International Journal of Production Research
Vol. 48, No. 4, 15 February 2010, 933–955

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

A machine learning approach to optimise the usage of recycled material in a remanufacturing environment

Purvin Shah[a], Abhijit Gosavi[b]* and Rakesh Nagi[c]

[a]*Cello Pack, 55 Innsbruck Drive, Buffalo, NY, USA;* [b]*Missouri University of Science and Technology, 219 Engineering Management, Rolla, MO, USA;* [c]*University at Buffalo (SUNY), Buffalo, NY, USA*

(*Received 6 September 2007; final version received 25 August 2008*)

Remanufacturing has acquired importance in recent years because of the increasing environmental concerns of manufacturing processes that deplete the Earth's resources. Some examples of remanufactured products are automobile parts, furniture, photocopiers, and computer printers. In a remanufacturing setup, raw materials are drawn from two sources: (i) 'cores', which are obtained from recycled products, and (ii) 'non-recycled' or unused materials, which are produced from minerals freshly mined from the earth. An important decision for the manager is to select material optimally from these two sources. Using cores has environmental benefits, and because they are cheap, they reduce manufacturing costs. However, their use generally increases the production time, because of the additional pre-processing usually needed, which can negatively impact service levels. When the supply of finished products is running low, to satisfy service levels, it makes sense to use unused material. This research focuses on identifying an optimal strategy of switching between the two sources of material. A reinforcement learning algorithm is used to solve the switching problem. The switching algorithm produced encouraging results, showing up to 65% cost improvements over a policy that uses only unused materials.

Keywords: process industry; process planning; processes; process economics; reconfigurable manufacturing

1. Introduction

Manufacturing industry has evolved significantly since the industrial revolution. This has led to a dramatic increase in the volume and variety of inexpensive products available for use in our daily lives. Unfortunately, this is juxtaposed with excessive consumption of minerals used in producing raw materials, which, if left unchecked, is bound to cause severe shortages. In addition, at the end of their useful lives, products generate enormous amounts of waste that are stressing landfills and resulting in hazardous consequences for the environment and human health. Under this backdrop, *recycling* of products has acquired increased importance.

The recycling of products returns them to their raw material state from which they can be reused. An example of recycling is melting a used gear, or bearing, to obtain steel, which is then reused. This helps to conserve materials. However, the labour and energy

*Corresponding author. Email: gosavia@mst.edu

ISSN 0020–7543 print/ISSN 1366–588X online

DOI: 10.1080/00207540802452157
http://www.informaworld.com

utilised in production are lost. The material value of any product is usually only about 5–10% of the value of the final product. The remaining value consists of processing raw materials (RIT Website 2006).

Remanufacturing involves disassembly of the products into individual components, upgrading the performance of the defective components (overhaul), and then re-assembling the components to reproduce the product. This not only conserves the material value of the product, but also conserves a considerable portion of the energy utilised in production of those components. This has considerable environmental benefits. Repairing only one or two defective components in the product is called *refurbishing*. See Figure 1 for a pictorial description of the differences between recycling, remanufacturing, and refurbishing. Remanufacturing involves not only the repair of all the defective components, but an overhaul and upgrade of the entire product assembly. It is clear that while recycling has smaller environmental benefits than remanufacturing, refurbished products, usually, do not have the quality of remanufactured products. Remanufactured products ideally match the customer requirements for a new product, which cannot be said of refurbished products.

According to Lund and Bollinger (1981), remanufacturing of an automobile can help to conserve 85% of its final product value. According to the RIT Website (2006), the annual energy saved world-wide because of remanufacturing is equivalent to the electricity produced by five nuclear power plants. Also, the amount of minerals saved by remanufacturing is equivalent to 155,000 railroad cars completely filled, forming a train as long as 1100 miles (RIT Website 2006). For the promotion of environmental conservation and green manufacturing, a number of state laws in the United States are giving incentives to companies involved in remanufacturing. The remanufacturing industry is a $53 billion industry that includes approximately 70,000 small and mid-size firms, and provides direct employment to 480,000 people (Lund and Bollinger 1981). This makes the economics of remanufacturing equivalent to the entire pharmaceutical industry in the US. Apart from the advantages of conservation of resources, the remanufacturing industry has the potential to grow rapidly, and, hence, can lead to more

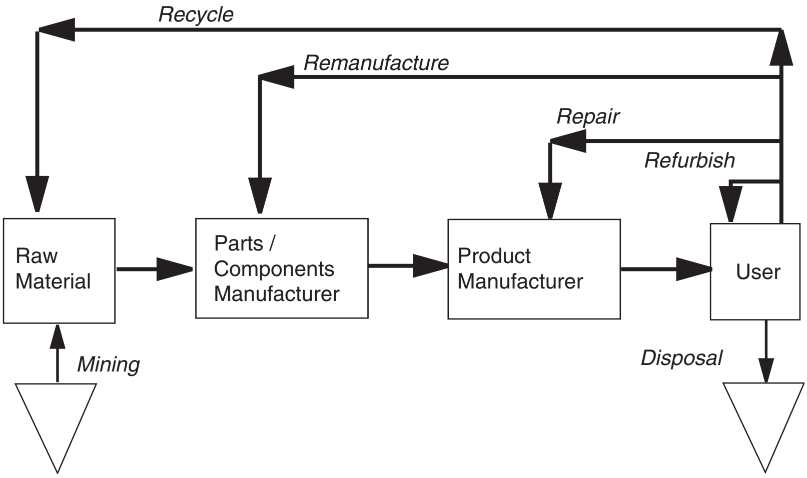

Figure 1. Differences between recycling, remanufacture, repairing, and refurbishing.

employment opportunities. The remanufacturing industry is expanding its horizons rapidly, and it is also becoming profitable (Guide *et al.* 2000). Statistics show that 50% of consumers prefer remanufactured products if they are 20% cheaper than non-remanufactured products (RIT Website 2006). Some prominent remanufactured products are: automobile parts (e.g., engines, gear boxes), printers, photo copiers, furniture, compressors used in freezers and air-conditioners, electrical apparatus, laser toner cartridges, musical instruments (e.g., keyboards, electric guitars), aircraft parts, bakery equipment, and gaming machines.

Remanufacturing facilities usually accept products that are ready to be scrapped. Even after the completion of their life, products often have reusable components. These parts, which can be used in assembling new products, will be referred to as *cores* in this paper. Cores are cleaned, inspected, reworked if necessary, and then used in the assembly of new products. Unfit cores are sold as *scrap* with low profit margins. Any product of which at least one component is a used part is called a remanufactured product.

According to the RIT Website (2006), parts in a given product can be termed 'remanufactured' if: (1) they can be cleaned or reworked so that they have the same characteristics as the new parts; and (2) they can be disassembled to an extent where they can be inspected thoroughly to determine their useful lives. Remanufactured products are almost of the same quality as products produced from fresh materials, but with a considerably lower cost of production. It should be noted that there is a considerable amount of uncertainty associated with obtaining cores and also in the quality of the cores obtained. It should also be noted that lower grades of cores require a considerable amount of time for reworking and inspection, which unused materials, procured from fresh minerals, do not need. Thus there is a trade-off, but in the end remanufacturing is a more environmentally responsible route, and it is expected that associated with it there are a number of hidden revenues.

While many of the issues of production planning in remanufacturing are similar to those of traditional manufacturing, it is a fact that there are several important distinguishing features that we now address (van der Laan *et al.* 1996, Ferrer and Whybark 2001, Guide and van Wassenhove 2001, Kongar and Gupta 2002, Ferrer and Ketzenberg 2004, Savaskan *et al.* 2004). The following features bestow on the management of remanufacturing systems a significant amount of complexity (Guide *et al.* 2000).

- *Uncertainty in timing and volume of returns*. The time between acquiring returns (cores) depends on how quickly the product deteriorates, its expected life, and how quickly technology changes. The actual volume of returns is also dependent on these factors. This has a significant influence on raw-material inventory control.
- *System balance and materials management*. A remanufacturing system has to ensure that the rate of core acquisition is roughly equal to that of the demand; again, this rate is dependent on technological innovation and the product's expected life. Also, if the outstanding demand is greater than what can be delivered with cores, one must switch to unused materials. This makes the materials management problem extremely complex because the system must remain balanced in order to meet customer deadlines.
- *Disassembly complexity*. Different cores require different degrees of disassembly, and hence firms in this business need a good understanding of the degree of disassembly needed and the methods needed for disassembly.

- *Core valuation*. The quality of the cores varies significantly. Identical-appearing items can return different sets of parts to be remanufactured. Also, cores have diverse properties and lives depending on their previous use. They need to be distinguished and graded properly, so that remanufacturing operations can be scheduled correctly. It is very difficult to identify desirable characteristics in cores, and extensive research is required on this subject.
- *Reverse logistics network*. How products are acquired from the users, incentives offered for them to return products, and transporting these products to the remanufacturing facility are issues that have to be dealt with by the managers of the reverse logistic networks. Some of these issues are quite challenging.
- *Matching parts*. Matching the parts obtained from cores to those in the products that are being remanufactured is a major challenge. It must be tackled comprehensively in order to be successful in meeting demand in terms of specifications and time.
- *Stochasticity*. Because of the uncertain nature of the cores, there is a great deal of uncertainty in the processing times and also in the actual routings on the shop floor. This is unlike the use of unused materials where this uncertainty can be minimised.

Classical manufacturing systems differ significantly from remanufacturing systems in the following areas.

- *Supply–demand logistics*. Traditional systems do not have any returns or reverse flows, and the flow is to a great extent dictated by demand for finished products, unlike in remanufacturing where it is also dictated by the supply of cores.
- *Shop floor control*. In traditional systems, raw materials do not have to be processed, they are not disassembled, and routings are to a great extent deterministic. As discussed above, the opposite is true of remanufacturing systems.
- *Inventory control*. In remanufacturing systems, the raw-material inventory is composed of cores, substitute parts, and some OEM parts; also, all part types must be tracked. In traditional systems, one tracks WIP and finished goods only.
- *Forecasting and purchasing*. Demand for both cores and product has to be predicted separately in remanufacturing systems, whereas only product demand needs to be forecasted in classical production planning. Also, purchasing in remanufacturing systems is a highly stochastic affair because of the randomness inherent in the quality and quantity of cores available, whereas it is a much more deterministic affair in the classical case.

As discussed above, the materials-management problem in a remanufacturing setting is a complex one. An important issue as discussed by Guide *et al.* (2000) is as follows: "A firm may need to manufacture new items to meet demands when too few cores are available ..." In this context, we must point out that since balancing the demand and the core availability is so challenging, in order to produce new items the manager must resort to using unused (new materials not derived from cores) raw materials when the desired cores are not available. The question that arises then is: when should the manager use cores, and when should he/she switch to unused materials? Since the system is very random, due to the random processing times of processing cores, an appropriate model must consider stochastic elements in the system. We will show that the problem that we discuss here can be set up near-exactly as a stochastic dynamic program. While heuristic solution methods can be developed, an exact approach is clearly more attractive. A traditional dynamic programming (DP) algorithm can be used to solve a stochastic dynamic program. However, DP algorithms require *accurate* transition probability functions, which are difficult to obtain for systems with a high degree of variability and randomness, such as the system we consider here.

It turns out that Reinforcement Learning (RL) is a methodology that allows us to generate near-optimal solutions to stochastic dynamic programs *without the need for transition probabilities*. This is the motivation for using RL as a solution technique to this problem. RL has a number of other features that make it very suitable for the problem domain we consider here. RL can be implemented in a discrete-event simulation setting, which turns out to be an appropriate modeling technique for systems with great randomness and modeling complexity. While RL has been extensively used in machine learning, to the best of our knowledge, this is the first use of RL in a remanufacturing setting. A proper use of RL in a remanufacturing setup can kick-start a new field of study, where RL could be applied to many other materials-management-related problems such as scheduling and production planning in remanufacturing systems.

Section 2 discusses some of the relevant previous literature on remanufacturing. Section 3 provides details of the mathematical model adopted here. The reinforcement learning and simulation approach, adopted to solve this problem, is discussed in Section 4, which is followed by numerical results in Section 5. Brief conclusions are presented in Section 6.

2. Literature review

We present a brief review of the literature on remanufacturing starting with some of the earliest papers and ending with some of the recent papers on the topic. Hoshino *et al.* (1995) proposed a theoretical model for understanding the fundamentals of optimal recycling plans. The main objectives of the model were to maximise the total profits and recycling carried out by a remanufacturing facility. This problem was addressed using a goal-programming approach. This was one of the first works related to optimisation in a remanufacturing facility. This was closely followed by the work of van der Laan *et al.* (1996), who proposed a so-called (s, Q) inventory model for remanufacturing. On the basis of these models, the authors developed a heuristic. The paper states that disposal of cores at disassembly is necessary in order to keep the inventory from increasing to very high levels.

Guide *et al.* (2000) presented an insightful paper that highlighted the differences between inventory control in traditional manufacturing processes and in ‘recoverable’ manufacturing systems. More importantly, they explored at least seven reasons for the uncertainty arising in the timing and quantity of returns (cores). Their focus is on what needs to be done by the manager in order to smoothen the management of remanufacturing systems.

Two important papers appeared in 2001: Guide and van Wassenhove (2001) and Ferrer and Whybark (2001). Guide and van Wassenhove (2001) proposed a framework for managing product returns in remanufacturing. Applying the core concepts of Economic-Value Analysis, this paper discusses the influence of reuse activities on operational requirements. They suggest that a market-driven approach where products returned are categorised is more beneficial for operational management than an approach where products are returned through the waste stream. Secondly, an economic-value analysis of reuse activities helps to justify the remanufacturing process and helps a manager in the decision-making process. Ferrer and Whybark (2001) presented a material management model for remanufacturing systems which is based on Material Resource Planning. The proposed model optimises the number of parts required for assembly, the disassembly schedule, and the number of cores procured as trade-in for finished products. Overall, this model provides an effective material management system for a remanufacturing facility. However, the model does not give us information on time and quantity of cores, and unused parts to be purchased.

Kongar and Gupta (2002) proposed an integer-goal-programming approach for balancing the supply chain in a remanufacturing setup. The integer-goal-programming proposed in this paper was primarily focused on providing a feasible solution for a desirable disassembly process plan. The paper discussed two different sets of approaches: (i) concentration on a cost revenue function, and (ii) concentration on environmental functions.

We cite three papers from 2004: Ferrer and Ketzenberg (2004), van der Laan and Teunter (2004) and Savaskan *et al.* (2004). It was suggested by Ferrer and Ketzenberg (2004) that the decision on using cores or unused parts in assembly depends on the yield of the disassembled parts at that time. The unused parts to be purchased may have a longer lead time and hence a decision may be needed before the disassembly yield is known. Their paper studies four different kinds of models to evaluate the above situation. van der Laan and Teunter (2004) proposed heuristic methods for near-optimal inventory control policies in remanufacturing facilities. The heuristics help in finding when the facility should start manufacturing instead of remanufacturing. Savaskan *et al.* (2004) explore a closed-loop supply chain model for the collection of cores.

We now discuss some of the more recent papers. Bhattacharya *et al.* (2006) present models for retailer order quantities in the context of remanufacturing under a number of scenarios. Zhou *et al.* (2006) present a kanban model for a remanufacturing system. Konstantarasa and Papachristos (2007) present an interesting model for a system in which both manufacturing and re-manufacturing options are allowed. They discuss a periodic review inventory model and explicitly model a switching period from remanufacturing to manufacturing. Geyer *et al.* (2007) "quantifies the cost-savings potential of production systems that collect, remanufacture, and re-market end-of-use products as perfect substitutes while facing the fundamental supply-loop constraints of limited component durability and finite product life cycles." They also show the need to coordinate a number of factors within production control to obtain cost savings. A very recent paper by Zikopoulos and Tagaras (2008) provides a mathematical model for measuring the benefits of sorting before disassembly in remanufacturing.

It is quite clear that remanufacturing is currently a topic of significant research interest, and because of environmental concerns is likely to become a vital aspect of the manufacturing industry. It is also clear that although there are a number of papers that examine logistics issues in environments where remanufacturing is already in place, not much work has been done on quantifying the benefits of using re-manufacturing in the context of production planning. Our paper seeks to fill this gap by examining the application of using recycled material in a production process where, usually, unused materials are used.

3. Problem description and mathematical modeling

The problem considered in this paper is related to material selection in a remanufacturing facility. Before discussing details of the mathematical model, the related remanufacturing system is described in some detail.

3.1 *The remanufacturing system*

Although a remanufacturing facility is similar to a typical manufacturing facility, it has several distinctive features. Unlike traditional manufacturing, remanufacturing involves different kinds of cores. The methods used in the procurement of cores depend on the kind of setup established in remanufacturing. Cores coming into the facility can either be purchased, like other raw materials, or may be traded-in with the customers for finished products.

It has been noted in the literature (Guide *et al.* 2000) that there is a great deal of uncertainty in obtaining cores of desirable quality. Hence in all our models we use the exponential distribution, which has a significant variance, to model the inter-arrival time between successive cores to the facility. Also, as discussed previously, all cores are not similar and they differ in their functional characteristics. Hence they need to be graded and marked into different categories for rework. It is assumed that while procuring the cores, they are sorted according to their merit, and stored as inventory in the facility. Also, the unused material, which may be required in case of a shortage of cores, is stacked as inventory in the remanufacturing facility. Hence, in one of our models, we have assumed that there are two types of cores available.

Figure 2 shows the different types of raw materials available for use, which are core 1, core 2, etc., and unused raw material. One can consider different bins of raw materials

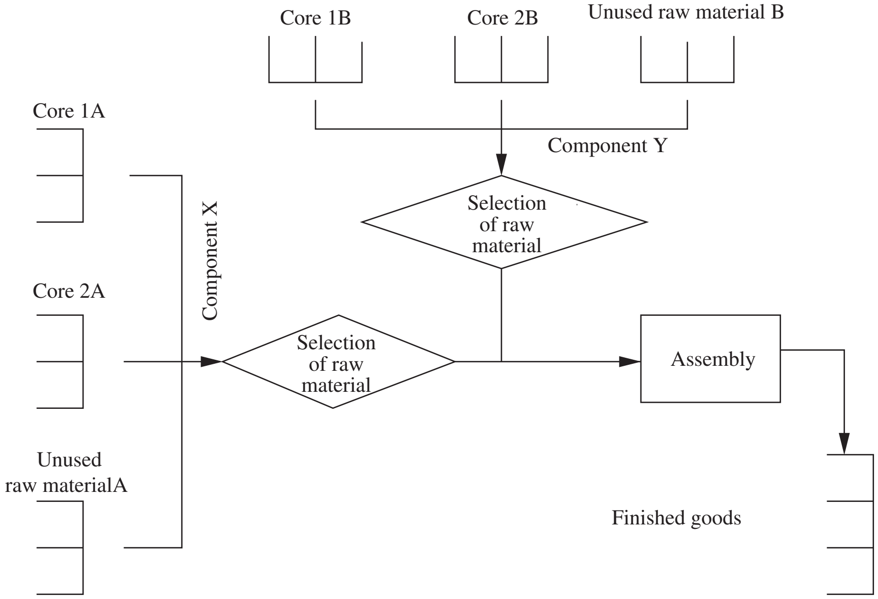

Figure 2. Remanufacturing environment with two parts (x and y) at the assembly station.

inventory in the remanufacturing facility. Any finished product consisting of cores is considered to be a *remanufactured* product. If no cores are available, manufacturing is carried out using unused raw materials. The decision related to selection of raw materials is not simple, and an incorrect selection of raw material can lead to reduction of profits.

3.2 *The material-selection problem*

Cores require time for rework and inspection. Hence cores can take additional production time, which is not needed with unused material. If the rate of arrival of demand for finished goods is high, then the additional time consumed by cores can result in cancellation of orders and revenue loss. It is wise to use unused material instead of cores if the finished goods inventory is at low levels. On the contrary, if finished goods inventory is sufficiently high, then using cores makes business sense. For the manager of a remanufacturing facility, one of the most important decisions required is to select the kind of raw material to be used – in order to ensure that profits are not diminished and, at the same time, cores are used at an optimal level.

There exists a trade-off between conservation of energy and profits of a company. The factors affecting this trade-off need to be addressed and controlled. The manager has to be aware of the profits associated with the selection of any given type of raw material. Selection of raw material should be based on factors affecting this trade-off. Monitoring the service level of the facility can help determine the rate of demand for finished goods, and this information can be utilised in the selection of raw material. Also, the inventory level of the final product provides information about the rate of demand. Monitoring inventory levels gives better control in evaluating the demand for the final product. It makes sense for the manager to determine the nature of the raw material to be used by observing the amount of finished-goods inventory.

This production system involves a number of random variables, such as (i) the time between arrivals of raw materials, (ii) the time required for repair operations, (iii) the time required for assembly operations, and (iv) the time between demand arrivals for finished products. All these random variables together make this system a complex one, which requires a sophisticated mathematical approach for analysis. In the next section, a semi-Markov decision model will be presented to solve the material-selection problem discussed above.

3.3 *A semi-Markov decision model*

A SMDP (semi-Markov decision problem) consists of five elements ($\mathcal{S}, p, \mathcal{A}, r, t$) (Bertsekas 1995), where $\mathcal{S}$ denotes the set of states, $\mathcal{A}$ denotes the set of actions, p denotes the transition probability function, r denotes the transition reward function, and t denotes the transition time function. Consider a stochastic process $(X, T) \equiv \{X_n, T_n : n \in \mathcal{N}\}$ where $X_n \in \mathcal{S}$ and $T_n \in \Re^+$ where $\Re^+ = [0, \infty]$. Here X_n denotes the state of the system in the nth decision epoch and T_n the time at the nth decision epoch.

The above process is said to satisfy the following semi-Markovian condition if (Bertsekas 1995)

$$\begin{aligned} &P\{X_{n+1} = j, T_{n+1} - T_n \le t \mid X_0, \ldots, X_n, T_0, \ldots, T_n\} \\ &\quad = P\{X_{n+1} = j, T_{n+1} - T_n \le t \mid X_n, T_n\}, \end{aligned}$$

where the notation $P\{\bullet \mid \bullet\}$ denotes a conditional probability in which the values of the variables in the condition are assumed to be known. Here, the finished goods inventory level is used as the state of the system.

In the SMDP, at every decision-making epoch, the decision maker (manager) has to select the action that controls the path of the stochastic process. For instance, in the remanufacturing problem, the manager of the remanufacturing facility must decide which raw material is to be used for the assembly process. A solution of a SMDP is called a policy. If π denotes a policy, then $\pi(i)$ will denote the action chosen in state i. The goal of the SMDP is to find an optimal policy that minimises some performance metric (measure), e.g. long-run average cost per unit time. We now present some details of a SMD model.

3.3.1 *Transition probability function* $p(\bullet, \bullet, \bullet)$

The transition probability denotes the chance of going from one state to another for a specific action. Now, $p(i, a, j)$ will denote the transition probability of going from state i to state j under the condition that action a is selected in state i. Transition probabilities of complex systems are difficult to find. This issue will be discussed in detail below.

3.3.2 *Cost function* $r(\bullet, \bullet, \bullet)$

On the successful completion of a transition from one state to another, the given action based on its performance is either rewarded or punished. In our example we have costs incurred while the transitions are in progress. Thus at the end of each transition, the accumulated costs are assigned to the specified action as fixed penalties. Now, $r(i, a, j)$ will denote the cost incurred from state i to state j under the condition that action a is selected in state i.

3.3.3 *Time function* $t(\bullet, \bullet, \bullet)$

One of the distinctive features of a SMDP that makes it different from a MDP (Markov decision problem) is that the SMDP incorporates the time function in its reward calculations. Here, $t(i, a, j)$ will denote the transition time in going from state i to state j given that action a is chosen in state i.

3.3.4 *Average cost*

Let i_k represent the kth decision-making epoch, and if policy π is pursued, $\pi(i_k)$ will denote the action taken in that epoch. Then, under a policy π with s as the starting state, the average cost of the SMDP, which is to be minimised, is given by Bertsekas (1995)

$$\rho^{\pi}(s) = \limsup_{T \to \infty} \frac{\mathrm{E}\left[\sum_{k=1}^{T} r(i_k, \pi(i_k), i_{k+1}) \mid i_1 = s\right]}{\mathrm{E}\left[\sum_{k=1}^{T} t(i_k, \pi(i_k), i_{k+1}) \mid i_1 = s\right]}.$$

It is well known that if the underlying Markov chains are recurrent and irreducible, the average cost is independent of s.

3.3.5 *Discounted cost*

The long-run discounted cost for policy π starting at state s is defined as

$$\sigma^{\pi}(s) = \limsup_{T\to\infty} \sum_{k=1}^{T} \mathrm{E}[(\psi_k)^{k-1} r(i_k, \pi(i_k), i_{k+1}) \mid i_1 = s],$$

where the continuous-time discount factor, ψ_k, in the kth epoch is defined as $\psi_k = \exp(-\lambda t(i_k, a_k, i_{k+1}))$, where a_k is the action chosen in the kth epoch. It is well known that, as the discount factor tends to 1, i.e. as $\lambda \to 0$, optimising with $\sigma^{\pi}(s)$ for all s is equivalent to optimising with ρ^{π}.

3.3.6 *The SMDP Bellman equation*

The well-known *Bellman's equation* for a SMDP is given as follows. For every state $i \in \mathcal{S}$, there exists a scalar, $v^*(i)$, that satisfies

$$v^*(i) = \min_{a\in\mathcal{A}(i)} \left[\sum_{j\in S} p(i,a,j) \left[r(i,a,j) + \int_0^{\infty} \exp(-\lambda t) v^*(j) F_{i,a}\,\mathrm{d}t \right] \right],$$

where $F_{i,a}(t)$ denotes the sojourn-time distribution for (i, a). It has also been established (Bertsekas 1995) that the policy defined as follows for all $i \in \mathcal{S}$,

$$d(i) \in \arg\min_{a\in A(i)} \left[\sum_{j\in S} p(i,a,j) \left[r(i,a,j) + \int_0^{\infty} \exp(-\lambda t) v^*(j) F_{i,a}\,\mathrm{d}t \right] \right],$$

is optimal.

3.3.7 *Value iteration algorithm*

The value iteration algorithm provides a mechanism to obtain approximate estimates of v^*. As mentioned before, the remanufacturing system is a complex one, the transition probabilities of which are hard to find. If it were possible to compute the transition probabilities, the value iteration algorithm would have been an excellent choice to solve the raw-material selection under the given conditions. However, since the transition probabilities are not available, an alternative algorithm that can do without transition probabilities must be used. As stated above, Reinforcement Learning (RL) is one method that does not require transition probabilities, but generates near-optimal solutions.

4. RL

RL is an artificial intelligence algorithm that can easily be incorporated into simulation models. Also, as stated before, it does not require the transition probabilities between the states. RL is often described as a technique used in teaching an agent how to act by rewarding and punishing it on a continuous basis (Bertsekas and Tsitsiklis 1996, Sutton and Barto 1998, Gosavi 2003). A popular algorithm called Q-learning is due to Watkins (1989). It has been extended to semi-Markov problems by Bradtke and Duff (1995), where they use a continuous cost rate; Gosavi (2003) consider a lump sum cost. We use the algorithm from Gosavi (2003) for semi-Markov problems in this paper.

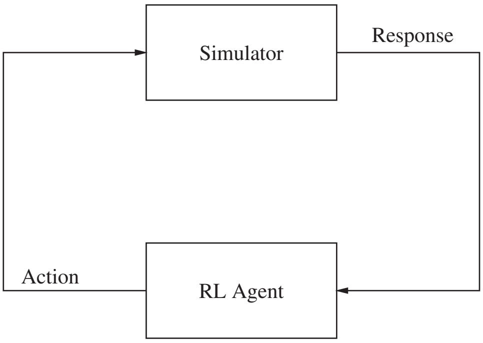

Figure 3. Schematic showing how RL works.

Within a simulation run of a RL algorithm, the system gradually learns (decides) which decision is the best. At the end of each state transition, the knowledge base of the algorithm is updated. The knowledge base of the RL algorithm is composed of *Q-factors*, which share the following relationship with the value function of dynamic programming. For every $i \in \mathcal{S}$,

$$v^*(i) = \min_{a \in \mathcal{A}(i)} Q^*(i,a),$$

where $Q^*(i,a)$ denotes the optimal Q-factor for state i and action a and $\mathcal{A}(i)$ is the set of actions allowed in state i. As shown in Figure 3, the agent learns via trial and error by visiting each state several times. At the end of the learning process, one has the optimal *Q-factors* for each state. From the optimal *Q-factors*, the optimal policy, π^*, can be determined as follows. For every i,

$$\pi^*(i) = \arg\min_{a \in \mathcal{A}(i)} Q^*(i,a).$$

4.1 *Algorithm description*

As discussed above, RL is a viable alternative to the value iteration algorithm. A RL algorithm for solving the discounted cost SMDP (introduced by Gosavi (2003) and convergence proved by Gosavi (2007)) is described below.

- Step I. Set the iteration count, k, to 0. Initialise action values $Q^k(i,u)=0$ for all $i \in \mathcal{S}$ and $u \in \mathcal{A}(i)$. Set α to a small number less than 1. Start system simulation.
- Step II. While $m < MAX\text{-}STEPS$ doIf the system starts at iteration k is $i \in \mathcal{S}$,

1: With a probability of $1/|\mathcal{A}(i)|$, choose an action $a \in \mathcal{A}(i)$ that minimises $Q^k(i,a)$.
2: Simulate the chosen action a. Let the system state at the next decision epoch be j. Also, let $t(i,a,j)$ be the transition time, and $r(i,a,j)$ be the immediate cost incurred in the transition resulting from taking action a in state i.
3: Update $Q(i,a)$ using

$$Q^{k+1}(i,a) \leftarrow (1-\alpha)Q^k(i,a) + \alpha\left[r(i,a,j) + e^{-\lambda t(i,a,j)} \min_{b \in \mathcal{A}(j)} Q^k(j,b)\right].$$

4: Set the current state i to the new state j. Also, decay α, and increment k by 1, then go to Step II(*1*).

The decay of α is done according to the rule $\alpha = 1/k$. For other rules, see Gosavi (2003). RL works in a simulator, and at the end of the learning process provides the optimal *Q-factors* for each state. From the optimal *Q-factors*, the optimal raw-material selection policy for each state of the remanufacturing environment can be obtained.

The RL algorithm is centred on updating the *Q-factor* values. The *Q-factors* are updated in a process that requires the reward function $r(i, a, j)$ and the time function $t(i, a, j)$. Note, however, the transition probabilities, $p(i, a, j)$, are not required. An alternative to RL is to use a 'brute-force' simulation model to determine the performance of a given switching policy. Clearly, this approach is feasible when the state space is small, so that the set of switching policies is small.

4.2 *Simulation models*

The simulator used here is designed using the ARENA software. The simulator consists of different modules, and each module has a specific function to perform. Three different classes of realistic scenarios are tested and analysed in this simulation. Different scenarios require different modules to obtain the final output from the simulator. These modules are explained next.

4.2.1 *Model X*

A small-size real-world problem is studied as Model X. The remanufacturing facility simulated is assumed to have only one kind of core in the facility in addition to unused material. The outputs from the RL algorithm simulation are the *Q-factors*, which can be collectively analysed to derive the optimal switching policy for selection of raw materials. *Model X* is a small-scale problem, and it can also be studied via brute-force simulation. Only two actions are possible in this situation: selecting cores as raw material and selecting unused material. In the brute-force approach, every possible switching policy is simulated. The simulation is run for a long duration of time until the average cost stabilises.

4.2.2 *Model Y*

In *Model Y*, the manager has to choose from *two* different grades of cores, named A and B, and, of course, unused material. Thus, here, the manager has additional options available for raw-material selection. With the increase in the available options, the size of the problem expands drastically. The RL algorithm can be implemented as before; however, due to the large size of the problem, validation with a brute-force simulation is not feasible.

4.2.3 *Model Z*

In the previous two models, only one component was required for the assembly process from the remanufacturing stream. In this model, *two* distinct types of component parts, named I and II, will be assumed to be needed for the assembly process from the remanufacturing stream. Each of these required components can be selected from their respective core stocks or their unused material stocks. Since the assembly cannot take place without the availability of both the components at the assembly station, selection of one component as a core or unused material may delay the final production, depending upon the repair time and their availabilities.

4.3 *ARENA modules*

The designed simulator has a number of modules depending on the scenario being simulated. A broad classification of the modules is as follows.

(1) Cores arrival module.
(2) Unused material procurement module.
(3) Finished goods demand module.
(4) Agent (decision-maker) module.
(5) Data collection module.

4.3.1 *Cores arrival module*

Cores are generally bought by trading-off finished products. The arrival of cores in the simulation is represented by the generation of entities. The inter-arrival time of cores is assumed to be exponentially distributed in the simulation. See Figure 4.

4.3.2 *Unused material procurement module*

Unused materials are procured from outside suppliers. These materials are procured using a periodic-review order-up-to-*R* policy for inventory control. See Figure 5.

4.3.3 *Finished-goods demand module*

The demand for the final product in the remanufacturing facility simulation is generated with a Poisson process. Demand generation and the rate of production are adjusted to

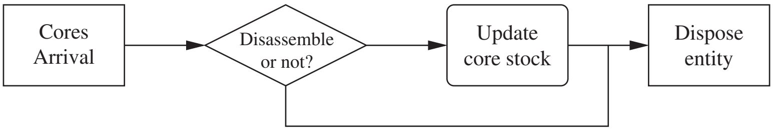

Figure 4. Core arrival module in ARENA.

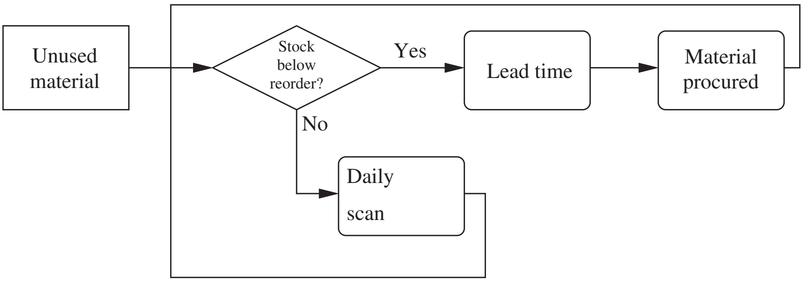

Figure 5. Unused material procurement module in ARENA.

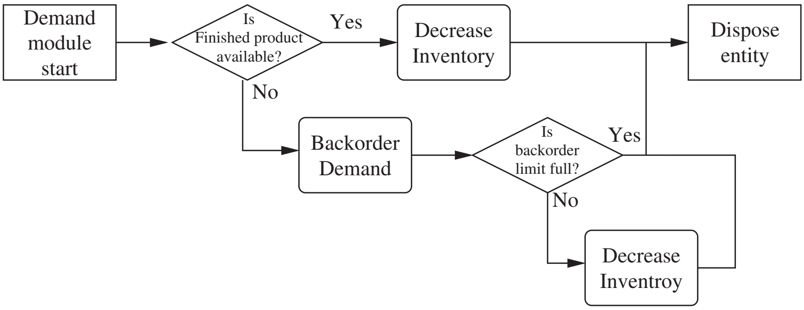

Figure 6. Demand generation for finished-goods module in ARENA.

have a stable system. As shown in Figure 6, as soon as the demand is created, a check is performed to determine whether this demand can be satisfied by available inventory of finished goods. If finished goods are available, the demand is met; otherwise the demand is back-ordered, and a penalty is imposed on the system. When the finished product becomes available, the back-orders are satisfied in the order in which they were recorded.

4.3.4 *Agent module*

Figures 7, 8, and 9 display the ARENA agent modules for Models X, Y, and Z, respectively. The decision of raw-material selection is incorporated into this module. The repair and assembly blocks have uniformly distributed time delays, which are different for the different systems studied. The *Q-factor* values for different polices and the average cost of simulation are also computed in this module.

4.3.5 *Data-collection module*

This module is required to collect the data generated during the simulation. It writes the *Q-factor* values into a file at the end of the simulation run. For the validation of the simulation in *Model X*, the average cost is also written into a file at the end of each simulation (Figure 10).

5. Numerical results

In this section, we describe the results of experiments performed with RL and a simulator of the remanufacturing environment. As described in the previous section, Models X, Y, and Z are simulated and analysed.

Table 1 describes the cases (systems) studied for Model X and also enumerates the optimal switching points for RL and a brute-force analysis. Here a periodic review inventory control for the unused material's procurement uses 1 day as the period and order-up-to-10 units as the policy. The switching points in the last two columns of Table 1 are from the core to unused material. Tables 2 and 3 list the cases for Models Y and Z.

Table 4 enumerates the *Q-factors* for Case 1 of Model X. A trend can be observed from the values of the *Q-factors*. The *Q-factors* for unused material have lower values than those of the cores *until* state 4. At state 5 and above, the *Q-factors* for unused material are higher than those for the cores. Thus the results of the RL algorithm suggest that a switching

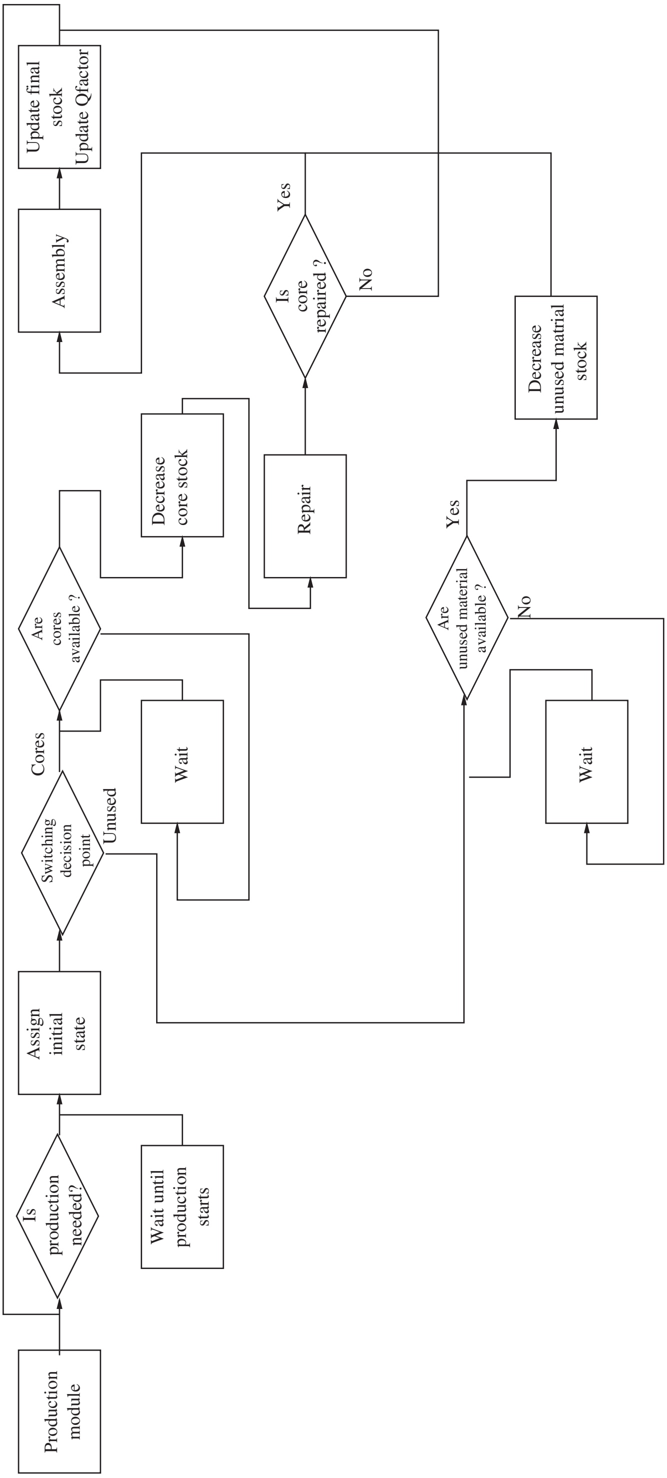

Figure 7. Production module in ARENA for Model X.

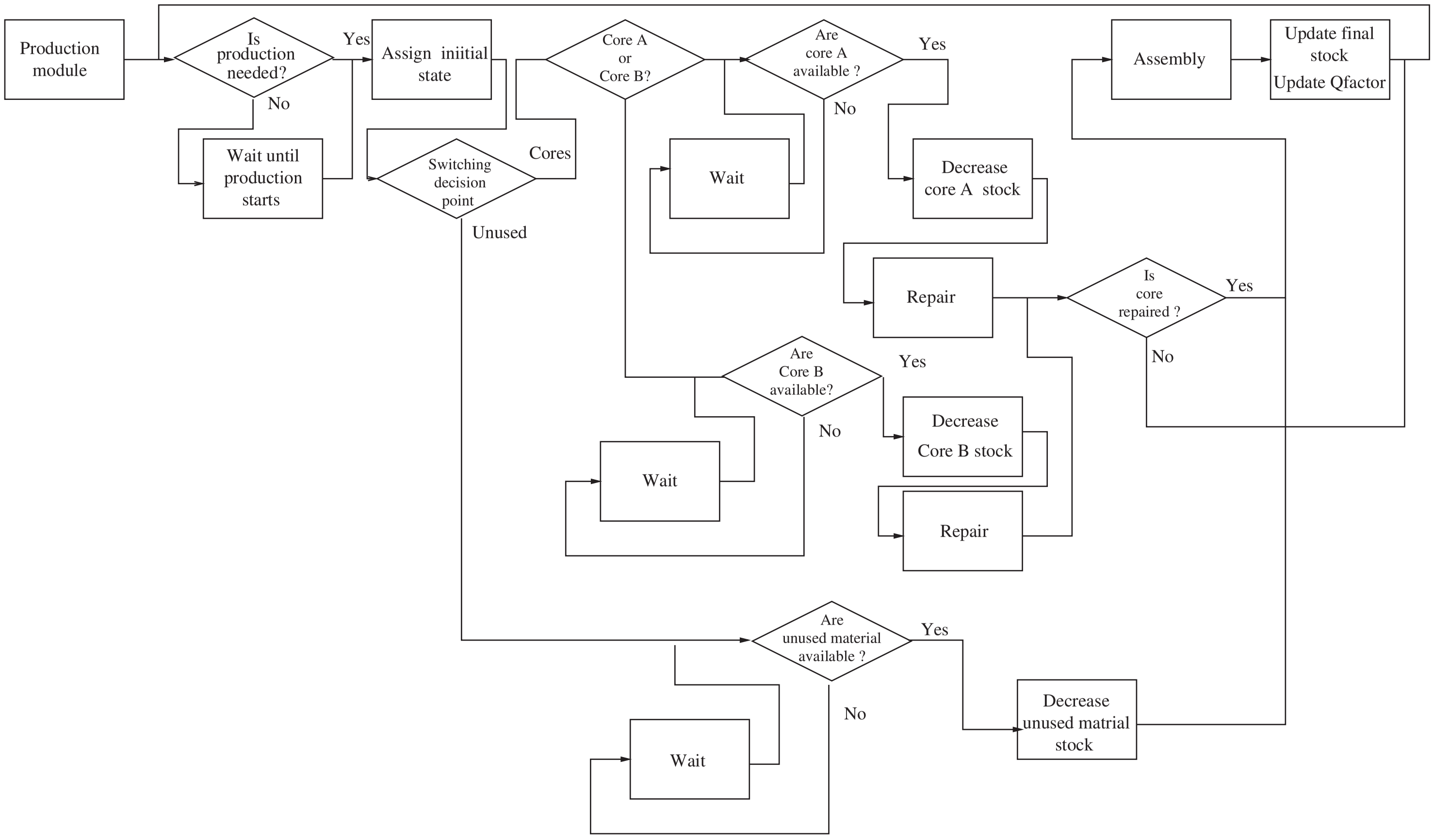

Figure 8. Production module in ARENA for Model Y.

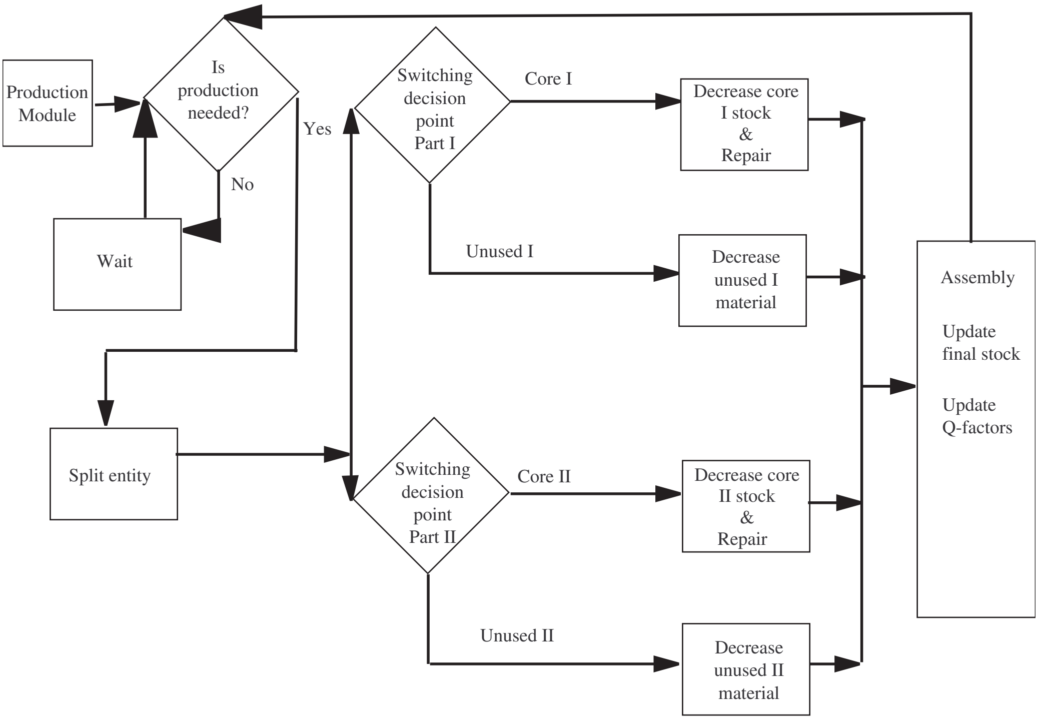
Production Module
Is production needed?
Yes
No
Wait
Split entity
Switching decision point Part I
Core I
Unused I
Decrease core I stock & Repair
Decrease unused I material
Switching decision point Part II
Core II
Unused II
Decrease core II stock & Repair
Decrease unused II material
Assembly
Update final stock
Update Q-factors

Figure 9. Production module in ARENA for Model Z.

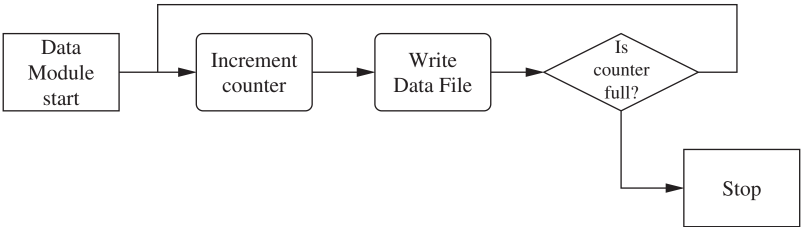

Figure 10. Data collection module in ARENA.

point exists at state 4. In other words, RL recommends switching from the core to unused material when the finished goods inventory falls to 4 or below. For every case in Table 1, a similar switching trend was observed; however, we do not show the actual Q-factor values for every case. Brute-force simulation is used to compute the average cost for all switching points. Figure 11 shows the average cost at each switching point for Case 1 obtained via brute force. From the graph it is obvious that the least average cost is for switching at state 6. Thus, clearly, RL does not produce an optimal policy. See, however, Table 5, which shows the gap between RL and the brute-force optimal solution. The optimality gap shown in Table 5 is defined as

$$\left[\frac{\rho_{\mathrm{RL}} - \rho^*}{\rho^*} \times 100\right].$$

Table 1. Parameters for the cases studied and simulation results for Model X. Here, expo(λ) denotes an exponentially distributed random variable with a mean of 1/λ. Time between successive core arrivals is expo(4).

Case	Demand interarrival time (days)	Core material cost ($)	Unused material cost ($)	Back-order cost ($)	Repair time (days)	RL switching point	Brute-force switching point
1	expo(5.8)	2	2.5	3.5	unif(4, 5)	4	6
2	expo(5.9)	2	3	17	unif(5, 6)	6	9
3	expo(5.1)	2	2.5	4	unif(4, 5)	7	9
4	expo(5)	3	4	7	unif(3, 5)	5	7
5	expo(5.8)	3	4	7	unif(3, 5)	5	8
6	expo(5.2)	2	3	4	unif(4, 5)	5	8
7	expo(5)	2	2.5	3.5	unif(4, 5)	6	9
8	expo(5.7)	3	4	7.5	unif(4, 5)	4	7
9	expo(5.9)	2	3	15	unif(5, 6)	6	9
10	expo(5.8)	2	4	6	unif(4, 5)	3	6

Table 2. Parameters for the cases studied and simulation results for Model Y. Legend: x is the level of inventory at which the RL policy switches from core A to core B; y is the same at which it switches from core B to unused. Time between successive core arrivals is expo(4).

Case	Demand interarrival time (days)	Core-A mat. cost ($)	Core-B mat. cost ($)	Unused material cost ($)	Back-order cost ($)	Repair time for A	Repair time for B	RL switching points (x, y)
1	expo(10)	2	5	9	20	unif(10, 15)	unif(2.5, 3)	(1, 4)
2	expo(11)	10	12	15	15	unif(10, 15)	unif(2.5, 3)	(1, 3)
3	expo(11)	10	12	15	35	unif(16, 17)	unif(2, 3)	(2, 4)
4	expo(11)	11	12	15	35	unif(16, 17)	unif(5, 6)	(2, 5)
5	expo(13)	10	12	15	15	unif(16, 17)	unif(2, 3)	(1, 4)
6	expo(6.5)	1	2	7	20	unif(5, 8)	unif(2, 5)	(2, 5)
7	expo(6.5)	1	2	7	10	unif(5, 8)	unif(2, 5)	(1, 4)
8	expo(12)	1	3	7	10	unif(15, 20)	unif(5, 8)	(2, 4)
9	expo(10)	8	9	15	30	unif(14, 15)	unif(2.5, 3)	(2, 6)
10	expo(10)	8	9	11	20	unif(14, 15)	unif(2.5, 3)	(3, 6)

Table 3. Average costs ($ per day) for Model Y. Improvement = $(\rho_{\text{Unused}} - \rho_{\text{RL}})/\rho_{\text{Unused}}$.

Case	RL policy	Unused material only	Improvement (%)
1	0.70284	1.3572	48.2
2	1.5530	1.9198	19.1
3	1.6702	2.2932	27.2
4	1.7739	2.2932	10.9
5	1.3955	2.2122	36.9
6	3.6529	3.7338	2.2
7	1.9403	2.2072	12.1
8	0.6518	1.9090	65.8
9	1.3206	2.1999	39.9
10	1.2340	1.5759	21.7

Table 4. Q-Factor values ($Q(i, a)$) for Case 1 of Model X.

State (i)	Core ($a=1$)	Unused ($a=2$)
−5	27.932965	25.794344
−4	27.888983	25.713552
−3	27.291605	24.982103
−2	26.459577	23.580912
−1	24.829261	21.554531
0	23.476489	19.288763
1	18.454407	15.707347
2	15.406330	13.920903
3	13.677106	13.266556
4	**12.865773**	**12.804720**
5	12.168334	12.286075
6	11.771908	11.978448
7	11.463808	11.622253
8	10.726801	10.500255
9	9.249244	8.228756
10	6.666473	4.835276

Note: The values in bold are for the state where switching must occur.

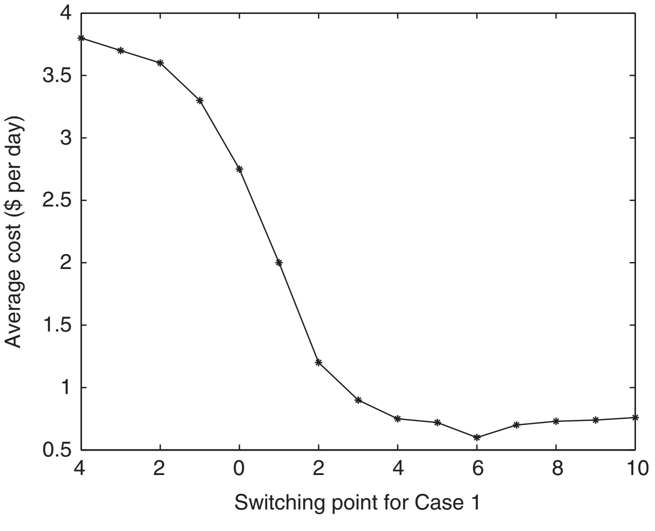

Figure 11. Average cost associated with different switching points obtained by brute force in Model X.

In the worst case, it is 10.9% and, in the best case, it is 1.4%. This provides sufficient encouragement that RL can produce near-exact solutions for these switching problems, and is likely to be useful where brute-force simulations are not possible because of problem size (Models Y and Z). In the context of the optimality gap, we need to point out that a change in the distributions of the repair time and the inter-arrival time for demands and

Table 5. Average costs ($ per day) for Model X.

Case	RL policy (ρ_{RL})	Unused material only	Brute-force optimal (ρ*)	Optimality gap (%)
1	0.4464	0.7141	0.4122	8.2
2	0.5028	0.6724	0.4716	6.6
3	0.4924	0.7935	0.4553	1.4
4	0.8470	1.1430	0.7631	10.9
5	0.6589	1.1426	0.6393	3.0
6	0.6004	1.1422	0.5529	9.6
7	0.5079	0.7144	0.4880	4.0
8	0.6719	0.6903	0.6622	1.5
9	0.4867	0.8560	0.4652	4.6
10	0.6810	1.2862	0.6480	5.1

Table 6. Simulation inputs for Model Z.

Case	Time between core arrivals I	Time between core arrivals II	Time between demands	Core I cost ($)	Unused I cost ($)	Core II cost ($)	Unused II cost ($)	Back-order cost ($)
1	expo(2)	expo(8)	17	8	10	5	8	24
2	expo(1)	expo(2)	11	5	7	3	5	10
3	expo(1)	expo(1)	11	5	7	3	8	10
4	expo(2)	expo(2)	12.5	5	7	3	8	10
5	expo(2)	expo(4)	17	5	7	10	15	15
6	expo(2)	expo(8)	17	9	10	10	12	24
7	expo(2)	expo(8)	16	9	10	10	12	24
8	expo(1)	expo(2)	12	5	7	3	5	20
9	expo(2)	expo(2)	17	8	10	5	8	24
10	expo(2)	expo(2)	18	8	10	5	7	30

cores can influence the optimality gap. However, our choice of these distributions is based on the following assumptions. There is a great deal of variability in the time between arrivals of cores and also demands, and hence the choice of the exponential distribution is made. Repair times are also likely to be random variables, but the variance is likely to be less. It was pointed out by a reviewer that the optimality gap could very well be a function of the choice of the distributions for these random variables. Hence, in order for a real-world manager to use these results, it is imperative that a thorough data-collection exercise be carried out in the beginning to determine the distributions in an appropriate manner. We must also point out that since the models we present are simulation-based, changing distributions in the simulator is not difficult.

Table 3 displays the results of the RL simulation for Model Y. In Model Y, there are two switching points: one from core A (cheaper core) to core B (costlier core), and the second from core B to unused material. Finally, the results of the product-mix model (Model Z) are displayed in Tables 6–8. Here, two different raw materials are required at the assembly facility from the remanufacturing stream of raw materials. The RL switching points for both parts are displayed in the results.

Table 7. Simulation inputs and results for Model Z.

Case	Repair time for I	Repair time for II	RL switching point (I)	RL switching point (II)
1	unif(18, 20)	unif(18, 20)	3	4
2	unif(12, 13)	unif(6, 7)	3	3
3	unif(12, 13)	unif(6, 7)	4	0
4	unif(12, 13)	unif(12, 13)	3	3
5	unif(19, 20)	unif(12, 13)	3	2
6	unif(18, 20)	unif(12, 13)	4	3
7	unif(18, 20)	unif(12, 13)	5	4
8	unif(6, 7)	unif(9, 10)	3	2
9	unif(18, 20)	unif(18, 20)	3	2
10	unif(18, 20)	unif(18, 20)	4	3

Table 8. Average costs ($ per day) for Model Z. Improvement $= (\rho_{\text{Unused}} - \rho_{\text{RL}})/\rho_{\text{Unused}}$.

Case	RL policy	Unused material only	Improvement
1	1.0064	1.3936	27.7
2	1.3393	1.7664	24.2
3	1.5785	2.0167	22.7
4	1.4613	1.6018	8.7
5	1.7196	1.9450	11.6
6	1.5147	2.6446	42.7
7	2.1173	2.7689	23.5
8	1.8186	2.2729	19.9
9	1.8634	2.3574	20.9
10	1.0122	2.0433	50.5

6. Conclusions

Remanufacturing is a philosophy that is gaining in popularity in the production world. It poses several new and exciting challenges to the production planner. As this philosophy becomes more visible, it is likely that the importance of studying logistics problems related to it will become more widely known. In this paper, we addressed an important problem of raw-material selection in a remanufacturing facility. If remanufacturing, which has tremendous environmental benefits, is to become economically viable, the raw-material selection problem described here has to be solved in a near-optimal fashion. This problem is complex and, as we have shown, it can be set up as a semi-Markov decision problem, the transition probabilities of which are not easily available for large-size real-world problems. Hence it cannot be solved easily via classical dynamic programming. An alternative to dynamic programming is to use brute-force simulation, which also is intractable for large-size problems. Hence, in this paper, we used a machine learning approach, namely reinforcement learning, which uses concepts of approximate dynamic programming within simulators, thereby bypassing the need for transition probabilities. The main attraction of RL is that it can scale up well to large problems. We demonstrate that, in general, the RL approach shows cost improvements over a strategy that uses unused material only. We believe that this is the first use of RL for environmentally conscious manufacturing.

A number of directions for future research can be envisaged. First, the problem studied here could be expanded to a much larger setting where scores of machines and multiple component types are involved. This will require a neural network to approximate the value function. Another possibility is to develop robust heuristics to determine the switching policy.

References

Bertsekas, D.P., 1995. *Dynamic programming and optimal control*. Belmont, MA: Athena, 2, 227–283.

Bertsekas, D.P. and Tsitsikis, J., 1996. *Neuro-dynamic programming*. Belmont, MA: Athena, 1, 245–251.

Bhattacharya, S., Guide, V.D.R., and van Wassenhove, L.N., 2006. Optimal order quantities with remanufacturing across new product generations. *Production and Operations Management*, 15 (9), 421–431.

Bradtke, S.J. and Duff, M., 1995. Reinforcement learning methods for continuous-time Markov decision problems, *Advances in neural information processing systems 7*. Cambridge, MA: MIT Press.

Ferrer, G. and Ketzenberg, M., 2004. Value of information in remanufacturing complex products. *IIE Transactions*, 36 (3), 265–277.

Ferrer, G. and Whybark, D., 2001. Material planning for a remanufacturing facility. *Production and Operations Management*, 10 (2), 112–124.

Geyer, R., van Wassenhove, L.N., and Atasu, A., 2007. The economics of remanufacturing under limited component durability and finite product life cycles. *Management Science*, 53 (1), 88–100.

Gosavi, A., 2003. *Simulation-based optimization: parametric optimization techniques and reinforcement learning*. Boston, MA: Kluwer Academic.

Gosavi, A., 2007. Adaptive critics for airline revenue management, *Proceedings of the 18th annual conference of the production and operations management society*. Dallas, TX.

Guide, V.D.R., *et al.*, 2000. Supply-chain management for recoverable manufacturing systems. *Interfaces*, 30 (3), 125–142.

Guide, V.D.R. and van Wassenhove, L.N., 2001. Managing product returns for remanufacturing. *Production and Operations Management*, 10 (2), 142–155.

Hoshino, T., Yura, K., and Hitomi, K., 1995. Optimization analysis for recycle-oriented manufacturing. *International Journal of Production Research*, 33 (8), 2069–2078.

Kongar, E. and Gupta, S., 2002. A goal programming approach to the remanufacturing supply chain model. *Proceedings of the SPIE international conference on environmentally conscious manufacturing*, 167–178.

Konstantarasa, I. and Papachristos, S., 2007. Optimal policy and holding cost stability regions in a periodic review inventory system with manufacturing and remanufacturing options. *European Journal of Operational Research*, 178 (2), 433–438.

Lund, R.T. and Bollinger, L., 1981. *Remanufacturing survey findings*. Cambridge, MA: MIT Center for Policy Alternatives, Technical report.

RIT Website, 2006. http://www.reman.rit.edu.

Savaskan, R.C., Bhattacharya, S., and Van Wassenhove, L.N., 2004. Closed loop supply chain models with product remanufacturing. *Management Science*, 50 (2), 239–252.

Sutton, R. and Barto, A.G., 1998. *Reinforcement learning: an introduction*. Cambridge, MA: The MIT Press.

van der Laan, E., *et al*., 1996. An (s, q) inventory model with remanufacturing and disposal. *International Journal of Production Economics*, 46 (1), 339–350.

van der Laan, E. and Teunter, R., 2004. *Simple heuristics for pull and push remanufacturing policies*. Rotterdam, Netherlands: Erasmus University Rotterdam, Econometric Institute Report.

Watkins, C.J., *Learning from delayed rewards*, Thesis. Cambridge: Kings College.

Zhou, L., *et al*., 2006. Dynamic performance of a hybrid inventory system with a Kanban policy in remanufacturing process. *Omega*, 34 (6), 585–598.

Zikopoulos, C. and Tagaras, G., 2008. On the attractiveness of sorting before disassembly in remanufacturing. *IIE Transactions*, 40 (3), 313–323.