Simulation based optimization of a stochastic supply chain considering supplier disruption: An agent-based modeling and reinforcement learning

A. Aghaie\textsuperscript{a,}* \textsuperscript{a}, M. Hajian Heidary\textsuperscript{b}

\textsuperscript{a}Department of Industrial Engineering, K.N. Toosi University of Technology, Pardis Street, Mollasadra Street, Vanaq Square, Tehran, 1999143344, Iran.

\textsuperscript{b}Corresponding Author: Aghaie@kntu.ac.ir, Tel: +982184063363, Fax: +982188674858

Abstract

Many researchers and practitioners in the recent years have been attracted to investigate the role of uncertainties in the supply chain management concept. In this paper a multi-period stochastic supply chain with demand uncertainty and supplier disruption is modeled. In the model, two types of retailers including risk sensitive and risk neutral, with many capacitated suppliers are considered. Autonomous retailers have three choices to satisfy demands: ordering from primary suppliers, reserved suppliers and spot market. The goal is to find the best behavior of the risk sensitive retailer, regarding the forward and option contracts, during several contract periods based on the profit function. Hence, an agent-based simulation approach has been developed to simulate the supply chain and transactions between retailers and unreliable suppliers. In addition, a Q-learning approach (as a method of reinforcement learning) has been developed to optimize the simulation procedure. Furthermore, different configurations for simulation procedure are analyzed. The R-net logo package is used to implement the algorithm. Also a numerical example has been solved using the proposed simulation-optimization approach. Several sensitivity analyzes are conducted regarding different parameters of the model. Comparison of the numerical results with a genetic algorithm shows a significant efficiency of the proposed Q-learning approach.

Keywords Supply chain management, simulation based optimization, reinforcement learning, demand uncertainty, supplier disruption.

1 Introduction

The Importance of uncertainties and consequent costs of ignoring them, caused a shift from deterministic configurations of the supply chain to the stochastic models. One of the most important problems in the stochastic supply chain ordering management is the newsvendor (NV) problem. The basic form of the NV problem consists of a buyer and a seller in which the buyer must decide on the amount of ordering from the seller while demand of the customers is not predetermined. In the basic form, the buyer only has an overall information about the customer demand such as the distribution function. Also the decision is made only at one period. The objective is to optimize the profit of the buyer. Two extensions of the problem have been done by the researchers: the multi-period NV Problem (MNVP), the NV Problem with Supplier Disruption (NVPSD). In the MNVP, the buyer(s) decides on the amount of ordering from the seller(s) at the beginning of each period. The buyer(s) decides on the amount of orders based on the uncertain demands of its customers and the inventory remained from the previous period. In the NVPSD (which is often consisted of one period) the buyer(s) decides on the amount of orders based on the uncertain customer demand and remained fixed capacities of the sellers. In the related literature of the NVPSD it is usually assumed that the network consists of many uncertain sellers and one buyer (e.g. [1], [2], [3]). On the other hand, in the literature of the MNVP, some researchers assumed many buyers and one seller in their network [4]. Thus, inspiring from [4] and the related literature of the NVPSD, a many-to-many relation has been defined in this paper. In the new configuration, each buyer decides on the amount of the order from uncertain seller at the beginning of each time unit. In addition, it is more practical to make decision for a contract period -which consists of several time units while demand varies during each time unit- instead of making decision at the beginning of each time unit. It was not elaborated in the context of the NV problem. Practical applications of the NVPSD specially arise in the decisions regarding the global sourcing. Following example clarifies the importance of the decisions of the buyers in the global sourcing with uncertain suppliers. An automotive component manufacturing had expected to save 4-5 million dollars a year resulting from sourcing of a product from Asia instead of Mexico. Port congestion and chartering aircraft to fly the products from Asia, caused a 20 million dollar loss [5]. This example and other practical applications of sourcing decision -specially when a contract is signed between a retailer and a supplier- highlights the importance of studding sourcing decisions in an uncertain supply chain (the MNVP and the NVPSD). Extension of the NVP to the MNVP or NVPSD makes the problem much more challenging. Suppliers with uncertain and limited capacities, and inventory positions of the retailers add some challenges to the basic NVP. To the best of our knowledge, there is no research on the combination of the MNVP and the NVPSD.
The combined problem is called MNVP. In addition, to avoid shortages, it is assumed that retailers have two options after the realization of the demand in each time unit: buying from a reserved supplier and if the amount of reservation is not sufficient to satisfy the demand, retailers have an option to buy from the spot market [1]. These options are common in the industries such as: semiconductors, telecommunications and pharmaceuticals. Details of the problem are discussed in the section 3. A two-stage decision making is needed to solve the problem in each time unit, in which an order must be issued before the realization of the demand and subsequent decisions regarding the ordering from the reserved supplier and the spot market must be made after the realization.

Solving a large size NVPSD is computationally not tractable ([1], [2], [3]). Also heuristic approaches are common tools to solve the MNVP [4]. Thus it could be concluded that solving the MNVPSD with an exact approach or by using a common optimization software is more difficult. In this regard and to consider autonomous retailers, an agent-based Q-learning is developed and implemented to solve the problem. Followings, the basics of the agent-based modeling and reinforcement learning are introduced.

1.1 Agent-based modeling

Agent-based modeling is a bottom-up approach among different simulation modeling approaches in which agents interact with each other and also with the environment [6]. Agent-based modeling facilitates simulation optimization loop of the related optimization of behavioral parameters [7]. An agent-based simulation model consists of a certain amount of agents and their behaviors which affect on their own and other actions and their environment.

Based on a research, different approaches to develop an ABMS could be divided into four categories [8]: Individual ABMS (agents have a prescribed behavior and there is no interaction between agents and the environment), Autonomous ABMS (agents have autonomous behavior and there is no interaction between agents and the environment), Interactive ABMS (agents have the same behavior as Autonomous ABMS but interaction between agents and environment is possible), Adaptive ABMS (behavior of the agents is same as Interactive ABMS but agents can change their behavior during the simulation). To make an intelligent network of agents, researchers usually add learning feature to their models. In this regard, reinforcement learning has been adopted in our modelling.

1.2 Reinforcement learning

Reinforcement learning (RL) is a machine learning approach and is a proper approach in order to optimize multi agent models [9]. Indeed, a RL algorithm is a learning mechanism to map the situations to actions [10]. In the RL there is a set of states (S), a set of actions (A) and a reward function (R). In the stochastic environments, a stochastic subset of the problem could be handled as a Markov or semi-Markov model [11]. Generally a Markov process is formulated as follow:

\[ \Pr(s_{t+1} = s^* \mid s_t, a_t, \ldots, s_0) = \Pr(s_{t+1} = s^* \mid s_t, a_t) \]

(1)

Above formula shows the memory less characteristic of the Markov process, which explains that the state and reward at time (t+1) only depends on the last time unit (t). RL is an algorithm with an ability to solve decision problems with Markov property. Basically the states which are defined in RL algorithm have to have Markov property, in the case that states don’t have Markov property, RL could represent a good approximation of the solution [10].

One of the most popular methods in order to implement RL and the optimal set of “action-state” is Q-learning (as a model-free algorithm). In this regards, a Q-function must be defined. A Q-function in RL algorithm could be defined as the expected value of the discounted reward gained from a specific set of states and actions:

\[ Q(s, a) = E(\sum_{\zeta=0}^{T-1} \beta^\zeta r_{s, a, \zeta+1} \mid s_t = s, a_t = a) \]

(2)

Since modeling all the dynamics of the system is not possible in most real world problems, usually an estimation of the Q-function is used to model the problem (e.g. by using an iterative Q-learning algorithm). At the end of the learning process, the action with largest value of Q-function is chosen for all the current states. In the section 4 the learning algorithm is described.

Remaining parts of the paper are organized as follows: In the next section related works are reviewed. In the section 3, the mathematical formulation of the problem is presented. In the section 4, based on the
formulation presented in the section 3, agent-based reinforcement learning approach is explained. In the section 5, results of applying the proposed framework on an illustrative example are shown and finally in the last section concluding remarks are presented.

2 Literature Review

The main focus of this research is to analyze the risk behavior of the retailers in the stochastic supply chain using simulation optimization approaches. To design a stochastic supply chain the MNVP is extended with multiple uncertain suppliers and the NVPSD is extended with multiple periods. Additionally, a simulation optimization approach is developed based on a multi agent system. The NV problem is a common problem in the inventory management. Many researchers studied this problem and developed it in different ways. According to the assumptions considered in this paper, related researches of NV problem which considered these two assumptions are reviewed: multi period modeling and unreliable suppliers.

In the past years, some of the researchers developed the newsvendor problem with one retailer and multi unreliable suppliers. There are a few papers regarding the supplier disruption in the newsvendor configuration [12].

Recently, some of the researchers focused more on the NV model with unreliable suppliers. Among them, Ray and Jenamani [2] proposed a one-period NV optimization model with one retailer and many unreliable capacitated suppliers. They solved the problem with a simulation optimization approach using discrete event simulation and genetic algorithm. They asserted that the problem is computationally not tractable by increasing the number of suppliers. Afterwards, Ray and Jenamani [3] proposed a heuristic approach to solve the problem which they developed in their previous work. They suggested that an important future extension of their problem is: “considering multiple periods in the modeling”. Merzifonluoglu and Feng [12] presented another important research regarding developing NV model with the unreliable suppliers. They proposed a heuristic approach to solve a one-period uncapacitated NV model. They suggested to use a risk sensitive (vs risk neutral) modeling. Afterwards, Merzifonluoglu [1] developed the model of Merzifonluoglu and Feng [12], with adding some assumptions such as option contracts. She also modeled the concept of the capacity reservation in the NV model ([13], [14]).

Based on the above researches, our assumptions regarding multiple unreliable capacitated suppliers were adopted from Ray and Jenamani [2], Merzifonluoglu [1] and also option contract assumption was adopted from Merzifonluoglu [1]. As suggested by Ray and Jenamani [3], we extended the problem of ordering from unreliable capacitated suppliers to multiple periods. Followings, related works are presented.

Developing multi period model for newsvendor problem is another extension to the common NV problem. In this regards Bouakiz and Sobel [15] by using utility function, performed a risk analysis on the multi-period NV problem. One of the main parts of the literature (relating to the multi period NV problem), is about the estimation of demand distribution with different approaches. Another main part of the literature is about modeling uncertainties in the NV problem. For example, uncertainty of the supplier capacity [16], uncertainty of the selling price [17], and uncertainty of the demand [4]. Additionally, Kim et al. [4] developed a multi-period NV problem with a distributor and many retailers. Hence, inspiring the extensions of Ray and Jenamani [2] and Merzifonluoglu [1], we mixed their assumptions and developed a multi-period NV model with many retailers and many unreliable capacitated suppliers considering option contracts. In addition it is assumed that retailers have a risk sensitive behavior.

One of the best tools to solve the complex decision making problem such as inventory replenishment problems is simulation optimization. Jalali and Nieuwenhuyse [18] reviewed and classified previous works on the simulation optimization technique in the inventory management. They classified related works into two categories: domain and methodology focused. Based on their classification, domain focused works are mainly contribute on the modeling of the inventory. Works focused on the methodology try to solve a simple problem with a new approach. They did not address agent-based simulation optimization works. Hence in this section those papers which their main focus is on the agent-based simulation optimization are reviewed.

Nikolopoulou and Ierapetritou [19] by using a MILP formulation developed an agent-based simulation optimization. They solved a small scale inventory problem with their proposed SimOpt framework. Kwon et al. [20] developed a hybrid multi-agent case based reasoning approach. A part of the literature surveyed ordering problem in the supply chain using reinforcement learning: [21], [22], [23].
In addition, Jiang and Sheng [24] developed a multi-agent reinforcement learning for a supply chain network with stochastic demand. Kim et al. [25] presented a multi-agent framework -considering a reward function- for an inventory management problem with the uncertain demand and the service level constraint. In the recent years, some papers applied reinforcement learning on the multi-agent simulation framework: [26], [27], [28].

As clarified in the previous sections and to the best of our knowledge, there is no research in the literature which modeled a multi-period newsvendor problem with many to many relationships and uncertain capacitated suppliers. In this research a new multi-agent reinforcement learning approach is developed to solve the model.

3 Problem description

Consider a supply chain with two echelons: retailers and suppliers. Retailers receive demands of customers at the beginning of each time unit and they have to satisfy these demands. In the case of shortage, they have to pay a certain amount of cost. In order to satisfy demands, retailers sign a forward contract with primary suppliers for a set of constant time units (which is called a contract period). In other word, at the beginning of each contract period, retailers have to decide on the amount of order from primary supplier for a contract period. Customer demands and supplier capacities are uncertain. Each supplier could sign a forward and an option contract with two different retailers. Hence, after demand realization, (as suggested by Merzifonluoglu [1]), retailers have two options respectively: 1- order from a secondary supplier up to the reserved capacity, 2- Buying from the spot market (with a spot price which increases with the increase in the excess demand). Indeed, if the forward contract is not enough, retailers could use these options respectively. In order to analyze the effect of the risk attitude in the decisions made by retailers, we assume that one of the retailers is risk sensitive and other retailers are risk neutral. The system is modeled for a certain contract periods (M).

Notations of the model are as follows:

Indices:
- $I$: Index of the retailers, $i \in \{1, \ldots, I\}$
- $J$: Index of the suppliers, $j \in \{1, \ldots, J\}$
- $T$: Index of the time horizon, $t \in \{1, \ldots, T_1, T_1+1, \ldots, T_T\}$

Variables:
- $l_{i,t}$: Inventory position of the retailer $i$ at time $t$
- $\alpha_{i,t}$: determines the risk sensitivity of the retailer $i$ at time $t$ (risk neutral retailers choose $\alpha_{i,t}$ equals to zero, the risk averse retailer chooses negative values, the risk taking retailer chooses positive values; values of $\alpha_{i,t}$ belong to $\{-0.6, -0.4, -0.2, 0.2, 0.4, 0.6\}$).
- $y_{i,j,t}$: The ordering amount from the secondary supplier $j$ by retailer $i$ at time $t$
- $z_{i,t}$: The ordering amount from the spot market by retailer $i$ at time $t$
- $\zeta_{i,t}$: The shortage amount of the retailer $i$ at time $t$

Random variables:
- $D_{i,t}$: The customer demand at time unit $t$, which is satisfied by the retailer $i$ (A random normal variable with mean $\mu_i$ and standard deviation $\sigma_i$)
- $\pi_{j,t}$: Percentage of the capacity of the supplier $j$ loses as a result of happening a disruption event at time $t$
- $x_{i,j,t}$: Ordering amount of retailer $i$ at time $t$ from supplier $j$ before realization of demand (risk attitude of the retailers has an effect on this variable)
- $\omega$: Spot market price (correlated with the amount of the excess demand not satisfied by primary and secondary suppliers)
Parameters

\( c^1_j \): The cost of ordering from the primary supplier \( j \)

\( c^2_j \): The cost of ordering from the secondary supplier \( j \)

\( f_j \): The cost of capacity reservation in the supplier \( j \) (as a secondary supplier)

\( p \): The revenue of selling product to customers

\( h \): The holding cost paid by retailers per product

\( \theta \): The shortage cost of retailers per product

\( Cap_{i,j}^1 \): A fixed nominal capacity dedicated to the retailer \( i \) by the supplier \( j \) during the contract period (which resets at the beginning of each time unit)

\( Cap_{i,j}^2 \): A fixed nominal capacity of the supplier \( j \) which could be reserved by the retailer \( i \) at the beginning of the contract period for a contract period with “\( g \)” unit times.

An important part of the model is the effect of the risk behavior of the risk sensitive retailer on the amount of his/her order as a primary contract. Because of the uncertainty of the demand, the risk neutral retailer places an order from the primary supplier based on \( N(\mu_i, \sigma_i) \), and the risk sensitive retailer places an order based on \( N((1 \pm \alpha_{i,t})\mu_i, (1 - \alpha_{i,t})\sigma_i) \). In other words, \( \alpha_{i,t} \) is the coefficient of the risk. For the risk neutral retailers \( \alpha_{i,t} = 0 \). We defined certain amounts of \( \alpha_{i,t} \) in this paper: \( \alpha_{i,t} \in \{-0.6, -0.4, -0.2, 0.2, 0.4, 0.6\} \). The risk sensitive retailer, uses a wider or tighter distribution than demand. For example suppose that the demand follows a normal distribution with mean 100 and standard deviation 20. Results of the numerical simulation showed that a retailer with extremely risk averse behavior (\( \alpha=0.6 \)) approximately in %95 of time, orders above the realized demand and a retailer with extremely risk taking behavior (\( \alpha=-0.6 \)) in %95 of time, orders under the realized demand. The risk attitude of the retailer towards the uncertain demand is depicted in the figure 1(a).

Please Insert Figure 1(a).

Chromosomes used in order to make a decision in different time units of contract periods are depicted in the figure 1(b). A simple numerical analysis (using 1000 random numbers) shows that the probability of ordering greater than the demand in different values of \( \alpha \) are as follows (values in parenthesis show the related probabilities):

\( \alpha = -0.6 \) (0.054), \( \alpha = -0.4 \) (0.253), \( \alpha = -0.2 \) (0.437), \( \alpha = 0.2 \) (0.557), \( \alpha = 0.4 \) (0.763), \( \alpha = 0.6 \) (0.952).

Additionally, the behavior of the risk sensitive retailer affects on the amount of the reserved capacity. In other words, the risk averse retailer prefers to order more from the primary supplier and less from the secondary supplier. The behavior of a risk averse retailer is defined as: large primary contract and small secondary contract. Likewise, the behavior of a risk taking retailer is: small primary contract and large secondary contract. These behaviors are defined using two parameters: \( \alpha \) (introduced before) and \( \beta \) (a percentage of \( Cap_{i,j}^2 \) that a retailer reserves in the secondary supplier). In the followings, details of the relations between \( \alpha \) and \( \beta \) are explained.

If the risk sensitive retailer decides to order based on \( \alpha = -0.6 \), the value of the parameter \( \beta \) is equal to 1. Likewise, for other values of \( \alpha \), the value of \( \beta \) would be: \( \alpha = -0.4 \) \( (\beta=0.8) \), \( \alpha = -0.2 \) \( (\beta=0.6) \), \( \alpha = 0.2 \) \( (\beta=0.4) \), \( \alpha = 0.4 \) \( (\beta=0.2) \), \( \alpha = 0.6 \) \( (\beta=0) \). For the risk neutral retailer \( (\alpha=0) \), the value of \( \beta \) is equal to 0.5.

As mentioned before, in this paper we are looking for the best decision of the risk sensitive retailer among other risk neutral retailers (agents). We define \( i^* \) as the index of the risk sensitive retailer. The objective function is considered as the maximization of the profit of retailer \( i^* \). Thus the profit function (consists of selling revenue and costs: holding cost, shortage cost, cost of purchasing and cost of reserving the capacity) is as follow:

\[
\psi_i = \sum_t pD_{i,t} - \sum_t \sum_j \beta_{i,j}Cap_{i,j}^2 f_j - \sum_t \sum_j c^1_j x_{i,j,t} - \sum_t \sum_j c^2_j y_{i,j,t} - \sum_t \theta z_{i,t} - \sum_t \alpha z_{i,t} - \sum_t h l_{i,t}
\] (3)
As a result of the disruption, in each time unit, available capacities of the suppliers ($Cap_{i,j,t}^1$, $Cap_{i,j,t}^2$) may be less than their nominal capacities. At the beginning of each contract period i.e. $(t \leq g)$, retailers have to decide on the amount of the forward contracts based on the updated capacities of the suppliers.

Let define $\phi_{i,j,t}^1$ and $\phi_{i,j,t}^2$ as two binary variables (respectively) relating to the forward/option contract of the supplier $j$ with the retailer $i$ at time $t$ ($t, t' \in T$).

$$Cap_{i,j,t}^1 = \phi_{i,j,t}^1 (1 - \pi_{j,t}) Cap_{i,j}$$

$$\sum_{i} \phi_{i,j,t}^1 = 1, \quad \forall j$$

$$Cap_{i,j,t}^2 = \phi_{i,j,t}^2 (1 - \pi_{j,t}) Cap_{i,j}$$

$$\sum_{i} \phi_{i,j,t}^2 = 1, \quad \forall j$$

$$\phi_{i,j,t}^1 + \phi_{i,j,t}^2 = 1, \quad \forall i, j$$

$$\phi_{i,j,t}^1 \phi_{i,j,t}^2 = \phi_{i,j,t}^2, \quad \forall t \in \left[ \frac{g}{g} \right] + 1 \quad \text{if } \phi_{i,j,t}^2 = 1$$

$$\phi_{i,j,t}^1 \phi_{i,j,t}^2 = \phi_{i,j,t}^2, \quad \forall t \in \left[ \frac{g}{g} \right] + 1 \quad \text{if } \phi_{i,j,t}^2 = 1$$

Above formulas ensure that a retailer only orders from a specific supplier (as a primary supplier) and reserves capacities in a different supplier (as a secondary supplier) during each contract period.

Based on $\phi_{i,j,t}^1$, the value of $x_{i,j,t}$ could be determined as follow:

$$0 \leq x_{i,j,t} \leq M \phi_{i,j,t}^1$$

Let define $\eta_{i,t}$ as the amount of satisfied order of the retailer $i$ from primary suppliers,

$$\eta_{i,t} = \sum_{j} \min(x_{i,j,t}, \text{Cap}_{i,j,t}^1)$$

In the case that $x_{i,j,t} < \text{Cap}_{i,j,t}^1$, suppliers add the remained capacity to their capacities as a secondary supplier ($\text{Cap}_{i,j,t}^2$).

Let define $t_{i,t} = (D_{i,t} - \eta_{i,t} - I_{i,t-1})$ as the unsatisfied amount of order of the retailer $i$ at time $t$ from primary the supplier ($\langle x \rangle^n$ is equal to $\max(x, 0)$).

Let define $\kappa_{i,t}$ as the amount of excess order of the retailer $i$ from secondary suppliers (in each time unit, primary suppliers add their remained primary capacity to their secondary capacity):

$$\kappa_{i,t} = \sum_{j} \min(t_{i,t}, \beta_{i,t} \text{Cap}_{i,j,t}^2 + \sum_{i} (\text{Cap}_{i,j,t}^1 - x_{i,j,t})^+)$$

Therefore,

$$0 \leq \sum \gamma_{i,j,t} \leq \kappa_{i,t}$$

Let define $\tau_{i,t} = (D_{i,t} - t_{i,t} - \kappa_{i,t})$ be unsatisfied order of the retailer $i$ which is unsatisfied at time $t$ (after receiving products from primary and secondary suppliers).

It is assumed that retailers will compare the cost of shortage with the cost of purchasing from the spot market and then decide on the amount of order from the spot market, indeed retailer agents examine different values for $\xi_{i,t} \in (0, 0.1, 0.2, ..., 1)$). Therefore the amount of the shortage will be: $z_{i,t} = \xi_{i,t} \times \tau_{i,t}$ and therefore the amount of order from the spot market will be: $z_{i,t} = (1 - \xi_{i,t}) \tau_{i,t}$.

The equation of on-hand inventory balance is as follow:

$$I_{i,t} = (I_{i,t-1} + \eta_{i,t} + \kappa_{i,t} + z_{i,t} - D_{i,t})^+$$

(14)
On-hand inventory is used as the states in the agent based model \( I_{i,0} = 0 \). Previous works in the area of NVPSD or MNVP, used a heuristic or a metaheuristic method to solve the problem. They also discussed about the computational complexity of the problems specially in the large sizes. Additionally, as mentioned before, the problem in this paper is MNVPSD and thus it is more complex than NVPSD or MNVP. Hence, an intelligent approach to solve the problem is necessary. Above formulations are modeled in a multi-agent simulation software (Netlogo 5.3.1) and then by using R-Netlogo package [29], optimization is done in cooperation with the simulation procedure. Detailed discussions are presented in the section 3.1.

3.1 Agent-based modeling

In this paper we are going to analyze a subsystem (among several subsystems of SCM such as transportation, financial, etc.) of the SCM as an agent-based system. The overall agent-based system (consists of the relations between agents, states and rewards) is depicted in the Figure 2. In this system, each agent is responsible for making decisions about the amount of forward and option contracts (autonomously) by interacting with other agents. The goal is to find the best behavior of the risk sensitive retailer during several contract periods with regards to the forward and option contracts, and based on the profit function. In the figure 2, based on the variables introduced in the previous section \((x, y, z)\), different flows (orders and goods) of the system are depicted. As explained in the above formulas, flows “\(y\)” and “\(z\)” are happened when “\(x + I_{i,j} < d\)”, hence we depicted \(y\) and \(z\) with dashed arrows. In addition to the direct arrows (orders), reverse arrows show the flow of goods towards retailers. In our agent based supply chain, environmental uncertainties consist of customer demand and supplier disruptions. As shown in the figure 2, each agent takes an action based on the environment state.

\[
\text{Figure 2.}
\]

The overall process occurred for each retailer (which orders from a primary supplier, and a secondary supplier or the spot market), is shown in the above figure. In the above agent-based model (considering the RL algorithm), agents are autonomous and interact with each other to satisfy constraints and to attain the optimal solution for the objective function. It is notable that a supply chain consists of different mechanisms but the main focus of this paper is on the ordering decisions of the risk sensitive retailer during a certain amount of contract periods.

Customer agents send their demands at the beginning of each time unit to the retailers, and retailers set their amount of fixed orders (primary contract) at the beginning of each contract period. If the resulting state (inventory position of retailer) satisfies the uncertain demand of the customer, \(y\) and \(z\) will be equal to zero. Otherwise, a retailer sends an order to the secondary supplier (reserved at the beginning of the current contract period). If again, reserving capacity does not satisfy the remained demand, a cost-benefit tradeoff is done to decide whether to order from the spot market or loss the excess demand and pay a certain amount of shortage cost. Supplier agents (when act as a primary or secondary supplier) are exposed to the disruption and may lose some parts of their capacity as a result of disruption. When supplier agents act as the secondary supplier, if they promised to reserve their capacity for a certain retailer, they satisfy that part of the excess demand that not exceeded from the predetermined capacity. Spot market agents could satisfy all the excess demands which are requested from them (i.e. their capacities are infinite) but they set their price according to the amount of excess demand requested from them. The correlation between the spot price and demand is explained in the section 5.

4 Simulation-optimization (SimOpt) approach

In the previous section, the procedure of decision making in the problem were discussed. In this section, in order to present a solution approach, all the procedures of decision making are mapped to a simulation-optimization algorithm.

4.1 Simulation procedure

First of all, the simulation procedure is explained. The overall procedure of the simulation is depicted in the figure 3. Note that in the simulation procedure, formulas (4)-(14) defined in the section 3, will be considered.

\[
\text{Figure 3.}
\]

Please Insert Figure 3.

TC means termination condition, \(S1\) denotes the primary supplier, \(S2\) denotes the secondary supplier. The simulation of the agent-based model is done by an agent-based simulation software (Netlogo 5.3.1). The
optimization part of the SimOpt algorithm is coded in R-studio in cooperation with Netlogo (R-Netlogo) which is discussed in the section 4.3.

4.2 Simulation based estimation
According to Liu et al. [30], we have performed some analyzes on the number of replications of the simulation. Indeed, in a two stage decision making occurs in each time unit (a decision is made before realizing the demand and a decision is made to order from reserved supplier and spot market) for each first stage decision variable, “R” replications are run and the profit function will be calculated based on the results of the replications. In other words, an estimation of the reward function related to \( x \) (i.e. \( f(x, \varepsilon) \) ) could be calculated as: 

\[
\bar{F}(x) = \frac{1}{R} \sum_{r=1}^{R} f(x, \varepsilon_r).
\]

Indeed, in the SimOpt algorithm, a same set of realizations (for demand and all other stochastic parameters) are used in each iteration toward the optimization. It was done by using R sets of seeds to generate different sequences of stochastic parameters in the replications. In this regard, four sample sizes are defined for the number of replications: 5, 10, 20, 50 and those are labeled with 1-4. Hence, we call four simulation optimization algorithms as: SimOpt-1-4. In the following section the RL algorithm is described.

4.3 RL algorithm
4.3.1 States
As mentioned before, although the states of the system have not Markov property, temporal dynamics (or dynamics happened in each step of the Markov process) make it possible (and give an appropriate approximations) to estimate the reward in the next step based on the current states and actions. States for different agents are defined as: 1- customer agents: \( S^i_t \), amount of the unsatisfied demand at time unit \( t \), 2- retailer agent: \( S^r_t \), the inventory position at time \( t \), 3- supplier agent: \( [S^{s,1}_t, S^{s,2}_t] \), the remained capacity at time \( t \) and remained reserved capacity at time \( t \), 4-spot market agent: \( S^{sp}_t \), the inventory position of the spot market. It is assumed that the capacity of the spot market is infinite, thus the state of the spot market always equals to infinite. Therefore, system state could be written as:

\[
S(t) = [S^c_t, S^r_t, [S^{s,1}_t, S^{s,2}_t], S^{sp}_t].
\]

In order to control the dimension of the above vector, a common approach is used to consider a limited set of cases for each member of the vector: e.g. for \( S^c_t \), \((-\infty, -1000) \equiv 1, [-1000, -500) \equiv 2, ...

It is worthwhile to note that in the simulation process (as mentioned in the previous section), fixed seeds are used in order to generate random numbers (specially regarding sampling from random variables), hence in each run of the simulation, the initial conditions will be same as other runs.

4.3.2 Reward
The reward function at time unit “\( t \)” is equivalent to the profit gained by the risk sensitive retailer at time unit \( t \). Therefore the reward function could be defined as:

\[
r_t = PD_{r,t} - \sum_j \beta_j Cap_{r,j}^2 f_j - \sum_j c_j x_{r,j,t} - \sum_j c_j^2 y_{r,j,t} - \sum_j \zeta_{r,j,t} \theta - \omega_{r,j,t} - h l_{r,t}.
\]

And ideally based on (2) the Q-function could be obtained. But since the values of the revenue and costs for the future periods could not be calculated, thus usually a Q-learning algorithm is used to estimate the value of the function. It is described in the forthcoming sections.

4.3.3 Actions
In the agent-based framework, for each agent a set of state-actions are defined. The states were explained before. In this section actions of different agents are explained: 1- Customer agents: demand based on the normal distribution, 2- Retailer agents: orders from suppliers, values of \( x \) and \( y \), 3-supplier agents: amount of satisfied demand as a primary and secondary supplier, 4- the spot market agent: satisfied excess demand of the retailer. A customer demand is defined as a random normal variable. The value of \( x \) depends on the risk attitude of the retailer (which was explained in section 3). The value of \( y \) is determined by the learning mechanism. The value of the supplier action depends on the constraints explained in the section 3. The value of an action of the spot market is equal to the amount of excess demand requested from the spot market. A decision between shortage and ordering from spot market is determined by learning mechanism.

4.3.4 Q-learning algorithm
In this section the proposed Q-learning algorithm is presented to estimate the Q-function. One of the most important challenges about the performance of reinforcement learning, is the efficient exploitation and exploration. In the initial steps more explorations are needed and in the further steps more exploitations must be occurred. The exploration and exploitation are defined in the algorithm by using the parameter $\Omega$.

**Inputs:** parameters $\lambda$, $\chi$, $\delta$

**Output:** Optimal action

1. **Initialization:** states and actions of the system, $Q(s,a)=0 \; \forall s,a$
2. **While** stopping criterion (iteration number > Max Iteration) is not met **do**
3. **Set the initial values for states**
4. Define the parameter $\Omega$ as: $\Omega = \frac{\text{Iteration number}}{\text{Max Iteration}}$
5. Run the simulation (figure 3)
6. Generate a random number between 0 and 1
7. If the random number is greater than $\Omega$, Select an action vector with maximum amount of $Q(s,a)$ among the current $Q(s,:)$, Otherwise take a random action. Calculate consequent reward ($r_{t+1}$)
8. Update $Q(s,a)$: $Q(s,a)+\chi \times (r_{t+1}+\lambda \times \max_a Q(s',a') - Q(s,a))$
9. Do action $a$
10. Update state set ($S(t)$)
11. **end**

After doing an action, the system enters into a new state. As a result of the performing Q-learning algorithm, $Q(s,a)$ matrix is formed for each set of state-action. The convergence of the RL algorithms was surveyed by a wide range of the researchers. In the above learning algorithm, $\chi$ is the learning coefficient. It is a usual coefficient and has a performance like the other similar uses of learning coefficients (e.g. same as the learning coefficient in the exponential smoothing forecasting). Indeed it gives a weight to the old estimations in contrast with the recent results. The stopping criteria of the problem is considered as the maximum iteration.

As showed in the figure 3, the above algorithm is a part of the SimOpt algorithm. Indeed in each time unit, all the states for the action made at the beginning of the contract period are calculated. At the end of the contract period, the best action is selected and again the simulation will run and states are calculated until the next contract period. The procedure will stop whenever the stopping criterion is met. In the next section, a numerical example is solved by using the proposed Q-learning algorithm and the results are compared with a genetic algorithm based SimOpt.

**5 Numerical example**

In this section results of the implementation of the proposed Q-learning algorithm are compared with another SimOpt algorithm in which Q-learning changed to a common genetic algorithm. Generally, we adopted our data from Merzifonluoglu [1]. Because of some additional assumptions in this paper in contrast with the base paper, some parts of data have been modified. Additional assumptions of our model are: multiple retailers, multiple periods, and time based disruptions. Details of the numerical example are as follows:

| Parameter | Value |
|-----------|-------|
| $D_{ij}$ | $N(1000, 100)$ |
| $\pi_{ij}$ | $N(\mu, \sigma^2)$ |
| $\mu_{ij}$ | $U(0.01, 0.03)$, $U(0.0001, 0.003)$ |
| $\sigma_{ij}$ | $N(250, 40)$ |
| $\omega$ | $U(196, 198), 165$ |
| $c_j$ | $40$ |
| $f_j$ | $300$ |
| $h$ | $10$ |
| $\theta$ | $10$ |
| $Cap^1_j$ (fixed capacity of primary suppliers) | $1000$ |
| $Cap^2_j$ (fixed possible reservation capacity) | $200$ |
Additionally, the probability of the disruption is considered as a uniform distribution between [0.01, 0.05]. The disruption effect (or the length of the disruption) is assumed as a uniform distribution between [0, 2] time units. The maximum number of disrupted suppliers is assumed as a uniform distribution between [0, J]. Same as Merzifonluoglu [1] it is assumed that the demand and the spot price are correlated with parameter $\rho$ ($\rho > 0$).

This coefficient of the correlation is assumed equal to 0.2. Different problem instances are defined based on the common NV problem. Problem instances are numbered with the number of suppliers and retailers. The basic problem instance in this paper is: NV10-10 in which the first number shows the number of suppliers and the second one shows the number of retailers. The number of contract periods and the number of time units in each contract period are: 20 and 11 respectively.

Values for $\lambda$, $\chi$ and $\delta$ by using the simulation were determined as 0.3, 0.2 and 0.4 respectively. Results were obtained by a PC with Intel(R) Corei7, 3.1 GHz CPU and 6 GB RAM.

As discussed in the section 4.2., we have defined four SimOpt algorithms with different replication numbers in the simulation procedure. Table 1 shows the results of applying these algorithms on the problem.

Please Insert Table 1.

Results showed that SimOpt-3 is the most proper SimOpt algorithm in terms of the accuracy and time. Thus in the remained parts of the paper, we only discuss about the results given from SimOpt-3.

The result of applying the algorithm (for 100 iterations) on the problem NV10-10-1 is depicted in the figure 4.

Please Insert Figure 4.

As shown in the figure 4, the algorithm converges to a near optimal profit of the risk sensitive retailer in 200 iterations. The best profit resulted from the applying the algorithm on the problem is 20028766.

To show the efficiency of the proposed RL algorithm, results are compared with another popular metaheuristic based on the simulation procedure. Genetic Algorithm (GA) is a meta-heuristic and evolutionary algorithm which has been used in the literature to optimize many complex problems. It works with some procedures like mutation and crossover which was originally inspired from the genetic science. Results of the SimOpt-RL are compared with a simulation based Genetic Algorithm (SimOpt-GA) applied on the problem. Hence, we used GA (instead of RL) to optimize the simulation procedure explained in section 4.1. The GA used in this paper is same as the algorithms used by the related works ([9], [2], [21]).

We defined NV10-10-1 as the problem with 10 suppliers and 10 retailers (i.e. 1 risk sensitive and 9 risk neutral retailers) which retailers have an option to buy from spot market. The problem NV10-10-2 is defined as a problem in which we neglect spot market option.

In addition, as proposed by Liu et al [30], to show the efficiency of the proposed SimOpt-RL algorithm, the results of the SimOpt are compared with a case in which all stochastic parameters are equal to their Expected Values, called Expected Value Method (EVM). Table 2 shows the comparisons.
Results show the impact of the correlation (decreasing), number of contract periods (increasing) and number of time units (decreasing) on the reward values. Additionally, Gap #1 and #2 represent the gap between best values of the SimOpt-RL and the SimOpt-GA (respectively) with the best value of the EVM. Gap #3 and #4 show the gap between the average values of the SimOpt-RL and the SimOpt-GA (respectively) with the average value of EVM. Also, table 3 shows the effect of different disruption probabilities on the objective function and fill rate (profit values in tables 3 and 4 were scaled out as previous tables).

A sensitivity analysis is done regarding the effect of the different values for deviation of the parameters: standard deviation of the demand (σ), standard deviation of the disruption effect (σ^ε) and standard deviation of the spot price (σ^σ). Table 4 shows different cases defined for the sensitivity analysis.

Table 5 shows the sensitivity results regarding the different values for the mentioned cases. Results show the decreasing effect of the wider deviations on the values of the profit and fill rate.

In the remained part of this section we discuss the resulted risk behavior of the risk sensitive retailer according to the best solution obtained.

Figure 5 shows the accumulated profit during 20 contract periods of the best solutions of two algorithms in NV10-10-1 problem.

Based on analyzes presented in tables 2, 3 and 5 and figures 4 and 5 the efficiency of the proposed SimOpt-RL algorithm was shown. Therefore in the remaining part of this section we discuss about the detailed results of the RL algorithm in different cases of the NV10-10-1. Based on the cases introduced in this section, consider following NV10-10-1 Problem Instances (NV10-10-1-PI) as shown in the table 6.

Figure 6 shows the results of the best and average solutions of the SimOpt-RL for these four problem instances.

According to the results shown in the figure 6, it could be concluded that decisions related to the risk attitude of the risk sensitive retailer have an important impact on the profit (reward) function. Additionally, more deviations of the demand and disruption intensity, result in more risk averse behavior. Furthermore, demand deviation has a greater effect on the risk averseness of the retailer rather than disruption intensity deviation. In the first two cases with high demand deviations, the retailer shows an extremely risk averse behavior in approximately 30% of the time, instead, in the last two cases with lower demand deviations the retailer is extreme risk averse in 5% of the time. Extreme risk taking behavior only obtained by the lower demand and disruption deviations. These results could help a decision maker in an uncertain environment (in both sides of the supply chain) to make a decision with an acceptable average reward.

6 Conclusions

The importance of decision making in an uncertain supply chain leads the researchers to develop intelligent approaches to solve the complex problems in an efficient manner. The NV problem is a popular problem which has been extended in many different ways in the past years. But, in the recent years, the NV problem with multiple unreliable suppliers is a type of problem which has been received many attentions: [2], [3], [12], [1], [13], [14]. The complexity of the resulted problem forced the researches to use heuristic or intelligent approaches. The main idea of our model was derived initially from previous mentioned works. A configuration proposed by Merzifonluoglu [1] consists of one retailer and many suppliers which are subjected to disruptions. In this configuration, retailers sign forward and option contracts before demand realization and can buy products from the spot market after the realization. These options are common in the industries such as: semiconductors, telecommunications and pharmaceuticals. In this paper a new model was developed based on this configuration. In addition to demand uncertainty and supplier disruptions, we developed a multi-period, multi-agent model with many-to-many relations between risk
sensitive retailers and capacitated suppliers. Additionally a reinforcement learning method (as an 
optimization approach) was presented to solve it. Different simulation configurations (with different 
realization number) were examined on different scales of the problem. Results showed an acceptable 
performance of the SimOpt algorithm in contrast with the non-stochastic algorithm. Moreover, results of 
the SimOpt-RL was compared with a SimOpt algorithm based on the genetic algorithm. Several sensitivity 
analyzes were done regarding different parameters (including number of contract periods, number of time 
units in each contract period, standard deviations of demands and disruptions). Additionally, details of the 
decisions were obtained based on a sample problem (NV10-10-1). For the future studies, considering 
multiple products in the problem would be an interesting idea and more challenging design. In addition, 
considering negotiation process and transportation assumptions are the other suggestions in order to 
extend the work presented in this paper.

Abbreviations

NV  
Newsvendor

NVP  
Newsvendor Problem

NVPSD  
Newsvendor Problem with Supplier Disruption

MNVPSD  
Multi-Period Newsvendor Problem with Supplier Disruption

ABMS  
Agent-based Modeling and Simulation

RL  
Reinforcement Learning

SimOpt  
Simulation Optimization

SimOpt-RL  
Simulation Optimization based on Reinforcement Learning

SimOpt-GA  
Simulation Optimization based on Genetic Algorithm

SimOpt- X  
Simulation Optimization approach with 5, 10, 20, 50 replications in each 
simulation run, X \in \{1, 2, 3, 4\}.

NV10-10-1  
A newsvendor problem with 10 suppliers and 10 retailers (i.e. 1 risk sensitive 
and 9 risk neutral retailers) in which retailers have an option to buy from 
the spot market.

NV10-10-2  
A newsvendor problem with 10 suppliers and 10 retailers (i.e. 1 risk sensitive 
and 9 risk neutral retailers) in which the spot market option is neglected.

NV10-10-1-PI, X  
Different Problem Instances defined based on the NV10-10-1, X \in \{1, 2, 3, 4\}.

EVM  
Expected Value Method

References

[1] Merzifonluoglu Y., “Risk averse supply portfolio selection with supply, demand and spot market volatility”, Omega 57, pp. 40–53 (2015)

[2] Ray P., Jenamani M., “sourcing under supply disruption with capacity-constrained suppliers”, Journal of Advances in Management Research, 10: 2, pp. 192 – 205, (2013).

[3] Ray P., Jenamani M., “sourcing decision under disruption risk with supply and demand uncertainty: A newsvendor approach”, Annals of Operations Research, 237:1, pp. 237–262, (2016).

[4] Kim G., Wu K., Huang E., “Optimal inventory control in a multi-period newsvendor problem with non-stationary demand”, Advanced Engineering Informatics, 29: 1, pp. 139–145, (2015).

[5] Chopra S., Meidt P., “Supply chain management: strategy, planning and operation”, Pearson, Sixth edition, USA, (2016).

[6] Chiacciio F., Pennisi M., Russo G., Motta S., Pappalardo F., “Agent-based modeling of the immune system: NetLogo, a promising framework”, BioMed Research International, 2, pp. 1-6, (2014).

[7] Humann J., Madni A.M., “Integrated Agent-based modeling and optimization in complex systems analysis”, Procedia Computer Science, 28, pp. 818–827 (2014).

[8] Macal C.M., “everything you need to know about agent-based modelling and simulation”, Journal of Simulation, 10, pp. 144–156, (2016).

[9] Avci M.G., Selim H., “A Multi-objective, simulation-based optimization framework for supply chains with premium freights”, Expert Systems with Applications, 67, pp. 95–106 (2017).

[10] Sutton R.S., Barto A.G., “Reinforcement Learning: An Introduction”, MIT press, Cambridge (1998).

[11] Gosavi A., “Reinforcement learning for long-run average cost”, European Journal of Operational Research, 155, pp. 654–674 (2004).
[12] Merzifonluoglu Y., Feng Y., “Newsvendor problem with multiple unreliable suppliers”, International Journal of Production Research, 52:1, pp. 221-242 (2014).
[13] Merzifonluoglu Y., “Impact of risk aversion and backup supplier on sourcing decisions of a firm”, International Journal of Production Research, 53: 22, pp. 6937-6961 (2015).
[14] Merzifonluoglu Y., “Integrated demand and procurement portfolio management with spot market volatility and option contracts”, European Journal of Operational Research, 258: 1, pp. 181–192 (2017).

Biographies

Abdollah Aghaie, is a Professor of Industrial Engineering at K. N. Toosi University of Technology in Tehran, Iran. He received his BSc from Sharif University of Technology in Tehran, MSc from New South Wales University in Sydney and Ph.D. from Loughborough University in the U.K. His main research interests are in Modeling and Simulation, Supply chain management, Social Networks, Knowledge Management and Risk Management.

Mojtaba Hajian Heidary, is a PhD student of Industrial Engineering at K.N. Toosi University of Technology, department of Industrial Engineering, Tehran, Iran. His main research interests concern Supply chain management, Computer simulation.
Figure Captions

Figure 1(a). Different risk attitude of the risk sensitive retailer
Figure 1(b). Procedure of decision making for two types of the retailers
Figure 2. The overall agent-based model (Up) and a schematic of the interactions in a RL algorithm (Down)
Figure 3. The overall simulation procedure
Figure 4. The progress of the learning through the SimOpt-3 algorithm (NV10-10-1)
Figure 5. Comparing best solutions of RL and GA SimOpt algorithm regarding accumulated rewards in NV10-10-1
Figure 6. Risk attitude of the risk sensitive retailer in the best (left) and average (right) solutions of the SimOpt-RL in different problem instances

Table Captions

Table 1. Results of the algorithms with different number of replications ($\times 10^6$)
Table 2. A comparison between different approaches to solve NV10-10 problem ($\times 10^6$)
Table 3. Sensitivity analysis of different disruption probabilities
Table 4. Different cases for sensitivity analysis
Table 5. Sensitivity results of $\sigma^\uparrow$, $\sigma^\downarrow$, $\sigma^\uparrow\downarrow$
Table 6. Different problem instances of NV10-10-1 problem
Figures

**Figure 1(a). Different risk attitude of the risk sensitive retailer**

| Risk sensitive retailer | $\alpha_1=0.6$ | $\alpha_2=0.6$ | $\alpha_M=0.6$ |
|-------------------------|----------------|----------------|----------------|
| $\alpha_1$             | 0.4            | 0.4            | 0.4            |
| $\alpha_2$             | 0.2            | 0.2            | 0.2            |
| $\alpha_3$             | -0.2           | -0.2           | -0.2           |
| $\alpha_4$             | -0.4           | -0.4           | -0.4           |
| $\alpha_5$             | -0.6           | -0.6           | -0.6           |

Risk neutral retailers

| Risk neutral retailers | $\alpha=0$ | $\alpha=0$ | $\alpha=0$ | $\alpha=0$ | $\alpha=0$ |
|------------------------|------------|------------|------------|------------|------------|
| $\alpha_1$             | 0          | 0          | 0          | 0          | 0          |
| $\alpha_2$             | 0          | 0          | 0          | 0          | 0          |

**Figure 1(b). Procedure of decision making for two types of the retailers**
Figure 2. The overall agent-based model (Up) and a schematic of the interactions in a RL algorithm (Down)
Figure 3. The overall simulation procedure
Figure 4. The progress of the learning through the SimOpt-3 algorithm (NV10-10-1)
Figure 5. Comparing best solutions of RL and GA SimOpt algorithm regarding accumulated rewards in NV10-10-1
Figure 6. Risk attitude of the risk sensitive retailer in the best (left) and average (right) solutions of the SimOpt-RL in different problem instances.
Table 1. Results of the algorithms with different number of replications ($\times 10^6$)

| NV instances | Best | Avg | Sec | Best | Avg | Sec | Best | Avg | Sec | Best | Avg | Sec |
|--------------|------|-----|-----|------|-----|-----|------|-----|-----|------|-----|-----|
| NV5-5        | 16.988 | 16.550 | 27  | 17.998 | 17.535 | 31  | 20.019 | 19.504 | 69  | 26.082 | 25.411 | 119 |
| NV10-5       | 17.049 | 16.540 | 61  | 18.046 | 17.507 | 91  | 20.040 | 19.442 | 146 | 26.021 | 25.244 | 357 |
| NV10-10      | 17.000 | 16.782 | 119 | 18.010 | 17.778 | 190 | 20.029 | 19.771 | 295 | 26.086 | 25.750 | 751 |
| NV15-10      | 17.022 | 16.662 | 174 | 18.031 | 17.650 | 281 | 20.050 | 19.626 | 451 | 26.106 | 25.554 | 1126 |
| NV20-20      | 17.039 | 16.775 | 250 | 18.043 | 17.763 | 385 | 20.050 | 19.739 | 585 | 26.072 | 25.667 | 1456 |
| NV50-20      | 17.148 | 16.749 | 368 | 18.156 | 17.733 | 560 | 20.171 | 19.701 | 878 | 26.217 | 25.606 | 2259 |
| NV50-50      | 17.066 | 16.858 | 513 | 18.081 | 17.861 | 847 | 20.111 | 19.866 | 1324 | 26.202 | 25.882 | 3341 |
| NV100-50     | 17.258 | 17.039 | 708 | 18.273 | 18.042 | 1156 | 20.304 | 20.047 | 1755 | 26.397 | 26.063 | 4464 |
| NV100-100    | 17.198 | 16.754 | 953 | 18.203 | 17.733 | 1498 | 20.212 | 19.690 | 2376 | 26.239 | 25.562 | 6002 |
| p   | Contract period | Time units | SimOpt-RL Best | SimOpt-RL Avg | SimOpt-RL sec | SimOpt-GA Best | SimOpt-GA Avg | SimOpt-GA sec | EVM Best | EVM Avg | EVM sec | Gap1% | Gap2% | Gap3% | Gap4% |
|-----|-----------------|------------|---------------|---------------|--------------|---------------|---------------|--------------|-----------|---------|---------|-------|-------|-------|-------|
|     |                 |            | SimOpt-RL Best | SimOpt-RL Avg | SimOpt-RL sec | SimOpt-GA Best | SimOpt-GA Avg | SimOpt-GA sec | EVM Best | EVM Avg | EVM sec | Gap1% | Gap2% | Gap3% | Gap4% |
| NV10-10-1 | 0.2 | 5 | 5 | 15.892 | 15.602 | 91 | 15.043 | 14.728 | 166 | 15.895 | 15.618 | 0.02 | 5.36 | 0.10 | 5.70 |
|          | 10 | 16.810 | 16.353 | 130 | 16.019 | 15.109 | 235 | 16.827 | 16.358 | 0.10 | 4.80 | 0.03 | 7.64 |
|          | 20 | 17.957 | 17.027 | 234 | 16.577 | 15.758 | 413 | 17.973 | 17.038 | 0.09 | 7.77 | 0.07 | 7.52 |
| 5     | 10 | 17.777 | 17.454 | 152 | 16.985 | 16.088 | 267 | 17.803 | 17.460 | 0.15 | 4.60 | 0.03 | 7.86 |
| 10    | 18.800 | 18.318 | 217 | 18.215 | 16.610 | 386 | 18.808 | 18.344 | 0.04 | 3.15 | 0.14 | 9.45 |
| 20    | 18.937 | 17.908 | 390 | 18.160 | 16.543 | 715 | 18.962 | 17.922 | 0.13 | 4.23 | 0.08 | 7.69 |
| 0.4   | 5   | 19.818 | 19.562 | 203 | 18.912 | 18.488 | 365 | 19.823 | 19.577 | 0.03 | 4.60 | 0.07 | 5.56 |
| 10    | 20.025 | 19.767 | 290 | 19.102 | 18.436 | 520 | 20.041 | 19.790 | 0.08 | 4.69 | 0.11 | 6.84 |
| 20    | 20.672 | 19.128 | 522 | 19.763 | 17.284 | 916 | 20.680 | 19.131 | 0.04 | 4.43 | 0.01 | 9.65 |
| 0.6   | 5   | 15.732 | 15.338 | 95 | 15.072 | 14.069 | 173 | 15.749 | 15.347 | 0.11 | 4.30 | 0.06 | 8.33 |
| 10    | 16.476 | 16.034 | 127 | 15.701 | 14.944 | 224 | 16.486 | 16.046 | 0.06 | 4.76 | 0.08 | 6.87 |
| 20    | 17.689 | 16.783 | 242 | 16.862 | 15.640 | 434 | 17.706 | 16.793 | 0.09 | 4.77 | 0.06 | 6.87 |
| 5     | 10   | 17.556 | 17.220 | 160 | 16.893 | 16.119 | 289 | 17.572 | 17.236 | 0.09 | 3.86 | 0.10 | 6.48 |
| 10    | 18.450 | 17.953 | 221 | 17.671 | 16.450 | 408 | 18.467 | 17.969 | 0.09 | 4.31 | 0.09 | 8.45 |
| 20    | 18.729 | 17.640 | 388 | 17.411 | 15.905 | 713 | 18.743 | 17.662 | 0.07 | 7.11 | 0.13 | 9.95 |
| 0.8   | 5   | 19.593 | 19.279 | 209 | 18.252 | 18.162 | 371 | 19.625 | 19.303 | 0.16 | 7.00 | 0.13 | 5.91 |
| 10    | 19.660 | 19.422 | 302 | 18.415 | 18.122 | 533 | 19.671 | 19.429 | 0.06 | 6.39 | 0.03 | 6.73 |
| 20    | 20.382 | 18.821 | 526 | 19.077 | 17.724 | 946 | 20.395 | 18.856 | 0.06 | 6.46 | 0.18 | 6.01 |
| NV10-2 | 5   | 15.193 | 14.867 | 104 | 14.475 | 13.526 | 183 | 15.495 | 15.168 | 0.09 | 6.28 | 0.10 | 8.84 |
| 10    | 16.013 | 15.526 | 118 | 15.366 | 14.427 | 212 | 16.026 | 15.543 | 0.10 | 4.89 | 0.08 | 5.46 |
| 20    | 17.159 | 16.364 | 244 | 16.281 | 15.134 | 442 | 17.176 | 16.375 | 0.10 | 5.21 | 0.07 | 7.57 |
| 5     | 10   | 16.965 | 16.708 | 176 | 15.733 | 15.700 | 312 | 16.975 | 16.724 | 0.06 | 7.32 | 0.09 | 6.12 |
| 10    | 17.957 | 17.386 | 207 | 16.798 | 16.114 | 379 | 17.962 | 17.389 | 0.03 | 6.48 | 0.02 | 7.33 |
| 20    | 18.197 | 17.252 | 369 | 18.880 | 15.995 | 654 | 18.200 | 17.262 | 0.01 | 7.19 | 0.06 | 9.64 |
| 0.8   | 5   | 19.025 | 18.656 | 219 | 17.861 | 16.865 | 401 | 19.038 | 18.680 | 0.07 | 7.13 | 0.13 | 9.72 |
| 10    | 18.981 | 18.786 | 316 | 17.694 | 17.162 | 567 | 19.006 | 18.804 | 0.13 | 6.91 | 0.10 | 8.73 |
| 20    | 19.673 | 18.259 | 528 | 18.975 | 16.754 | 955 | 19.693 | 18.266 | 0.10 | 6.35 | 0.04 | 8.27 |
|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
Table 3. Sensitivity analysis of different disruption probabilities

| Disruption Probability | RL  | GA  | RL  | GA  |
|------------------------|-----|-----|-----|-----|
|                        | Profit | Fill rate (%) | Profit | Fill rate (%) | Profit | Fill rate (%) | Profit | Fill rate (%) |
| U(0.001, 0.005)        | 25.924 | 100.65 | 25.224 | 83.89 | 93.35 | 100.65 |
| U(0.005, 0.01)         | 24.033 | 95.21 | 23.527 | 82.45 | 92.01 | 85.21 |
| U(0.01, 0.03)          | 22.003 | 90.00 | 20.603 | 81.20 | 90.00 | 79.20 |
| U(0.03, 0.05)          | 16.003 | 85.00 | 14.603 | 79.20 | 85.00 | 73.00 |

Table 4. Different cases for sensitivity analysis

| Case | Lower | Low | High | Higher |
|------|-------|-----|------|--------|
| σ   | 10    | 50  | 150  | 250    |
| σₚ  | U(0.0001, 0.0005) | U(0.0005, 0.001) | U(0.003, 0.005) | U(0.005, 0.01) |
| σₚₚ | 10    | 25  | 50   | 60     |

Table 5. Sensitivity results of σ, σₚ, σₚₚ

| Case | Lower | Low | High | Higher |
|------|-------|-----|------|--------|
| σ   | 23.105 | 100.65 | 22.852 | 83.89 | 93.35 | 100.65 |
| σₚ  | 20.906 | 95.94 | 20.629 | 83.45 | 92.01 | 79.21 |
| σₚₚ | 18.509 | 89.13 | 18.183 | 81.20 | 90.00 | 79.20 |

Table 6. Different problem instances of NV10-1 problem

| Case | NV10-1-1,1 | NV10-1-1,2 | NV10-1-1,3 | NV10-1-1,4 |
|------|------------|------------|------------|------------|
| σ   | Higher     | Higher     | Higher     | Higher     |
| σₚ  | Lower      | Lower      | Lower      | Lower      |
| σₚₚ | Higher     | Higher     | Higher     | Higher     |