ELSEVIER

Int. J. Production Economics 78 (2002) 153–161

international journal of
production
economics

www.elsevier.com/locate/dsw

Inventory management in supply chains: a reinforcement learning approach

Ilaria Giannoccaro, Pierpaolo Pontrandolfo*

Dipartimento di Ingegneria Meccanica e Gestionale , Politecnico di Bari, Viale Japigia 182, 70123 Bari, Italy

Received 8 August 2000; received in revised form 5 September 2000

Abstract

A major issue in supply chain inventory management is the coordination of inventory policies adopted by different supply chain actors, such as suppliers, manufacturers, distributors, so as to smooth material flow and minimize costs while responsively meeting customer demand. This paper presents an approach to manage inventory decisions at all stages of the supply chain in an integrated manner. It allows an inventory order policy to be determined, which is aimed at optimizing the performance of the whole supply chain. The approach consists of three techniques: (i) Markov decision processes (MDP) and (ii) an artificial intelligent algorithm to solve MDPs, which is based on (iii) simulation modeling. In particular, the inventory problem is modeled as an MDP and a reinforcement learning (RL) algorithm is used to determine a near optimal inventory policy under an average reward criterion. RL is a simulation-based stochastic technique that proves very efficient particularly when the MDP size is large.

Keywords: Supply chain; Inventory management; Markov decision processes; Reinforcement learning

1. Introduction

A supply chain (SC) is a network of organizations that are involved in the different processes and activities that produce value in the form of products and services in the hands of the ultimate consumer [1]. Such activities are mainly the procurement of materials, the transformation of these materials into intermediate and finished product, and the distribution of finished products to the end customer. Supply chain management (SCM) is concerned with the integrated management of the flows of goods and information throughout the supply chain, so as to insure that the right goods be delivered in the right place and quantity at the right time.

The SCM literature covers different areas, such as forecasting, procurement, production, distribution, inventory, transportation, and customer service, under several perspectives, i.e. strategic, tactical, and operational. supply chain inventory management (SCIM), which is the main concern of this paper, is an integrated approach to the planning and control of inventory throughout the entire network of co-operating organizations, from the source of supply to the end user. SCIM is focused on the ultimate customer demand and aims at

* Corresponding author. Tel.: +39-80-5962-763; fax: +39-80-5962-788.

E-mail address: pontrandolfo@poliba.it (P. Pontrandolfo).

0925-5273/02/$ - see front matter
PII: S0925-5273(00)00156-0

improving customer service, increasing product variety, and lowering costs [2].

An effective management and control of the material flow across the boundaries between companies and their customers is vital to the success of companies, but is a difficult task due to the demand amplification effect, known as 'Forrester effect' [3]. The latter depends on factors such as the supply chain structure, the time lags involved in accomplishing actions (e.g. from the order release to fulfillment), and the poor decision making concerning information and material flows. Recent empirical studies [4] demonstrate that inventory management policies can have a destabilizing effect due to the increase in the volatility of demand as it passes up through the chain. For example, Towill [5] claims that the demand amplification experienced across each business interface is about 2:1.

Lee et al. [6] describe the Bullwhip effect occurring in supply chains as the considerable increase of the order variability relative to the variability of buyers' demand. They identify the main mechanisms that destabilize supply chains, i.e. order batching, price fluctuation, capacity shortfalls that lead to over-ordering and cancellation, and the updating of demand forecast.

A tight coordination among inventory policies of the different actors in the supply chain can reduce the ripple effect on demand. To this end an appropriate information infrastructure is necessary that allows all the actors within a SC make decisions synchronized and coherent among each other. Such an infrastructure is referred to as networked inventory management information systems (NIMISs) [2]. However, the exploitation of the NIMISs requires the adoption of suitable inventory management policies. For instance, Kelle and Milne [7] provide quantitative tools to study the effect of an (s, S) policy on the supply chain and show that small frequent orders and the cooperation among the SC partners can reduce demand variability. Towill [5] investigates the impact of different strategies, such as JIT, vendor integration, and time-based management, on the reduction of demand amplification. Wikner [8] stresses that the Forrester effect is lowered through the fine tuning of existing ordering policies, the reduction of delays, the removal of the distribution stage in the SC, the change of local decision rules, and a better use of the information flow through the supply chain. Johnes and Riley [9] and Hoekstra and Romme [10] address the optimal positioning of stocks in the chain and suggest the use of strategic stocks to de-couple push from pull operations. Stalk and Hout [11] and Blackburn [12] focus on time compression and the integration of operations with both customers and suppliers.

Studies on supply chain inventory management generally identify three stages, namely supply, production, and distribution [13], yet the focus is usually put on the coordination between only two of them [14,13]. Coherently, Thomas and Griffin [15] classify the models for coordinated supply chain management into buyer–vendor coordination, production–distribution coordination, and inventory–distribution coordination.

To our knowledge, there are only a few approaches that simultaneously analyze inventory decisions at more than two stages under an operational perspective. In this paper, we propose an approach to coordinate inventory management in a supply chain made up of three stages, i.e. supply, production, and distribution. A model based on Markov decision processes (MDPs) and reinforcement learning (RL) is proposed to simultaneously design the inventory reorder policies of all the SC stages. After a brief description of MDPs and RL algorithm (Sections 2 and 3), in Section 4 we define the considered supply chain and the attendant MDP model. Results obtained by the inventory policy determined through the proposed approach are discussed in Section 5.

2. Markov decision processes

A Markov decision process is a sequential decision-making stochastic process characterized by five elements [16]: decision epochs, states, actions, transition probabilities, and rewards. An agent (decision maker) controls the path of the stochastic process. In fact, at certain points in time in the path, this agent intervenes and takes decisions which affect the course of the future path. These points are called decision epochs and the decisions are called actions. At each decision epoch, the system

occupies a decision-making state. This state may be described by a vector. As a result of taking an action in a state, the decision-maker receives a reward (which may be positive or negative) and the system goes to the next state with a certain probability which is called the transition probability. A decision rule is a function for selecting an action in each state, while a policy is a collection of such decision rules over the state-space. Implementing a policy generates a sequence of rewards. The MDP problem is to choose a policy to maximize a function of this reward sequence (optimality criterion). Possible choices for these functions include the expected total discounted reward or the long-run average reward.

In this article we use the average reward criterion. The *average reward* or *gain* of a stationary policy π, starting at state i and continuing with policy π, is defined as follows:

$$g^{\pi}(i) = \lim_{N\to\infty} \frac{1}{N} \sum_{i}^{\pi} \left\{ \sum_{t=1}^{N} r(X_t, Y_t) \right\},$$

where $r(X_t, Y_t)$ represents the reward received when using action Y_t in state X_t, Y_t being the action prescribed by policy π in state X.

MDPs have been widely applied to inventory control problems [17]. For example, they can be used for determining optimal reorder points and quantities. In such a case decision epochs occur periodically, according to an inventory review policy, and the system state is a function of inventory position at the review time. In a given state, actions correspond to the amount of stock to be ordered (with "not ordering" being a possible action). The transition probabilities substantially depend on the ordered quantity, the supply rate, and the demand process until the next decision epoch. A decision rule specifies the quantity to be ordered at the review time, while a policy consists in a mapping of the replenishment orders onto the possible inventory positions. Inventory managers (the decision makers) seek the optimal policy, namely a policy that maximizes a profit index (e.g. revenues minus ordering costs and inventory holding costs) over the decision-making horizon.

Semi-Markov decision processes (SMDPs) extend MDPs. In fact, differently from MDPs where decisions are allowed only at predetermined discrete points in time, in SMDPs the decision maker can choose an action any time the system state changes. Moreover, SMDPs model the system evolution in continuous time, and the time spent by the system in a particular state follows a probability distribution. In SMDPs, the action choice not only determines the joint probability distribution of a subsequent state, but also the time between decision epochs. In general, the system state may change several times between decision epochs, but only the state at the decision epochs is relevant to the decision maker. What happened between two subsequent decision epochs provides no relevant information to the decision maker. Therefore, two processes can be distinguished: (1) the *semi-Markov decision process* represents the evolution of the system state at the decision epochs, and (2) the *natural process* describes the evolution of states continually throughout time. The two distinct processes coincide at decision epochs. The reward function associated with SMDPs is more complex. When the decision maker chooses action a in state s, first he receives a lump sum reward, further he accrues a reward at a rate $c(j, s, a)$ as long as the *natural process* occupies state j.

For $i \in S$, when action $a \in A_i$ is chosen (for any state i, A_i denotes the set of possible actions that can be taken in i), and if the next state is j, let $r(i, j, a)$ represent the reward obtained and $t(i, j, a)$ represent the time spent, during the state transition. Also let i_k represent the state visited in the kth epoch and μ_k represent the action taken in that epoch. Then the average reward (gain) of an SMDP starting at state i and continuing with policy π can be given as

$$g^{\pi}(i) = \frac{\lim_{N\to\infty} \{E[\sum_{k=1}^{N} (r(i_k, i_{k+1}, \mu_k | i_0 = i_1))]/N\}}{\lim_{N\to\infty} \{E[\sum_{k=1}^{N} (t(i_k, i_{k+1}, \mu_k | i_0 = i_1))]/N\}}.$$

Modeling inventory control problem through SMDPs rather than MDPs presents several advantages. It allows inventory policies to be considered in which review time intervals are not required to be constant as well as makes it possible to have the system accrue rewards (or incur costs) between decision epochs depending on the *natural process* (inventory holding and pipeline costs are examples of such costs).

3. Reinforcement learning

Traditional approaches to solve MDPs and SMDPs, such as value iteration, policy iteration, modified policy iteration, and linear programming, become very difficult to be applied as system space and action space grow, due to the huge computational effort required.

Reinforcement learning [18,19] is an artificial intelligent technique that has been successfully utilized for solving complex MDPs that model realistic systems. This technique is a way of teaching agents the optimal control policy [20], which is based on simulation and value iteration, the latter being a traditional method to solve MDPs and SMDPs.

The RL model is based on the interaction of two elements, i.e. the *learning agent* and the *environment*, and two mechanisms, namely the *exploitation* and the *exploration*.

The learning agent selects the actions by trial and error (exploration) and based on its knowledge of the environment (exploitation). The environment responds to these actions by an immediate reward, which is called the *reinforcement signal*, and evolves in a different state. A good action either results in a high immediate rewards directly or leads the system to states where high rewards are obtainable. Using this information (i.e. the reward received), the agent updates its knowledge of the environment and selects the next action. The agent knowledge consists of an R-value for each state-action pair: each R-value is a measure of the goodness of an action in a state. The updating algorithm (which is based on value iteration) ensures that a good environmental response, obtained as a consequence of taking an action in a state, results in increasing the attendant action-value while a poor response results in lowering it. Thus, as the good actions are rewarded and the bad actions are punished over time, some action-values tend to grow and others tend to diminish.

When a system visits a state, the learning agent chooses the action with the highest action value. Sometimes the learning agent chooses a random action. This is called exploration, which ensures that all actions are taken in all states. The learning phase ends when a trend appears in all R-values such that it is clear which is the best action in each state. The vector that maps every state into the associated optimal action, represents the learned optimal policy.

3.1. SMART algorithm

Semi-Markov average reward technique (SMART) can be implemented with a simulator of the system [20]. The environmental response for each action is captured from simulating the system with different actions in all states. Information about the response is obtained from the immediate rewards received and the time spent in each transition from one decision-making state to another. The updating of the knowledge base, which has to happen when the system moves from one decision-making state to a new decision-making state, basically means changing the action value of the action taken in the old state (this process is called learning). To implement this change apart from the response one also needs to use a variable called learning rate which is gradually decayed to 0 as the learning progresses. The probability of exploration is also similarly decayed to 0. The decaying scheme may be as follows: $a_m = M/m$ where a_m is the value of the variable (learning rate or exploration probability) at the mth iteration and M is some predetermined constant. Typically M is about 0.1 for exploration probability and 0.01 for learning rates. Fig. 1 depicts the steps of the adopted algorithm.

4. Supply chain inventory problems and reinforcement learning

In this section it is shown how reinforcement learning can be used to address supply chain inventory problems. First we describe the considered model of a supply chain, which includes the main stages identified in the literature (supply, production, and distribution), as well as the logic of the attendant material and order flows. Then we code the described model into an SMDP that can be solved through the SMART algorithm.

1. Let time step $m = 0$. Initialize action values $R(i, a) = 0$ for all $i \in S$ and $a \in A_i$. Set the cumulative reward $C = 0$, the total time be $T = 0$, and the reward rate $G = 0$. Start system simulation.
2. While $m <$ MAX_STEPS do

   If the system state at the time step m is $i \in S$,

   (a) Decay α_m and p_m according to some scheme.

   (b) With probability 1- p_m, choose an action $a \in A_i$ that maximizes $R(i, a)$, otherwise choose a random (exploratory) action from the set $\{A_i \setminus a\}$.

   (c) Simulate the chosen action. Let the system state at the next decision epoch be j. Also let $t(i, j, a)$ be the transition time, and $r(i, j, a)$ be the immediate reward earned as a result of taking action a in state i.

   (d) Change $R(i, a)$ using:

   $R(i, a) \leftarrow (1\text{-}\alpha_m) \cdot R(i, a) + \alpha_m \cdot \{r(i, j, a) - G \cdot t(i, j, a) + \max_b R(j, b)\}$

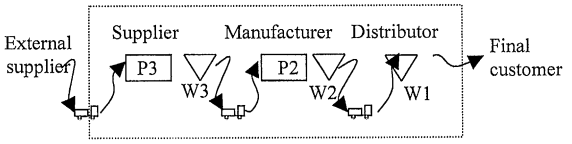
   (e) In case a non-exploratory action was chosen in step 2.(b)

   - Update total reward $C \leftarrow C + r(i, j, a)$
   - Update total time $T \leftarrow T + t(i, j, a)$
   - Update average reward $G \leftarrow C/T$

   Else, go to step (f).

   (f) Set current state i to new state j, and m ← m+1.

Fig. 1. The SMART algorithm.



Pi = production plant at the i-th stage
Wi = stock point at the i-th stage
i = supply chain stage (i = 1, 2, 3)

Fig. 2. The supply chain model.

4.1. A supply chain inventory model

A simple supply chain model is considered to show the way in which the proposed approach can be utilized to determine a near-optimal inventory order policy.

The supply chain model consists of three stages, namely supply, production and distribution (Fig. 2). It is assumed that a decision maker (actor) exists at every stage, who has the responsibility of managing inventory at that stage.

At fixed time intervals, each actor reviews the stock at his stage and, according to a certain inventory policy, decides whether to issue an order to the upstream stage. Once the order is placed, the delivering process begins as long as upstream stock is sufficient to cover the request. Otherwise the order is backordered and waits until the upstream stock reaches the ordered quantity. In particular, backordering customer demand at the distribution stage involves penalty costs, which grows with the waiting time. Even though the time interval at which periodic inventory reviews occur does not change from stage to stage, actors may adopt diverse decisions in terms of both *how much* and even *when* to order: in fact, any actor may decide not to order at a certain point in time even if the others do order.

The inventory management process is characterized by time and cost variables as reported in Table 1. Production costs have not been considered as in the model they do not depend on the specific inventory policy. Similarly, production times are almost not influenced by the inventory policy and in any case they do not substantially differentiate the performance of one policy from another.

Each time actor i issues an order, she incurs the ordering cost Co_i, which includes the transportation cost. Ordering costs are assumed to be independent on the order size. The Inventory holding cost h_i is the cost of keeping a stock unit per time unit at stock point i. The pipeline cost Cp_i is the cost per time unit of a stock unit in transit to W_i.

Table 1
Cost and time variables

Cost variables	
Unit price P	1000
Ordering cost Co_i ($i = 1, 2, 3$)	80, 80, 80
Unit inventory cost (per t.u.) h_i ($i = 1, 2, 3$)	10, 5, 3
Unit pipeline cost (per t.u.) Cp_i ($i = 1, 2, 3$)	10, 5, 3
Unit penalty cost for late delivery (per t.u.) Cb	50
Time variables	
Stock review time interval (constant) It	10
Transportation time (uniform distribution) T_i	1–3
Demand	
Mean interval time (exponential distribution) d	1

When final demand cannot be immediately satisfied, the system incurs a penalty cost Cb times the waiting time until demand is fulfilled. Although estimating the penalty cost is often difficult, it proves crucial when responsiveness to market demand is a key performance, which is especially true in time-based competition.

4.2. The SMDP model

The discussed SC inventory management process, resulting from a given inventory policy, is a stochastic process. This section explains the way in which the inventory process has been mapped into an SMDP and solved through the proposed approach.

The SMDP definition involves the choice of the reward function to be maximized. This choice is linked to the hypotheses on the cost structure of the inventory model. With this regard, two basic options are available, namely averaging vs. discounting the cost. When the time value of money is considered, cost must be discounted rather than averaged. In the considered case, the average reward criterion has been chosen as this is more frequently used in common inventory models (both single stage and multi-echelon). Also, such a criterion simplifies the hypotheses because there is no need to assume a discount factor. Furthermore, averaging is more appropriate when the performance is analyzed over a time horizon that is theoretically infinite. Finally, the reward function has to be associated with the whole supply chain performance, given that an integrated inventory management policy has to be determined. Therefore, the *average reward* or *gain* over the long run has been utilized, which is defined as follows:

$$\rho = (\text{Total Reward})/(\text{Total Time}),$$

where

$$\text{Total reward} = \text{price Tot sell} - \text{Tot Cost},$$

$$\text{Tot sell} = \text{products sold at Total Time},$$

$$\text{Tot Cost} = \text{total costs incurred at Total Time}.$$

To complete the mapping of the inventory management problem into the SMDP, the decision epochs, the system state variable, and the possible actions in every system state have to be identified.

A decision epoch occurs at each stock review time interval, when the actors at all the three stages make decision on inventory. Decision agents are three, as many as the actors in the supply chain.

As a decision agent must make decision solely based on the system state, the state variable must provide him with any information that is relevant to an integrated inventory management. In particular, she needs to know the inventory position of the whole supply chain, being inadequate that of her own stage only. Also, an integrated inventory management requires that the three decision makers share a unique reward function, given that the performance index must refer to the supply chain as a whole. Thus, the system state variable is given by the following vector:

$$(IP_1, IP_2, IP_3),$$

which describes the global SC inventory position as the inventory position IP_i at every stage i.

The inventory position IP_i at a given stage depends on schedule receipts (SR_i), on-hand inventory (OH_i), and backorders (BO_i) as follows:

$$IP_i = OH_i + SR_i - BO_i.$$

From the above equation it follows that IP_i is not bounded, which would imply an infinite size of the associated MDP. Therefore, every IP_i has been coded so as to let it assume a limited number of values (Table 2).

At every decision epoch, all decision agents must select the replenishment order quantity for their

Table 2
Actual and coded inventory positions

Actual IP_i	< -8	$[-8;-6[$	$[-6;-4[$	$[-4;-2[$	$[-2;0[$	$[0;2[$	$[2;4[$	$[4;6[$	$[6;8[$	$\geqslant 8$
Coded IP_i	1	2	3	4	5	6	7	8	9	10

own stage, i.e. each must take an action that ranges from ordering nothing up to a maximum equal to the stock point capacity plus the current backorder plus the estimated consumption during the transportation lead time minus the stock on hand. We have assumed a value of 30 for such a maximum, which simulation has shown to be as much high to be never reached. The needed capacity can be determined by measuring the maximum stock level that is reached at each stage, by simulating the system under the learnt inventory policy.

The action space size and the state space size, respectively equal to 29,791 and 1000, yield 29,791,000 action values. Their estimated values define the near optimal inventory policy.

Even though decision epochs occur at predetermined points in time, the considered decision process is an SMDP. In fact, the system state may change as well as cost (rewards) are incurred (accrued) between two subsequent decision epochs.

The SMDP has been solved by the SMART algorithm. The learning phase has been simulated by a commercial simulating package ARENA [21]. The learning process, which has required a length of 1,500,000 time units, has taken about 2 h on a PC Pentium II 450.

5. Results

A near-optimal supply chain inventory policy, which will be referred to as SMART policy, has been determined through the proposed approach. This policy, which can be thought of as an (s, S) policy where both s and S vary with the system state, is relatively simple to be implemented, as it requires the knowledge of just the optimal action to be taken in each of the system states.

The effectiveness of the SMART policy has been evaluated against a periodic order policy that adopts an integrated perspective, as it is based on the minimization of the total SC costs. Such a policy is defined by two vectors that specify the stock review time intervals (T_1, T_2, T_3) and the target levels (S_1, S_2, S_3) at each stage. At the stock review time interval T_i, an order is placed to raise the inventory position up to the target level S_i.

In particular, the vector (T_1, T_2, T_3) has been determined by solving the non-linear programming problem that minimizes the average SC cost, subject to the following relaxed constraints [22]:

$$T_i \geqslant T_{i-1} \geqslant 0 \quad \text{for} \quad i = 1, 2, 3.$$

The echelon inventory concept is utilized for computing holding costs, so that the average SC cost is given by

$$C_{\mathrm{t}} = \sum_{i=1}^{3} \left[\frac{\mathrm{Co}_i}{T_i} + \frac{1}{2} \mathrm{d} H_i T_i \right]$$

the echelon unit holding cost H_i at the ith stage being the incremental holding cost of the ith stage with respect to the upstream stage $(i-1)$th $(H_i = h_i - h_{i+1})$.

The vector (6, 8, 8) has been obtained as solution for (T_1, T_2, T_3).

The target stock S_i has been defined equal to the demand during the reorder interval time T_i $(T_i d)$ plus the stock necessary to cover the demand during the transportation lead time (LT_i):

$$S_i = (T_i + \mathrm{LT}_i)d.$$

The safety stock (SS) has been added at the last stage to cope with customer demand uncertainty.

Therefore, the order quantity OQ_i at every stage is given by

$$\mathrm{OQ}_1 = (S_1 + \mathrm{SS}) - \mathrm{IP}_1; \quad \mathrm{O\,Q}_2 = S_2 - \mathrm{IP}_2;$$

$$\mathrm{O\,Q}_3 = S_3 - \mathrm{IP}_3.$$

Table 3
SMART policy vs. the benchmark policy

Gain	SMART	Benchmark	Δ (gain)
$k = 1$ (CV = 100%)	856	826	3.63%
$k = 5$ (CV = 45%)	881	852	3.40%
$k = 10$ (CV = 32%)	884	855	3.39%

The performance achieved by the two policies have been measured through simulation runs with a time length of 100,000 time units. In particular, three demand patterns have been considered, all characterized by Erlang distributions with same demand rates but different variance. The three patterns are indeed characterized by three diverse values of the k parameter of the Erlang distribution. As known, the relationships between the k parameter, the mean μ, the variance σ, and the coefficient of variation CV of demand are as follows:

$$\mu = k\beta, \quad \sigma^2 = k\beta^2, \quad \mathrm{CV} = 1/k^{1/2}.$$

The results are depicted in Table 3.

The benchmark policy has been adapted to the demand variance by adjusting the safety stock SS. On the contrary, the inventory policy learned for the $k = 1$ case has been used for the other demand patterns ($k = 5$ and 10). This has allowed the robustness of the proposed approach to be verified.

It can be observed that the performance increases with k for both the SMART and the benchmark policies, which was expected, given that when k increases demand uncertainty diminishes. Less obvious is that the SMART policy performs better than the benchmark does even for $k = 5$ and 10, namely when demand is different from that experienced during the learning ($k = 1$). In fact, while the learned policy can surely deal with the same demand pattern used during the learning phase, it could show a performance decline for new patterns. Based on the results, we can then conclude that, not only is the SMART policy more efficient, but is also robust as long as demand undergoes slight changes.

The higher efficiency of the SMART policy is mainly due to the fact that the decision rule, on which basis actions (i.e. replenishment orders) are taken, is more sophisticated. In fact, there is neither a unique reorder point nor a unique reorder quantity. Rather, replenishment orders are placed as complex functions of inventory position: differently from the benchmark policy, reorder points as well as ordered quantities vary with the global inventory position.

Furthermore, the SMART policy considers the stochastic nature of the environment, namely demand and lead time variability, whereas the benchmark policy is determined based on a deterministic demand equal to the average and copes with uncertainty through a safety stock.

6. Conclusions

In this paper the SCM problem has been addressed with particular emphasis on inventory management. Supply chain management is widely recognized as a vital source of competitive advantage, yet SCM techniques, especially in the inventory area, are very difficult to be put into practice, given the high need of information communication and processing involved. To this end many efforts have been lately devoted to the design of appropriate networked inventory management information systems (NIMISs).

Despite the efforts focused on the implementation of NIMISs, relatively less attention has been given to define an appropriate logic for managing inventory, so missing the opportunity of exploiting the potential of such information systems. In particular, integrated approaches to manage inventory decisions at all stages of the supply chain need to be developed.

In this paper an approach has been proposed, which addresses this problem. It is based on three techniques, namely Markov decision processes, reinforcement learning, and simulation. MDPs make it possible to model sequential decision-making problems under uncertainty. RL and simulation allow MDPs to be solved in a wider range of cases than conventional methods (e.g. dynamic and linear programming) do.

The approach has been tested on a supply chain model consisting of the supply, manufacturing, and distribution stages. The integrated inventory policy determined through the proposed approach (SMART policy) outperforms a centralized

periodic order policy, which has been used as a benchmark. Also, the SMART policy proves quite robust with respect to slight changes in demand.

It is expected that the superiority of the SMART policy would be greater for more complex cases. In fact, centralized but simpler policies (such as the POQ based utilized as a benchmark) cannot adapt to complex environments as the SMART policy does. This depends on (i) the ability of simulation modeling of capturing detailed features of the system as well as (ii) the capability of MDPs of describing time dependencies between decisions.

Further research should address the issue of having the supply chain actors actually implementing the optimal policy determined through the proposed approach. This is quite a difficult task, given that the supply chain actors are likely to belong to diverse firms. Therefore, having them actually share a unique reward function needs a way (e.g. appropriate incentive mechanisms) to fairly split the higher rewards that the optimal policy would guarantee.

References

[1] M. Christopher, Logistic and Supply Chain Management, Pitman Publishing, London, 1992.
[2] M. Verwijmeren, P. Van der Vlist, K. van Donselaar, Networked inventory management information systems: Materializing supply chain management, International Journal of Physical Distribution and Logistics Management 26 (6) (1996) 16–31.
[3] J.W. Forrester, Industrial Dynamics, MIT Press, Cambridge, MA, 1961.
[4] M.P. Baganha, M. Cohen, The stabilizing effect of inventory in supply chains, Operations Research 46 (3) (1998) S72–S73.
[5] D. Towill, Industrial dynamics modeling of supply chains, Logistics Information Management 9 (1996) 43–56.
[6] H.L. Lee, V. Padmanabhan, S. Whang, The bullwhip effect in the supply chains, Sloan Management Review 38 (3) (1997) 93–102.
[7] P. Kelle, A. Milne, The effect of (*s, S*) ordering policy on the supply chain, International Journal of Production Economics 59 (1999) 113–122.
[8] J. Wikner, D.R. Towill, M. Naim, Smoothing supply chain dynamics, International Journal of Production Economics 22 (1991) 231–248.
[9] T.C. Jones, D.W. Riley, Using inventory for competitive advantage through supply chain management, International Journal of Physical Distribution and Materials Management 17 (2) (1987) 94–104.
[10] S. Hoekstra, J. Romme, Integral Logistics Structures: Developing Customer-Oriented Goods Flows, McGraw-Hill, London, 1992.
[11] G.H. Stalk, T.M. Hout, Competing against Time, How Time-Based Competition Is Reshaping Global Competition, Free Press, New York, 1990.
[12] J.D. Blackburn, Time-based Competition: The Next Battleground in American Manufacturing, Irwin, Homewood, IL, 1991.
[13] S. Erengüc, A.J. Vakharia, Integrated production/distribution planning in supply chains, European Journal of Operational Research 115 (1999) 219–236.
[14] C. Forza, Achieving superior operating performance from integrated pipeline management: An empirical study, International Journal of Physical Distribution and Logistics Management 26 (9) (1996) 36–63.
[15] D.J. Thomas, P.M. Griffin, Coordinated supply chain management, European Journal of Operational Research 94 (1) (1996) 1–15.
[16] M. Puterman, Markov Decision Processes: Discrete Stochastic Programming, Wiley Interscience, New York, 1994.
[17] E. Porteus, Stochastic inventory theory, in: D.P. Heyman, M.J. Sobel (Eds.), Handbooks of Operations Research, North-Holland, Amsterdam, 1990.
[18] R.L. Sutton, A.G. Barto, Reinforcement Leaning – An Introduction, MIT Press, Cambridge, MA, 1998.
[19] D. Bertsekas, J. Tsitsiklis, Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.
[20] T.A. Das, A. Gosavi, S. Mahadevan, N. Marchalleck, Solving semi-Markov decision problems using average reward reinforcement learning, Management Science 45 (4) (1999) 560–574.
[21] W.D. Kelton, R.P. Sadowski, D.A. Sadowsky, Simulation with Arena, McGraw-Hill, New York, 1998.
[22] J.A. Muckstadt, R.O. Roundy, 1993, Analysis of multistage production systems, in: S.C. Graves, A.H.G. Rinnooy Kan, P.H. Zipkin (Eds.), Handbooks in Operations Research and Management Science, Vol. 4, North-Holland, Amsterdam, 1993.