Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

A General Model for Maintenance of Complex Systems

P. de Waal

Department of Operations Reasearch, Statistics, and System Theory

BS-R9201 1992

A General Model for Maintenance of Complex Systems

Peter de Waal
CWI
P.O.Box 4079, 1009 AB Amsterdam, The Netherlands

This paper describes a semi-Markov reward model for a process system or installation with components that are subject to failure. The installation is represented as a directed graph with the nodes representing the components and the arcs representing the material flows between the components. Each component may have several failure modes. For each component two actions can be specified: a production and a maintenance decision. The production decision specifies the production level of the component, and this action depends on the current mode of the component. The maintenance decision specifies whether maintenance or repair is to be performed on the component, and this action depends on the current component's mode plus the current maintenance manpower allocation. The state of the system together with the decisions determine the dynamic behaviour and the rewards, the latter of which is split into maintenance costs and production rewards. Some examples are included.

AMS Mathematics Subject Classification: 60K15, 60K20, 90B25

Keywords & Phrases: Semi-Markov reward model, reliability, maintenance, production

Note: This author is supported by a grant from Koninklijke/Shell Laboratorium, Shell Research B.V., Amsterdam

1 Introduction

In this note we describe a general stochastic model for a process system or installation with components that are subject to failures. The model is meant as a framework for further research on the relation between maintenance manpower and the performance of a production plant.

In the literature on reliability and maintenance little attention seems to have been paid to the relation between production and maintenance. In many practical situations, however, it seems that these two topics have a significant impact on each other. The decisions that concern the production management of each component, like for instance the production level, will influence the failure dynamics. A machine that is operating at its maximum capacity may show a failure behaviour that is different from a component that is producing at half speed. Maintenance decisions on the other hand also have an impact on production. A machine that is undergoing a major repair is generally not available for production.

In this paper we present a semi-Markov reward model for a production plant that shows the above mentioned interaction. We describe, in this order, the state space formulation, the control variables, the reward structure and the dynamic behaviour. The *state space* comprises

Report BS-R9201
ISSN 0924-0659
CWI
P.O.Box 4079, 1009 AB Amsterdam, The Netherlands

all the possible values that the state variables can take. The *state variables* carry all the information of the current quality or mode of the components *plus* the current maintenance manpower allocation. The *decision variables*, which are split in production and maintenance variables, represent the actions that are undertaken. The set of available actions depends on the current state of the process. The state variables and the actions together determine the *revenue* that is incurred. This revenue is split into rewards associated with production and costs associated with maintenance. The state and actions also determine the *dynamic behaviour* of the process. These dynamics may be caused by completion of repairs or changes in the state of a component due to tear or breakdown.

This paper is organised as follows. In Section 2 we describe the characteristics of a production plant. In Section 3 we introduce a Semi-Markov Reward Process for such a plant. In particular we introduce the state space description, the control structure, the reward and cost structure and the dynamic behaviour of the process. In Sections 4 and 5 we discuss some implications that arise from the Semi-Markov modelling approach. We present some examples in Section 6 and conclude with some remarks and suggestions for future research topics in Section 7. An selected overview of related literature is given in Appendix A.

2 Description of a production plant

A production plant is a set of N components. The index set of components is denoted as $\mathcal{C} = \{1, \ldots, N\}$. A component can either be a *production unit (pu)* or a *buffer*. We let $\mathcal{C}_P$ and $\mathcal{C}_B$ denote the sets of pu's and buffers, respectively. Of course $\mathcal{C}_P \cap \mathcal{C}_B = \emptyset$ and $\mathcal{C}_P \cup \mathcal{C}_B = \mathcal{C}$. Buffers are passive items, which means that they cannot break down, nor can we apply any control on how they behave. They function solely as storage space of materials and intermediate products. An example of a production plant is depicted in Figure 1. The pu's and buffers are represented as circles and triangles, respectively.

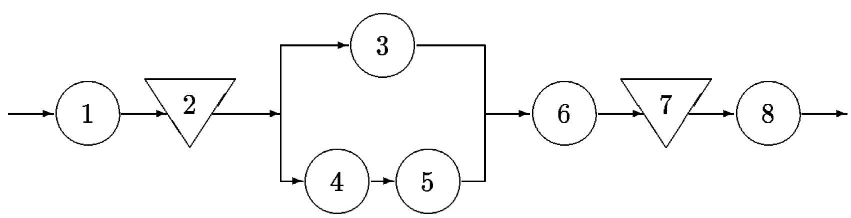

Figure 1. Example of a production plant

The plant layout is captured in a directed graph. The nodes in the network represent the components and the directed arcs correspond to the flow of products in the plant. The values of the flows over the links in the network are determined by the flows through the production units – i.e. the production rates of the pu's – plus the constraints that arise from the layout of the plant and the contents of the buffers. An algorithm to compute these values is very problem dependent and we shall elaborate on this in Section 5.

3 A Semi-Markov Reward Model

In this section we shall introduce a semi-Markov reward (SMR) model to describe the behaviour of a production plant.

3.1 State space description

The state space of the model must reflect the "state" or of the plant. In the setting of SMR models this means that the state must provide all information to determine the available control actions, and the control action together with the state will determine the future behaviour of the system and the direct costs and revenues.

Recall that the plant is a connected set of components and that the behaviour of the plant is determined by the behaviour of the individual components and the way in which the components are connected. The state of the SMR model must thus reflect the state or mode of all the components. Since the current allocation of the repair facilities will also influence the direct costs, they must be included in the state description of the model.

Notation 3.1 *The state variable of component $i \in \mathcal{C}$ is denoted as s_i and the set of possible values of s_i as $\mathcal{S}_i$. The state of the repair facilities is denotes as $s_{\mathcal{M}}$ and it must lie in some abstract set $\mathcal{S}_{\mathcal{M}}$. The state of the SRM model is $s = (s_1, \ldots, s_N, s_{\mathcal{M}})$, and the state space of the plant is thus $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_N \times \mathcal{S}_{\mathcal{M}}$.*

As we mentioned in Section 2 we distinguish between two types of components, viz. production units (pu's) and buffers. The state of a buffer represents the actual contents of the buffer. The value of s_i of a buffer is allowed to be in an interval $[0, L_i]$ for some $L_i \in \mathbb{R} \cup \{\infty\}$.

The state variable s_i of a pu i is a two-dimensional vector (s_i^1, s_i^2). The first component s_i^1 represents the state or production quality of unit i and it is an indication of how well the unit is able to perform. We shall use the phrase *mode* to denote this first component s_i^1. Its value must lie in a finite discrete set S_i. Some value may for instance represent a failure and another value may correspond to a brand new unit. The second component $s_i^2 \in \mathbb{R}$ is the time that the unit's production quality has been in the current mode s_i^1. It is introduced to allow non-exponential lifetime and repairtime distributions.

The mode s_i^1 of a pu must not necessarily have a direct relation with a real life technical situation. As an example s_i^1 can lie in the set $\{0, 10, 20, \ldots, 100\}$ and the mode represents the percentage of the full capacity that is actually available for production. In another example the state space might be the abstract set $\{\mathtt{down}, \mathtt{bad}, \mathtt{good}\}$. The interpretation can be that in the `down` mode no production is possible, in the `bad` and `good` mode a full production level is available, but the pu is more likely to deteriorate (get into the `down` mode) from the `bad` mode.

The mode of a pu is thus an indication of the production quality (or maximally available production level). The production rate of the unit will be determined by the state *and* the production level control variable (to be described in Section 3.2).

The state $s_{\mathcal{M}}$ of the repair facilities is an abstract representation of the allocation of manpower for repair. The exact form is very problem dependent. In a model with mobile repairmen the variable $s_{\mathcal{M}}$ may for instance represent the current location of all the repairmen. In the classical machine repair problem (see for instance Tijms [34, Chapter 4] and Kelly [20, Chapter 4]) $s_{\mathcal{M}}$ may represent the order of the machines in the queue for the repair shop.

3.2 Control variables

At any moment in time we must be able to take actions to influence the behaviour of the plant. We shall take into account two ways of controlling the plant, viz. by controlling the production levels and by allocating the maintenance manpower. The purpose of this control structure is twofold. First it allows us to influence the future behaviour of the plant: performing maintenance on a pu will probably bring it into a better mode, or choosing a high production level might increase the wear. Second it permits us to identify certain costs and rewards with decisions: we may have alternatives for maintenance strategies with varying costs, or revenues may differ with the chosen production levels.

In Section 3.3 we shall describe how the current state of the plant together with the actions will determine the plant revenues and costs. The influence of state and control on the evolution of the state shall be described in Section 3.4. Since buffers are considered passive resources, control actions are defined only for pu's.

If the state of a pu $i \in \mathcal{C}$ is $s_i \in \mathcal{S}_i$, then the *unit production level* p_i is allowed to take values in a set of allowable actions $\mathcal{U}_p^i(s_i)$. Like in the representation of unit modes the control variable p_i need not have any physical interpretation. The control variable p_i represents the chosen production level and it determines, together with the state s_i of the unit, the production rate of the unit *in isolation*. This is not necessarily the actual production rate of the unit in the plant, since this depends also on the plant layout and the capacity of other units. As an example consider units 4 and 5 in Figure 1. We assume that the production rate of both units must be equal. Now suppose that unit 4 has failed and unit 5 is in a brand new state. The production level decision at unit 4 can only be not to produce, while at unit 5 we can choose any production level that we want. However, the actual production rate in the "subplant" of unit 4 and 5 will always be zero, since this is imposed by the failure in unit 4. In effect this means that it is possible to make insensible production level decisions, but a controller that is trying to maximize revenues will not choose such a decision.

The second type of control action concerns the allocation of maintenance manpower. If the state of a unit $i \in \mathcal{C}$ is $s_i \in \mathcal{S}_i$, then the *unit maintenance effort* m_i is allowed to take values in a set $\mathcal{U}_m^i(s_i)$. Just like the state variable $s_{\mathcal{M}}$ (the current maintenance manpower allocation) the variable m_i does not necessarily have any physical interpretation; it is an abstract representation of maintenance manpower allocation. The control action m_i determines, together with the state of the unit, the dynamics of the unit. Just like in the production level control, the set of available decisions $\mathcal{U}_m^i(s_i)$ of unit i depends only on the state of the unit, and not on the states of the components in the rest of the plant. Limitations in the total available maintenance manpower can, if necessary, be implemented by overlaying the product set $\mathcal{U}_m^1(s_1) \times \cdots \times \mathcal{U}_m^N(s_N)$ with extra constraints. An alternative and more realistic approach to include this kind of restrictions is to put high costs on maintenance allocations that exceed the available manpower. For the same reason there is no influence of the current maintenance manpower allocation, as given by the component $s_{\mathcal{M}}$ in the state variable, on the possible choice of the control variable m_i. Maintenance decisions that are undesirable because of the current allocation, should be discouraged by imposing high costs on them.

NOTATION 3.2 *If the state of the plant is* $s = (s_1, \ldots, s_N, s_{\mathcal{M}})$, *then the plant control variable is represented as* $u = (u_1, \ldots, u_N)$ *with* $u_i = (p_i, m_i) \in \mathcal{U}_p^i(s_i) \times \mathcal{U}_m^i(s_i)$. *The set of allowable control actions in state* s *will be denoted as* $\mathcal{U}(s) = (\mathcal{U}_p^1(s_1) \times \mathcal{U}_m^1(s_1)) \times \ldots \times$

$(\mathcal{U}_p^N(s_N) \times \mathcal{U}_m^N(s_N))$. *For notational convenience we also introduce* $u_{\mathcal{P}} = (p_1, \ldots, p_N)$ *and* $u_{\mathcal{M}} = (m_1, \ldots, m_N)$.

With respect to control of the production plant, we shall restrict our attention to *deterministic state feedback policies*, meaning that we only allow non-randomised decisions that are based on the current state of the process.

3.3 Cost and revenue

We already mentioned that the control actions together with the state of the plant will determine the direct costs and revenues. The direct costs are usually related to the maintenance and repair actions, while the direct revenues are associated with production level decisions. In the remainder of this paper we shall restrict ourself to the use of revenues, with the convention that costs are revenues with a negative value.

Assume that the state of the plant is $s \in \mathcal{S}$ and the decision $u \in \mathcal{U}(s)$ is chosen, then the total revenue per time unit will be denoted as $c(s, u)$. It can be decomposed into $c(s, u) = c_p(s, u) + c_m(s, u)$, the revenues per time unit from production and maintenance, respectively. In the functions c_p and c_m we can include immediate or lump rewards (cf. Ross [30, Chapter 7]), but we shall not elaborate on this to avoid notational complexity.

The production revenue $c_p(s, u)$ in state $s \in \mathcal{S}$ under action $u \in \mathcal{U}(s)$ represents the value that is added to the materials and products by the processing in the pu's. In particular this means that the production revenue will be a function of the flows of products and materials in the plant. In order to determine the value of $c_p(s, u)$ we need to compute the value of the flows on all the arcs in the production network. Algorithms for this computation will be discussed in Section 5.

In the function c_p we can incorporate "restrictions" on the control decisions, that were not modelled in the control action sets $\mathcal{U}(s)$. Consider for instance the decision to have a machine operate at full speed and at the same time have it undergo repair. If this is an impossible combination, then the revenue for this decision should have a large negative value. In a situation where a controller is responsible for optimizing revenue, then clearly such a decision is very unlikely to be taken. This mechanism thus allows us to implement restrictions that arise from practical considerations, in an indirect way. Via the function c_p we also include the restrictions arising from the physical layout of the plant. Using the example that was mentioned in Section 3.2, the production revenue should reflect the uselessness of operating unit 5 when unit 4 has broken down. This is possible by making c_p independent of p_5 when the production decision p_4 at unit 4 is "do not produce". Moreover, by making the pu 5 wear out faster, when it is decided to produce in this situation, we can force an optimizing controller to avoid this decision.

The function $c_m(s, u)$ is the (presumably negative) revenue for the manpower allocation u in state s. It is intended to model only the costs involved with the decision $u_{\mathcal{M}}$ — the maintenance allocation — and it should not depend on the production level decisions $u_{\mathcal{P}}$. Strictly speaking c_m should thus be a function only of the current allocation $s_{\mathcal{M}}$ and the new allocation $u_{\mathcal{M}}$, but this is merely a question of formulation. The function c_m models the real cost of maintenance allocation, but it is also used to induce restrictions on allocation decisions. Consider for instance the situation that there are 10 repairmen, who can each perform maintenance on one pu at a time. By imposing a huge (or infinite) cost on any allocation decision $u_{\mathcal{M}}$ that assigns more than 10 repairmen to do repair, an optimizing controller will never make such a decision. The advantage of inducing this kind of restrictions indirectly, as opposed to hardwiring them into the action sets $\mathcal{U}$, is the flexibility that it offers. Restrictions in practice are always a little looser than they seem at first; in the above mentioned example we can adjust the model to make it possible to hire an extra repairman externally, possibly at a substantially higher cost than an ordinary repairman. The dependency of the cost c_m on the current maintenance allocation $s_{\mathcal{M}}$ is important, because it enables us to model extra costs when the current and the new allocation differ significantly. As an example we can think of a model for two production platforms, where there is a high cost involved in transferring a repair crew from one platform to another. In a model for a plant with only one production site, this approach can be used to include switching costs, when changing the manpower allocation.

3.4 Plant dynamics

The dynamics of the model should reflect how the state of the plant evolves in time, given the current state and the chosen control actions. Recall that the state of the plant was represented by the states of the different production units and buffers, plus the current maintenance manpower allocation.

The state of a production unit represents the actual production quality of the unit plus the time spent at that quality level. Assume that a given unit $i \in \mathcal{C}_P$ is in state $s_i \in \mathcal{S}_i$ and we apply control action $u_i \in \mathcal{U}_p^i(s_i) \times \mathcal{U}_m^i(s_i)$. The time until the quality level or mode of the unit changes (taken from the entrance time to that level), has a distribution function $F_i(s_i, u_i) : \mathbb{R}_+ \to [0,1]$ and the next quality level of the unit will be $s_i{}'$ with probability $p_{s_i,s_i{}'}(u_i)$. A change of mode can occur due to wear of the unit or completion of a maintenance action.

The second component of the state variable of a pu represents the time that the unit has been at the current quality level. Since this is a continuous variable, we are in principle allowed to constantly change the control variables. For practical purposes we shall allow changes in the decisions only when this timer exceeds some level, and for each component we shall allow only a finite set of levels. This is a restriction, but it still allows us, for instance, to admit age replacement type strategies.

The state of a buffer represents the actual contents of the buffer and this will change constantly if there is an imbalance in inflow and outflow. This means that the actual production flows in the plant can stay the same until one of the buffers becomes empty or full. The time until this event can be computed from the buffer contents and the flows. At that time we might get blocking or starvation and the actual production flows will have to be recalculated. After adjustment of the flows, the pu states may continue to evolve as if nothing has changed, since the distribution function $F_i(s_i, u_i)$ does not depend on the flows. An alternative is to allow a change of decisions when some of the buffer levels change. An example is to change the production level decisions when a certain buffer becomes 90% full. For each buffer only a finite set of such buffer levels may be considered.

The last component of the state variable, $s_{\mathcal{M}}$, changes as follows. If the current allocation is $s_{\mathcal{M}}$, and the new allocation decision is $u_{\mathcal{M}}$, then the time until the allocation changes — taken from the time that the decision is made — has distribution function $F_{\mathcal{M}}(s_{\mathcal{M}}, u_{\mathcal{M}})$.

After the transition the new value of $s_{\mathcal{M}}$ should be such that it corresponds to the manpower allocation decision $u_{\mathcal{M}}$.

4 Remarks

In this section we shall explain some of the choices that we have made in the model.

Remark 4.1 For each component i we have chosen the action sets $\mathcal{U}_p^i$ and $\mathcal{U}_m^i$ to depend only on the local state s_i. Note that this does not mean that we cannot use global state information in the decisions. One reason for this choice is to allow the decision maker to make unprofitable decisions. In the case of $\mathcal{U}_m^i$ this means that we can have "infeasible" allocations bring about high costs. An infinite cost would then represent a really impossible allocation; a cost that is considerably larger than other allocations could represent the extra expenses of hiring manpower externally.

The action sets $\mathcal{U}_p^i$ for production level decisions do not use global state information, since the restrictions that are caused by the plant layout and the buffer contents are very difficult (or practically impossible) to calculate beforehand. It also would require that the plant would have to be controlled globally, i.e. all the decisions p_i of all the components would have to be determined at the same time, thus prohibiting distributed control.

Remark 4.2 In Section 3 there is no explicit description of the possibility that components may queue for maintenance. One aspect of queueing – the availability of servers – is modelled implicitly by the maintenance allocation costs. Queueing disciplines can be implemented by an appropriate choice of the maintenance allocation strategy.

5 Product flows

In this section we discuss some aspects of the computation of product flows. In Section 3.3 we described that in order to determine the product revenue $c_p(s, u)$ for a given state $s \in \mathcal{S}$ and decision $u \in \mathcal{U}(s)$, we have to determine the flows on all the arcs in the network. Unfortunately it is impossible to give a general exact algorithm for this computation, because it is very dependent on the way the plant's physical restrictions are implemented. We shall illustrate some of these aspects using the plant model of Figure 1 as an example.

Consider a pu $i \in \mathcal{C}_P$. The first step in the algorithm is to determine the product flow through this pu if it would have an infinite supply at the input side and an unlimited buffer at the output side. The value of this flow is uniquely determined by the state s_i of the pu and the production level decision p_i.

The layout of the plant may impose restrictions on the actual flows in a number of ways. First there may be a production unit j with its output connected to the input of i. If the production flow through j is smaller than the flow through i, and there is no buffer in between the two pu's, then the actual production flow through i will be reduced to that through j. In such a situation we say that i is *starving*. As an example consider components $j = 4$ and $i = 5$ in Figure 1. The actual flow through i has to be reduced to the flow through j. This situation can also occur when there is an empty buffer between i and j. Note that after this reduction we may also have to adjust flows through other components downstream of i.

A second type of conflict may arise when the output of i is connected to the input of j and i is producing faster than j. If there is no buffer in between the two components, or if there is a buffer but it is full, then the flow rate through i has to be adjusted. In such a situation we say that i is (partially) *blocked*. One way to resolve the flow imbalance is to let i produce at the chosen rate and let the excess flow be lost. The production revenue function $c_p(s, u)$ should account for this waste. Note that in this approach there is an imbalance in the flow into and out of the component. Another approach is to force the flow rate through i to reduce to j's flow rate. This approach differs from the first one, since it may bring about further adjustments upstream of component i.

Conflicts can also arise when several parallel branches of the production network join their outputs into the input of one component. In Figure 1 we find such an example in components 3, 5 and 6. Here we may encounter situation similar to the ones described above. If for instance the total flow out of 3 and 5 is larger than the flow of 6, then either 3 or 5 or both have to be slowed down (or have their outputs lost). The algorithm for this adjustment should be part of the design of the plant.

6 Examples

In this section we describe some examples to illustrate the formulation described in the preceding sections. The first example considers a system of two production units with a buffer, and the second example concerns a k-out-of-n system with c repairmen.

6.1 *Example 1: Two production units and a buffer*

The first example concerns a serial production line, consisting of two production units and a buffer (see Figure 2). In this model intermediate products from pu 1 (component 1) are stored in a buffer (component 2), which provides input for component 3.

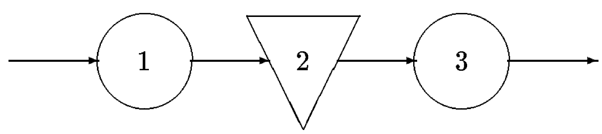

Figure 2. A serial production line with a buffer

The modes of components 1 and 3 represent the production quality of the units. For both components we distinguish three levels of quality, viz. down, bad and good. With respect to component 2 we assume that the buffer has an infinite capacity, so $\mathcal{S}_2 = [0, \infty)$.

The state of the repair facilities is described by the vector $s_{\mathcal{M}} = (s^1_{\mathcal{M}}, s^3_{\mathcal{M}}) \in \{0,1\} \times \{0,1\}$. A value of $s^i_{\mathcal{M}} = 1$ means that repair or maintenance is being performed on component i, $i = 1, 3$, while $s^i_{\mathcal{M}} = 0$ means no repair.

The action spaces for the production levels are constructed as follows. If a pu is down, then no production is possible. We denote this as $U^1_p(\texttt{down}) = U^3_p(\texttt{down}) = \{0\}$. If a pu is in mode bad, then we can decide to let the unit produce (action $p_i = 1$) or to let the unit idle (action $p_i = 0$), so $U^1_p(\texttt{bad}) = U^3_p(\texttt{bad}) = \{0, 1\}$. The same set of actions is available for pu 3 if its mode is good, i.e. $U^3_p(\texttt{good}) = \{0, 1\}$. For pu 1 we have an extra production action available, namely to produce at half speed, so $U^1_p(\texttt{good}) = \{0, 0.5, 1\}$.

The action sets for maintenance manpower allocation are identical for both pu's. If a pu is in mode good, then the only choice is not to repair (action $m_i = 0$). For the down and bad modes we have a choice to repair (action $m_i = 1$) or to do nothing (action $m_i = 0$). Summarizing we thus have $U_m^i(\texttt{good}) = \{0\}$ and $U_m^i(\texttt{down}) = U_m^i(\texttt{bad}) = \{0, 1\}$, $i = 1, 3$.

The production revenues represent the values that are added to the (intermediate) products by the pu's. We assume that the production flows through the pu's are equal when they are both operating at full speed and that the product revenue is not affected by the maintenance manpower allocation. Note that no value is added at a pu when the flow through that pu is 0. First we consider component 1. We assume that the revenue gained per time unit is proportional to its production speed and this revenue is normalized to 1 when operating at full speed. For pu 3 we assume that a value of 2 is added to the intermediate products per time unit when it is operating. (Recall that for pu 3 we can only decide to operate or to idle.) However, in the computation of the actual product revenue of pu 3 we have to account for the possibility of an empty buffer. If the buffer is not empty ($s_2^1 > 0$), then a revenue of 2 is incurred per time unit when $s_3^1 \in \{\texttt{bad}, \texttt{good}\}$ and $p_3 = 1$. If the buffer is empty, then the flow through pu 3 is bounded by the output flow of pu 1. Summarizing, in state $s = (s_1, s_2, s_3, s_{\mathcal{M}})$ the production revenue in pu 1 under action $u = (u_1, u_3)$, with $u_i = (p_i, m_i)$, $i = 1, 3$, is equal to p_1 (provided the unit is not down) and the production revenue in pu 3 is

$$\begin{cases} 2 * p_3, & \text{if } s_2^1 > 0 \text{ and pu 3 is not } \texttt{down}, \\ 2 * p_1 * p_3, & \text{if } s_2^1 = 0 \text{ and pu 1 and pu 3 are not } \texttt{down}, \\ 0, & \text{otherwise}. \end{cases} \tag{1}$$

The maintenance costs are calculated as follows. We assume that there is one repairman who can work on one pu at a time. We can hire an extra repairman, but this gives a considerable extra cost. The maintenance manpower revenues are

$$c_m(s, u) = -1(m_1 + m_3 = 1) - 10 * 1(m_1 + m_3 = 2). \tag{2}$$

We conclude this example with a description of the dynamic behaviour of the plant. If a pu is down and no repair is undertaken, then the unit stays in that mode. If a pu is in mode down or bad and repair is being done, then the unit will become operational again (good) after 10 time units. A unit in mode bad with control chosen not to repair, will go down after a time that has a Weibull distribution with shape parameter 2 and mean 100. A unit in mode good under production will stay in that mode for a time that has a Weibull distribution with shape parameter 3 and mean 80. After that time it will become bad with probability 0.9 or down with probability 0.1.

6.2 Example 2: A k-out-of-n system with c repairmen

The second example concerns a plant that consists of 20 pu's, denoted as $i = 1, \ldots, 20$. Each pu i can be in any of 5 different modes numbered $s_i = 0, \ldots, 4$. The mode 0 corresponds to a breakdown, while in the other modes the component is capable of making products. The quality of the pu is an increasing function of the value of its state variable, so $s_i = 4$ corresponds to a unit being brand new, while a value of $s_i = 0$ means that the unit has broken down. The maintenance allocation state variable is represented as $s_{\mathcal{M}} = (s_{\mathcal{M}}^R, s_{\mathcal{M}}^W)$. Here $s_{\mathcal{M}}^R$ represents the set of indices of the pu's that are under repair. We assume that there are at most 5 repairmen available, so $s^R_{\mathcal{M}}$ must be a subset of $\{1,\ldots,20\}$ with at most 5 elements. The second component of the state variable, $s^W_{\mathcal{M}}$ represents the pu's that are waiting for repair. It is a vector $s^W_{\mathcal{M}} = (s^W_{\mathcal{M}}(1),\ldots,s^W_{\mathcal{M}}(l))$, where l is the number of pu's waiting and $s^W_{\mathcal{M}}(i)$ is the index of the i-th pu in the queue.

The action sets for the production decisions are quite simple: if a pu is capable of production ($s^1_i \in \{1,\ldots,4\}$), then the available decision p_i is either to produce ($p_i = 1$) or to let the machine idle ($p_i = 0$). If the pu is down ($s^1_i = 0$), then there is no choice: we have to stop production ($p_i = 0$).

The action sets for maintenance decisions are identical for all components: If a pu is in mode 0, 1 or 2, then can we choose to either move it to the repair shop ($m_i = 1$) or to do nothing ($m_i = 0$). In modess 1 and 2 a decision of $m_i = 1$ corresponds to preventive maintenance. If a pu is mode 3, 4 or 5, then the only maintenance decision for the pu is to do nothing ($m_i = 0$).

The production revenues represent the values of the goods that are produced by the pu's. The value does not depend on the quality of the machine, so the total revenue per time unit is proportional to the number of operational machines. However, when there are less than 10 machines operational *and* producing, then we incur extra costs because of contractual obligations. More specifically, if the state of the system is $s = (s_1,\ldots,s_{20},s_{\mathcal{M}})$, then the total production revenue under action $u = (u_1,\ldots,u_{20})$ with $u_i = (p_i,m_i)$, $i = 1, \ldots, 20$, is

$$c_p(s,p) = \sum_{i=1}^{20} 1(s^1_i > 0) * p_i - K_1 * 1\left\{\Big(\sum_{i=1}^{20} 1(s^1_i > 0) * p_i\Big) < 10\right\} \tag{3}$$

where $K_1 \in \mathbb{R}$ is the penalty cost for not having the required number of 10 operational machines.

The maintenance costs are split in a lump cost that is paid at the decision moment plus a cost that is paid per time unit until the next decision epoch. The lump costs are start up costs and they have to be paid when all the repairmen are idle and one of the pu's is sent to the repair shop. This lump cost can be expressed in the state variable $s_{\mathcal{M}} = (s_1,\ldots,s_{20},s_{\mathcal{M}})$ and the decision variables $u = (u_1,\ldots,u_{20})$ with $u_i = (p_i,m_i)$, $i = 1, \ldots, 20$, as follows:

$$K_2 * 1\Big(\#(s^R_{\mathcal{M}}) = 0\Big) * 1\Big(\sum_{i=0}^{20} m_i > 0\Big). \tag{4}$$

where $\#(s^R_{\mathcal{M}})$ is the number of elements of the set $s^R_{\mathcal{M}}$ and K_2 is the overhead cost for repair. The "continuous" part of the maintenance costs are paid per time unit for each operational repairman, so for state s and decision u they are proportional to

$$\min\Big\{5, \sum_{i=1}^{20} m_i\Big\} \tag{5}$$

The dynamics of the plant can be summarized as follows. Repair of a broken down unit requires an amount of time (excluding waiting) that is uniformly distributed over the interval $[9.0, 11.0]$. Preventive maintenance on a a pu in mode 1 or 2 takes 5.0 time units with probability 0.8 and 10.0 time units with probability 0.2. When repair of a pu completes then a pu is fetched from the waiting queue in FIFO order, if any. Otherwise the repairman waits until a pu is sent to the queue. A pu that is not broken down and not under preventive maintenance. (i.e. $s_i > 0$ and $m_i = 0$) can only change to another mode when it is producing; if it is left idling then it stays in the same mode. A pu that is actually producing (i.e. $s_i > 0$ and $p_i = 1$), stays in that mode for an Erlang distributed time, after which it will change to an inferior mode. The parameters of the "lifetime" distributions per mode and the transition probabilities are given in Table 1.

	Erlang lifetime parameters		transition probability from s_i to s_j			
state	# phases	mean	$s_j = 3$	$s_j = 2$	$s_j = 1$	$s_j = 0$
$s_i = 4$	4	10.0	0.8	0.1	0.1	
$s_i = 3$	4	8.0		0.7	0.2	0.1
$s_i = 2$	3	5.0			0.6	0.4
$s_i = 1$	3	5.0				1.0

TABLE 1. Lifetime distributions and transition probabilities

7 Topics for future research

We already mentioned in the Introduction that the model described in this note is meant as a framework for future research on the relation between maintenance manpower and performance of production plants. In this section we introduce some proposals for research topics.

PROPOSAL 7.1 A model that consists of two production units with a buffer in between can be considered. The performance of several heuristics for control should be evaluated, preferably by exact analysis. Research topics should include the influence of lifetime and service time distributions, buffer size, maintenance strategies, etc.

PROPOSAL 7.2 A study can be made on the effect of switching costs for maintenance manpower allocation. As a basis for this we can consider the model of ASSAF AND SHANTHIKUMAR [5], which can be extended to include two production sites with one repair crew. The switching of the manpower allocation may have its effect on both the costs structure and the dynamic behaviour: transportation of the crew from one platform to another costs both money and time.

PROPOSAL 7.3 For a small plant, e.g. two production units with one buffer, different optimisation criteria may be formulated. Both the optimal production and maintenance strategies can be investigated. Special attention should be paid to the influence of the optimisation criteria and the influence of the cost and reward structure.

PROPOSAL 7.4 An investigation should be made of algorithms for exact and approximate analysis of complex plants. The idea is to provide a set of algorithms that allow a quick and easy evaluation of the performance of different operational strategies. The algorithms could form the core of a decision support system to be used in the operation of a real production plant.

A Literature overview

This appendix presents a literature overview of maintenance and repair models in which there is some kind of interaction between production and maintenance. In most papers this interaction is due to a limitation in manpower for repair and maintenance or due to a limited number of spare parts.

A.1 General machine repair models

The standard machine repair model consists of a finite number of machines that are subject to failures. A finite number of repairmen are available to perform repair. The main difference with machine replacement models is the fact that repairs take time as opposed to replacement, which might be considered immediate repair.

A.1.1 Performance analysis model

Albright and Soni present in [3] computational algorithms for the performance analysis of a machine repair model with one machine type, several repairmen and exponential lifetime and repairtime distributions. Included in the formulation is the probability that a failure of a machine may be non-repairable. New machines are ordered according to a state dependent ordering policy.

In [13] Gross and Ince present an approximation method for a machine repair model with two machine types, one FCFS server and exponential lifetime and repairtime distributions. The approximation is based on state aggregation combined with brute force calculation of the equilibrium distribution.

In Foster and Garcia-Diaz [11] the performance analysis is presented of some specific Markov models of machine repair problems. It is shown that the steady state distribution has a product form in the case of exponential lifetime and repairtime distributions for a series system, a k-out-of-n system and a k-out-of-n system with multiple repairmen.

In van der Heijden and Schornagel [19] an approximation method is introduced for the computation of the stationary interval uneffectiveness distribution of a k–out–of–n system. This distribution is approximated by the distribution of a two–state single unit, with Gamma type life time and repair time distributions.

A.1.2 Optimal stochastic control models

Albright considers in Albright [2] an optimal control problem for a machine repair model. The system consists of several machines of one machine type and failure and repair time distributions are exponential. The service rate, failure rate and number of servers are considered to be controllable. The main result is the derivation of monotonicity properties of the value function with respect to the optimal decision variables. The cost function here is a combination of service costs, machine operation costs and revenue from production. The monotonicity properties include for instance the fact that the number of repairmen to use increases when the number of broken down machines increases.

In Crabill [10] an optimal stochastic control problem is discussed for a machine repair problem where there is one server that can operate with different service rates. Lifetime and repairtime distributions are exponential. Results include structural properties of optimal policies.

GOHEEN in [12] presents an optimal stochastic control problem for a machine repair model with a multiserver repair shop. Lifetime and repairtime distributions are allowed to be of Erlang type. It is shown how the control problem can be reformulated as a nonlinear program.

WINSTON describes in [35, 36] two optimal stochastic control problems for machine repair models. In both papers the optimality of simple repair policies is established.

In [31] SCHORNAGEL presents an optimal stochastic control approach to the problem of finding the best preventive maintenance strategy in a complex system with interdependent equipment units. The model allows performance correlation (e.g. repair queueing) or economic or statistical correlation. The approach taken in these papers is to formulate the model as a generalised Markov decision process (cf. DE LEVE et al. [24, 25]).

A.2 Multiechelon systems

Multiechelon systems were first introduced by SHERBROOKE [32]. He designed the so-called METRIC (Multi-Echelon-Technique-for-Recoverable-Item-Control) for the US Air Force as a method for determining stock sizes. A multiechelon system is a stochastic model of a production system where there are more than one production sites (often called bases) and repair depots. Usually these systems have a two-level hierarchy: there are a number of operating bases, having their own repair shop, plus a central depot for repair of items coming from the different bases. MUCKSTADT [28] extended the class of models to allow multi-item, multi-echelon and multi-indenture inventory systems (the so called MOD-METRIC models).

A.2.1 Performance analysis models

In GROSS et al. [14] an approximation is presented for the steady-state analysis of a multiechelon inventory model for repairable items. The approximation is based on an iterative method involving aggregation at base level. It uses algorithms from the performance analysis of product form queueing networks.

The same authors discuss in [15] an approximation method for large Markov models of multiechelon repairable item inventory systems. The method uses state space truncation and randomisation techniques to analyse the transient behaviour of these models.

In GROSS AND MILER [16] randomisation is used to perform a transient analysis of a multiechelon inventory model with time-varying parameters. Lifetime and repairtime distributions are exponential.

A.2.2 Optimal stochastic control models

In [17] GROSS et al. describe a constrained optimisation problem for a queueing network model of a multiechelon repairable item inventory system. The optimisation problem is to minimize the operational costs with a constraint on the availability of the system. The lifetime and repairtime distributions are exponential and the service discipline at the different repair depots are all FCFS. The optimisation problem is solved using the product form equilibrium distribution together with some monotonicity results.

In HATOYAMA [18] an optimal stochastic control problem is discussed for a machine repair model. The model has an operating machine, several spare machines and a repair facility. The decision maker has to decide to repair or leave the operating machine, and to close or open the FCFS queue repair facility. The machine can be in a number of (observable) states, and the time spent in each state has a state dependent exponential distribution. Repair times are exponentially distributed. Hatoyama shows that under some conditions the optimal policy for this two-dimensional queueing system is of control limit type.

A.3 Series-Parallel systems

There are a number of papers by Japanese researchers, that present the performance analysis of different variants of a series-parallel system. The system is a production line consisting of a tandem of different components in series with a parallel part of identical components. The failure times are exponentially distributed, while the repair times may have a general distribution. KODAMA in [21] presents expressions for the reliability and availability of (small) systems with a preemptive repeat priority discipline. In NAKAMICHI et al. [29] three different repair disciplines — preemptive repeat, preemptive resume and head-of-the-line priority — are considered. In TAKAMATSU [33] the system has two different components in the parallel part.

A.4 Reviews

CHO AND PARLAR [9] is a survey on maintenance models for multi-unit systems.

A.5 Coherent systems of Multi-state components

An important aspect of the model that we introduced in this paper is the multi-mode nature of the components. Components can be in more different states than just "up" or "down". A number of papers have been published on coherent systems in which the components can be in more than one state, but most of the results in these paper are of a theoretic nature. They concern the translation of reliability concepts like for instance the reliability function, duality and critical elements from the binary to the multistate case. Examples of papers are BLOCK AND SAVITS [1], BARLOW AND WU [6], BAXTER [7, 8], and MONTERO *et al.* [27].

A.6 Miscellaneous

In ALBRIGHT AND SONI [4] an approximation method is presented for the calculation of the steady-state distribution of large multidimensional Markov processes. They illustrate the method with an example of a repairable item inventory model with returns.

LEHTONEN gives some stochastic ordering results for repair policies in machine repair problems in [22, 23].

In LI [26] a machine repair for a serial production line with buffers is discussed. The model is interesting, but the paper is badly written and the analysis is based on dubious assumptions.

REFERENCES

[1] H.W. BLOCK ABD T.H. SAVITS (1982). A Decomposition for Multistate Monotone Systems. *J. Appl. Prob.* 19, 391–402.

[2] S.C. ALBRIGHT (1980). Optimal Maintenance-Repair Policies for the Machine Repair Problem. *Naval Res. Logist. Quart.* 27, 17–27.

[3] S.C. ALBRIGHT AND A. SONI (1984). Evaluation of Costs of Ordering Policies in Large Machine Repair Problems. *Naval Res. Logist. Quart.* 31, 387–398.

[4] S.C. ALBRIGHT AND A. SONI (1988). An Approximation to the Stationary Distribution of a Multidimensional Markov Process. *IIE Trans.* 20, 111–118.

[5] D. ASSAF AND J.G. SHANTHIKUMAR (1987). Optimal Group Maintenance Policies with Continuous and Periodic Inspections. *Management Sci.* 33, 1440–1452.

[6] R.E. BARLOW AND A.S. WU (1978). Coherent Systems with Multi-state Components. *Math. Oper. Res.* 3, 275–281.

[7] L.A. BAXTER (1984). Continuum Structures I. *J. Appl. Prob.* 21, 802–815.

[8] L.A. BAXTER (1986). Continuum Structures II. *Math. Proc. Camb. Phil. Soc.* 99.

[9] D.I. CHO AND M. PARLAR (1991). A Survey of Maintenance Models for Multi-Unit Systems. *European J. Oper. Res.* 51, 1–23.

[10] T.B. CRABILL (1974). Optimal Control of a Maintenance System with Variable Service Rates. *Oper. Res.* 22, 736–745.

[11] J.W. FOSTER AND A. GARCIA-DIAZ (1983). Markovian Models for Investigating Failure and Repair Characteristics of Production Systems. *IIE Trans.* 15, 202–209.

[12] L.C. GOHEEN (1977). On the Optimal Policy for the Machine Repair Problem When Failure and Repair Times Have Erlang Distribution. *Oper. Res.* 25, 484–492.

[13] D. GROSS AND J.F. INCE (1981). The Machine Repair Problem with Heterogeneous Populations. *Oper. Res.* 29, 532–549.

[14] D. GROSS, L.C. KIOUSSIS, AND D.R. MILLER (1987). A Network Decomposition Approach for Approximating the Steady-State Behavior of Markovian Multi-Echelon Reparable Item Inventory Systems. *Management Sci.* 33, 1453–1468.

[15] D. GROSS, L.C. KIOUSSIS, AND D.R. MILLER (1987). Transient Behavior of Large Markovian Multiechelon Repairable Item Inventory Systems Using a Truncated State Space Approach. *Naval Res. Logist.* 34, 173–198.

[16] D. GROSS AND D.R. MILLER (1984). Multiechelon Repairable-Item Provisioning in a Time-Varying Environment Using the Randomization Technique. *Naval Res. Logist. Quart.* 31, 347–361.

[17] D. GROSS, D.R. MILLER, AND R.M. SOLAND (1983). A Closed Queueing Network Model for Multi-Echelon Repairable Item Provisioning. *IIE Trans.* 15, 344–352.

[18] Y. HATOYAMA (1977). Markov Maintenance Models with Control of Queue. *J. Oper. Res. Soc. Japan* 20, 164–181.

[19] M.C. van der Heijden and A. Schornagel (1988). Interval Effectiveness Distribution for a k–out–of–n multistate reliability system with repair. *European J. Oper. Res.* 36, 66–77.

[20] F.P. Kelly (1979). *Reversibility and Stochastic Networks.* Wiley, New York.

[21] M. Kodama (1976). Probabilistic Analysis of a Multicomponent Series-Parallel System under Preemptive Repeat Repair Discipline. *Oper. Res.* 24, 500–515.

[22] T. Lehtonen (1983). Stochastic Comparisons for Many Server Queues with Nonhomogeneous Exponential Servers. *Opsearch* 20, 1–15.

[23] T. Lehtonen (1984). On the Optimal Policies of an Exponential Machine Repair Problem. *Naval Res. Logist. Quart.* 31, 173–181.

[24] G. de Leve, A. Federgruen, and H.C. Tijms (1977). A General Markov Decision Method I: Model and Techniques. *Adv. Appl. Prob.* 9, 296–315.

[25] G. de Leve, A. Federgruen, and H.C. Tijms (1977). A General Markov Decision Method II: Applications. *Adv. Appl. Prob.* 9, 316–335.

[26] K-F. Li (1987). Serial Production Lines with Unreliable Machines and Limited Repair. *Naval Res. Logist.* 34, 101–108.

[27] J. Montero, J. Tejada, and J. Yáñez (1991). Multivalued Continuum Systems. Technical report, Dept. Statistics and OR, Fac. Mathematics, Complutense University, Madrid.

[28] J.A. Muckstadt (1973). A Model for a Multi-Item, Multi-Echelon, Multi-Indenture Inventory System. *Management Sci.* 20, 472–481.

[29] H. Nakamichi, J. Fukuta, S. Takamatsu, and M. Kodama (1974). Reliability Consideration on a Repairable Multicomponent System with Redundancy in Parallel. *J. Oper. Res. Soc. Japan* 17, 39–48.

[30] S.M. Ross (1970). *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco.

[31] A. Schornagel (198?). Optimum Preventive Maintenance for Complex Systems with Interdependent Equipment Units. Tech. rep., Mathematical Centre, Amsterdam.

[32] C.C. Sherbrooke (1968). METRIC: A Multi-Echelon Technique for Recoverable Item Control. *Oper. Res.* 16, 122–141.

[33] S. Takamatsu, J. Fukuta, M. Kodama, and H. Nakamichi (1975). Reliability Analysis of a Complex System with Dissimilar Units under Preemptive Repeat Repair Discipline. *J. Oper. Res. Soc. Japan* 18, 1–15.

[34] H.C. Tijms (1986). *Stochastic Modelling and Analysis: A Computational Approach.* Wiley, Chichester.

[35] W. WINSTON (1977). Optimal Control of Discrete and Continuous Time Maintenance Systems with Variable Service Rates. *Oper. Res.* 25, 259–268.

[36] W. WINSTON (1978). Optimality of Monotonic Policies for Multiple-Server Exponential Queueing Systems with State-Dependent Arrival Rates. *Oper. Res.* 26, 1089–1094.