume on each link based on Equation 9 and 10. We estimated the vehicular flow on normal road network, the pre-disaster condition, with 300 trials. This is because the network environment faced by O-D with the different departure timing changes, resulting in the difference in the amount of traffic on each link. We utilized the total average loads as the standard for evaluating human mobility recovery to mitigate the variability. Therefore, we identify the degree of human mobility recovery rate in each damage roads and in overall.

Chapter 4. STUDY AREA AND DATASET

4.1 WESTERN JAPAN FLOODING

In 2018, heavy rain for almost a week (from June 28 to July 8) resulted in widespread and destructive floods and landslides. Weathernews Co., Ltd. conducted a hearing survey of approximately twenty thousand local residents. Approximately 80% of all high-risk area might have been flooded or corrupted based on these result [65].

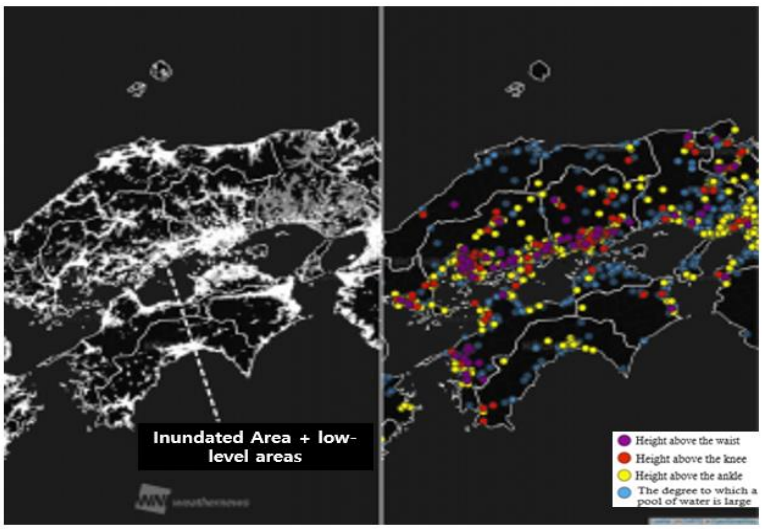


Figure 10. The flooding damage situation in Western Japan Flooding [65]

Figure 10 illustrate the flood damage situation reflecting their survey. Further, five hundred and eight-two road section in Hiroshima and Okayama Prefecture were disrupted by this flooding. It took from as low as four days to as high as 80days to allow transit of vehicles. Among many damage roads, six sections were expressways, 56 sections were national roads, and the rest were designated city streets and prefectural roads [66]. Figure 12 shows the extend of the road damage in Western Japan.

Figure 11. the extent of the damage along national roads 2 and 31

Transportation networks are the backbone of critical infrastructures because they provide accessibility to the other system and rescue operation after flooding and during restoration processes [12]. This flooding caused clogged road and resulted in the isolation of numerous regions. Further, it took more than a week for commonly used roads in the daily lives of citizens after the occurrence of flooding [66] to have their basic function re-established. Additionally, a year was required to remove all the expressway restrictions [66]. People and industrial parties had some problems with a shortage of daily necessities, due to mobility constraints and unstable procurement.

4.2 DATA COLLECTION / PROCESSING

We utilized three types of dataset to determine the optimal reconstruction strategies with the consideration of human mobility.

4.2.1 *Mobile phone GPS dataset*

GPS data of mobile phone from the Agoop Co., Ltd. is collected by the users who give the agreement of providing their location information. This data basically consists of four types of data: 1) user ID, 2) longitude, 3) latitude, and 4) timestamp. Their locational information is procured whenever the individuals' position was changed. Table 2 show the detail of GPS data we utilized. The number of users who provide mobile GPS data is approximately 0.3% of the population in Hiroshima Prefecture (Hiroshima, Higashi-Hiroshima, Kure) and Okayama Prefecture (Kurashiki, Soja). The GPS log amounted to 102,821, and the period of observation is from June 1 to June 30.

Table 2. The detail of mobile phone GPS data from Agoop Co., Ltd.

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 (ave. 27 logs/user)

The transportation mode assumed in this model is the automobile, so 1km grid is defined as a stay point detection unit. We could identify the stay point and the timing of departure and arrival of each stay point. With time-periodic location data, travel demand, the representative of human mobility, is estimated by using the concept of Origin-destination matrix. Through trajectory analysis, we identify three thousand three hundred twenty O-Ds passed through the road section afflicted by the Western Japan Flooding.

4.2.2 *Local geographic information*

Japan government provide local geographic information which is the combination of numeric information and geographical factors. These datasets include population, disaster damage, land-usage and so on. Furthermore, they make general road network's information. Japan Digital Road Map Association has reproduced real road-network based on the 1:25,000 topographic maps and been updating this map every year [67].

To embody the change of human mobility with the road reconstruction, we acquire actual road network information and estimate the hourly traffic volume of target O-Ds. We utilize the number of commuting population and residents of each 1km grid which is provided from the Statistic Bureau and Digital road map. In addition, sediment disaster alert areas and inundation depth rank are also utilized to estimate the extend of disaster damage.

4.2.3 *Information of disrupted roads*

Citizens and industrial entities need the information related to real recovery operation situation to secure their basic mobility with reasonable and safe route. The Municipality and Ministry of Land, Infrastructure, Transport and Tourism (MLIT) provide daily situation information. Road recovery situation is presented: 1) the name of target road, 2) damaged stretch, and 3) state of restoration. The restoration process is divided into three stages: Road closed, one-way traffic and the completion of recovery operation. We could know how long it takes to reopen each disrupted road from this information. However, the short period does not mean that the disaster damage is small. This is because that the restoration work of specific road type (e.g.,

expressway, highway) would be carried out on 24hour system for rapid reopening. Current operation system makes it difficult to estimate the relationship between damage risk and recovery period.

Table 3. The duration of reopening with the past damage level [68]

Damage Level	Time of reconstruction	Damage Level	Time of reconstruction
Minor	4hr	Large	12hr
Medium	6hr	Extra large	24hr

It is important to consider the risk level because the amount of required workload and the period of reconstruction is determined by the degree of road damage. In other words, the worse the road damage, the more operation and longer recovery period are needed. We estimate the risk level of target damage roads by overlapping the road location map and the hazardous areas (e.g., sediment disaster alert areas, inundation depth rank), and determine the workload weight. Ohkubo K et al [68] conducted the analysis on 2,373 disaster cases in the "Record of rainfall disaster history" by Japan Highway Public Corporation from 1993 to 2004.

Table 4. Reconstruction weight with the past damage level

Inundation Depth	Reconstruction Weight	Sediment disaster area	Reconstruction Weight
No damage in the past	1	No damage in the past	1
0.5m ~ 1m	1	Alert areas	1.5
1m ~ 2m	1.5		
2m ~ 5m	2	Special alert area	3
5m over	3		

As Table 3 shown, they identify the relationship between the extends of damage which is categorized by the volume of collapsed soil and the duration of road's reopening. We utilize their standards for setting the workload weight with estimated damage level. Table 4 shows the reconstruction weight. We recalculate the weighted workload of each disrupted road with the consideration of disaster damage risk.

Chapter 5. EXPERIMENT

In this chapter, we carry out the experiments from single agent Deep RL to multi-agents Deep RL in the same framework described in Chapter 3. The goal is that agent find the optimal strategies for achieving human mobility recovery rate over 70%.

5.1 SINGLE AGENT RL AND MULTI AGENT RL

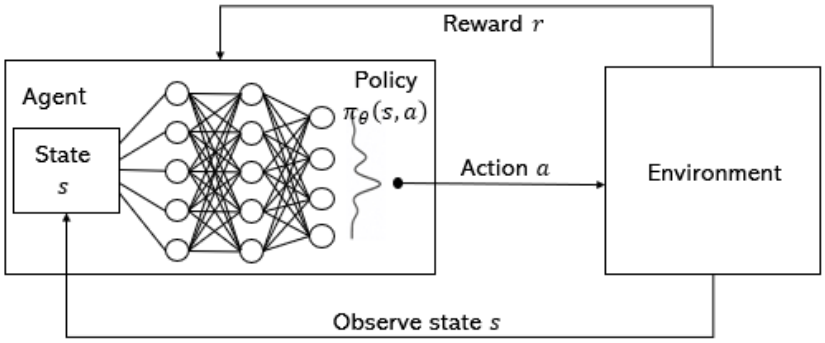


Figure 12. Framework of Deep Reinforcement Learning

Deep Reinforcement learning basically is based on the interaction between the agent and environment. The agent selects the action and then receive the information about the effect of its behavior, corresponding state and a reward informing if it has achieved the objectives. The purpose is to identify action value function maximizing the expected total sum of discount reward. The agent in Deep RL utilize neural network to estimate this value function. With neural network, the representations of state are derived efficiently from high-dimensional input layers.

The environment in single agent RL is assumed to be stationary and the agent's behavior only causes the current state to change the next state. On the other hand, the environment of multi-agent Deep RL is dynamic. As we mentioned in Chapter 3, the information related to other agents is also included in environment. Unlike other things

in environment, the action of other agents at each time step is changed and causes significant variations in each agent's environment. Therefore, the problem in multi agent Deep RL is extremely more complex and quite difficult to identify the optimal policy for the objective.

People in real life usually face many challenges. For achieving their own goal, they build their own strategies through competition, cooperation, and communication with others. In multi agent system, the agents utilize these behavioral strategies for their own optimal policy. There are many previous studies on multi-agent RL system where this phenomenon occurs. Ardi T et al [69] described how competitive and collaborate behavior occurred using the agent's reward setting. Tan M [70] suggested how make the agent cooperative with three methods: 1) sharing sensation, 2) sharing experiences and 3) sharing the parameters of learned policy function. Yang J et al [71] proposed a two-level hierarchical multi-agent RL and made them perform soccer skill with fully cooperativity. Niranjan B et al [72] focused on making agents learn collaboration with specialization and evaluated agents' learning result with four methodology: parameter sharing, concurrent learning, counterfactural method, the utilization of communication protocol. Their result suggested that agents with communication could identify their own policy the best. Furthermore, there have been also many studies on multi-agent RL that is partially competitive and cooperative through communication signals or partial information sharing.

5.2 SINGLE-AGENT BASED DEEP Q-LEANING

5.2.1 *Outline of single-agent DQN*

5.2.1.1 Definition of Action space

Under section 13 of the Road Act, road reconstruction projects shall be conducted by MLIT for designated road sections, and other parts shall be handled by the Prefecture government. The primary road section types managed by MLIT are expressway and highway. In contrast, prefectural administrations manage national and prefectural roads. In addition, MLIT performs recovery project in cooperation with each municipal authority, and the target roads are the disrupted road sections in the corresponding area.

Based on disrupted roads provided by two management entities, we could estimate 95 road sections as the candidates of agent's action. The number of damage roads that should be managed by MLIT or Hiroshima prefecture government is approximately 30 roads in case on Western Japan flooding. Furthermore, the number of disrupted roads in each municipal area in Hiroshima Prefecture is at least 15. In other words, one operation crew (the agent) should be able to identify the optimal policy with at least 15 actions. We choose 15 road sections for single agent RL. Figure 13 describes the details damage roads subjected single agent's targets (actions).

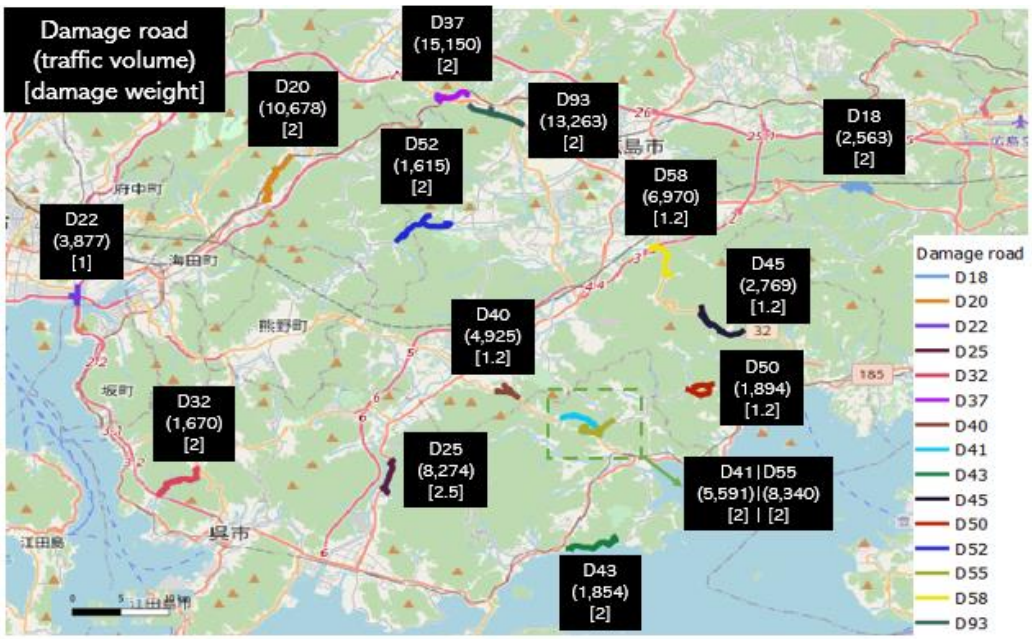


Figure 13. Damage roads subjected agent's action with traffic volume

5.2.1.1 Reward setting

The reward at each time step plays a vital role in finding the optimal strategies with the agent's objective. When the agent achieves its own goal, it is necessary to give the agent a large reward as the signal of the goal compared to the basic reward. With the basic reward setting (Figure 14), we provide the agent the greatest reward, +100, when human mobility recovery rate is over 70%.

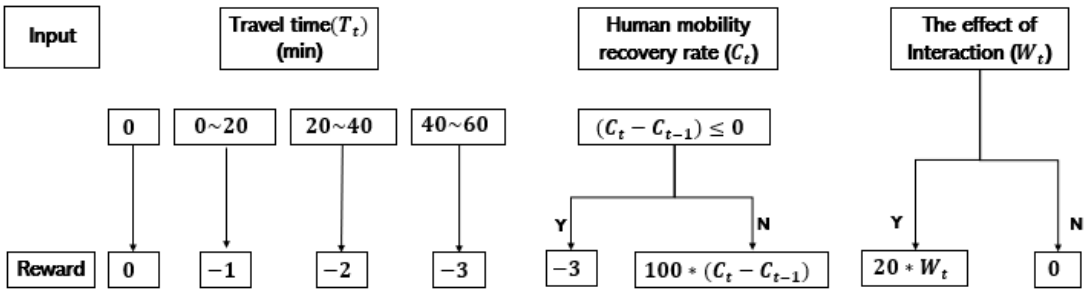


Figure 14. Reward setting in single agent RL

There are two primary rules that the agent needs to learn. First, the agent should learn which road sections have a significant impact on human mobility recovery and

tend to prioritize highly influential road. Second, we basically want the agent to move to another damage road after a certain amount of work is completed at the current place. In addition, the agent is recommended to go to other place with better accessibility or connectivity. We utilize the elements with the values of variables and constants for reward settings.

We want the agent to learn which roads have the impact on human mobility recovery while choosing its own action, not giving prior knowledge. Accordingly, the change of human mobility recovery rate is utilized as the major reward factor and convert the percentage to integer form. Furthermore, we hope that when the agent chooses its behavior, it prioritizes the connectivity and recovery effects over moving cost. We set the score for the longest travel time group as the negative number of the largest value of the connectivity to ensure that the reward of the connectivity offset the reward related to travel time.

5.2.1.2 Definition of State space

The agent utilizes the current states, S_t, when choosing an action at Step t. It is strongly advised to include factors having a relationship with the agent's objective directly or indirectly into the state space. The state space in this model is as follows:

$$S_t = \{W_t^{r_1}, \cdots, W_t^{r_n}, TR_t^{r_1}, \cdots, TR_t^{r_n}, M_t, RR_t, a_{t-1}, I_t\} \quad (12)$$

The agent's fundamental goal is to restore traffic volume to a certain level as fast as possible. There are three elements related to recovery goal. $TR_t^{r_n}$, RR_t and $W_t^{r_n}$ represent each damage road (r_n) 's recovery rate, human mobility recovery rate and

the cumulative progress rate of the kth action at Step t respectively. RR_t notifies the road crew of how far human movement recover overall and implies the target value of the agent's goal. $TR_t^{r_n}$ refers to the impact of $W_t^{r_n}$ on recovery rate. That is, two elements are expected to change in a similar direction at each time step. We want the agent to identify which action has the great recovery effect with the aspects of changes in these two values.

We add three additional factors to convey the considerations of sequent action choices. M_t is the travel time between the starting point and the current workplace. We have calculated the shortest travel time between these two points in advance. a_{t-1} and I_t represent the action at step $t-1$, the connection between the action at $t-1$ and at t. These three elements describe the operation cost and the relationship among damage roads subjected to the agent's action.

5.2.2 *Result*

5.2.2.1 The change of human mobility recovery rate and travel time with learning trend

Figure 15 presents the final human mobility recovery rate of total episode. The title of each graph means the starting point of the agents. With these graphs, we could grasp the performance of the model and the agent's learning ability. We could confirm that the mobility recovery rate when the agent of the model finds the optimal policy is approximately 25% better on average than the lowest mobility recovery rate when working randomly.

The agent with the optimal policy stably could reach its own target value in the latter part of the learning. In the case of our model, we could confirm whether the agent could find out the optimal policy or not with the final human mobility recovery rate. We could identify that the agent with any starting point succeeds in finding out the optimal policy. Because the final recovery rate of each epoch in later part would be equal to or greater than the simulation goal.

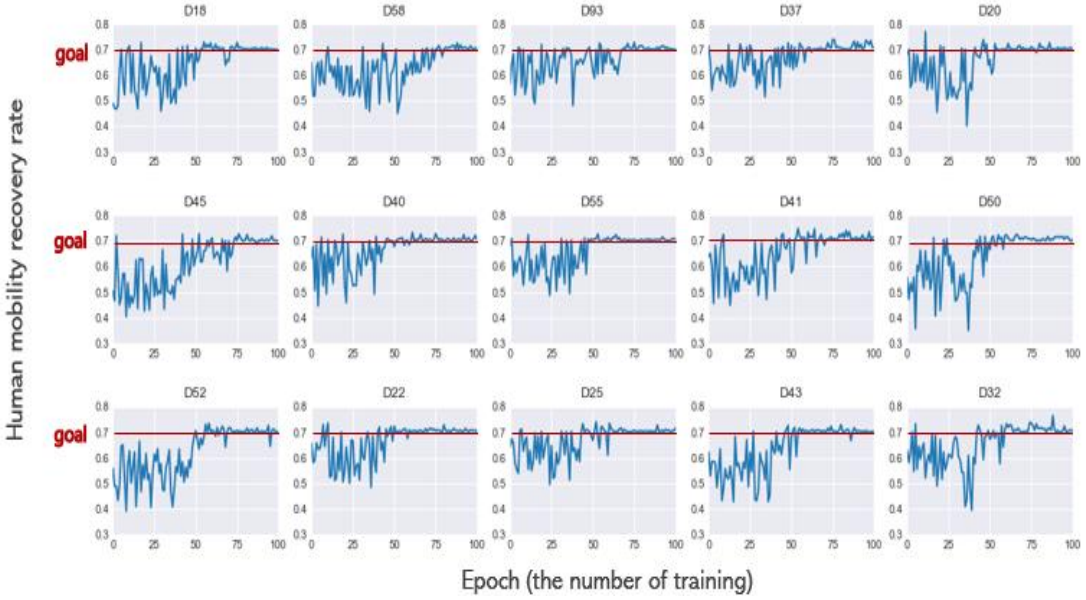


Figure 15. the relationship of recovery rate and learning trend

As we mention in 3.2.3, we want that the agent considers accessibility with the selection of actions. Figure 16 describes the change in total travel time with learning process. After learning, the total travel time is decreased. In other words, the agent might tend to work continuously in one damage road rather than moving frequently around the workplace. This is because the agent in RL choose behaviors that could achieve the high sum of rewards, and frequent transition makes the total rewards smaller.

Figure 16. the change in total travel time with learning

5.2.2.2 The Visualization of the Change of Traffic volume with Reconstruction

The road crew with good learning result is anticipated to prioritize road section restoration with high effect of human mobility recovery effect. In this sub-chapter, we want to check the priority order with the visualization of the change of traffic volume of each road-link. Figure 17 represents the change of traffic volume under reconstruction process. The agent first recovers the disrupted roads near the urban district. These road sections located near this district have larger traffic volume on normal days than other roads (Figure 13). Therefore, we could identify that the agent recognized which damage road has higher impact of human mobility recovery and established the strategy to restore the road with large traffic volume preferentially.

We could confirm O-Ds' behavior according to the agent's operation. Some road sections often have the traffic exceeding the maximum expected traffic. O-Ds seemed to use generally alternative routes rather than original routes that they usually pass on normal days. In fact, traffic volume on detour route increased more than five times, and travel time increased more than 1.8 time as usual after Western Japan

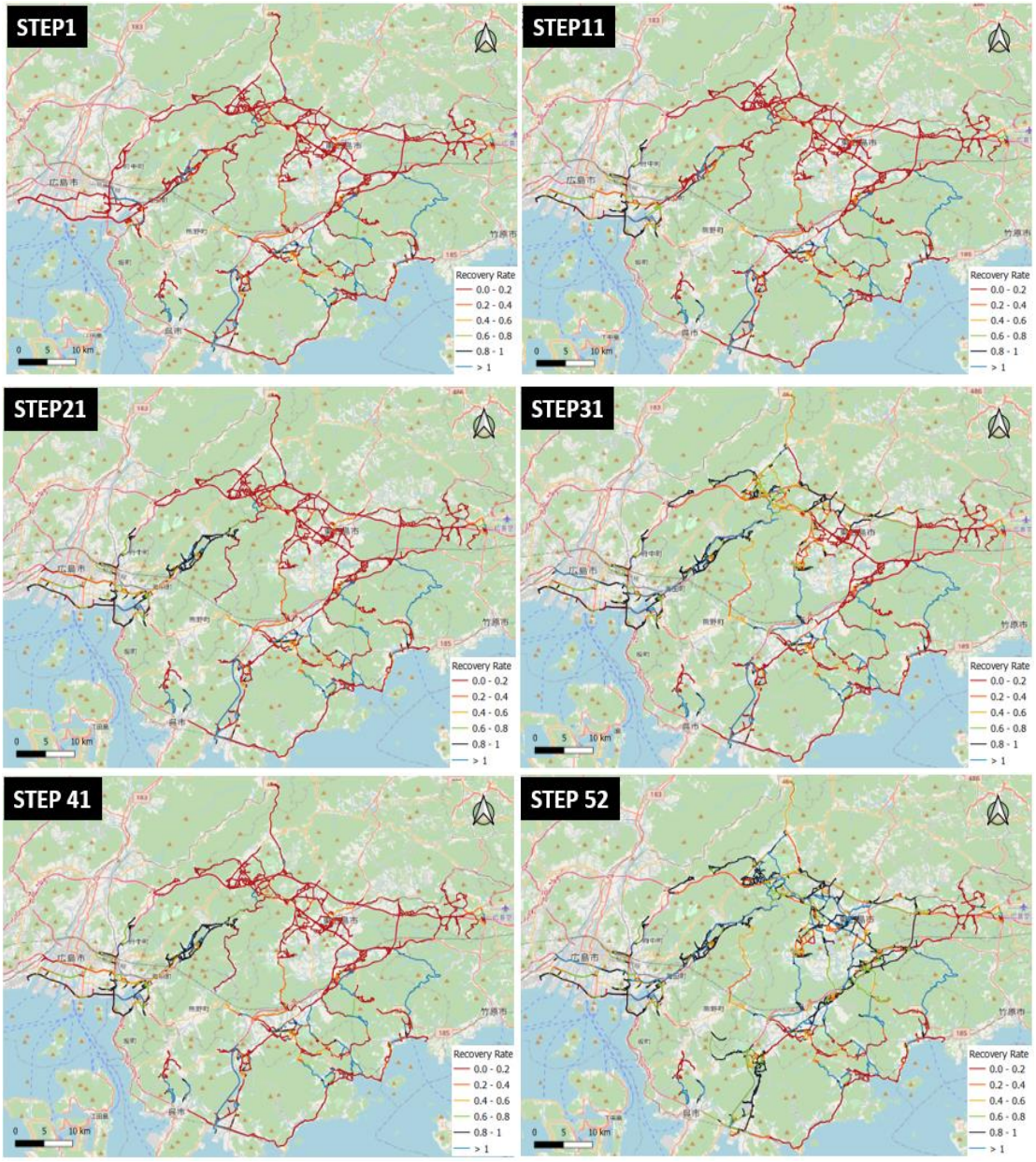


Figure 17. the change of road usage with the sequence of reconstruction operation

flooding occurred. So, we could conclude that O-Ds in our model might be the representatives of real citizen. It is possible for administration to check the change of road usage with the sequence of reconstruction operation with the result of traffic volume's visualization.

5.2.3 *The analysis of the agent's learning framework*

The agents in RL found the optimal policy, automatically updating their own action value function based on sample data which is obtained through interaction with the environment. However, they only mention the answer to the optimization problem, but do not explain why this answer come out. Inexplicable RL causes the users to question the reliability of the model's result and is the obstacle to the adoption of RL.

5.2.3.1 Analysis of influencing factors of the agent's learning

We tried to identify the relationship between the input data and the recovery operation order which is the result of this model. The input layers in this model mainly consist of three factors: 1) The change of human mobility recovery rate with the change of each road's progress rate, 2) travel time, and 3) the estimated traffic volume at each time step. For analysis of influencing factors, we extracted general priority order of each episode using the operation order and the selection frequency of each action. Furthermore, we defined the representative of these input factors. This is because each value that constitutes the input layer varies with the agent's action and environment at each time step.

Human mobility recovery rate is determined by these two factors: 1) the traffic volume of disrupted road selected by the agent at each time step and 2) the change of

operational progress rate of current action. We define the recovery effect of each damaged road as representative indicator for human mobility recovery rate and estimate this effect value of each damage road using the change of human mobility recovery rate and the change of operational progress rate at each time step. Recovery effect of each disrupted road refers to how much the reconstruction of each road affects human mobility recovery. Recovery effect of road x is calculated with this equation:

$$RE(x) = E\left[\frac{\Delta\ human\ mobility\ recovery\ rate}{\Delta\ work\ progress\ rate}\right] \quad (13)$$

Traffic volume which is the other input factor is estimated under updated transport network with the agent's action at each time step. The process of traffic allocation is a subset of network analysis, exploring the relationship between nodes and links. Accordingly, we could expect that the agent could identify the characteristics of each damage road and inter-connectivity using the estimated traffic volume. We utilize the centrality index for representing the feature of each damage road. In network analysis, indicators of centrality identify the importance of vertices within graph. We focus on two types of centrality: Betweenness centrality, Closeness centrality. The detail of two indicators is as follows:

- Betweenness Centrality is the number of these shortest paths that pass through the vertex. High betweenness vertices have the potential to disconnect graphs if these vertices are removed

- Closeness Centrality is the sum of the length of the shortest paths between the nodes and all other nodes. The high closeness centrality, the more central the road is to the road network.

Heat map is the data visualization technique that presents how specific phenomenon is clustered or varies using the intensity of color in two dimensions. We select this visualization technique to identify how the feature of the damage roads chosen by the agent change with the learning trend.

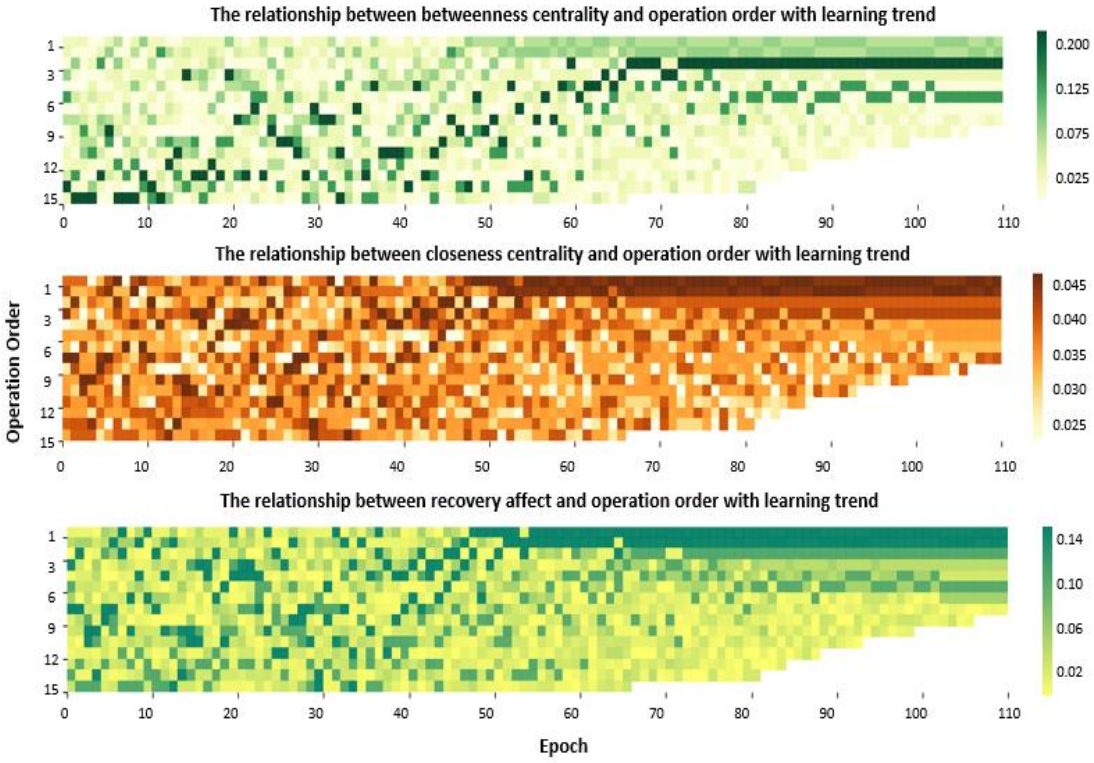


Figure 18. The relationship between road factors and operation order with learning trend

We could confirm that the agent has the tendency to select disrupted roads with central roles in transportation networks on the preferential basis. In addition, the agent could recognize which damage road have high effect on human mobility recovery through the learning process. In summary, the reward of the change of human mobility recovery rate makes the agent know the impact of each road on achieving its goal. With

the use of network analysis's result as input data, it is possible that the agent could identify the connectivity between roads and the importance of those roads within road network. In other words, the agent in our model might make the decision about reconstruction with the consideration of the characteristic of each damage roads considered in previous studies.

5.2.3.2 Sensitivity Analysis

The agent determined the action value function utilizing the accumulated information of the state, the action and immediate feedback, which is the result of the interaction with environment. In other words, the reward refers to the value of each action for specific situation information. The appropriate reward setting helps the agent derive the optimal policy for its objective, but if it goes wrong, the agent fails to determine its own policy. Therefore, it is necessary to find out how changes in reward setting affect the learning of the agent. In original reward setting (Figure 14), there are three components: 1) travel time from current workplace to specific starting point, 2) the change of human mobility recovery rate, and 3) the degree of inter-connectivity between current workplace at time step t and the past workplace at time step $t-1$. More specially:

- The reward of travel time has negative values from 0 to -3 and is divided 20 minutes intervals.
- The change of human mobility recovery rate and the degree of inter-connectivity are represented as the percentage. To convert this value to integer value, we defied the integer conversion weights.

The integer conversion weights for human mobility recovery and connectivity which is expressed as the percentage are 100 and 20, respectively. Table 5 described the detail of sensitivity analysis. Each integer weight related to the above two factors fluctuates between -50% ~ 50%. For the reward of travel time with negative values, either double the reward of each group (20 minutes intervals) or subdivide groups of 20 minutes intervals into groups of 10 minutes intervals.

Table 5. The detail of sensitivity analysis

	Travel time (min)	Reward of travel time	The weight of recovery rate	The weight of road-connection
Case 1	1 ~ 20	**-2**	100	20
	20 ~ 40	**-4**		
	40 ~ 60	**-6**		
Case 2	1 ~ 10	**-1**	100	20
	10 ~ 20	**-2**		
	20 ~ 30	**-3**		
	30 ~ 40	**-4**		
	40 ~ 50	**-5**		
	50 ~ 60	**-6**		
Case3	1 ~ 20	-1	**50, 75, 100, 125, 150**	20
	20 ~ 40	-2		
Case 4	40 ~ 60	-3	100	**10, 15, 20, 25, 30**

To confirm the change in the learning results with the change in reward, we estimated the average human mobility recovery rate, success rate, and number of steps for achieving its goal using the result of 70^{th} ~ 110^{th} episode. This is because the agent selects the action at each time step based on action value functions during these episodes. Figure 19 present the result of sensitivity result. With the change of weight about human mobility recovery rate and the connectivity, the agent could determine optimal policy overall. The higher weight of two factors makes the agent reach the goal

faster than the original form. In other words, it is easier for the agent to identify behaviors that have a great influence on achieving its own goal.

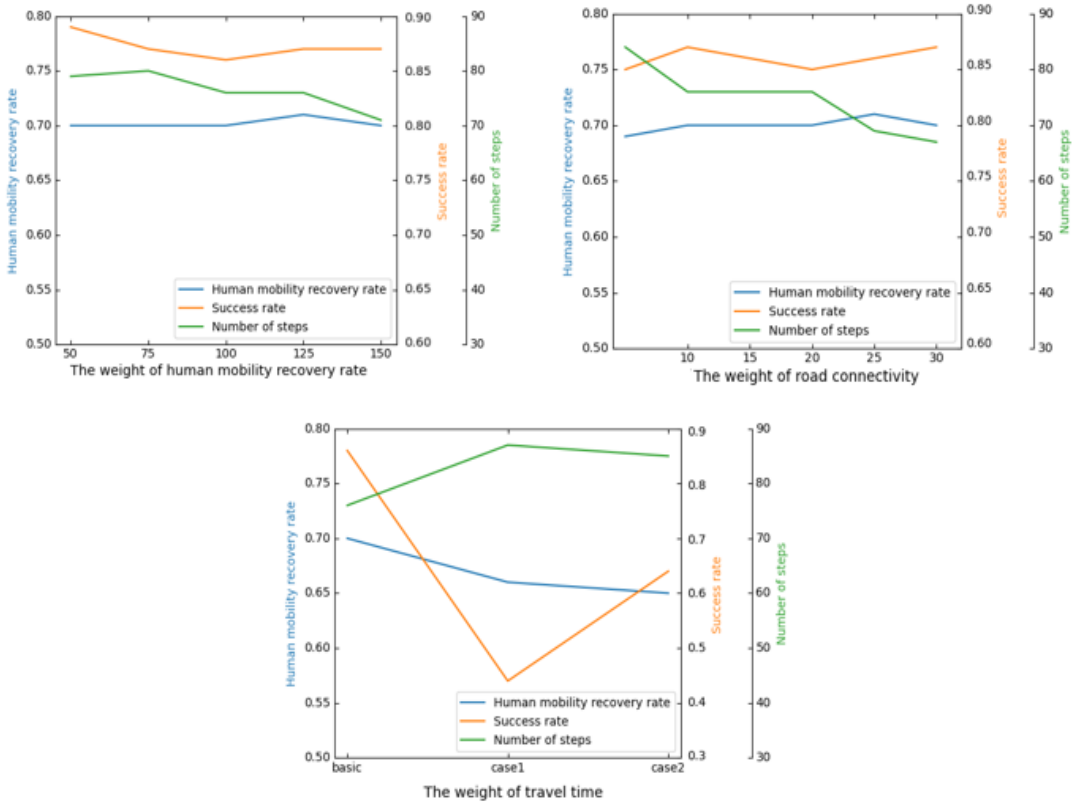


Figure 19. The result of sensitivity analysis of the change of reward setting

The agent fails to find out the optimal policy with the change of travel time reward. This is because the impact of reward related to road connectivity and mobility recovery rate decreases as the negative reward increases. This means that the agent receives negative reward for action which has a significant impact on mobility recovery, and derived policies that are far from achieving goals. Therefore, the agents select the disrupted roads located near the starting point to maximize the expected total sum of discount rewards.

We suggest the reward setting that make the agent to learn and converge optimal policy. First, the agent could quickly converge on policies as rewards related to goal attainment are transformed to larger than existing ones. Even with the basic reward setting, the agent finds out optimal policy sufficiently. Second, penalties could be used to constrain the agent's action choices, leading to the consideration of certain factors such as travel time. However, it more than -3 penalties are imposed, it prevents the agent from learning optimal policy.

5.2.4 *Comparative Analysis with Present Method*

Comparative analysis should be done to ensure that our model's result works and be more effective than that previous method. We choose two present method, travelling salesman problem (TSP) algorithm and government's method, and do comparative analysis with the result of these two methods. There are two assumptions for comparison:

- If the agent in two present methods select one damage road, the recovery operation is keep going until operation progress rate will be 100%.
- We define one selection equal to the number of steps required until the agent in our model finished work on the selected damage road.

5.2.4.1 Comparison with Traveling Salesman Problem

TSP is the classical example of a NP-hard combinatorial optimization problem [74]. Many scheduling problems could be reduced to simple concept that there is a salesman who must travel from city to city, visiting each city once and returning to the home city [75]. Various heuristics and approximation algorithm are used for solving

the optimization problem, the shortest travel time. Accordingly, we assumed that the agent in TSP algorithm only consider accessibility for their own operation order and estimate the agent's operation trajectories with given starting point by using genetic algorithm.

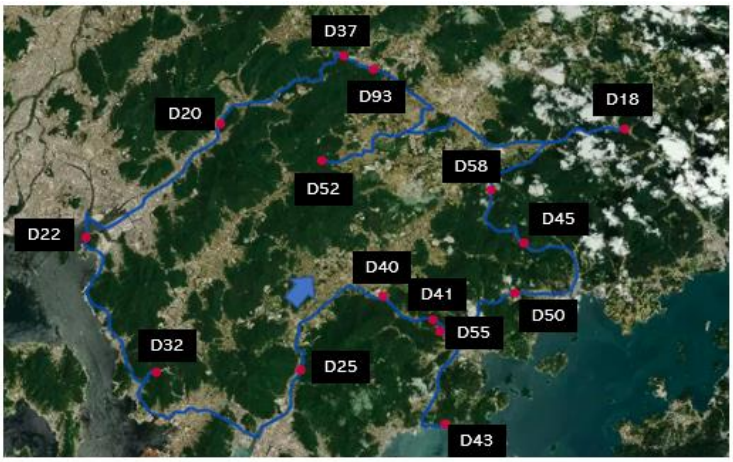


Figure 20. Agent's trajectory in TSP (start point: D25)

Figure 21 presents the change in human mobility recover rate at each time step. The x-axis indicates the step and the y-axis is human mobility recovery rate. The range

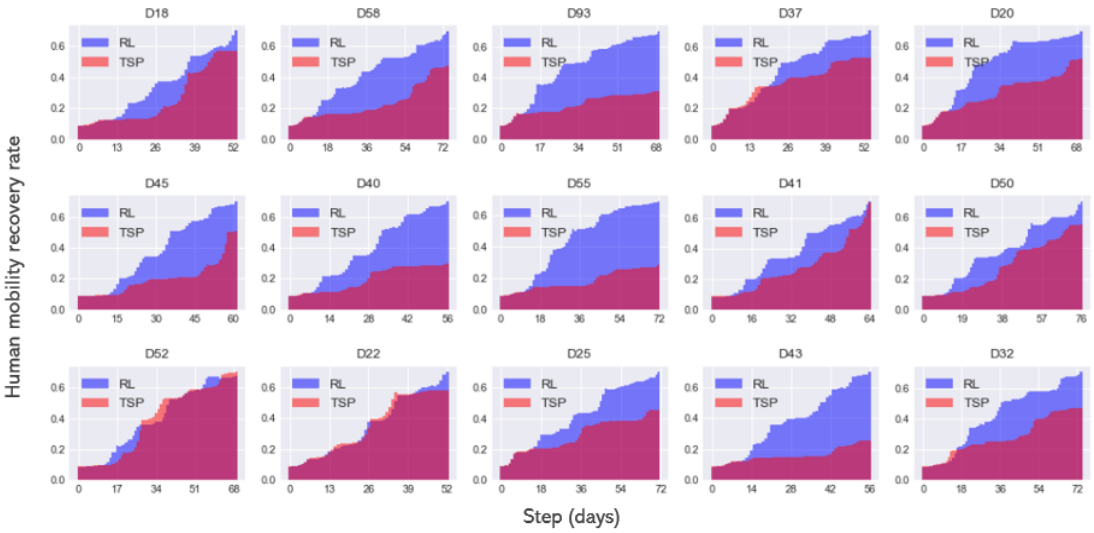


Figure 21. The comparison result between RL and TSP

of x-axis is the number of steps it takes the agent in this model to achieve the recovery rate of more than 70%.

The difference in the result of two models, RL and TSP, varies depending on the given starting point. But the operation order in this model could generally recover human mobility faster than TSP algorithm. The agent in this model considers traffic volume with travel time. In other words, the agent selects other damage road with high effect of human mobility recovery even if travel time is a little longer.

5.2.4.2 Comparison with Government's Standard

As we mentioned, officials utilize scoring method for setting the priority order with three classification and reconstruct the high-scoring road first. These three elements are sorted in order of importance as follows: 1) Hazard risk, 2) road importance and 3) stability. The group of hazard risk includes inspection score, degree of damage, and progress of displacement. These features affect the possibility of secondary damage and the long duration of reconstruction, so authorities evaluate is as the most important thing.

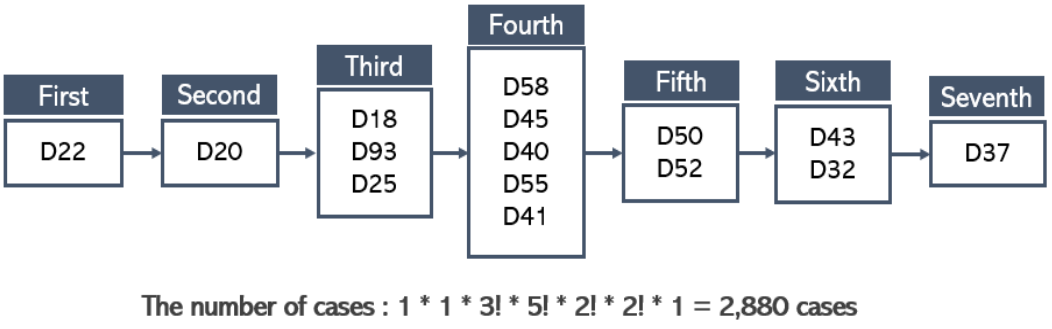


Figure 22. The result of setting priority group

It is difficult for us to estimate hazard risk. So, we assumed that the hazard risk of all damage roads is the same and only consider road importance which include traffic volume, road classification for the priority order. Figure 22 shows the result of setting priority group. The number of cases in Greedy algorithm is 2,880. Among them, we select one case with the shortest travel time. The comparison result is as follows:

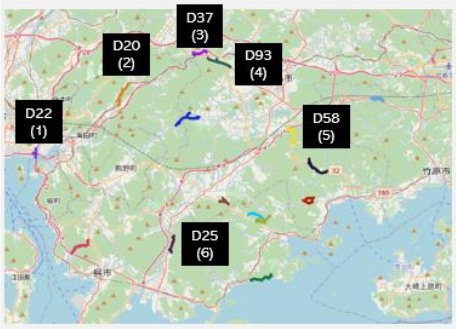
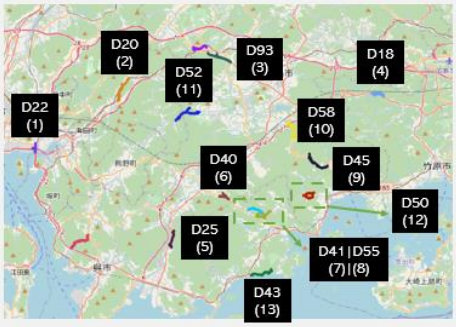
	Deep Reinforcement Learning	Government's standard
Starting point	D22 (Road classification : Urban expressway)	
steps (days)	53steps (days)	116steps (days)
Operation order		

Figure 23. The comparison result between RL and government standard

Government think that intercity connection is more important than inner-city. So, high-level road which is connected to other city has the high priority. On the other hand, GPS data we used is based on human mobility generated from Hiroshima Prefecture. Actually, the reconstruction of neighborhood roads was delayed for a week after Western Japan flooding. We think that the difference between results in our model and government's standard represent the above real situation.

5.2.5 *Modification of basic model with time-periodic objective*

As we mentioned in Chapter 2, the purpose of road-network's reconstruction and road users depends on recovery operation's stage. The current model in this paper would be judged to find out the optimal policy of the recovery operation. However, the

effective reconstruction plan needs to respond flexibly to each phase's objective. Therefore, it is necessary to make the agent determine optimal strategies with the consideration of time-periodic goals. We define the initial and mid-to long-term objectives: 1) Initial objective is to secure at least one lane of the damage roads which O-Ds with no alternative route pass through, and 2) mid-to long-term goal is to make human mobility recovery rate be over 70%.

5.2.5.1 Outline of modified single-agent DQN

The agent's information and all disrupted road sections are the same as the model settings in Chapter 5.1. We determine the damaged roads which needed the initial recovery operation using the path analysis with two types of road-network: normal days and immediately after disaster. As Figure 24 shown, O-Ds which pass through the disrupted roads in the orange box do not have alternative routes. This means that these O-Ds are expected to be isolated in the event of disaster.

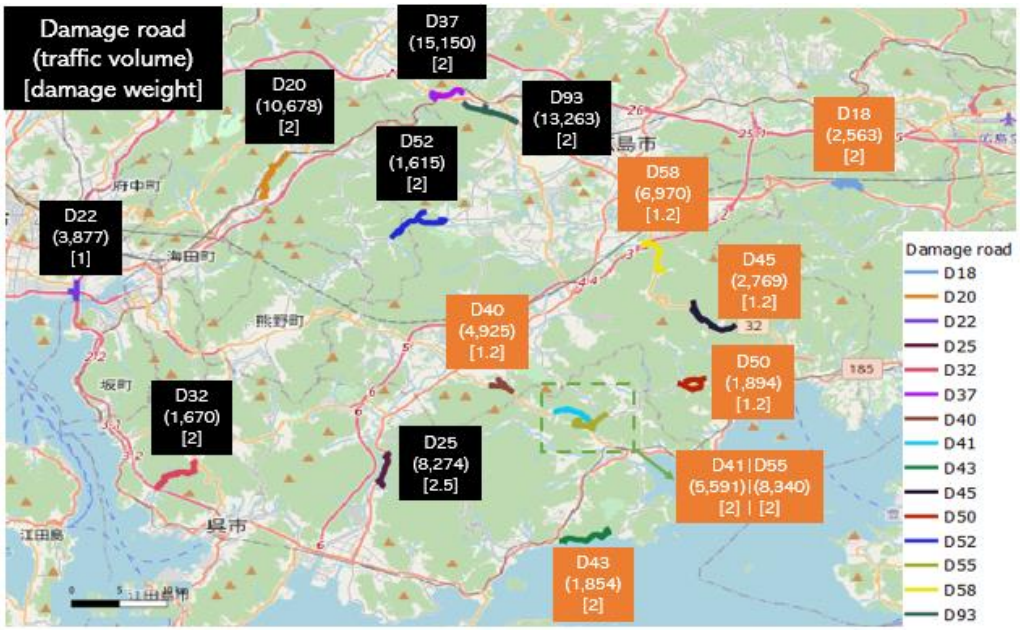


Figure 24. The disrupted roads in model with time-periodic goals

The agent in modified model should do reconstruction operation for achieving human mobility recovery rate above 70% after securing one lane of all disrupted roads requiring the contingency measurement. We modify the reward setting so that the agent could recognize the actions to be taken according to time-periodic objectives. Figure 25 depicts the modified reward setting for agent's action at each time step. The agent could learn behavior's difference in particular state with reward and punishment system.

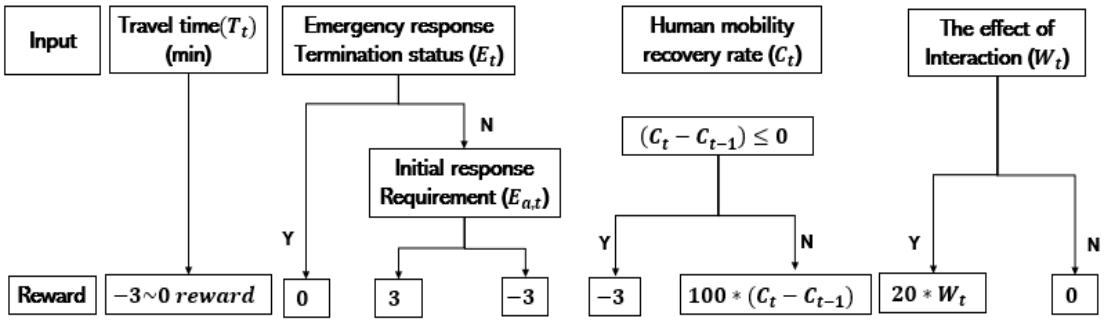


Figure 25. The reward setting with time periodic goal

In this model, the agent should recognize which roads required emergency recovery to achieve both goals that vary over the period of reconstruction. The agent would be penalized when choosing other damage roads, not the target places before the emergency relief process is over. This punishment allows the agent to identify the timing of selecting each damage road. In recovery operation process, the reward for the agent's behavior follows the basic reward setting. This value could directly inform how much each damage road effects human mobility recover and make the agent recognize effective action to mid-to long-term objectives.

5.2.5.2 Learning Result

We checked the final human mobility recovery rate of each episode and the number of required steps until the end of the emergency relief phase to evaluate the agent's learning results. Figure 26 presents the final human mobility recovery rate of all episodes with pre-determined starting point. We could confirm that the agent

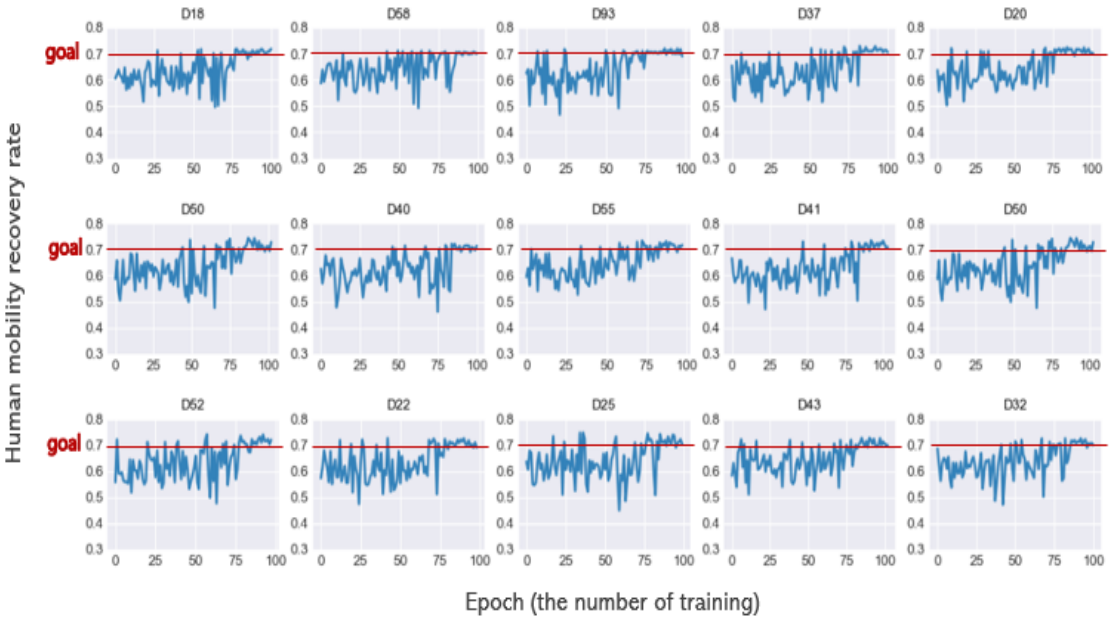


Figure 26. The human mobility recovery rate with learning trend

reliably achieves mid-to long-term goals because they achieve stably the recovery rate over 70% after learning.

The minimum number of steps required by the end of emergency response in this model is 38 steps. We think that the agent first selects damage roads in this group if the agent well recognizes which disrupted roads require initial response. In other words, the required number of steps for finishing the emergency operation stage would be close to the minimum number of steps.

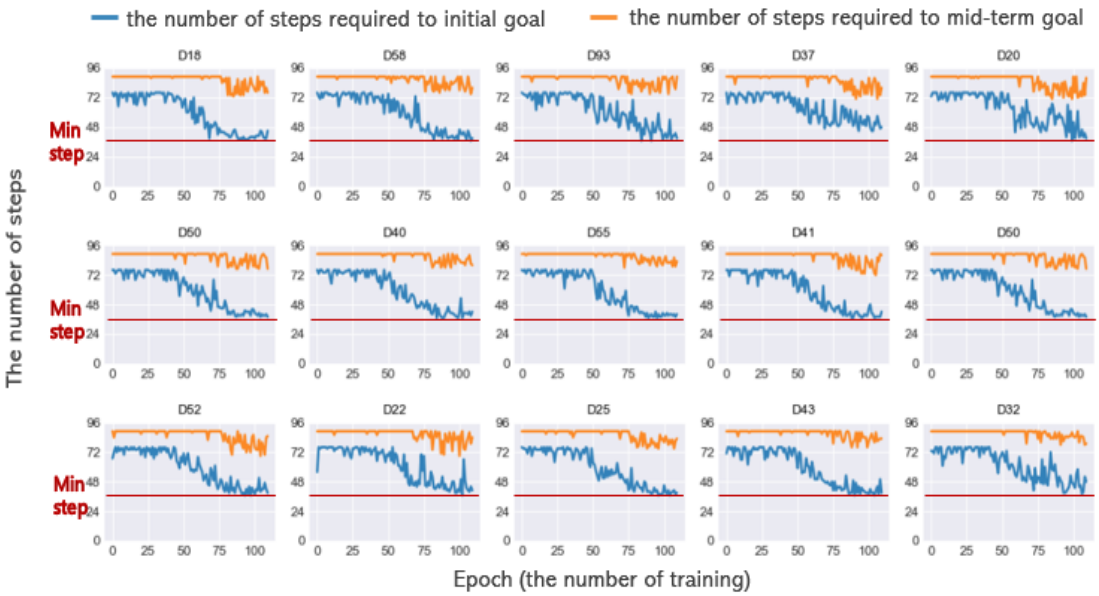


Figure 27. The required number of steps for initial and mid-term goals

As Figure 27 shown, we could identify that the number of required steps for the initial objective gradually decreases and approaches the minimum level as the learning progresses. With the comparison between Figure 17 and Figure 28, O-Ds seem to recover their own mobility from the bottom part where the damage roads needed for contingency operation are located. After the emergency operation might be completed which is after 45 steps, the agent moved to the disrupted roads with heavy traffic on normal days. This is because the agent might be induced to recover from damage that requires initial work preferentially. We could conclude that the agent considers time-

periodic objectives and determines the optimal policy for achieving both goals concurrently with modified reward setting.

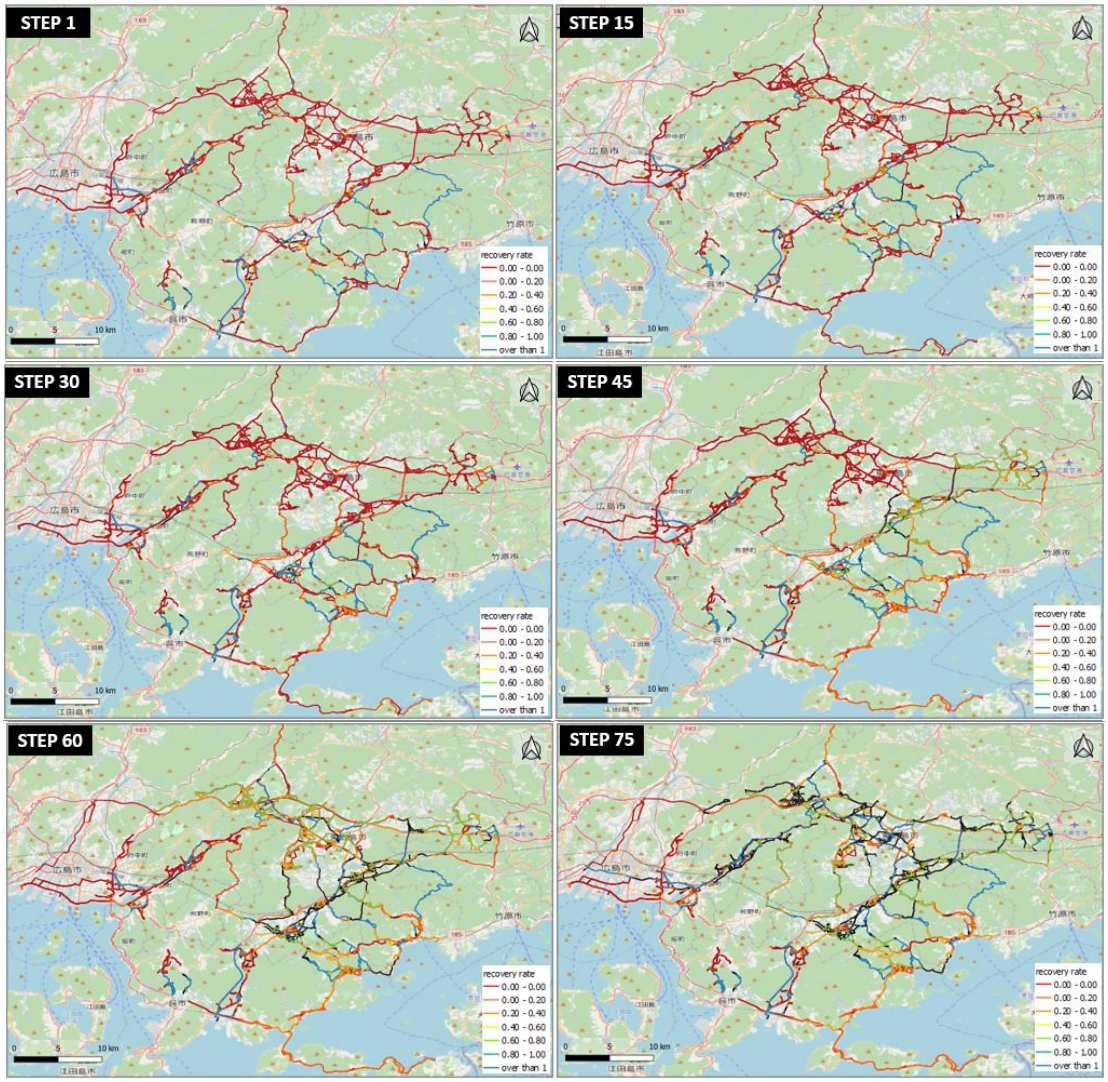


Figure 28. the change of traffic volume with time-periodic objectives

5.3 MULTI-AGENTS BASED DEEP Q-LEARNING

After the flooding, several operation crews would be deployed and cooperate with others depending on government's strategies. From the decision-maker's perspective, we might designate the government as the centralized controller and make this controller map states of all operation crews to a collection of all crews' actions [76]. With this framework, a strong partnership could be established between all operators involved in the reconstruction operation. However, this method might be impractical with the collaborating of many agents. This is because the central manager deals with all crews' state and action. There are exponential increase of state space and action space. The centralized agent with huge action and state space has a hard time converging its own optimal policy.

We suggested multi agent RL system using decentralized method. The agent in this method resides in the same environment with other agents and identify its own policy with its own observation. In addition, we could place the responsibility on each operation crew and make it learn the cooperation or the collaboration with others using communication protocol, partially information. Accordingly, we applied the method suggested by Jakob N.F et al [77] to make multi-agents find out their own optimal policy for shared objective. The agent in their system is partial cooperative and share important information using specific protocols. It is possible to treat others the part of environment with partial observed state. They could achieve the cooperation through their own network different from others [77].

5.3.1 *Outline of multi-agent Deep Q-learning*

5.3.1.1 Definition of Agent and Action space

We defined three types of operation crew with 15 damaged roads which are included in each type crew's action space. There are 45 disrupted roads that could be considered simultaneously with multi-agent RL system. Figure 29 describes the target

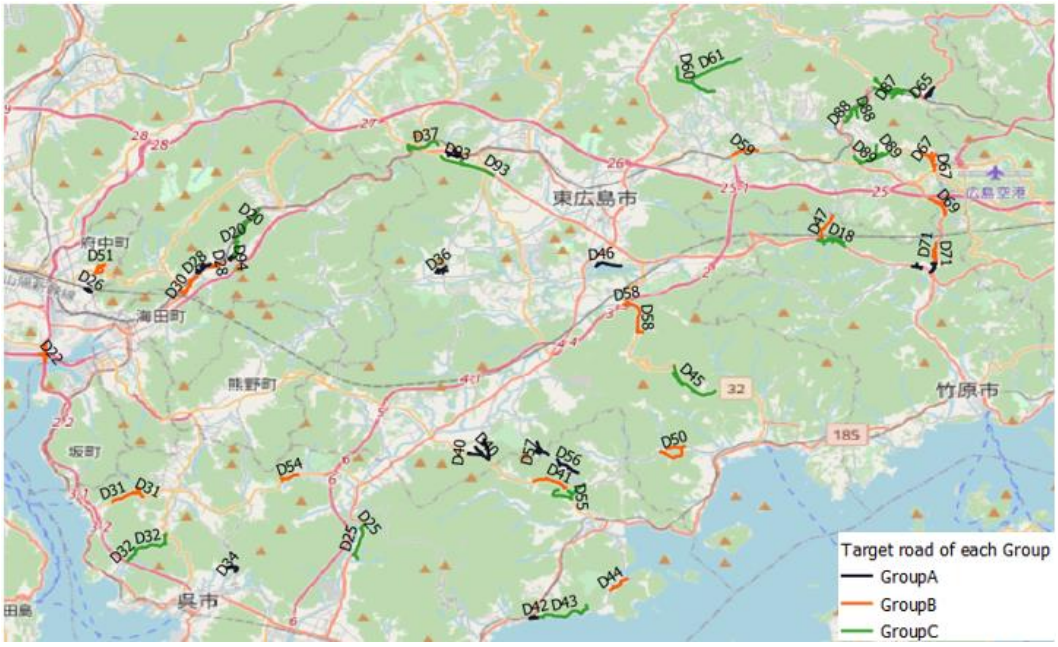


Figure 29. The damage road with multi agent system

damage roads subjected to each group's action. In addition, we defined that the number of workers in each group is four, eight, and seventeen workers respectively based on the workload considering the risk level.

5.3.1.2 The definition of cooperation and state space

As we mentioned in 3.3.2.2, the recovered traffic volume of each O-D is determined by the minimum operational progress rate of the damage roads they pass on normal days. In other words, all relevant roads should be restored in a certain amount to recover the traffic of O-Ds passing through multiple damage roads. We divided O-

Ds related each disrupted road into two classes: 1) O-Ds passing through only one damage road and 2) O-Ds passing through multiple damage roads.

Table 6. Traffic on each disrupted road on normal days

Group A			Group B			Group C		
Name	A*	B**	Name	A*	B**	Name	A*	B**
D21	7,652	8,572	D58	1,697	538	D18	0	1,780
D46	4,012	77	D59	5,316	1,542	D88	956	3,270
D36	2,078	427	D67	74	388	D89	3,506	2,206
D39	603	1,417	D50	380	622	D86	162	2,965
D40	115	2,336	D69	323	1,000	D87	56	1,962
D57	705	616	D51	1,230	260	D45	820	538
D56	56	951	D22	634	0	D60	339	328
D26	6,423	260	D30	76	514	D55	1,160	2,992
D28	256	1,152	D54	3,350	515	D25	5,843	592
D34	1,128	0	D31	465	0	D93	1,696	4,734
D42	481	468	D44	679	0	D37	1,715	6,050
D65	0	654	D41	54	2,736	D20	3,441	3,017
D93	0	3,017	D47	0	780	D32	999	0
D72	0	1,000	D71	0	1,000	D43	143	468
D90	0	1,000	D95	0	2,944	D61	300	0

A*: Traffic volume which pass through only this damage road

B**: Traffic volume which pass through other disrupted roads other than the road

The traffic volume of O-Ds over multiple disrupted roads accounts for approximately 52% of the total traffic considered in this model. These agents in this system need to select and recover disrupted roads with connectivity based on the usage of road-users at the same time step for their shared objective. For example, we assumed that some O-Ds passed through D21 and D36 concurrently. If one agent restores D21

and another agent restores D36 at same time step, there are three types of traffic that could be recovered with the operation of both agents: 1) O-Ds passing only D21, 2) O-Ds passing only D36, and 3) O-Ds passing both of D21 and D36. Accordingly, we could conclude that the concurrent restoration of roads with connectivity could help recover human mobility rapidly. We define cooperative behavior in this model to solve their own challenge. The meaning of cooperative behavior defined in this paper is as follows:

- **Definition 1.** The cooperation means that two or more agents choose each target road that has connectivity at same time step.

It is necessary to provide communication protocol to the agents for coordinating their action and achieving the shared objective. Communication protocol refers to numerical message related to other agents' action. These protocols are provided as the input layers on the next time step. The agents each could discretize and identify the cooperation with another agent through learning process.

We defined the information about the collaboration with specific another agent utilizing the traffic flow. Let denoted by $\mathcal{T}_c^e$ traffic volume concurrently passing through 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 15:

$$CE_t^{\mathcal{A}_o} = \frac{\mathcal{T}_c^e}{\sum_{d \in \mathcal{R}, g \in \mathcal{R}} \mathcal{T}_d^g}, c \in \mathcal{R}, e \in \mathcal{R} \tag{15}$$

where $CE_t^{\mathcal{A}_o}$ means the effect of cooperation at step t assuming that agent $\mathcal{A}$ selects damage road c and the agent $\mathcal{A}_0$ selects damage road e.

As we described in 5.1.2.2, the basic state space consists of four factors: 1) the operation progress rate of each damage road ($W_t^{r_n}$) in action space, 2) human mobility recovery rate ($TR_t^{r_n}$) of each damage section, 3) travel time (M_t) and 4) the average of human recovery rate (RR_t). With these basic components, we add the effects of cooperation with other agents ($CE_t^{\mathcal{A}_o}$) and impact of selected damage road (α_{r_t}). Each agent could deal with other agents' action as the one element of environment adding the cooperation protocol in the state space and identify its own policy with the consideration of collaboration and shared goal.

$$S_t = \left\{W_t^{r_1}, \cdots, W_t^{r_n}, TR_t^{r_1}, \cdots, TR_t^{r_n}, M_t, RR_t, CE_t^{\mathcal{A}_1}, \cdots, CE_t^{\mathcal{A}_m}, \alpha_{r_t}\right\} \qquad (14)$$

With partially observation and corresponding reward, the agents each determine Q-function (action value function) having different parameters. This method makes the agents each select different kind of actions and do its own responsibility. In other words, the agent recognizes which damage road among its own operation places has the cooperative relationship or is efficient for common goal in their respective ways.

5.3.1.3 Reward

The common objective in multi agent RL system is to recover human mobility up to 75% within the pre-determined steps (35 steps). These agents with shared goal often receive the same global reward regardless of the effect of their own action on the shared goal. With the same global reward, some agents choose behaviors that help their

goals, while other agents get lazy selecting weak effect actions [78]. And then, lazy agents might interfere with achieving their goal.

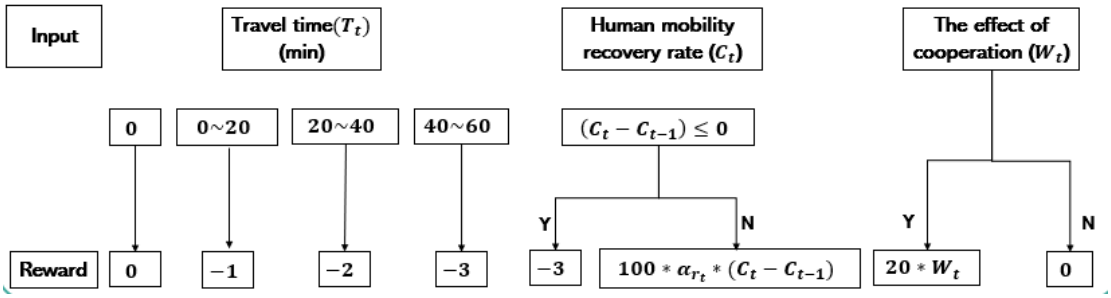


Figure 30. reward setting in multi-agent RL system

We utilized the change of human mobility recovery rate as the one factor of reward. Unlike the single agent model, the human mobility recovery rate of each time step derived through a multi-agent system is the result of the mixture of effects of multiple actions. So, it is difficult for each agent to identify the impact of recovery on individual damage roads. As a result of giving the reward of the change of human mobility recovery rate derived at each time step as a global reward, some agents have become lazy and an adverse effect on achieving shared goal. We utilize the traffic weight (α_{r_t}) indicating how much disrupted road (r_t) they selected 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.

It is also necessary to map communication protocols with other agents into the state space as well as provide appropriate reward to accurately interpret and act on cooperative behavior. As Figure 30 shown, we convert the sum of cooperation effect (W_t) to integer reward. In other words, the extra positive reward is arising when the agent chooses the cooperative action. We could expect that the agent might have the tendency to select disrupted roads with much relationship of other roads for

maximizing the sum of discount reward. And then, the traffic volume of O-Ds which passed through several damage roads could be recovered in a shorter time than single agent model. These policies help the agents achieve their common goals.

5.3.2 *Learning Result*

We graph the sum of reward each agent received at each episode to verify the learning result of each agent. We could identify that all agents obtained higher reward as the learning process progresses. With learning process, human mobility recovery rate would be over 75% stably and the required number of steps for achieving their goal has been decreased. Therefore, we could conclude that all agents could determine their own optimal policy.

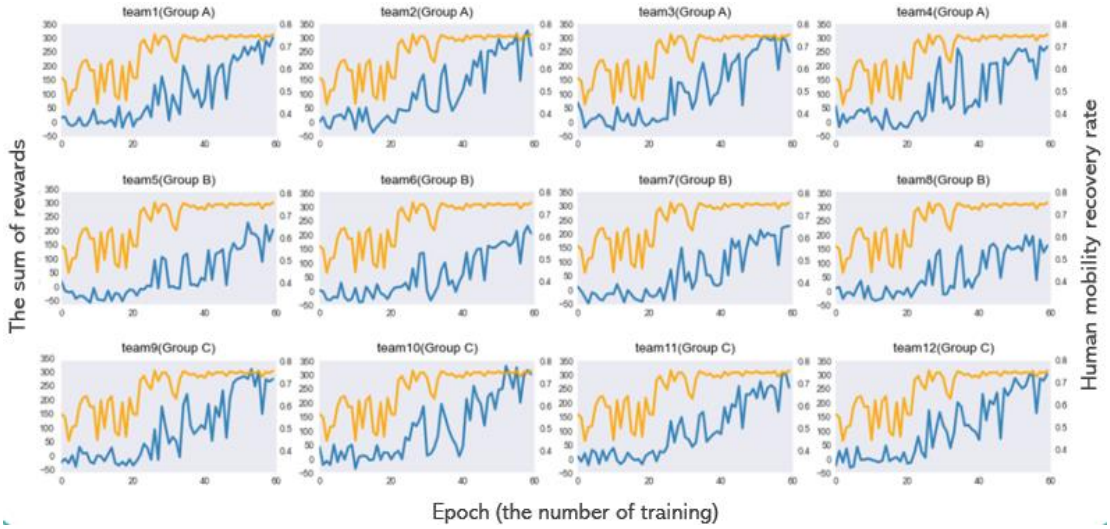


Figure 31. The change of reward and recovery rate with learning trend

As Figure 31 shown, there are the differences among the agents' reward at each epoch. The agents in this framework observes other agents' state partially and makes their own optimal policy respectively. Furthermore, we induced all agents to learn the concept of cooperation. However, they focus more on achieving their own responsibility than on collaborating with other agents. Competition may also appear

among agents belonging to the same group. That is, some of the agents that share the same action space might try to select effective action first, resulting in differences in rewards that agents obtain.

5.3.3 *Verification*

We expected that the agent in this system could learn the cooperation through learning process. We estimate the probability of doing cooperative action. The detail is as follows:

- At each time step, there are 11 opportunities (the number of other agents) for each agent to engage in cooperative action.
- This probability is calculated using the total number of cooperative actions in each epoch divided by the total number of opportunities.

Figure 32 presents the probability that each agent does the cooperation with learning trend. We could confirm that all agents recognize which roads have interrelationships with other roads and tend to restore these type roads for traffic recovery. Therefore, we could conclude that each agent could know the meaning of cooperative operation through partial observed information and corresponding reward.

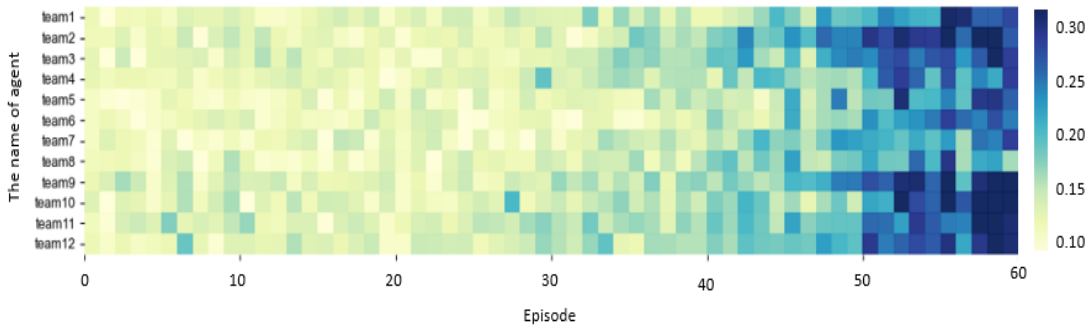


Figure 32. The probability of cooperation with learning trend

We provided each agent in multi-agents RL with the information of network analysis's result and the reward of human mobility recovery rate. As we confirmed in 5.2.3.1, the agent could recognize the meaning or importance of damage roads in the road network with the agent's state and reward setting. It is necessary to make sure that the agent in multi-agent system could also learn about it.

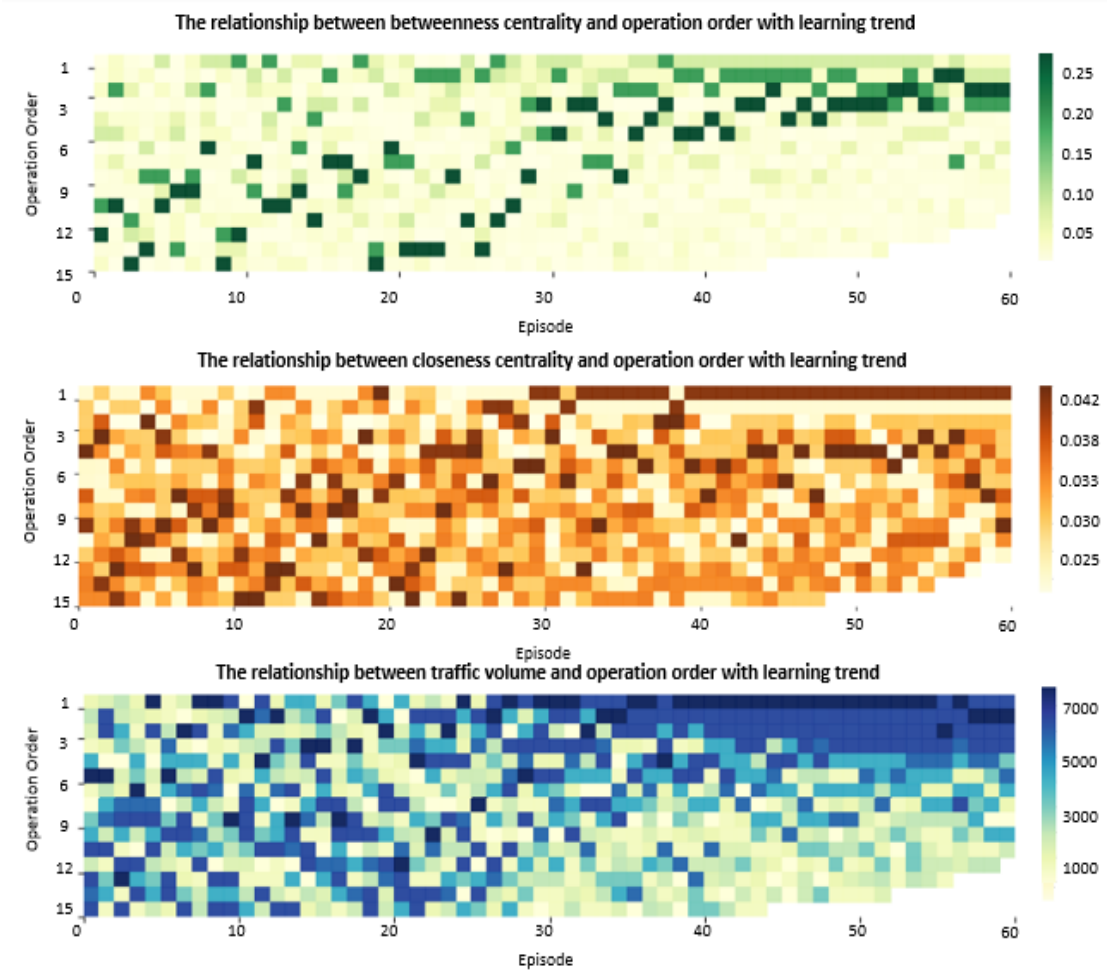


Figure 33. The relationship between road factors and operation order (Group C)

We make the heat map presenting the relationship between damage road's characteristic and general operation order of each group. As Figure 33 shown, agents in each group tend to preferentially select roads with high traffic or high importance, although there are some differences depending on the characteristics of the roads that

are subject to each group's behavior. In addition, we confirm the change of traffic volume with recovery operation in multi agent RL system (Figure 34). The agents seemed to be performing their own recovery operation in close proximity at each time step. This is because they could learn the connectivity among damage roads based on the usage of road-users.

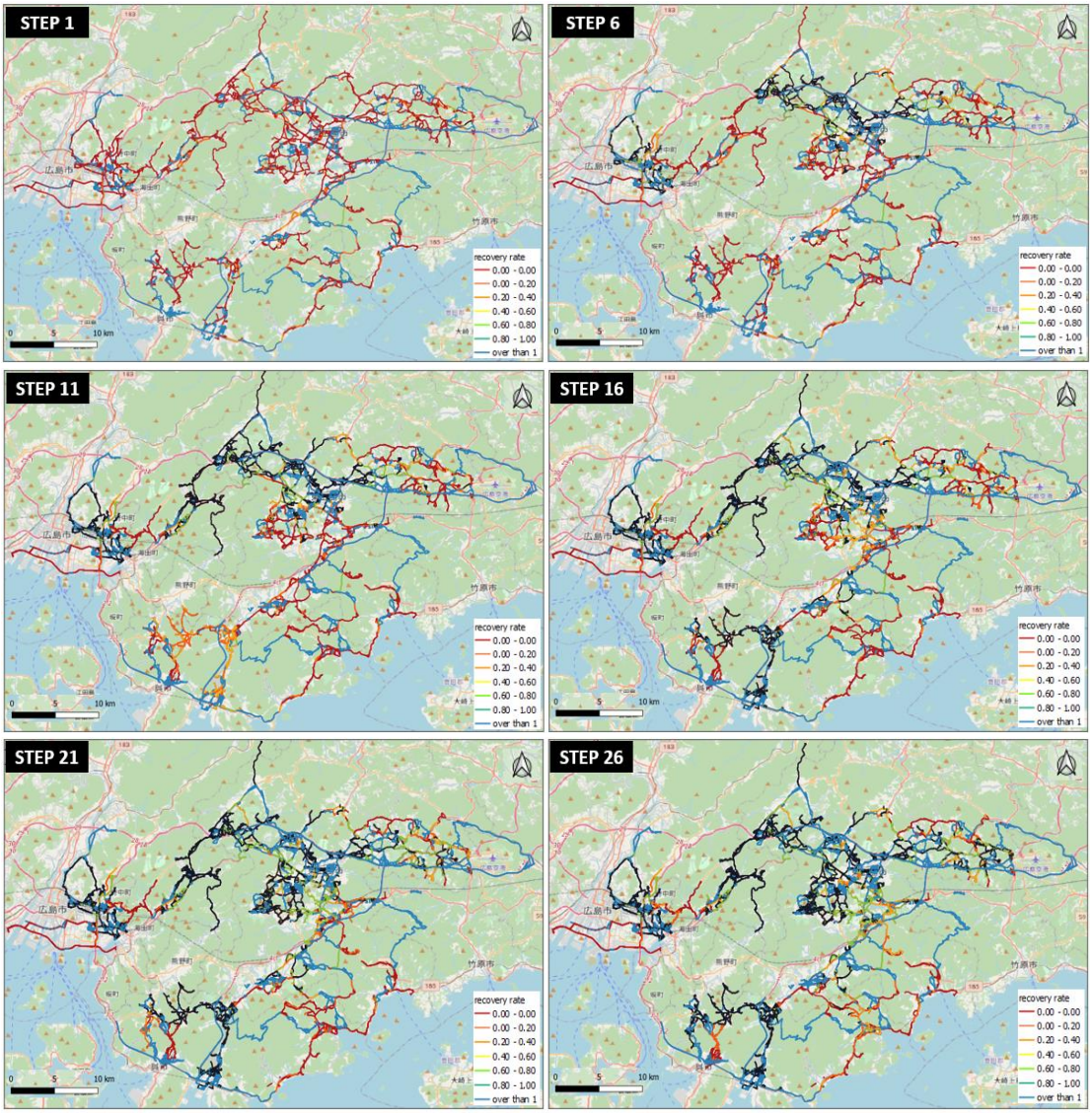


Figure 34. the chage of traffic volume in multi agents RL system

Chapter 6. CONCLUSION

Disrupted road network with large-scale disaster often loses its own fundamental ability to assure basic mobility services and manage emergent situation. Post-disaster situation has the lack and confusion of information and the occurrence of abnormal traffic. Road network is fluid, changing with the context of each road link. The consideration of changed mobility is necessary to recover the utility of road users after disaster.

This study suggested the efficient road reconstruction plan for speed human mobility recovery with the application of single agent Deep RL and multi agent Deep RL. We utilize digital road map and Origin Destination pairs from mobile phone GPS data to estimate the change of human movement and evaluate the degree of recovery according to successive recovery operation.

We provided the reward and the state related to the recovery effect, inter-connectivity with traffic allocation's result and operational progress rate. We could confirm that the agent in these two frameworks identify the optimal policy with 15, 45 damage roads respectively. The agent in this model might work preferentially on the effective damage road to its goal and learn the meaning of target roads under road network. Furthermore, we could induce the agent to cooperate with other agents using the interconnection based on O-Ds' road usage.

We could estimate some information related to the whole operational procedure: 1) Sequence of operations, and 2) the traffic volume and the degree of human mobility recovery. We identified the human mobility change under the agent's reconstruction.

The result of movement's visualization would make the government check the congestion or the abnormal situation and do additional measurements.

This study address road condition and recovery resource dynamically combining Deep RL and human mobility data on optimal recovery strategies for road network. The final human mobility recovery rate with their optimal policy is 25% better on average than the lowest recovery rate when working randomly. We performed 300 simulations based on a single agent model. The number of times an agent has reached a recovery rate of more than 70% is 239 out of 300 simulations. We could say that the accuracy of this model is 0.79.

Approximately 1,000 kinds of O-D pairs were used to estimate mobility. We estimated not just the shortest path but all passable routes to consider realistic human mobility. The number of cases in single agent RL is $6.81 * 10^{43}$. Single agent model takes about 2hour 30 min to get its own optimal policy with 15 damaged roads. Multi agent model takes about 3 hours with 45 damage roads. The number of cases that need to be explored to solve the optimization problem in this paper is 10^7 times than the number of cases in previous studies, but computation time for getting the solution is similar to previous studies.

We suggest future research topic to improve the proposed system. First, it is necessary to consider not only the location of O-Ds on weekdays, but also the change in location of evacuation or on weekends. Second, environmental information around each damage road is also one of important factor for setting priority order. We need to how these factors are utilized as the state of the agent. Third, we estimated the risk level of disrupted roads based on the past disaster record. We could not mention that the

aspect of disaster damage is always the same to the past. Accordingly, further review of damage prediction on disrupted roads and cooperation with some researchers regarding damage estimation is required to improve the more realistic model. Fourth, we would believe that the improved model could be devised to link the restoration process among roads, railways, and subways with the consideration of the changes in the mode of transportation of O-Ds. Lastly, we need to review the application for different RL models and compare the efficiency with other model's result. In detail, we could consider these two versions: 1) the comparison of recovery results for large-scale single agent which has the same size of multi agent, and that for multi-agent, and 2) the efficiency analysis between the multi-agent sharing all information completely and our agents with partial observation.

REFERENCE

[1] Ministry of the Environment, Ministry of Education, Culture, Sports, Science and Technology, Ministry of Agriculture, Forestry and Fisheries, Ministry of Land, Infrastructure, Transport and Tourism, Japan Meteorological Agency, *Climate Change in Japan and Its Impacts,* Synthesis Report on Observations, Projections and Impact Assessments of Climate Change, 2018.

[2] Japan Meteorological Agency, *Weather Forecasts and Analysis,* Available at: https://www.jma.go.jp/ [accessed 1 November 2020]

[3] Nippon.com, *Japan Suffers Record High ¥2.15 Trillion in Flood Damage in 2019,* 2020. 09.18, Available at: https://www.nippon.com/en/japan-data/h00812/ [accessed 31 October 2020]

[4] Statistics Bureau of Japan

[5] Cabinet Office, Japan. *Disaster Management in Japan,* 2011

[6] Gajanayake A, Khan T, and Zhang K.G, *Post-Disaster Decision Making in Road Infrastructure Recovery Project – An Interview Study with Practitioners in Queensland,* 2019, Australian & New Zealand Disaster & Emergency Management Conference.

[7] Chang Y, Wilkinson S, Potangaroa R and Seville E, *Managing resources in disaster recovery projects,* 2012, Engineering, Construction and Architectural Management, 19(5):557-580

[8] Lyons M, *Building back better: the large-scale impact of small-scale approaches to reconstruction,* 2009, World Development, 37(2): 385-398

[9] Le Masurier J, Rotimi J.O, and Wilkinson S, *Comparison between routine construction and post-disaster reconstruction with studies from New Zealand,* 2006, Disaster Prevention and Management: An International Journal, 15(3): 396-413

[10] Hayat, Ezri and Amaratunga, Dilanthi, *Road Reconstruction in Post – Disaster Recovery; Challenges and Obstacles,* 2011, In: International Conference on Building Resilience: Interdisciplinary approaches to disaster risk reduction, and development of sustainable communities and cities, Kandalama, Sri Lanka

[11] Chang, S. E. *Disasters and transport systems: loss, recovery and competition ate the Port of Kobe after the 1995 earthquake,* 2000, Journal of Transport Geography, 8: 53-65

[12] Aydin N.Y., Duzgun H.S., Heinimann H.R., Wenzel F. and Gbyawli K.R, *Framework for improving the resilience and recovery of transportation networks*

under geohazard risks, 2018, International Journal of Disaster Risk Reduction, 31: 832-843

[13] Bothale V.M, Khobragrade A.N, and Srivastav N.T, *Role of geoinformatics in development of disaster management information system*, 2015, Journal of Homeland Security and Emergency Management, 12(3): 571-602.

[14] Kawasaki S, *Issues and Lessons Learned from the Great East Japan Earthquake,* 2011, XXIV World Road Congress Mexico.

[15] Wako R, Sekimoto Y, Kanasugi H, and Shibasaki R, *Analysis of people's route and destination choice in evacuation using GPS log data*, 2014, Infrastructure Planning and Management, 70(5): 681-688 (in Japanese)

[16] Nguyen L.H, Yang Z, Zhu J, Li J and Jin F, *Coordinating disaster emergency response with heuristic reinforcement learning,* , 2018, arXiv preprint arXiv:1811.05010.

[17] Yamada Y, Iemura H, Noda S, and Izuno K, *Evaluation of the optimum restoration process for transportation system after seismic disaster,* 1986, Journal of JSCE , 1986(368): 355-362

[18] Chang S.E, *Transportation planning for disasters: an accessibility approach,* 2003, Environment and Planning A, 35(6): 1051-1072

[19] Balal E, Valdez G, Miramontes J and Cheu R.L, *Comparative evaluation of measures for urban highway network resilience due to traffic incidents,* 2019, International Journal of Transportation Science and Technology, 8(3):304-317

[20] Masafumi H, *Study on Priority Level of Restoration in Road Network Based on Possibility of Using Roundabout after Natural Disaster,* 1998, Infrastructure Planning Review, 15:337-344 (in Japanese)

[21] Gonzales, M.C., Hidalgo, C.A., and Barabasi, A.L, *Understanding individual human mobility patterns,* 2008, Nature, 453(7196).

[22] Wang, Q., and Taylor, J. E, *Quantifying human mobility perturbation and resilience in Hurricane Sandy,* 2014, PLoS one, 9(11).

[23] Lu, X., Bengtsson, L., and Holme, P, *Predictability of population displacement after the 2010 Haiti earthquake*, 2012, Proc. National Academy of Sciences, 109(29):11576-11581

[24] Song, X., Zhang, Q., Sekimoto, Y., Horanont, T., Ueyama, S., and Shibasaki, R, *Modeling and probabilistic reasoning of population evacuation during large-scale disaster,* 2013, Proc. ACMSICKDD.

[25] Yabe, T., Tsubouch, K., Sudo, A., and Sekimoto, Y, *A framework for evacuation hotspot detection after large scale disaster using location data from smartphones; case study of Kumamote earthquake,* 2016, SIGSPATIAL'16.

[26] Saravi S, Kalawsky R.S, Joannou M.R, Casaado G.F and Meng F, *Use of artificial intelligence to improve resilience and preparedness against adverse flood event,* 2019, Water, 11(5): 973.

[27] Su Z.P, Jiang J.G, Liang C.Y and Zhang G.F, *Path selection in disaster response management based on Q-learning,* 2011, International Journal of Automation and Computing, Issue 1.

[28] Yang S, Ogawa Y, Ikeuchi K, Akiyama Y and Shibasaki R, *Firm-level behavior control after large-scale urban flooding using multi-agent deep reinforcement learning,* 2019, Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.

[29] Wisetjindawat W, Ito H, Fujita M, and Mideshima E, *Modeling Disaster Response Operations including Road Network Vulnerability*, 2013, Journal of the Eastern Asia Society for Transportation Studies, 10:197-214

[30] Japan International Cooperation Agency (JICA), *The study of reconstruction processes from large-scale disasters – JICA's Support for Reconstruction* -, 2013

[31] Chen X.Z, Lu Q.C, Peng Z.R, and Ash J.E, *Analysis of Transportation Network Vulnerability Under Flooding Disaster,* 2015, Transportation Research Record, 2532:37-44

[32] Dilek T.A and Linet O, *A mathematical model for post-disaster road restoration: Enabling accessibility and evacuation,* 2014, Transportation Research Part E, 61: 56-67

[33] Chang S.E and Nojima N, *Measuring post-disaster transportation system performance: the 1995 Kobe earthquake in comparative perspective,* 2001, Transportation Research Part A: Policy and Practice, 35(6): 475-494

[34] Sohn J, *Evaluating the significance of highway network links under flood damage: an accessibility approach,* 2006, Transportation Research Part A, 40: 491-506

[35] Toshihiro A and Ei-Ichi O, *A Construction Method of Resilient Network Considering Betweenness Centrality – Application to Restoration of Road Network Damaged by Disaster* -, 2015, The 29th Annual Conference of the Japanese Society for Artificial Intelligence (in Japanese)

[36] Pauline G, Angelo F, and Nour-Eddin E.F, *Road network resilience: How to identify critical links subject to day-to-day disruptions,* 2018, Transportation Research Record, 2672(1):1-12

[37] Djamel B, Jacques R, Monia R and Angel R, *Transportation in disaster response operations,* 2012, Socio-Economic Planning Sciences, 46(1): 23-32.

[38] Osawa S, Nakayama S, Fujiu M and Takayama J.I, *A study on decision method for restoration priority rank in road network based on accessibility index after natural disaster,* 2017, Journal of Construction Engineering and Management, 73(5):281-289 (in Japanese)

[39] Sugimoto H, Tamura T, Arimura M and Saito K, *The restoration model of the damaged road network based on the cooperation of the improvement teams,* 1998, Journal of Construction Engineering and Management, 625: 135-148 (in Japanese)

[40] David R and Hillel B.C, *Long-term scheduling for road network disaster recovery*, 2019, International Journal of Disaster Risk Reduction, 42:101353

[41] Sakamoto J and Nishiuchi H, *Proposal of a Decision Method for Road Recovery Considering Recovery Capacity after a Large-Scale Disaster,* 2018, Journal of the City Planning Institute of Japan, 53(3): 859-866 (in Japanese)

[42] Hori M, Yugeta K, Ichimura T and Wijarthne L, *On development of multi-agent simulation for recovery process of lifeline damaged by earthquake,* 2011, Journal of Construction Engineering and Management A1, 67(1): 165-176

[43] Bhatia U, Sela L and Ganguly A.R, *Hybrid Method of Recovery: Combining Topology and Optimization for Transportation Systems,* 2020, American Society of Civil Engineers, 26(3)

[44] Yerukola A, Pokle A, and Jhunjhunwala M, *Deep Reinforcement Learning for Long Term Strategy Games*, 2012, University of Stanford.

[45] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou D. Wierstra, and Riedmiller M, *Playing Atari with Deep Reinforcement Learning,* 2013, NIPS Deep Leaning Workshop.

[46] Schmidt J, Mario R, Marques G, Botti S, and Marques M, *Recent Advances and Applications of Machine Learning in Solid State Materials Science,* 2019, NPJ computational materials, 83.

[47] Shrestha A and Mahmood A, *Review of Deep Learning Algorithms and Architectures,* IEEE Access, 7

[48] Sutton R.S and Barto A.G, *Reinforcement learning: an introduction; Adaptive computation and machine learning series; Second edition,* The MIT Press, Nov. 2018.

[49] Stuart J.R, and Peter N, *Artificial Intelligence: A Modern Approach,* 2010, Third Edition, Prentice Hall.

[50] Thomas M.M, Joost B, and Catholijn M.J, *Model-based Reinforcement Learning: A Survey,* arXiv:2006

[51] Chow J.H, Wu F.F, and Momoh J.A, *Applied Mathematics for Restructured Electric Power System: Optimization, Control, and Computational Intelligence: Load Forecasting,* 2005, New York: Springer.

[52] TOKIC, Michel, *Adaptive ε-greedy exploration in reinforcement learning based on value differences,* 2010, In: Annual Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 203-210.

[53] Kanerviso A, Scheller C, and Hautamaki V, *Action Space Shaping in Deep Reinforcement Learning,* 2020, In: IEEE Conference on Games 2020.

[54] "Standard Specifications for Civil Construction Management", Hiroshima Prefecture, 2015, 04. (in Japanese)

[55] Thrun S and Schwartz A, *Issues in using function approximation for reinforcement learning,* 1993, In: Proceedings of the 1993 Connectionist Models

[56] Zahavy T, Haroush M, Merlis N, Mankowitz Daniel J, and Mannor S, *Learn What Not to Learn: Action Elimination with Deep Reinforcement Learning,* 2018, Advances in Neural Information Processing Systems, 3566-3577

[57] Adam D.L, *Theory and application of reward shaping in reinforcement learning,* 2014, University of Illinois at Urbana-Champaign

[58] Van M.K, Drugan M.M and Nowe A, *Scalarized multi-objective reinforcement learning: Novel design technique,* 2013, IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 191-199

[59] Tim B, Harutyunyan A, Vrancx P, Taylor M.E, Kudenko D and Nowe A, *Multi-objectivization of reinforcement learning problem by reward shaping,* 2014, IEEE, International joint conference on neural networks (IJCNN), 2315-2322

[60] Skomlj N.O and Radujkovic M, *S-curve modeling in early phases of construction projects,* 2012, Gradevinar, 64(8): 647-654

[61] Federal Highway Administration, *Traffic assignment,* 1973, U.S. Department of Transportation

[62] Kim C.K, *Application of user equilibrium traffic assignment in evacuation modelling,* 1991, Virginia Polytechnic Institute and State University

[63] Luisa D.M, Musolino G and Vitetta A, *Traffic Assignment Model in Road Evacuation,* 2012, WIT Transactions on Ecology and the Environment, 155: 1041-1051

[64] Meng Q, Lam W.H, and Yang L, *General stochastic user equilibrium traffic assignment problem with link capacity constraints*, 2008, Journal of Advanced Transportation, 42(4): 429-465

[65] Weathernews, *Analysis of 20,000 flood damage reports in Western Japan Flooding*, Weathernews, 2018.07.10. Available at: https://jp.weathernews.com/news/23807/ (accessed by 25 April 2020)

[66] Asahi Weekly, *582 sections of roads nationwide are closed and traffic jams occur*, Asahi Weekly, 2018.07.15. Available at: https://www.asahi.com/articles/ASL7H35D0L7HPTIL00H.html (accessed 25 May 2020).

[67] Japan Digital Road Map, *What is the DRM Database?* Available at: http://www.drm.jp/english/drm/database/structure.html (accessed 22 May 2020)

[68] Ohkubo K, Kamemura K and Hamada M, *On the business continuity planning of expressway based on the case analysis of embankment damage due to natural disaster*, 2013, Journal of Construction Engineering and Management F5, 69(1): 1-13

[69] Ardi T, Tambet M, Dorian K, Ilya K, Kristjan K, Juhan A, Jaan A and Raul V, *Multiagent cooperation and competition with deep reinforcement learning*, 2017, PLoS ONE 12(4): E0172395

[70] Ming Tan, *Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents*, 1993, Proc. of the 10th International Conference on Machine Learning, 330-337

[71] Yang J, Borovikov I and Zha H, *Hierarchical Cooperative Multi-agent Reinforcement Learning with Skill Discovery*, 2020, International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)

[72] Balachandar, N., Dieter J and Ramachandran G.S, *Collaboration of AI Agents via Cooperative Multi-Agent Deep Reinforcement Learning*, 2019, arXiv preprint arXiv: 1907.00327

[73] Jakob N.F, Yannis M.A, Nando D.F and Shimon W, *Learning to Communicate with Deep Multi-Agent Reinforcement Learning*, 2016, NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2145-2153

[74] Razali N.M and Geraghty J, *Genetic Algorithm Performance with Different Selection Strategies in Solving TSP*, 2011, In: Proceedings of the World Congress on Engineering

[75] Lawler E.L, Lenstra J.K, Rinnooy Kan A.H.G, and Shmoys D.B, *The Traveling Salesman Problem*, 1985, John Wiley & Sons Ltd.

[76] Balachandar, N., Dieter J and Ramachandran G.S, *Collaboration of AI Agents via Cooperative Multi-Agent Deep Reinforcement Learning*, 2019, arXiv preprint arXiv: 1907.00327

[77] Jakob N.F, Yannis M.A, Nando D.F and Shimon W, *Learning to Communicate with Deep Multi-Agent Reinforcement Learning*, 2016, NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2145-2153

[78] Mao H, Gong Z and Xiao Z, *Reward design in cooperative multi-agent reinforcement learning for packet routing*,2020, arXiv preprint arXiv: 2003.03433