ISE
Industrial and Systems Engineering

Supply Disruptions with Time-Dependent Parameters

Andrew M. Ross
Ying Rong
Lawrence V. Snyder
Lehigh University

Report No. 05T-009

LEHIGH
University

200 West Packer Avenue
Bethlehem, PA 18015

Supply Disruptions with Time-Dependent Parameters

Andrew M. Ross [*], Ying Rong and Lawrence V. Snyder [†]

Abstract

We consider a firm that faces random demand and receives product from a single supplier who faces random supply. The supplier's availability may be affected by events such as storms, strikes, machine breakdowns, and congestion due to orders from its other customers. In our model, we consider a dynamic environment: the probability of disruption, as well as the demand intensity, can be time dependent. We model this problem as a two-dimensional non-homogeneous continuous-time Markov chain (CTMC), which we solve numerically to obtain the total cost under various ordering policies. We propose several such policies, some of which are time dependent and others of which are not. The key question we address is: How much improvement in cost is gained by using time-varying ordering policies rather than stationary ones?

We compare the proposed policies under various cost, demand, and disruption parameters in an extensive numerical study. In addition, motivated by the fact that disruptions are low-probability events whose non-stationary probabilities may be difficult to estimate, we investigate the robustness of the time-dependent policies to errors in the supply parameters. We also briefly investigate sensitivity to the repair-duration distribution. We find that non-stationary policies can provide an effective balance of optimality (low cost) and robustness (low sensitivity to errors).

Keywords: inventory, supply disruption, time dependent, CTMC

[*] Corresponding author. phone (610) 758-4039, fax (610) 758-4886

[†] Industrial and Systems Engineering Dept., Lehigh University, Bethlehem, PA 18015 USA. {amr5,yir204,lvs2}@lehigh.edu

1 Introduction

1.1 Motivation

There is a growing realization that supply chain managers need to plan for disruptions. Some types of disruptions, such as strikes, terrorism, and machine breakdowns, seem equally likely to happen at any time of year. Other types, such as storms and congestion due to other orders, are more likely to happen in some parts of the year and less likely in others. For example, hurricanes are more likely in the summer and fall than in the winter and spring. Similarly, when the manufacturer's production scale is large, small retailers may find that the manufacturer has no supply at times due to a surge in demand from end customers, or to randomness in manufacturing times. The demand for consumer electronics, toys, and other products is highly seasonal, with greater demand near the end of a year, and such seasonality may cause order congestion. From the downstream customer's perspective, congestion at the manufacturer looks identical to a supply disruption, since no product is available. Thus, the disruption risk itself is seasonal.

One strategy for protecting against disruptions is to hold additional inventory. If the disruption risk is non-stationary, then it seems reasonable that the inventory policy should be non-stationary as well. If inventory managers are reluctant to change their ordering policies based on seasonal factors, they may either carry too much extra inventory during lulls in the disruption cycle or lose revenue during disruption cycle peaks.

One problem with trying to make real-time decisions about the ordering policy is that it requires estimates of disruption frequencies, times to resolve disruptions, and demand rates. Like any estimates, these may be inaccurate, and thus we should ask how possible inaccuracies will affect the overall cost.

In this paper, we develop several reasonable inventory policies that might be used under non-stationary supply and demand processes. We compare the performance of these policies empirically, focusing especially on the differences in performance between stationary and non-stationary policies. In addition, we evaluate the robustness of non-stationary policies by examining their performance even when the disruption parameters have been estimated incorrectly.

1.2 Problem Setting

We consider a single firm that sells a single product. Demands for the product occur according to a Poisson process, possibly with a cyclic arrival rate. The firm has a single supplier and uses a zero-inventory ordering (ZIO) policy, which means that the firm orders only when the on-hand inventory level hits zero. The supplier might be able to fill the firm's order immediately (zero lead time), which we call the "up" state. Or, the supplier may be impaired by a storm, strike, machine failure, etc., and will be unable to fill the order immediately. We call this the "down" state. The firm knows the state of the supplier at all times. When the supplier comes back up, the firm may revise its order quantity (with no penalty). The time until the supplier comes back up is called the repair time. Unmet demands (i.e., those that occur when the firm has no inventory and the supplier is down) are lost.

The time spent in the up and down states have phase-type distributions, whose parameters may vary in time. Phase-type distributions occur as the time until absorption in a CTMC, and are treated in detail by Neuts [13] in the constant-parameter setting. Each state in the defining CTMC is called a phase or a stage. A phase-type distribution with n phases has $n^2 + n - 1$ free parameters (n^2 exponential rates and $n - 1$ probabilities). Phase-type distributions are a generalization of the exponential, Erlang, Coxian, and hyperexponential distributions. We use them because they allow us to change the variability of the repair time relative to its mean, which we could not do if we used only an exponential distribution. They also allow us to formulate our problem as a CTMC, which makes it amenable to numeric solution. Unfortunately, the word "phase" must have two meanings in this paper: one meaning in the setting of phase-type distributions, and one meaning pertaining to the phase offset of a sinusoid wave. The two meanings are not particularly related. We shall endeavor to keep the two uses clear.

Although our model and numerical solution procedures can use phase-type distributions, for the most part we will use the special case of exponential distributions. This is because our focus is on the cyclic behavior of the system. The use of exponential distributions makes the failure process a Poisson process, often a non-homogeneous one. In Section 6.3.2 we will examine the impact of non-exponential repair times.

The firm faces three types of cost: an inventory holding cost, a fixed cost for orders, and a shortage cost for lost demands. In our solution procedures, each of these costs may vary with time without causing any difficulty. However, in our experiments we keep them constant for the sake of simplicity.

We propose various ordering policies, varying in complexity, for this problem. Some are simple and have no parameters for fine-tuning, while others permit some optimization of parameters. We compare these policies to each other and examine the robustness of a time-dependent ordering policy when there is a mismatch between the estimated and true parameters of the disruption process.

The remainder of this paper is structured as follows. In Section 2, we review the literature on supply chain disruptions. We also discuss the literature on queueing systems whose arrival rates undergo daily cycles, since the yearly cycles seen in our problem are somewhat similar. We formulate our model in Section 3. In Section 4, we propose several ordering policies, set up a CTMC model, and discuss our method for finding its time-varying probabilities. In Section 5, we discuss the time-dependent behavior of the system, to build our intuition about how it behaves. In Section 6, we analyze the performance of various policies under different scenarios and investigate the robustness of some of the policies. We summarize our results and make recommendations in Section 7.

2 Literature Review

2.1 Inventory

The classic Economic Order Quantity (EOQ) model does not consider the possibility of supply disruptions. Parlar and Berkin [14] provide several models for the economic order quantity under disruptions (EOQD). They assume that demand is deterministic, that the firm uses a ZIO policy, and that demands that occur when the firm has no inventory and the supplier is down are lost. Berk and Arreola-Risa [2] show that Parlar and Berkin's cost function is incorrect and provide a correct model.

The ZIO assumption is relaxed in several subsequent papers. Parlar and Perry [15] use a (Q, r) model instead of an EOQ model. They consider supply uncertainty in the form of both disruptions and yield uncertainty (i.e., the amount received by the firm may differ from the order quantity by a random amount). Parlar and Perry [16] consider the EOQD with one, two, or multiple suppliers, allowing the order quantity to depend on the states of the suppliers. They show that, as the number of suppliers increases, the model reduces to the classical EOQ; i.e., a ZIO policy becomes optimal. Moinzadeh and Aggarwal [11] consider an unreliable system, analogous to the classical economic production quantity (EPQ) model, with a finite production rate and a fixed cost to start production; they propose a continuous-review (s, S) policy for this system. Bielecki and Kumar [3] show that, under certain conditions, a ZIO policy may be optimal even for a supply system subject to disruptions.

The addition of supply uncertainty to classical inventory models renders most of them analytically intractable in the sense that closed-form solutions are rarely available. Nearly all of the models in the papers discussed thus far must be solved numerically. However, Snyder [17] provides an approximate cost function for Berk and Arreola-Risa's [2] EOQD model; his approximation can be solved in closed form and has a number of analytical properties reminiscent of the classical EOQ. The approximate EOQD cost function plays a role in several of the inventory policies we propose below.

To our knowledge, only two papers consider non-stationary disruption probabilities. Tomlin and Snyder [18] consider several up-states, called "threat levels," each of which entails a different disruption probability. The state transitions among threat levels, and between up and down states, according to a Markov process. They show that a state-dependent base-stock policy is optimal and demonstrate that the benefit of such a policy over a state-independent policy increases as the difference among the threat levels increases. Li, Xu, and Hayya [10] allow the disruption probability to vary based on the length of time since the last disruption; such a model is appropriate for machine breakdowns, for example. They prove that an age-dependent base-stock policy is optimal and provide bounds on the optimal base-stock levels. These two papers model the non-stationarity of the disruption probability as a stochastic process, while we consider cyclic, deterministic changes to the disruption and demand rates.

2.2 Queueing

Cyclic variations in demand rates are common in queueing applications, where arrival rates often vary by time-of-day. Eick, Massey, and Whitt [5, 6] provide fundamental descriptions of the results. The main intuition is that multiserver queues act like low-pass filters: rapid variations in the arrival rate are damped out. Also, the peak occupancy in the system occurs after the peak arrival rate (this is referred to as "lag").

The standard technique for solving a time-of-day queueing system is to set up a Continuous-Time Markov Chain (CTMC), then numerically solve its Kolmogorov Ordinary Differential Equations (ODEs). Because this procedure is somewhat laborious, it is common to start by looking at approximations. The two approximations that are most relevant here are the

SSA : Simple Stationary Approximation, where we consider only the daily average arrival and service rates, then treat the system as always in steady-state with those average rates, and the

PSA : Pointwise Stationary Approximation, where we estimate the performance measures at each timepoint t by taking the arrival and service rates at time t, then using those to find the steady-state performance measures.

Green, Kolesar, and Svoronos [9] consider typical time-of-day cases where the average service duration is short compared to the demand cycle duration. They demonstrate that when the peak arrival rate is as little as 10 percent above the daily average arrival rate, the SSA is "likely to produce poor estimates of performance." Green and Kolesar [7] demonstrate that the PSA generally does well, and its accuracy increases as the arrival and service rates increase (which was confirmed by Whitt [19]). Green and Kolesar [8] then improved the PSA accuracy by incorporating the lag effect.

Zipkin [20] considers inventory problems with changing rates. He recommends using long-run averages if the rates change quickly (essentially, an SSA), and current rates if they change slowly (essentially, a PSA). This matches common intuition in the queueing literature. It is the middle ground, where rates change too slowly for an SSA to be accurate but too quickly for a PSA to be accurate, that we investigate in this paper.

In call centers, a common heuristic for initially setting the number of servers throughout the day is to determine how many servers are needed at each timepoint to keep the PSA prediction of waiting times within management goals. Thus, we will use this kind of heuristic below as one possible policy when setting order quantities. We will also allow a lag adjustment.

3 Model Formulation

3.1 Input Parameters

When discussing cycles in disruptions or demand, it is useful to think of the relative amplitude (RA) of the cycle. This is defined as the distance from the average level to the peak, divided by the average. For example, if the average demand rate is 100 items/year and the peak demand rate is 150 items/year, then the relative amplitude is 0.50 or 50 percent.

Since our focus is on exploring the effect of seasonality in demand and disruption risk and on developing effective but approximate inventory policies, we consider a simple functional form to describe the demand and failure rates. We will use one that is common in the queueing literature in which the cycle length is 1 time unit, which we set to be a year. The function that we use to describe the demand rate is given by

$$\lambda(t) = \bar{\lambda} \cdot (1 - RA_d \cdot \cos(2\pi(t + \phi_d)))$$

See Figure 1 for a diagram. Here, $\bar{\lambda}$ is the average demand rate over the whole cycle, RA_d is the relative

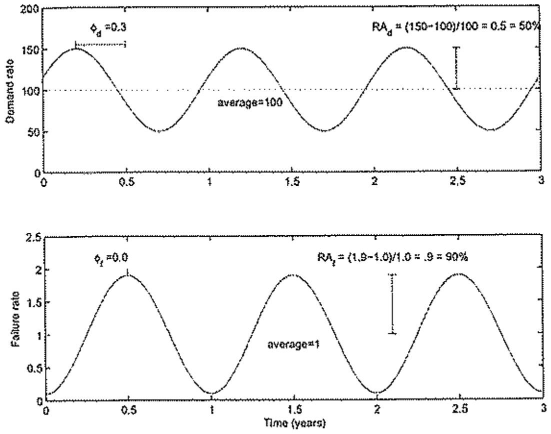


Figure 1: Example demand rate and failure rate functions

Table 1: Input Parameter Notation

Parameters	Definition	Basic Setting
K	Fixed cost	31/order
h	Holding cost	1/item/year
p	Stockout cost	11/item
r	Repair rate	12/year
$\bar{f}$	Average failure rate	1/year
RA_f	Relative amplitude of failure rate	0.9
ϕ_f	Phase of failure rate	0 years
$\bar{\lambda}$	Average demand rate	100/year
RA_d	Relative amplitude of demand rate	0
ϕ_d	Phase of demand rate	0 years

amplitude of the demand cycle, and ϕ_d is the phase shift. Using this form, we are restricted to $RA_d \leq 1$ (otherwise $\lambda(t)$ may be negative for some t). The failure rate function is defined similarly:

$$f(t) = \bar{f} \cdot (1 - RA_f \cdot \cos(2\pi(t + \phi_f)))$$

and we must have $RA_f \leq 1$. Our basic settings have $\phi_f = 0$, so this function hits its minimum at $t = 0, 1, 2, \ldots$ and hits its maximum at the mid-year marks.

We could similarly define a cyclical repair rate, but we will limit our analysis to the constant-repair-rate case for convenience. As mentioned above, the cost parameters could also fluctuate, but we will focus on situations where they are constant as well. Table 1 summarizes our notation for the input parameters and provides sample values that constitute our basic problem settings. Our numerical procedures do not depend on the sinusoid form of the cyclic functions; any cyclic function may be substituted, though sharply discontinuous functions may require more care in computing.

3.2 Cost Equation

Let $IL(t)$ be a random variable representing the inventory level at time t, and let $S(t)$ be a random variable representing the state of the supplier at time t. The state of the supplier may be a symbol denoting which phase of time-to-fail or time-to-repair it is in, but for cost computations we are concerned mostly with whether it is up or down, which we denote as U or D. We define the time-dependent joint distribution of these two random variables as

$$\Pr(x, s; t) \equiv \Pr(IL(t) = x \text{ and } S(t) = s \text{ at time } t). \tag{1}$$

The marginal distribution of the inventory level is then

$$\Pr(x; t) = \Pr(x, U; t) + \Pr(x, D; t) \tag{2}$$

Also, let $L(t) \equiv E[IL(t)]$ be the expected number of items in inventory at time t.

The main decision we have to make is how much to order at time t; we call this function $Q(t)$. Since the firm follows a ZIO policy, $Q(t)$ is only needed at times t when an order occurs; since we don't know *a priori* what times these will be, we define $Q(t)$ for all t. This function applies to orders placed when the inventory level hits zero and the supplier is up and to orders placed at the end of a down-state when the supplier comes back up. Since demands are lost (rather than backordered) in the latter case, the system state is identical in the two cases, and the same value of $Q(t)$ is appropriate for both. The choice of a $Q(t)$ curve determines the probability distributions given in (1) and (2), which then are integrated into the overall cost function as follows.

First, let us consider the instantaneous expected cost due to holding inventory. At time t, it is simply $L(t)h$.

Next, let us consider the expected cost due to stockouts. Let $\alpha(t) \equiv \Pr(0, D; t)$ be the probability that the system is empty at time t. If the system is empty, then lost demands occur at rate $\lambda(t)$, and each incurs the stockout cost p. Thus, our instantaneous expected cost due to lost demands at time t is $\alpha(t)\lambda(t)p$.

Now, let us consider the expected cost due to ordering (that is, paying the fixed cost). If we were in state $(1, U)$ at time t (which happens with probability $\Pr(1, U; t)$) then we would order once a demand occurs, which happens with rate $\lambda(t)$. Or, if we were in state $(0, D)$ (which happens with probability $\Pr(0, D; t)$) we would order once the supplier completes repairs, which happens with rate r. Thus, our instantaneous order rate at time t is $\beta(t) \equiv \Pr(1, U; t)\lambda(t) + \Pr(0, D; t)r$. The instantaneous expected cost due to ordering is then $\beta(t)K$.

We now have expressions for the expected cost at each time point. The cost over a cycle is simply the integration of these values. For example, the cost for the year between $t = 2$ and $t = 3$ would be

$$\int_2^3 [L(t)h + \alpha(t)\lambda(t)p + \beta(t)K]\, dt \tag{3}$$

4 Methods

4.1 EOQ and EOQD Review

The traditional assumptions for the EOQ model are

- Deterministic and constant demand with rate λ
- Zero lead time
- No stockouts allowed
- Fixed cost of K per order
- Holding cost of h per unit per year

Under these assumptions, it is well known (see, e.g., Nahmias [12]) that the optimal order quantity is given by $\sqrt{2K\lambda/h}$, with total optimal cost $\sqrt{2K\lambda h}$.

The EOQD model is a modified version of the EOQ in which the supplier faces disruptions. The assumption that stockouts are prohibited can no longer be enforced. Therefore, the EOQD assumes lost

Table 2: Summary of Policy Types

	Time-Independent	Real time
Non-Tunable	EOQ EOQD	EOQD-PSA EOQD-PSA-t
Tunable	Q-nt (1 param)	EOQD-PSA-ph (1 param) Q-t (3 params)

sales whenever there is no inventory to satisfy demand. Suppose the disruption rate is given by γ and the repair rate is given by μ. (Therefore, up-states last $1/\gamma$ years and disruptions last $1/\mu$ years on average.) Also, let p be the cost per lost demand. Using the approximation given by Snyder [17], the approximate optimal order quantity is

$$Q^* = \frac{\sqrt{(\rho\lambda h)^2 + 2h\mu(K\lambda\mu + \lambda^2 p\rho)} - \rho\lambda h}{h\mu}, \tag{4}$$

where $\rho = \frac{\gamma}{\gamma+\mu}$ is the long-run probability that the supplier is down.

The EOQ and EOQD both assume deterministic demand. They can be treated as good approximations under stochastic demand when the standard deviation of demand is small compared to its mean. Under our basic setting, demand is Poisson with rate equal to 100 items/year, so the coefficient of variation is 0.1. Therefore, we may use these two models as approximations for our basic setting.

4.2 Policies

In our time-dependent setting, it is unlikely that one could determine the true optimal order quantity $Q(t)$ for all t. Instead, we focus on a number of heuristics that trade off simplicity and flexibility. We can categorize these policies in two ways. First, does the policy depend on time, or does it instead prescribe a constant order quantity? Second, does the policy involve parameters that can be tuned to reduce the overall cost? Policies that have no such parameters are substantially easier to implement, but are of course more costly. Table 2 summarizes these categories and gives the names of the policies, which we describe in greater detail next.

First, we define the policies that prescribe an order quantity that does not change in time. The first two use the SSA approach, while the third is neither an SSA nor a PSA.

EOQ: Use the optimal Q derived from the ordinary EOQ model under the average value for the demand rate and ignoring disruptions; that is, $Q(t) = EOQ(t) = \sqrt{2K\bar{\lambda}/h}$.

EOQD: Use the optimal Q derived from the approximate EOQD model under the average values for demand and failure rates; that is, use Q^* from (4), replacing λ with $\bar{\lambda}$ in (4) and γ with $\bar{f}$ in the definition of ρ.

Q-nt: Numerically find the optimal constant Q value, evaluating the overall cost using the time-varying demand and failure rates.

The Q-nt policy is not a PSA because it evaluates the Kolmogorov ODEs to determine the cost, rather than using an approximate cost. Even though it uses a constant order quantity, it is not an SSA because it chooses that quantity taking into account the full time-varying behavior rather than just the averages.

Next we define the policies that depend on time but have no tunable parameters. These use the Pointwise Stationary Approximation system.

EOQD-PSA: Use the Q derived from the EOQD model under the values of the failure and demand rates at the current time. Let $EOQD_{PSA}(t)$ be the resulting order quantity function:

$$EOQD_{PSA}(t) \equiv \frac{\sqrt{(\rho(t)\lambda(t)h)^2 + 2hr \cdot (K\lambda(t)r + \lambda(t)^2 p\rho(t))} - \rho(t)\lambda(t)h}{hr}, \tag{5}$$

with $\rho(t) \equiv f(t)/(f(t)+r)$.

In the next policy, we will use the high and low points of the EOQD-PSA policy, so we will define the following notation:

$$EOQD^{lo}_{PSA} = \min_{0 \leq t \leq 1} \{EOQD_{PSA}(t)\}$$
$$EOQD^{hi}_{PSA} = \max_{0 \leq t \leq 1} \{EOQD_{PSA}(t)\}$$

EOQD-PSA-t: This essentially fits a sinusoid to the EOQD-PSA function.

$$\begin{aligned} Q(t) =& EOQD_{PSA-t}(t) \\ \equiv& \frac{EOQD^{hi}_{PSA} + EOQD^{lo}_{PSA}}{2} \left(1 - \frac{EOQD^{hi}_{PSA} - EOQD^{lo}_{PSA}}{EOQD^{hi}_{PSA} + EOQD^{lo}_{PSA}} \cos(2\pi t)\right) \end{aligned}$$

We use the EOQD-PSA-t policy to determine the importance of the precise shape of the PSA curve as compared to a pure sinusoid. This policy fits a pure sinusoid to match the peak and valley of the PSA curves exactly; other fits are possible.

Lastly, we define two policies that depend on time and that have tunable parameters:

Q-t: a sinusoid of the form

$$Q(t) = \bar{Q} \cdot (1 - RA_q \cos(2\pi(t + \phi_q)))$$

where we then tune $\bar{Q}$, RA_q, and ϕ_q to minimize the total expected cost.

EOQD-PSA-ph: This policy lets us tune a single parameter, the phase shift, to compensate for the lag effect discussed earlier:

$$Q(t) = EOQD_{PSA}(t + \phi_q)$$

The Q-t policy has the most flexibility of any of the policies we consider. The EOQD-PSA-ph policy is attractive because it strikes a balance between the simplicity of a plain PSA policy and the flexibility of adjusting for the lag. This can make a substantial difference, as we will see.

One might also define a policy, call it Q-t-F (the F standing for Fourier), that consists of a sum of many sinusoids of progressively higher frequencies. Each sinusoid would have an RA and a phase that we could tune. Such a policy should outperform the Q-t policy since it is more flexible. However, because it involves more dimensions of parameters to optimize, we do not consider this policy in this paper. We suspect that the benefit of increasing the number of sinusoids diminishes rather quickly.

Since Snyder [17] shows that the EOQD outperforms the EOQ under disruptions, we will not consider EOQ-type policies except the non-time-dependent EOQ itself (which may occur in practice if a firm is ignoring the possibility of failures).

4.3 CTMC Model

As mentioned in the problem setting, we consider Poisson demand, and phase-type times to failure and repair. This allows us to define a non-homogeneous CTMC to model the problem. The state space has two dimensions. The first is the inventory level $IL(t)$, which is a non-negative integer due to the lost-sales assumption. The second is the phase of the current time-to-failure or time-to-repair process, denoted $S(t)$. Figure 2 diagrams the various types of possible transitions. Since the lead time is 0 when the supplier is up, the states with $IL = 0$ cannot be reached when the supplier is up. As we mentioned earlier, we will use only exponential distributions for the time-to-failure, and mostly exponential distributions for time-to-repair.

Our order policies are not restricted to integer order quantities. To implement fractional order quantities in our CTMC model, we impose a 2-value version of random yields, concentrated on the two integers adjacent to the fractional order quantity. For example, an order quantity of 10.2 will result in a shipment of 10 items 80 percent of the time, and 11 items 20 percent of the time.

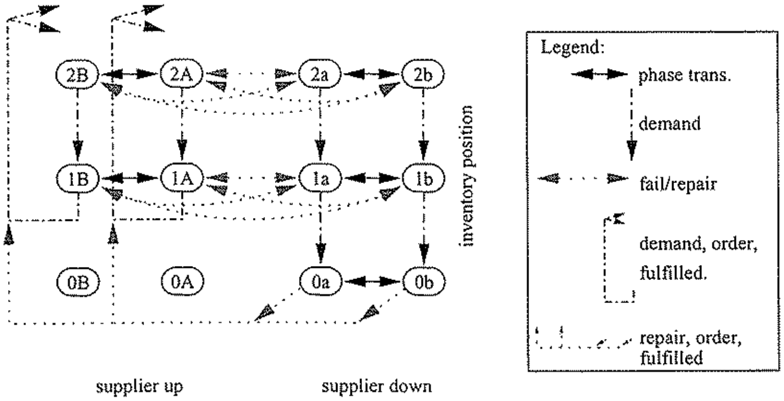


Figure 2: CTMC diagram

4.4 ODE Solution

In order to derive the joint distribution of $IL(t)$ and the failure/repair phase, we must solve the Kolmogorov ODEs numerically. To do this, we use the workhorse ODE solver in Matlab, the `ode45` function. It is a single-step solver, based on the Runge-Kutta formulas [4] of 4th and 5th order.

We need an initial distribution on the system status (set of $\Pr(x, s; t = 0)$ values) to start the ODE solver. Ideally, we want something as close to cyclic steady-state as possible, to reduce the amount of time needed to "warm up" the system (let the initial transient behavior fade away). The easiest method is to let the system start empty, but that is also perhaps the least accurate method. Instead, we use a PSA at time 0 to find the joint distribution of $IL(0)$ and the failure/repair status, and use this as our initial condition.

Another computation-related issue is the choice of a suitable warm up time. To choose an appropriate time, we plot $L(t)$, $\alpha(t)$ and $\beta(t)$ from time 0 to 10 for the basic settings under the EOQD policy (that is, constant order quantity, determined by EOQD formula) in Figure 3. Ideally, these performance measures would stabilize rather quickly, but Figure 3 suggests that there is still some variability in the height of the peaks even at time 10. However, as shown in equation (3), we actually only need $\int L(t)dt$, $\int \alpha(t)dt$ and $\int \beta(t)dt$ over one steady-state cycle to compute the long-run average cost. From the figure, we can see the average values of $IL(t)$, $\alpha(t)$ and $\beta(t)$ per cycle are almost the same from year to year, even from time 0. So we use 1 cycle to warm up and compute the average cost based on periods 2 and 3 to get our results. A pilot study showed that this method produced results that were very close to the results obtained from 9 years of warm-up. Our method is more attractive than using a long warm-up period because any real system would almost certainly see a change in its parameters in the many years it apparently would take to reach a true steady-state, so a focus on the near future better matches reality.

The run time to solve the ODEs depends on how much fluctuation the order quantity exhibits. For the basic settings of the parameters, if we use a time-independent ordering policy, the run time is approximately 5.5 seconds for $Q = 50, 100, 200, 400$. If we use a time-dependent order policy with average order quantity equal to 100, the run times are approximately 5.5, 15.1, 20.7, 27.1, 35.6, and 45.2 seconds for RA_q = 0, 0.1, 0.2, 0.3, 0.4, 0.5, respectively. The computations were performed on a Dell Dimension 4600 P4 with a 2.8GHz processor and 512MB RAM running under Windows XP.

5 System Dynamics

Our system with cyclic rates can exhibit some surprising behavior. In this section we discuss how the system dynamics can evolve. Here, we are not concerned with the costs, only with the number of items

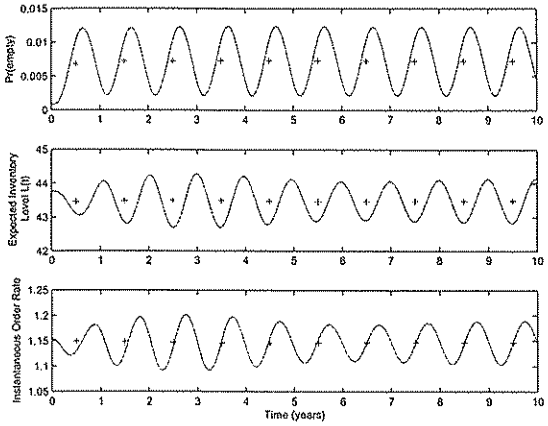


Figure 3: Warm-up behavior of performance measures. Yearly averages are shown with "+" symbols at mid-year marks.

in the system.

5.1 Stationary Systems

First, we consider stationary systems. It is well known that the inventory level under the ordinary EOQ model without failures has a uniform distribution. This continues to hold in the discrete demand setting, with the small exception that non-integer order quantities can cause $\Pr(IL = \lceil Q \rceil)$ to be different than all the others. Furthermore, if the demand rate is changed without changing the order quantity, the same uniform distribution still holds. This is because there is only one rate in the model, so changing it is similar to changing the time units.

In the stationary EOQD model, the Uniform distribution does not hold, as shown in Figure 4. We see

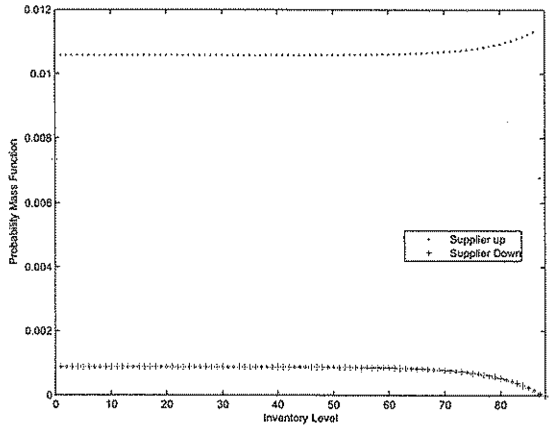


Figure 4: Example joint distribution of inventory level and supplier status

that, given that the supplier is up, we are more likely to have a large inventory than a small inventory. Given that the supplier is down, the opposite holds. This is because when an order arrives, we know the supplier is up, and it takes a while (as inventory drops) for probability mass to drift over to the supplier-down states. With two exceptions, the marginal distribution of IL (that is, $\Pr(x, U) + \Pr(x, D)$) is still Uniform. The exceptions are the $\Pr(0, d)$ state and the states just after an order arrives, if the order quantity is non-integer. Also, if we look at the marginal distribution of supplier status, we can calculate it without considering inventory level, since the supplier status is an alternating renewal process that is not affected by the inventory level.

In contrast to the behavior of the model without failures, if the demand rate is changed but not the order quantity, the distribution changes. This is because the demand rate is not the only rate in the model (the others are the failure and repair rates). Changing the demand rate without changing these rates changes the probability that the supplier will be down when an order is requested.

5.2 Cyclic Systems, Inventory Level Distribution

Now we move from stationary to cyclic systems. The easiest case to analyze is an EOQ model with a constant order quantity and a cyclic demand rate. The cyclic-steady-state distribution of inventory level remains uniform, as we might expect from knowing that it does not change in steady-state if the demand rate changes. If we add failures to the model, then cyclic demand rates cause non-uniform distributions, again as we might expect from the steady-state case with different demand rates.

Building in complexity, we next look at an EOQ system (no failures) with an order quantity that changes in time. In practice, we would only do this if the demand rate varied in time, because it is known that stationary order quantities are optimal for the stationary EOQ problem. However, to keep our examination of system dynamics uncluttered, we will set the demand rate to be a constant and examine the impact of cyclic order quantity policies.

A first question is: is this a linear system? In particular, if the input is sinusoidal, is the output sinusoidal? This is the case for the queueing system analyzed by Eick, Massey, and Whitt [5, 6]. However, it is not the case for our inventory system. Sinusoidal order quantity policies can give markedly non-sinusoidal curves for $L(t)$, among other performance measures. Thus, our system is not as easily described as we might hope. For example, the amount of lag can vary not only with cycle frequency but also with cycle amplitude.

A second question is: does the inventory level still have a uniform distribution? Let us take an example using our basic settings. Our order quantity function will be sinusoidal with $\bar{Q} = 86.614$, $RA_q = 0.1$, $\phi_q = 0$. Here, the average order quantity comes from the EOQD policy, but we have chosen a different RA_q and ϕ_q. Figure 5 shows the distribution at a few time points in a cyclic-steady-state order quantity cycle. The relative amplitude of the order quantity is only 10 percent. The distributions are markedly non-uniform at all time points. We have added vertical lines to indicate the $L(t)$ value for each curve

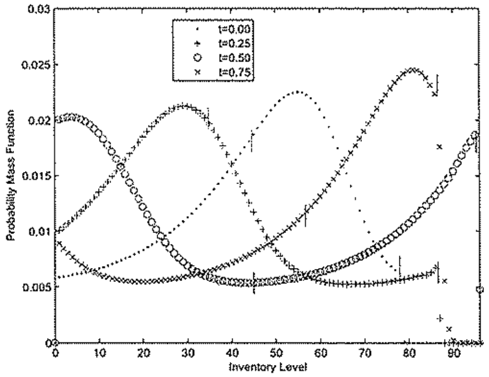


Figure 5: Distribution of inventory level at various time-points

(between 30 and 60 items) and to indicate the current order quantity at the right-hand edge of each curve. At values above this quantity, the curves drop off rapidly. The dropoff is more rapid for the curves for time points when the order quantity is rising. This figure is essentially a series of stills from a movie, available as an Internet supplement to this article.

5.3 Cyclic Systems, Equilibrium Behavior

Now we turn from the distribution of the inventory level to its mean $L(t)$ and its interactions with the instantaneous order rate $\beta(t)$. This will give us an idea of when orders tend to get placed. We will plot examples between years 7 and 10 to show 3 full cycles, but we will refer to times in the cycles using only the fractional part of the cycle. Our demand rate is 100 units/year.

If the order quantity is small, there are many orders per year, and the fluctuation in $L(t)$ and $\beta(t)$ are not large. This is demonstrated in Figure 6, where there are an average of 12 orders per year. We can see a small lag in the peak $L(t)$ values relative to the peaks of the order quantity curve, and the instantaneous order rate curve is roughly a half-cycle out of phase. This essentially says that when the order quantity is small, orders are frequent, and when they are large, orders are infrequent, which matches our intuition.

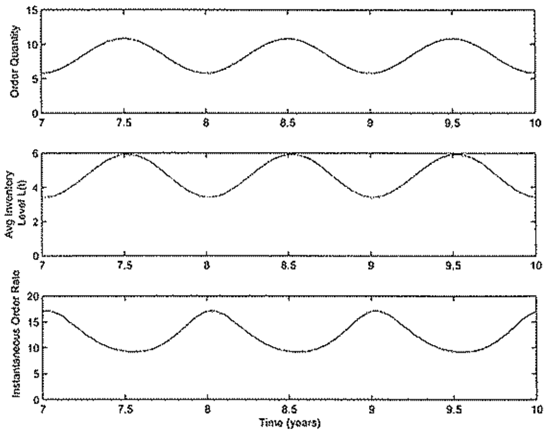


Figure 6: Order quantity $Q(t)$, Expected inventory $L(t)$, and instantaneous order rate $\beta(t)$ using an average of 12 orders per year, and $RA_q = 0.3$

However, when the average number of orders per year is small, interesting effects occur. If we make the order size match the yearly demand, so we average one order per year, we can get different results depending on the amplitude of the order quantity curve. For a low RA of 10 percent, Figure 7 shows a phase-lock phenomenon. Our $L(t)$ curve has more of a sawtooth shape, similar to what we are used to seeing in the traditional EOQ model. Orders are much more likely to happen around the $t = .75$ point in the cycle (September and October) than at other times. This is because the order quantity cycle crosses its average then. If we ordered exactly at $t = 0.75$, we would order 100 units (the yearly average demand), which would last us roughly until the same time next year. That is, ordering at this time is an equilibrium. We might ask why a similar equilibrium does not seem to occur at $t = 0.25$ (late March/early April), when the order quantity curve is also at 100 units. In fact, an equilibrium does occur there, but it is unstable. For $t = 0.75$, if the order is a little early, it will be a little larger and will last longer, bringing the next order point closer to $t = 0.75$ in the cycle. Similarly, if the order is a little late, it will be a little smaller and not last as long, making $t = 0.75$ a stable equilibrium. However, the opposite behavior happens around $t = 0.25$, making it unstable.

If we increase the RA of the order quantity to 50 percent, we are met with another surprise. Figure 8 shows that the relative variation in $L(t)$ actually goes down, and orders are more spread through the

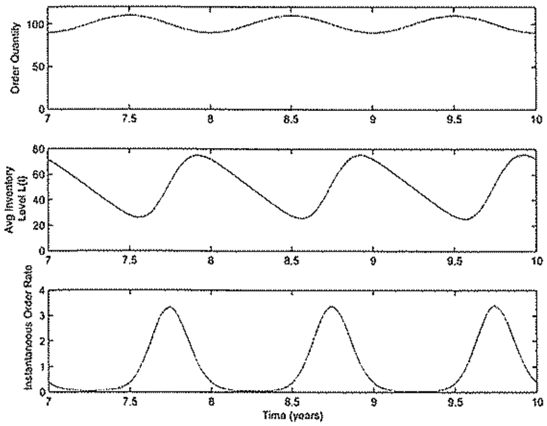


Figure 7: Order quantity $Q(t)$, Expected inventory $L(t)$, and instantaneous order rate $\beta(t)$ using an average of 1 order per year, and $RA_q = 0.1$

year. Consider an order near mid-year, when the order quantity curve is highest at 150 items. This will

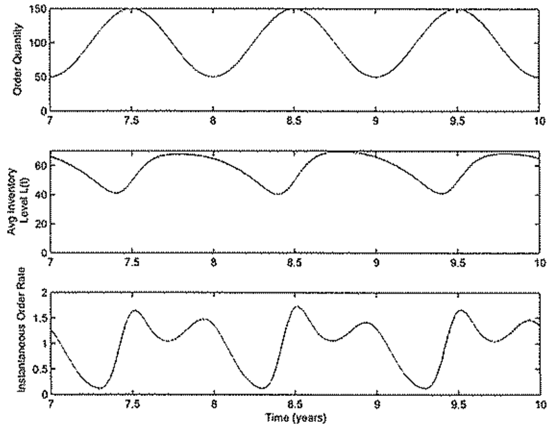


Figure 8: Order quantity $Q(t)$, Expected inventory $L(t)$, and instantaneous order rate $\beta(t)$ using an average of 1 order per year, and $RA_q = 0.5$

last roughly 18 months, causing another order when the order quantity curve is lowest at 50 items, and this order will last 6 months. Thus, we still average 1 order per year, but we alternate between long and short inter-order intervals. There is also a stable equilibrium around $t = 0.75$ as before.

Now that we have seen how the system can behave under fluctuating order quantities without regard to cost, we turn to how the various policies affect overall cost.

6 Results

6.1 Central Example

Before we explore the general behavior of our policies under a variety of parameter settings in Section 6.2, we start by focusing on a single set of parameters. We compare the EOQ, EOQD, Q-nt and Q-t policies

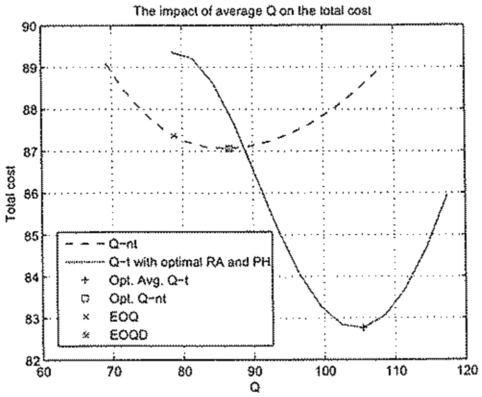


Figure 9: Cost versus order quantity

for our basic settings described in Table 1. As expected, the optimal Q-t policy returns the lowest cost. This is intuitive since this policy is both time dependent and tunable. In Figure 9, we plot the cost functions as we vary the average order quantity, under the Q-nt and Q-t policies. Here, we have fixed $RA_q = .2$ and $\phi_q = .7$, which are the optimal values for the Q-t policy. We can see that the optimal average order quantity for the Q-t policy is greater than the order quantity for the optimal Q-nt policy. If we used the optimal Q-nt policy rather than Q-t, our cost would increase by 5 percent. Later examples will show larger differences.

Our Q-t policy has three decision variables: average order quantity, relative amplitude, and phase shift. It would be difficult to visualize the objective function in these three dimensions, so we will look at cross-sections. In Figure 10, we plot contours of the objective function for the Q-t policy, with each subplot fixing one of the variables at its optimal value. In each subplot, the optimal solution is marked with a star.

6.2 The benefits of the Q-t policy

In order to investigate the benefit of the Q-t policy, we vary, one by one, the fixed cost K (Figure 11), stockout cost p (Figure 12), average failure rate $\bar{f}$ (Figure 13), repair rate r (Figure 14), RA of the failure rate RA_f (Figure 15), RA of the demand rate RA_d (Figure 16), and phase of the demand rate ϕ_d (Figure 17). Each figure shows the actual cost curves on the left, and the percent increase above the Q-t cost on the right. Compared to those constant-order-quantity policies, the benefit of the Q-t policy increases with stockout cost, average failure rate and its relative amplitude, and the relative amplitude of demand rate.

From these plots, it is clear that the Q-t policy generally outperforms the other proposed policies. The second best policy is EOQD-PSA-ph, which is clearly better than EOQD-PSA; that is, the ability to adjust the phase in response to the system lag is important, as we mentioned when we first introduced EOQD-PSA-ph. Generally, EOQD-PSA-ph stays within 5 to 10 percent of the Q-t policy. However, Figure 16 shows that the difference can climb toward 15 percent when the demand rate has a large relative amplitude.

The EOQD-PSA and EOQD-PSA-t curves are close in many cases; this shows that the exact shape of the curve (the near-sinusoid PSA versus a pure sinusoid) is not so important. The exception is when the demand phase is not synchronous with the failure rate phase. Then, the EOQD-PSA is not nearly a sinusoid, and so there are larger differences in cost between it and the EOQ-PSA-t policy. One surprise is that these time-dependent EOQD-PSA and EOQD-PSA-t policies are actually substantially worse than the constant-order-quantity policies Q-nt and EOQD (and even EOQ) in many cases. This reflects the

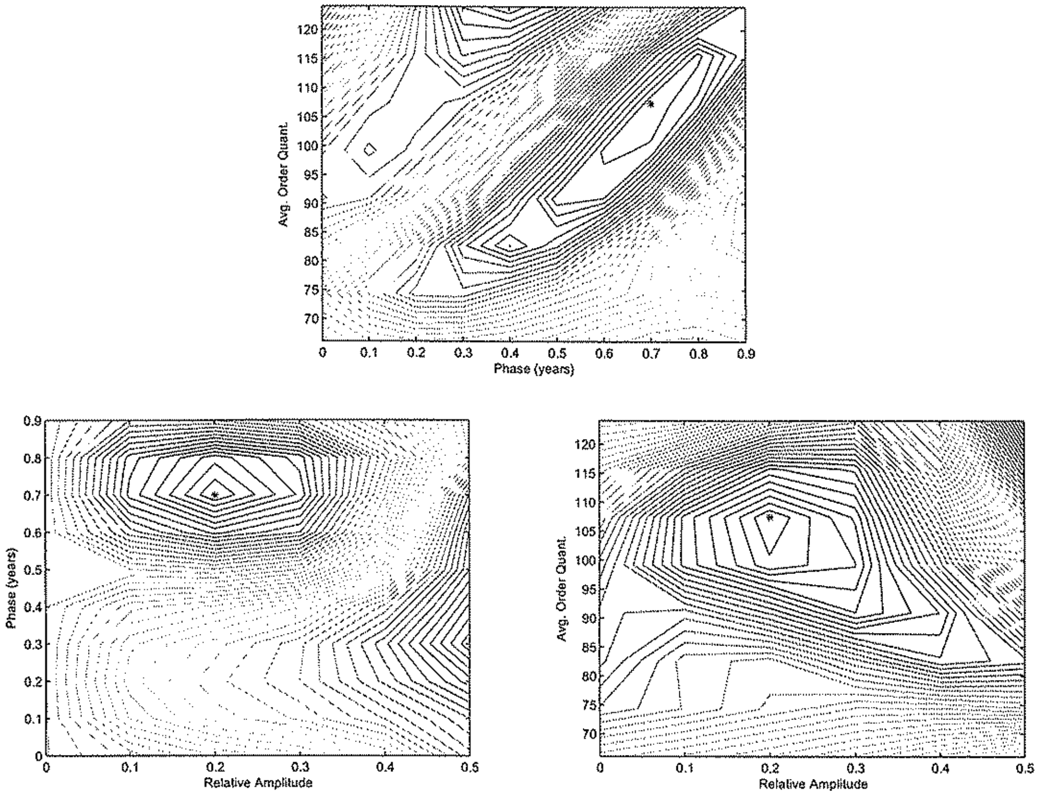
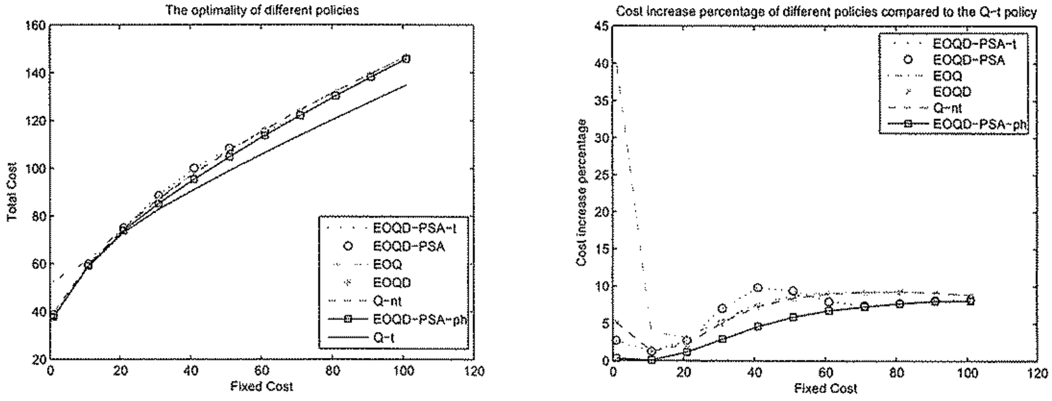


Figure 10: Contour plot near the optimal solution for the Q-t policy


The optimality of different policies

Cost increase percentage of different policies compared to the Q-t policy

Figure 11: Total cost under different policies when the fixed cost changes

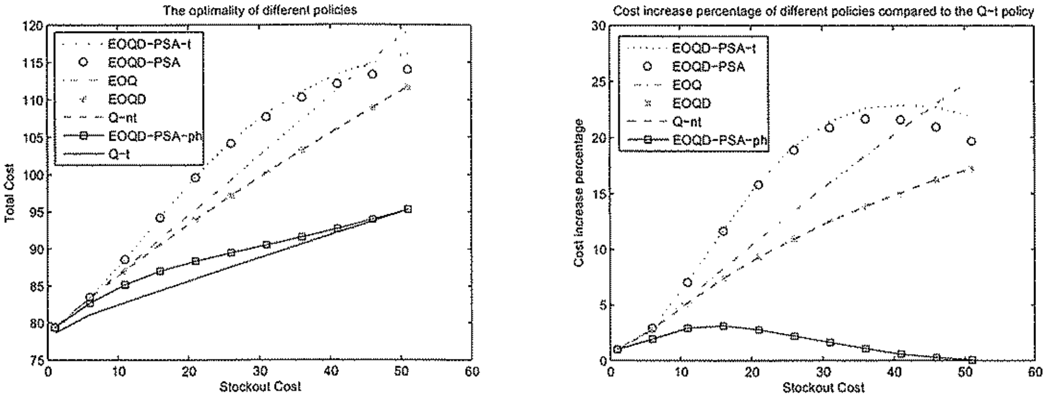


Figure 12: Total cost under different policies when the stockout cost changes

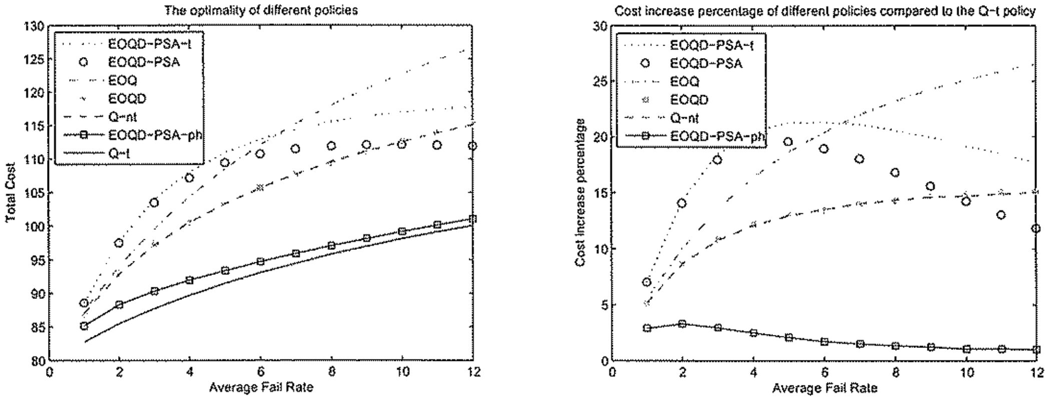


Figure 13: Total cost under different policies when the average failure rate changes

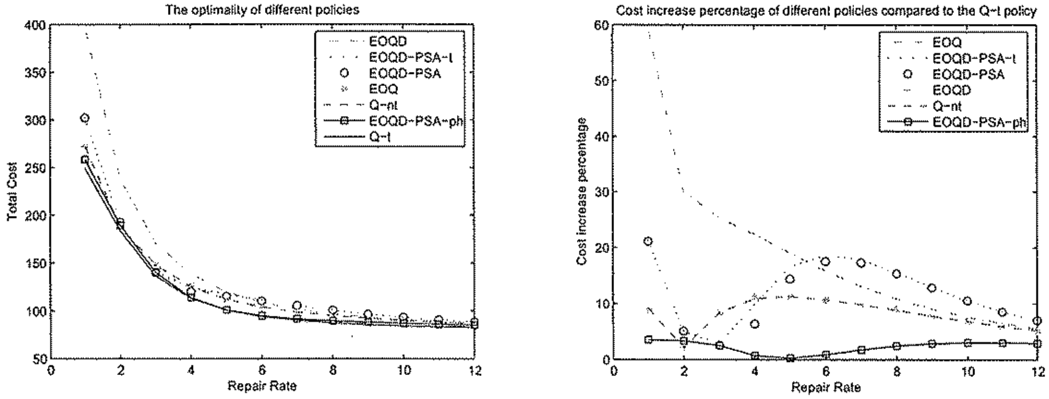


Figure 14: Total cost under different policies when the repair rate changes

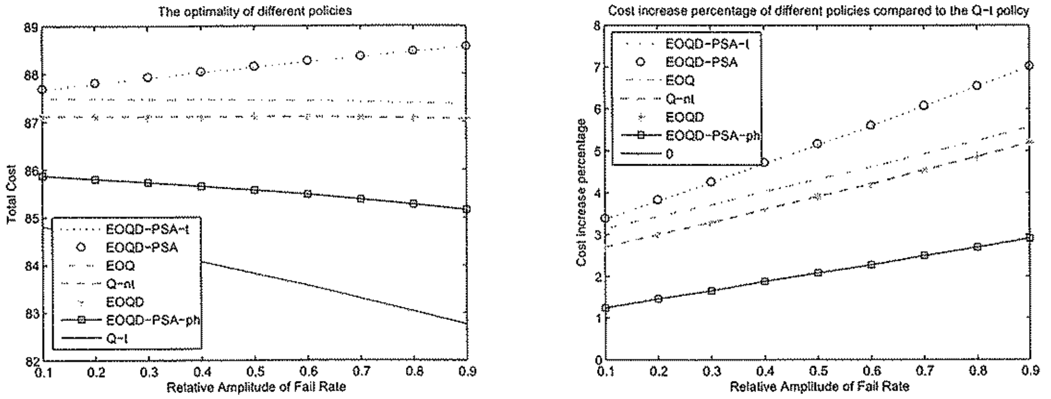


Figure 15: Total cost under different policies when the relative amplitude of the failure rate changes

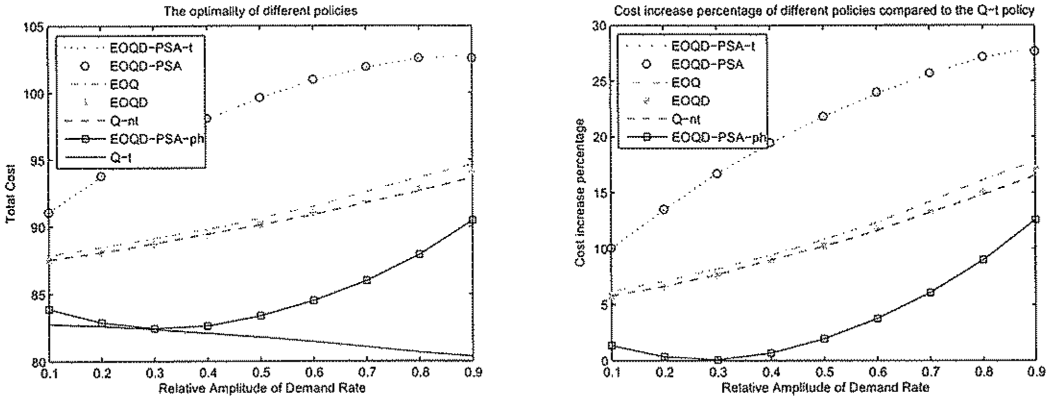


Figure 16: Total cost under different policies when the relative amplitude of the demand rate changes; here, $\phi_d = 0$, so the demand peak coincides with the failure rate peak.

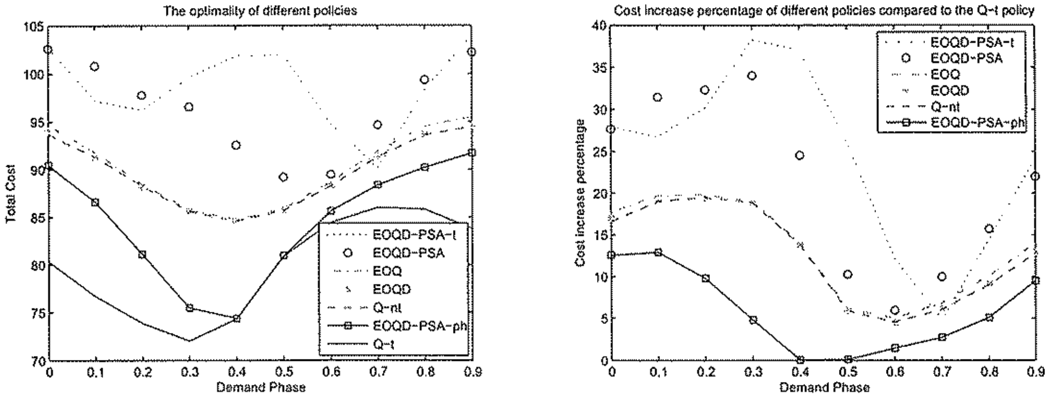


Figure 17: Total cost under different policies when the phase of the demand rate changes; here, we have $RA_d = 0.9$ instead of our basic setting of $RA_d = 0$.

importance of accounting for the system lag.

The constant-order-quantity policies EOQD and Q-nt are very close in almost all cases. Thus, there is little benefit to be obtained by optimizing a constant order quantity instead of simply using the SSA EOQD policy. The EOQ is even close to (but always worse than) the EOQD in some cases. A phenomenon of minor interest is shown in Figure 15: policies with constant order quantities are almost entirely insensitive to changes in RA_f, as long as the demand level does not fluctuate at all.

When the failure rate amplitude increases (Figure 15), we are met with another surprise: the costs for Q-t and EOQD-PSA-ph go down rather than up. Inspection of the optimal policies reveals that they tend to concentrate their orders during the time of low failure rate, and avoid the peak failure rate. This illustrates the benefits of a real-time ordering policy, as opposed to a constant order quantity.

When the demand amplitude increases rather than the failure rate amplitude (Figure 16), the Q-t policy still saves more money. This is also surprising, because in many fields and models (call centers, electric power generation, etc.) it costs more to meet non-constant demand than steady demand. The optimal Q-t policies in these cases tend to place their orders as the demand curve is increasing, so holding costs are reduced.

6.3 Robustness

In most cases, the overall probability of disruption is low, which makes it difficult to estimate the rates needed for the CTMC. Also, using an exponential distribution to approximate the repair process may not be accurate, as an exponential distribution has an unchangeable variability relative to its mean. One might expect variability to be larger if disruptions vary widely in severity, or smaller if the supplier is equipped to handle a small number of foreseeable disruptions.

Suppose we obtain the optimal Q-t policy using the assumptions introduced earlier. We would like to evaluate its performance under "real" settings as compared to the optimal Q-nt policy for those settings. Therefore, we study the tradeoff between the savings by employing a time-dependent control system and the robustness obtained by using a time-independent control rule.

6.3.1 Supplier Parameters

We start by investigating the effect of errors in the relative amplitude and phase of the failure rate. Starting with our basic settings from Table 1 as the "true" parameters, we consider the optimal Q-t policy and its cost, as was done in Section 6.1. Then, we suppose that the firm has mis-estimated RA_f and ϕ_f, and compute the Q-t and Q-nt policies they would have found. We then take these policies and evaluate them using the true parameters, to determine the resulting cost increase.

We are using both Q-t and Q-nt because one might think that it could be better to use a constant-order-quantity policy in the presence of uncertainty about the behavior of the failure rate. If the failure rate is the only fluctuating rate (as it is in our basic settings), then the Q-nt policy is entirely insensitive to ϕ_f because it amounts only to a change of time origin.

In our experiment, we let the firm's estimates of RA_f and ϕ_f vary as follows: $RA_f = 0, 0.1, \ldots, 1$ and $\phi_f = 0, 0.1, \ldots, 1$. This would result in $11 \cdot 11 = 121$ data points, with the exception that some data points with $RA_f = 1.0$ result in numeric difficulties because the failure rate hits zero, and thus are excluded (this is also common in queueing models). Each of these has a cost for the misguided Q-t used in the true-parameter model, and for the misguided Q-nt used in the true-parameter model. From these roughly 121 cost increases for the two policy types, we examine the maximum, minimum, and average (with all combinations given equal weight). Note that, as we saw in Figure 15, the Q-nt policy experiences almost exactly the same cost regardless of the value of RA_f when $RA_d = 0$; thus, its average, minimum, and maximum are essentially the same.

Intuition would predict that the worst-case cost increase occurs when the failure rate's relative amplitude, RA_f, is large and the phase, ϕ_f, is mis-estimated by half a yearly cycle. This is indeed the case. However, when the relative amplitude is large, more data will be available concerning when the peak occurs, because failures will be more concentrated there. Thus, the worst case of mis-estimation is unlikely to occur. In that sense, giving equal weight to all possibilities of mis-estimation in taking the average is being conservative.

Figure 18 shows the resulting total costs as various system parameters change. In these plots, the demand rate is constant. We only plot the average cost for the Q-nt policy, because its minimum and maximum are negligibly different, as previously noted. The figure shows that there are some cases where using the Q-nt policy guards against the worst case of mis-estimation when using Q-t. However, even in those cases, it is still better on average to use the Q-t policy. In other cases, the Q-t policy calculated based on the most-wrong estimates is still better than the corresponding Q-nt policy.

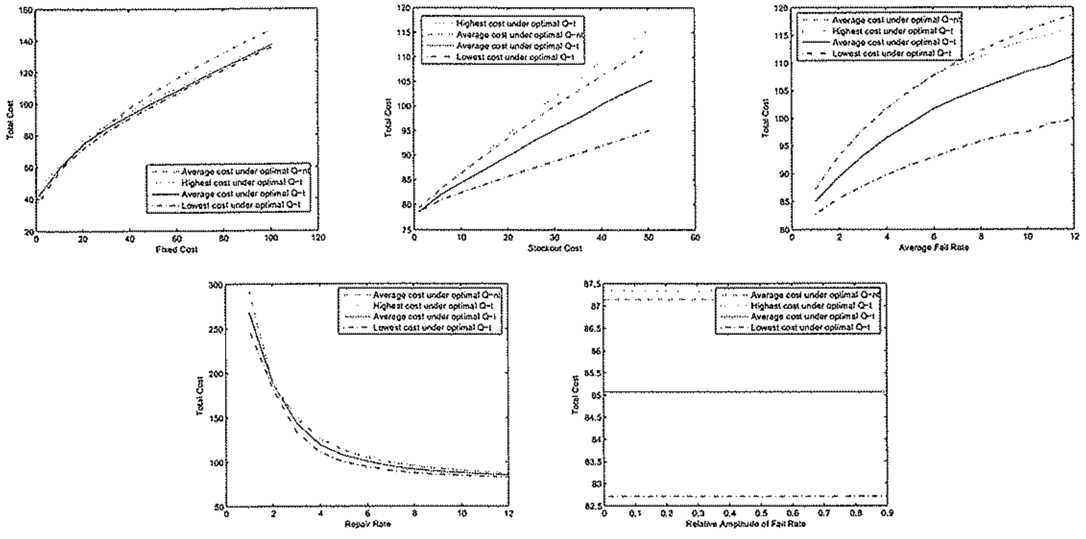


Figure 18: Robustness plot when the demand rate is constant over time

Next, we turn to the case where the demand rate varies in time, shown in Figure 19. We are supposing that the demand rate function is certain; it is the failure rate function that is uncertain, as before. This is reasonable, because the firm should have plenty of data on demand rates, since demand is nowhere near

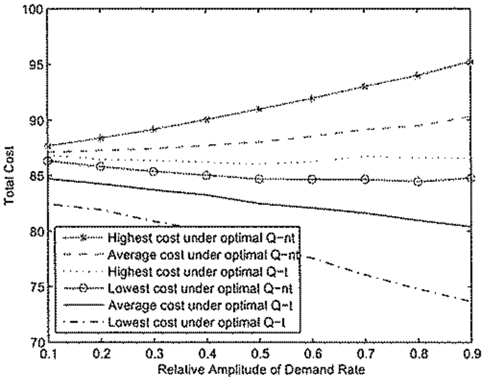


Figure 19: Robustness plot when the relative amplitude of demand increases

as rare as failures. In the fluctuating-demand scenario, the Q-nt policy costs are not essentially constant as they were in the previous figure. We see that the Q-nt policy is, on average, worse than the worst-case scenario for the Q-t policy. Thus, we still should use the Q-t policy rather than the Q-nt policy.

6.3.2 Repair Distribution

Because failures and their subsequent repairs are not particularly common, it is difficult to estimate the distribution of the time it takes to repair the system. Estimates of anything beyond the mean (variance, skewness, etc.) would be rough at best. Thus, we should investigate the effects of assuming a simple (i.e. exponential) distribution when the true distribution is different.

As mentioned in Section 1.2, we will use phase-type distributions. For a distribution that is less variable than exponential, we will use an Erlang distribution with 2 phases, abbreviated E2. Its coefficient of variation (standard deviation divided by mean) is $1/\sqrt{2} \approx 0.707$. For a distribution that is more variable than exponential, we will use a hyperexponential with two phases (abbreviated H2). We have chosen its coefficient of variation to be 2. We have used a balanced-means hyperexponential, which is a common choice (see Allen [1]).

From our basic settings, we change the exponentially distributed repair time to an E2 distribution to make repair times less variable. We find numerically that the optimal Q-t policy changes only by an insignificant amount. Thus, there is practically no harm in using an exponential distribution for repair times if the true distribution is E2. We conjecture that higher-order Erlang distributions (Erlang-3, etc.), which become less and less variable, may induce somewhat larger differences. The number of states in our CTMC, and thus the run-time of our solution procedures, goes up approximately linearly with the number of stages in the repair distribution. Thus, there is a disincentive to use Erlang distributions with a large number of stages.

If we change the exponentially distributed repair time to the H2 distribution that we have chosen, and find the new optimal Q-t policy (knowing that the H2 distribution holds), its true cost is 90.82. If we apply the old Q-t policy determined under the exponential repair time assumption, but evaluate it using H2 repair times, its cost is 90.88, a very small increase. The policies are mildly different: the new (H2-optimal) policy has an average order quantity 5.7 percent higher than the old (exponential-optimal) policy, and the phase shift increases by 0.05 to 0.75.

7 Conclusions

We said in Section 2.2 that we would explore the middle ground between PSA and SSA applicability. We have seen that the plain PSA-type policies (EOQD-PSA and EOQD-PSA-t) are worse than the SSA-type policies (EOQD, Q-nt) in most of the cases we evaluated. So, it would seem that our parameter choices have tended toward the SSA side of the middle ground. However, we have also seen that the Q-t policy (which is decidedly not an SSA) is much better than the SSA policies. Also, the EOQD-PSA-ph policy is second best, and it is also neither a true PSA nor an SSA. Thus, we conclude that our parameter choices have indeed established an interesting middle ground.

The effect of lag in the system is important, and in many cases it is better to use a plain stationary policy (such as EOQD or Q-nt) than to use a nonstationary policy that ignores lag effects (such as EOQD-PSA).

Overall, we see that the EOQD-PSA-ph policy balances cost savings with ease of computation, since only one parameter must be optimized, as opposed to 3 parameters for the Q-t policy. However, if the demand fluctuations are medium-large, only the Q-t policy can take better advantage of them and is probably worth the extra effort.

When we looked at the effects of mis-estimating the failure rate phase and relative amplitude, we found that the benefits of using a time-dependent policy outweigh the virtual insensitivity of a non-real-time policy in almost every case. Thus, we conclude that our time-dependent policy is robust in the face of uncertainty about these parameters. While it may suffer some cost increase, that cost increase is preferable to what the firm would suffer if a too-simple policy was used. We also found that the system was robust against the changes in the repair-time distribution that we tested.

It would be straightforward to include a variety of other features in the model and still be able to formulate it as a CTMC. Such features include phase-type demand arrivals (to model smooth or bursty demand), phase-type random lead times even when the supplier is up, backordered demand, non-zero reorder points, batch demand arrivals of random sizes, and random yield.

References

[1] Arnold O. Allen. *Probability, statistics, and queueing theory : with computer science applications, 2nd Ed.* Harcourt Brace Jovanovich, 1990.

[2] Emre Berk and Antonio Arreola-Risa. Note on "Future supply uncertainty in EOQ models". *Naval Research Logistics*, 41:129–132, 1994.

[3] T. Bielecki and P. R. Kumar. Optimality of zero-inventory policies for unreliable manufacturing systems. *Operations Research*, 36(4):532–541, 1988.

[4] J.R. Dormand and P.J. Prince. A family of embedded Runge-Kutta formulae. *J. Comp. Appl. Math.*, 6:19–26, 1980.

[5] Stephen G. Eick, William A. Massey, and Ward Whitt. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science*, 39(2):241–252, February 1993.

[6] Stephen G. Eick, William A. Massey, and Ward Whitt. The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742, July-August 1993.

[7] Linda V. Green and Peter J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37(1):84–97, January 1991.

[8] Linda V. Green and Peter J. Kolesar. The lagged PSA for estimating peak congestion in Markovian queues with periodic arrival rates. *Management Science*, 43(1):80–87, Jan 1997.

[9] Linda V. Green, Peter J. Kolesar, and Anthony Svoronos. Some effects of nonstationarity on multi-server Markovian queueing systems. *Operations Research*, 39(3):502–511, May–June 1991.

[10] Zhaolin Li, Susan H. Xu, and Jack Hayya. A periodic-review inventory system with supply interruptions. *Probability in the Engineering and Informational Sciences*, 18:33–53, 2004.

[11] Kamran Moinzadeh and Prabhu Aggarwal. Analysis of a production/inventory system subject to random disruptions. *Management Science*, 43(11):1577–1588, 1997.

[12] Stephen Nahmias. *Production and Operations Analysis.* McGraw-Hill/Irwin, 5th edition, 2005.

[13] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Dover, 1981.

[14] M. Parlar and D. Berkin. Future supply uncertainty in EOQ models. *Naval Research Logistics*, 38:107–121, 1991.

[15] Mahmut Parlar and David Perry. Analysis of a (Q, r, T) inventory policy with deterministic and random yeilds when future supply is uncertain. *European Journal of Operational Research*, 84:431–443, 1995.

[16] Mahmut Parlar and David Perry. Inventory models of future supply uncertainty with single and multiple suppliers. *Naval Research Logistics*, 43:191–210, 1996.

[17] Lawrence V. Snyder. A tight approximation for a continuous-review inventory model with supplier disruptions. working paper, Lehigh University, 2005.

[18] Brian T. Tomlin and Lawrence V. Snyder. Inventory management with advanced warning of disruptions. working paper, Lehigh University, 2006.

[19] Ward Whitt. The pointwise stationary approximation for Mt/Mt/s queues is asymptotically correct as the rates increase. *Management Science*, 37(3):307–314, Mar 1991.

[20] Paul H. Zipkin. *Foundations of Inventory Management.* McGraw Hill, 2000.