World Academy of Science, Engineering and Technology
International Journal of Economics and Management Engineering
Vol:11, No:1, 2017

A Reinforcement Learning Approach for Evaluation of Real-Time Disaster Relief Demand and Network Condition

Ali Nadi, Ali Edrissi

***Abstract*—**Relief demand and transportation links availability is the essential information that is needed for every natural disaster operation. This information is not in hand once a disaster strikes. Relief demand and network condition has been evaluated based on prediction method in related works. Nevertheless, prediction seems to be over or under estimated due to uncertainties and may lead to a failure operation. Therefore, in this paper a stochastic programming model is proposed to evaluate real-time relief demand and network condition at the onset of a natural disaster. To address the time sensitivity of the emergency response, the proposed model uses reinforcement learning for optimization of the total relief assessment time. The proposed model is tested on a real size network problem. The simulation results indicate that the proposed model performs well in the case of collecting real-time information.

***Keywords*—**Disaster management, real-time demand, reinforcement learning, relief demand.

I. Introduction

NATURAL disasters kill many people around the world every year. Emergency responses in the case of natural disaster have to be carried out perfectly to decrease death tolls. The effectiveness of an emergency response directly depends on the available information relating to the relief demand, network condition, and vulnerable population. This information unfortunately is not available unless a natural disaster happens. In most studies, relief demand is predicted based on previous information and regional features. These features include structural instability, materials and network reliability [1]. The main major concern of researchers in the field of natural disaster management is the uncertainty in demand and network conditions. This uncertainty and stochastic environment causes the prediction methods used for predicting relief demand to be over or under estimated, and also a lack of information or missing data may affect the prediction accuracy. This inaccuracy may lead the emergency response operation to fail, resulting in a higher death toll. For this purpose, the aim of this paper is assessing real-time relief demand and transportation network availability is suggested at the onset of a natural disaster. However, the time consuming nature of assessing relief demand in such a time sensitive situation may affect the immediate emergency response, and thus, the proposed dynamic scheduling model provided in this paper addresses the time sensitivity problem by minimizing the total relief assessment time. An artificial intelligent and machine learning technique is employed to solve this problem dynamically. In this proposed model, reinforcement learning and Markov decision process is used to formulize the problem. Reinforcement learning is an unsupervised learning technique that allows an agent chooses an action in a stochastic environment and obtains a reaction from that environment. By experiencing different scenarios in a stochastic environment, that agent learns how to react in an unexpected condition. The main question that the proposed model is will address in this paper is: How can the assessment process be scheduled to improve the information available during the emergency response process to minimize a time consuming operation? The implementation of the proposed model shows that the real-time demand and network information can be obtained as soon as possible at the onset of a natural disaster.

Ali Nadi and Ali Edrissi are with the Civil Engineering Department of K.N. Toosi University of Technology, No. 1346, Vali Asr. Street, Mirdamad Intersection, 19697 Tehran, Iran (phone: +98-913-309-4880; e-mail: anadi@mail.kntu.ac.ir, edrisi@kntu.ac.ir).

II. Related Works

Demand prediction has a challenging point in the uncertainty of relief demand information in large scale natural disasters. Many researches proposed time series and autoregressive models to predict dynamic relief demand [2]. Sun et al. [3] proposed a fuzzy rough set approach for emergency demand prediction to overcome the inaccuracy and incomplete information. These models commonly use historical information to forecast time varying demand. Lack of information, missing historical values, unreliable and outlier information of relief demand may place its pattern recognition into certain troubles; therefore, time-series based models seem to be unsuitable for real-time demand. Sheu [4] considered that the real-time demand information such as the number of survivors and missing people comes from diverse sources in the affected region. Although his model considers information reliability and accuracy by using frequently updated information, an accurate and reliable assessment process, especially for transportation infrastructure availability, is still a problem.

There are a few studies that have considered optimization in urban relief assessment operations. Huang et al. [5] introduced the assessment routing problem and proposed a continuous approximation approach to solve it. Considering the time sensitivity of emergency response, the time consuming assessment relief operation can increase the death toll. Beside relief demand, critical transportation links are important in emergency response; this increases the uncertainty in disaster logistics. Edrissi et al. [6] proposed an emergency reliability measure that incorporates both zonal travel time and the level of supply and demand in each zone. They also proposed a heuristic algorithm to solve real size network. Their research showed that an increase in investments in the network improvement reduces the death toll more than a higher budget increment. Fiedrich et al. [7] have introduced a dynamic operation model that finds the best assignment of resources to the affected zone. Rennemo et al. [8] considered a three stage mixed integer programing model for emergency response planning containing the opening of local distribution facilities, initial location of supplies and last mile distribution of aid. This model considers vehicle availability, the infrastructure state and demand uncertainty. Cavdur et al. [9] minimized the total distance traveled, the unmet demand and the total number of facilities. They allocate facilities by considering the potential difficulties to access the supplies. Their model considers relief distribution in the second stage with minimization of total travel distance and unmet demand. They customized a scenario-based approach to evaluate the sensitivity of the model to uncertainties. Nevertheless, the five scenarios they prepared may not be sufficient compared with the capability of life span training dynamic scenarios which are proposed in this paper.

There is a pool of research using relief demand information for emergency response and operations. Minimizing the number of fatalities is the main aim of search and rescue operations. Chen and Hooks [10] routed urban search and rescue teams using multistage stochastic programming based on the column generation method. Their objective was the maximization of the total number of survivors.

Delivery of relief to the affected region is the final stage of the disaster relief chain. Therefore, different extensions of the vehicle routing problem have been modeled due to dynamic time varying relief demand to solve this problem in previous researches. Wohlgemuth et al. [11] proposed dynamic optimization for the pickup and delivery problem in consideration of varying travel times, link availability and unknown demand. Ozdamar et al. [12] also developed a solution to the last mile pickup and delivery problem. The hierarchical optimization model that they proposed has the goal of minimizing total travel time. They used hierarchically clustered nodes for routing with respect to vehicle and supply availability. In this model, the optimal allocation for cluster centers is found and then the routing problem within each cluster's sub-network is solved. A sound review of routing problems solved to deliver goods and services within disaster affected regions has been presented by Luis et al. [13].

Different objective functions have been taken into account in the logistic operation. Barbarosoglu and Arda [14] used the total cost of deliveries with respect to satisfying all demands. An integrated multi commodity network and vehicle routing problem to model mixed pick-up and delivery is proposed by Ozdamar and Demir [15]. Yi and Kumar [16] also proposed an ant colony optimization to minimize the sum of unsatisfied demands on all commodities as well as the unsaved people in each node. Ahmadi et al. [17] considered road destruction probability and standard relief time for humanitarian logistic operations. They develop multi-depot location routing problem to minimize the total distribution time of humanitarian relief. They considered the standard amount of relief goods needed for each person and proposed a variable neighborhood search algorithm. Their model was tested in large scale GIS data. This model showed that the standard relief time window and link failure increases the penalty cost of unsatisfied demand. Therefore, humanitarian logistic needs a higher number of local depots and vehicles, rather than commercial logistics. Minimization of transportation cost, minimization of unsatisfied demand and minimization of unserved injured people are the three conventional objective functions in disaster response research. These objectives do not have same priority in practice. Najafi et al. [18] proposed a multi hierarchical objective robust optimization model that manages the logistics of both commodities and the injured population in the response phase. Huang et al. [19] also integrated resource allocation with emergency distribution. The point comes from their model is considering the lifesaving effectiveness, human suffering and fairness in the objective function.

The common attitude of all these researches is that they all involve dynamic time varying relief demand in emergency response. These models mostly used dynamic programing to handle dynamic demand; however, the real value of this demand is not clear until the responder arrives in the affected area.

III. Methodology

To evaluate real-time relief demand and transportation network conditions, a Markov Decision Process (MDP) model is formulated and proposed in this paper. In a MDP model, an agent has a set of state S (instance for affected region) and in each state there exists a set of actions A (links). In time *t,* the agent is in state s_t and chooses the action a_t. The environment gets reward $r_t(s_t, a_t)$ and the agent moves to state $s_{t+1} = \delta(s_t, a_t)$. The r_t and δ are from the environment and are not known in advance. The $r_t(s_t, a_t)$ and $\delta(s_t, a_t)$ depend only on the current state and action of the system, and are independent from previous actions and states; in this problem, $r_t(s_t, a_t)$ and $\delta(s_t, a_t)$ are probabilistic. This environment is known as the nondeterministic Markov decision process. To formulate the scheduling problem of relief and network condition assessment, suppose T is a set of time epoch of a system. The A_s is a set of action in each state. The initial system state is $s_0 = (1,0,0,0,\dots,0)$ and the final state is $s_I = (1,1,1,1,\dots,1)$, the final state is the situation in which the agent returns to depot after fully serving all demands. The final state is also an absorbing state; that is the only available action in this state is

staying in it, the notation 1 in s_t instance for a visited region. For example, if choosing an action in state s_0 directs an agent to region 3, the next state will be $s_1 = (1,0,0,1,0,\dots,0)$. In each state, a set of actions is available. These actions are of two types. 1- Choosing a link that leads to an unvisited region. For example, in state s_0, supposes the agent chooses $a_{0\to3}$ among all available actions, this agent may be in region 3 with probability of p and may stay in current state with probability of 1-p. This means that action $a_{0\to3}$ may not be available due to a destroyed link probability. The p can be obtained from reliability and risk analysis of transportation roads. The transition probability can be defined as:

$$P(s'|s,a) = \begin{cases} p & if\ \delta(s,a) = s' \\ 1-p & if\ \delta(s,a) = s \end{cases} \quad (1)$$

In this study, the reward obtained from the environment is defined to be proportionate to the process time of relief assessment which contains travel time and working time. The agent should learn an optimum policy such as π:S→A that specifies the next action a_t from state s_t. This could happen when the agent maximizes the expected discounted cumulative reward during the time horizon of the system.

$$\pi^*(s) = \underset{a}{\operatorname{argmax}}[E[r(s,a)] + \gamma E[V^*(\delta(s,a))]] \quad (2)$$

Assuming the dynamic nondeterministic value for r(s, a) and δ(s, a), Q is a recursive function that is defined in (3) to solve the maximization problem mentioned in (2).

$$Q(s,a) = E[r(s,a)] + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a') \quad (3)$$

For solving stochastic problems, the enumeration and evaluation of all policies is needed. For more details, in the remaining section, the Q-learning algorithm proposed by Watkins and Dayan [20] is described to solve this optimization problem. Reliable optimum policy can be gained by this iterative Q-Learning algorithm.

A. Q-Learning Algorithm

Function $Q(s,a)$ estimates the value that maximizes the discounted cumulative reward for each *s* and *a* in the first step of the algorithm. A matrix $\hat{Q}(s,a)$with the *s* and *a* value is assumed to be the approximation of $Q(s,a)$; the $\hat{Q}(s,a)$ fills with an initial random value of *s* and *a*. In each step, the agents look to the state and choose the action *a* and receive the nondeterministic value of $r(s,a)$ and observe the next state $s' = \delta(s,a)$ with probability of $P(s'|s,a)$. The agent updates the value of $\hat{Q}(s,a)$ as:

$$\hat{Q}(s,a) \leftarrow (1-\alpha_n)\hat{Q}_{n-1}(s,a) + \alpha_n[r(s,a) + \gamma \max_{a'} \hat{Q}_{n-1}(s',a')] \quad (4)$$

$$\alpha_n = \frac{1}{1+N_{s_n(s,a)}} \quad (5)$$

where n is the steps of algorithm and $N_{s_n(s,a)}$ is the number of visited *(s, a)* until step *n*. In this algorithm, the agents do not need any information about δ and *r* to learn the optimum policy. It just does the action and observes the reward.

This algorithm works under two conditions [21].

1) The reward should be limited to a value like c as $|r(s,a)| \le c$.
2) Each pair (s, a) should be evaluated by an agent repeatedly and more than once.

To satisfy condition 2, there is a method called Softmax action selection. This method makes an agent choose an action with the probability of $q(a_i|s)$. This probability is defined with (6):

$$q(a_i|s) = \frac{e^{\hat{Q}(s,a_i)/T}}{\sum_j e^{\hat{Q}(s,a_j)/T}} \quad (6)$$

where T is the counterbalance between exploration and exploitation.

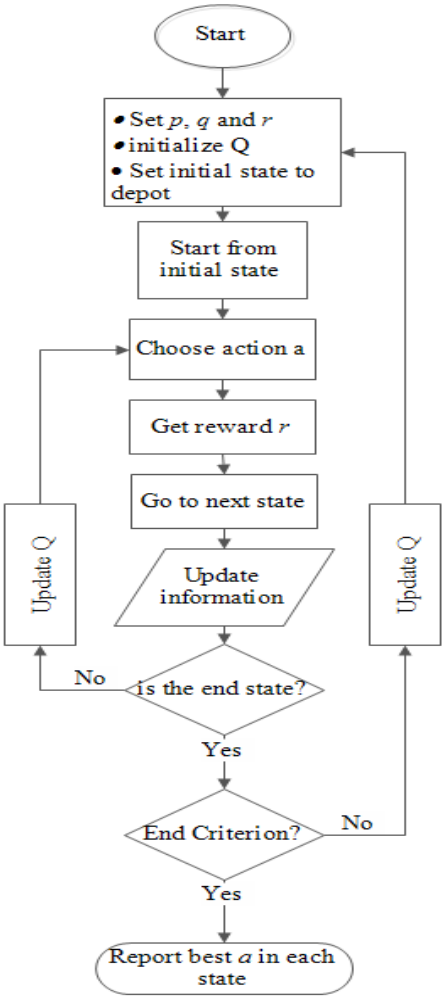

Fig. 1 Procedure of proposed model

The procedure of the proposed model is illustrated in Fig. 1. In the first step of the algorithm, which is called the

initialization phase, a scenario is characterized by setting the value for r, p, q and network definition. An initial value of 0 also is defined for the $\hat{Q}$ matrix. The predefined agents start from the depot, which sets the initial state of the system to depot.

Each agent chooses an available link as an action with the probability defined in (6). The system immediately obtains a reward (or punishment) based on the network condition and assessment process time of the next region in which the agent is directed to by choosing this action. When the agent arrives in the next region with the probability defined in (1), the information relating to the relief demand and network conditions will be updated.

If the next state is the final state $s_I = (1,1,1,1,\dots,1)$, which means all the regions are visited and evaluated, the process will be continued until the end criterion satisfied. It means that this systems can test a lot of scenarios in the offline mode and use the final $\hat{Q}$ in the online mode as an initial value of $\hat{Q}$. Otherwise the agent is continuously updating $\hat{Q}$ and chooses an action in the current state.

After final $\hat{Q}$ is obtained from the whole process, the probability of choosing the optimum action becomes very high. Thus, in online mode, this $\hat{Q}$ can be used in the initial phase, so that if the agent encounters an unknown network condition or demand level, the $\hat{Q}$ immediately converges to the Q, and therefore, the best actions in each state are revealed to reroute the assessment teams.

In the next section, a real size network of Isfahan province is prepared to test the proposed model.

IV. Case Study: Isfahan Province

Fig. 2 shows the Isfahan province in Iran, which is divided into 47 zones and consists of a total number of 105 cities. Isfahan as the capital city is supposed to be the affected region, and consists of 14 cities. The instability ratio is provided by the average building ages and their construction materials. The failure probability of transportation links are also estimated using parameters such as the length of the corridor and the presence of specific infrastructure such as bridges [22]. In this paper, the vulnerable population is defined as the nominal demand and Kolmogorov-Simonov Test proved the Gaussian distribution of the vulnerable population. For this problem, this distribution function is used to generate real-time demand as well as rewards. The attribute of this affected region is shown in Table I.

The link information of the affected region in Table II consists of the transportation link failure probability and travel times. The travel times are obtained from Google Earth.

One other demand attribute that is needed to generate the problem scenario is the probability density functions of on-site service times. As it is difficult to acquire this additional required data, simulated data were generated from discrete uniform distributions for these factors based on limited real information, including site locations, building uses and damage severity. This model is simulated and coded in MATLAB 2015a using a 2.40 GHz core i7 laptop with 4 GB RAM. The number of agents assessing the relief demand and network condition is set to four. Fig. 2 shows that the algorithm is well optimized in the defined iteration number.

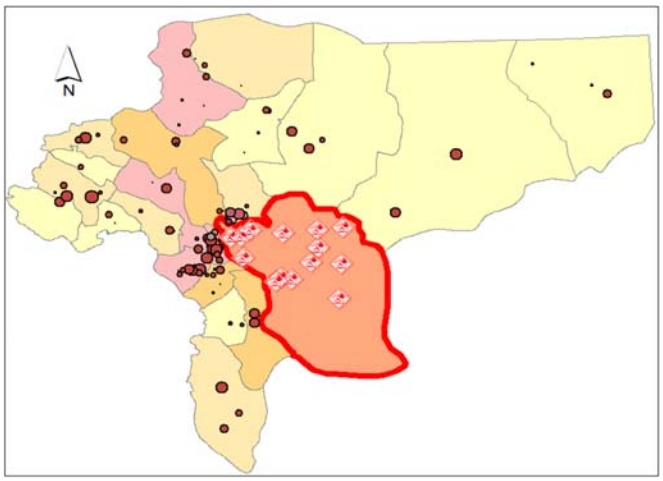

Fig. 2 Affected region in Isfahan province

TABLE I
Attribute of Affected Region in Isfahan Province

No.	City	Population	Instability	Vulnerable
1	Isfahan	1756126	0.06	105367
2	Khorasgan	97167	0.1	9716
3	Baharestan	61647	0.04	2465
4	Varzaneh	11924	0.17	2027
5	Ghahjavarestan	7906	0.14	1106
6	Harand	7108	0.15	1066
7	Nasrabad	6176	0.15	926
8	Sejzi	4698	0.15	704
9	Kohpayeh	4587	0.12	550
10	Mohammadabad	4549	0.15	682
11	Nikabad	4303	0.16	688
12	Hasanabad	4267	0.15	640
13	Toodeshk	4229	0.15	634
14	Ejyeh	3481	0.15	522

TABLE II
Attribute of Affected Region in Isfahan Province

Start node	End node	Travel time (min)	Failure probability
1	2	30	0.15
1	3	39	0.05
1	5	32	0.15
1	27	83	0.00
2	5	19	0.05
2	8	23	0.00
3	10	42	0.00
3	11	42	0.00
6	4	27	0.10
8	9	23	0.00
9	6	15	0.00
9	13	14	0.00
10	7	12	0.00
10	11	13	0.00
11	12	34	0.00
14	6	16	0.10
14	4	25	0.00

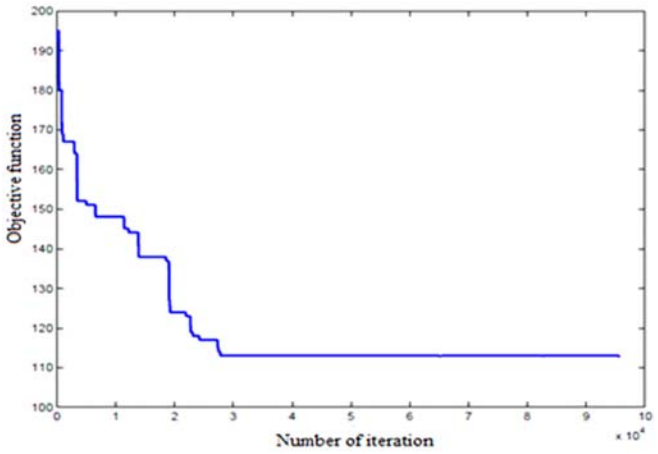

Fig. 3 Objective function value in each iteration

The results of the proposed model are shown in Table III and Table IV. In the Table III, the values of $\hat{Q}(s,a)$ for each state are compared with the deterministic values of $Q(s,a)$.

TABLE III
THE VALUE OF ESTIMATED $\hat{Q}$ COMPARED TO THE DETERMINISTIC VALUE OF Q

Zone	$\hat{Q}(s,a)\times 10$	$Q(s,a)\times 10$
1	81.419	81.419
2	79.354	79.351
3	73.726	73.928
4	67.21	66.987
5	65.83	65.834
6	61.73	61.728
7	62.36	62.379
8	55.36	55.3
9	58.21	57.87
10	48.12	48.238
11	44.39	44.41
12	45.17	45.17
13	41.36	41.36
14	39.88	39.878

The percentage of each agent's failure is indicated in Table IV. This failure occurs when an agent encounters an unexpected unavailable link. The mean absolute error of the algorithm to estimate $\hat{Q}(s,a)$ is also presented in Table IV. The results show that the maximum completion time of the assessing process is 114 minutes. This means that in less than two hours, the real time true vulnerable demand and network condition can be evaluated with the proposed model.

TABLE IV
THE RESULT OF PROPOSED MODEL

Agents	Failure (%)	MAPE	Completion time (min)
1	1.36	0.839	114
2	0.98	0.741	97
3	1.271	1.05	85
4	0.87	0.585	73

Although consuming 114 minutes for the assessment process may increase the death toll, true information about the network conditions and real amount of vulnerable demand can significantly decrease the death toll due to a more accurate emergency response. To show the impact of true real-time demand and network condition, the emergency response problem is solved to compare the death toll under the conditions of with and without the assessment process. The term 'nominal demand' is defined to show that the uncertain amount of vulnerable demand that is expected to be in affected region. Fig. 3 shows the effect of the proposed model in reducing the death toll.

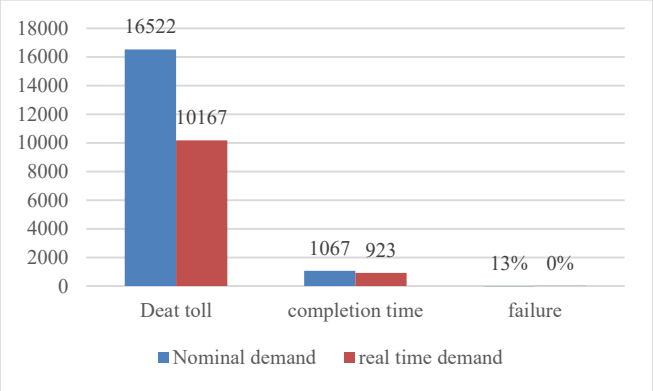

Fig. 4 Comparison of emergency response with and without proposed assessing process

As it can be observed from the chart in Fig. 3, the death toll decreased due to the availability of true and real-time demand. Although the completion time of the emergency response was expected to be increased, it was observed that it is less than the completion time of an emergency response with nominal demand. That is because in nominal demand in which there is no certainty in the network condition and vulnerable demand, the failure of the emergency process is about 13%. This failure rate means that in 13% of all journeys, the emergency responders encounter to an unexpected failed transportation link or unexpected demand which it takes more time to reroute or satisfy that demand.

V. CONCLUSION

Although vulnerable demand and transportation network conditions in natural disaster management are essential to every response process, the true amount of this information is not available in-hand until the disaster happens. In this paper, a stochastic decision processing model is proposed based on reinforcement learning to schedule the real-time assessment process of relief demand and network conditions. The results showed that although the assessing process takes more time and that this may lead to an increased death toll, the entire completion time of the emergency response is decreased; this is because of the emergency response failure using nominal demand and network information. For future researches, the integration of emergency response and relief assessment teams in a way that they work with each other interactively and simultaneously (not consecutively), may lead to a major reduction in death tolls.

REFERENCES

[1] Liu, W., G. Hu, and J. Li, Emergency resources demand prediction using case-based reasoning. *Safety Science*, 2012. 50(3): p. 530-534.
[2] Aviv, Y., A time-series framework for supply-chain inventory management. *Operations Research*, 2003. 51(2): p. 210-227.
[3] Sun, B., W. Ma, and H. Zhao, A fuzzy rough set approach to emergency material demand prediction over two universes. *Applied Mathematical Modelling*, 2013. 37(10): p. 7062-7070.
[4] Sheu, J.-B., Dynamic relief-demand management for emergency logistics operations under large-scale disasters. *Transportation Research Part E: Logistics and Transportation Review*, 2010. 46(1): p. 1-17.
[5] Huang, M., K.R. Smilowitz, and B. Balcik, A continuous approximation approach for assessment routing in disaster relief. Transportation Research Part B: Methodological, 2013. 50: p. 20-41.
[6] Edrissi, A., M. Nourinejad, and M.J. Roorda, Transportation network reliability in emergency response. *Transportation research part E: logistics and transportation review*, 2015. 80: p. 56-73.
[7] Fiedrich, F., F. Gehbauer, and U. Rickers, Optimized resource allocation for emergency response after earthquake disasters. *Safety science*, 2000. 35(1): p. 41-57.
[8] Rennemo, S.J., et al., A three-stage stochastic facility routing model for disaster response planning. *Transportation research part E: logistics and transportation review*, 2014. 62: p. 116-135.
[9] Cavdur, F., M. Kose-Kucuk, and A. Sebatli, Allocation of temporary disaster response facilities under demand uncertainty: An earthquake case study. *International Journal of Disaster Risk Reduction*, 2016. 19: p. 159-166.
[10] Chen, L. and E. Miller-Hooks, Optimal team deployment in urban search and rescue. *Transportation Research Part B: Methodological*, 2012. 46(8): p. 984-999.
[11] Wohlgemuth, S., R. Oloruntoba, and U. Clausen, Dynamic vehicle routing with anticipation in disaster relief. *Socio-Economic Planning Sciences*, 2012. 46(4): p. 261-271.
[12] Özdamar, L., E. Ekinci, and B. Küçükyazici, Emergency logistics planning in natural disasters. *Annals of operations research*, 2004. 129(1-4): p. 217-245.
[13] Luis, E., I.S. Dolinskaya, and K.R. Smilowitz, Disaster relief routing: Integrating research and practice. *Socio-economic planning sciences*, 2012. 46(1): p. 88-97.
[14] Barbarosoğlu, G. and Y. Arda, A two-stage stochastic programming framework for transportation planning in disaster response. *Journal of the operational research society*, 2004. 55(1): p. 43-53.
[15] Özdamar, L. and O. Demir, A hierarchical clustering and routing procedure for large scale disaster relief logistics planning. *Transportation Research Part E: Logistics and Transportation Review*, 2012. 48(3): p. 591-602.
[16] Yi, W. and A. Kumar, Ant colony optimization for disaster relief operations. *Transportation Research Part E: Logistics and Transportation Review*, 2007. 43(6): p. 660-672.
[17] Ahmadi, M., A. Seifi, and B. Tootooni, A humanitarian logistics model for disaster relief operation considering network failure and standard relief time: A case study on San Francisco district. *Transportation Research Part E: Logistics and Transportation Review*, 2015. 75: p. 145-163.
[18] Najafi, M., K. Eshghi, and W. Dullaert, A multi-objective robust optimization model for logistics planning in the earthquake response phase. *Transportation Research Part E: Logistics and Transportation Review*, 2013. 49(1): p. 217-249.
[19] Huang, K., et al., Modeling multiple humanitarian objectives in emergency response to large-scale disasters. *Transportation Research Part E: Logistics and Transportation Review*, 2015. 75: p. 1-17.
[20] Watkins, C.J. and P. Dayan, Q-learning. Machine learning, 1992. 8(3-4): p. 279-292.
[21] Melo, F.S., Convergence of Q-learning: A simple proof. *Institute Of Systems and Robotics*, Tech. Rep, 2001.
[22] Isfahan Atlas, 2013<http://new.isfahan.ir/Index.aspx?lang=1&sub=105> (Accessed: August, 2016).