TWO-STAGE PRODUCTION PLANNING IN A DYNAMIC ENVIRONMENT

Stephen C. Graves, Harlan C. Meal
Sriram Dasu, Yuping Qui

August 1985 1698-85

INTRODUCTION AND OVERVIEW

In this document we report on research to develop and study mathematical models for production smoothing in a dynamic production environment. This effort was part of a larger study whose goal was to investigate the production planning practices for an electronic equipment manufacturing firm, and in particular to explore possible mechanisms for improvement. To motivate the presentation of our research, we need to indicate the nature of this production environment.

The production process can be viewed as a multi-stage system. In its most aggregate form there are two stages: component fabrication and equipment assembly. Customer demand has significant uncertainty and variability over the cumulative manufacturing lead time. Nevertheless, customers demand a high level of service. The manufacturing process is not totally reliable; in particular, there is significant yield uncertainty in component fabrication. Finally, production capacity for both component fabrication and assembly (including final test) is very capital intensive. As a consequence, it is expensive and/or difficult to adjust the production level.

As a consequence of these characteristics, one must consider a make-to-stock system that would smooth production and provide short and reliable delivery times to customers. We considered a hierarchical planning system (Hax and Meal, 1975) with two levels: one to set the aggregate production level (e.g 400 lots/month) and one to schedule the daily lot starts (e.g. one lot of item j, two lots of item k ...). Our study then focused on the design of such a make-to-stock system.

Key questions in the design of a make-to-stock system are what items should be stocked, at what point(s) in the production process should stock be accumulated, and what stock level is needed to provide satisfactory service and permit adequate production smoothing. Our research addresses these questions. To answer these questions, we must specify the rules or algorithms that will set the aggregate output rate and determine the daily production starts. We develop mathematical models, termed production smoothing models, to set the aggregate output rate. (For a review of production smoothing models, see Silver 1967.) We then give a mechanism for "disaggregating" the aggregate output into production starts for individual items. The analysis of these models determines the stock level(s) necessary for a desired service level. In addition, from these models one can quantify the tradeoff between the smoothness of aggregate production and the stock level.

In the next section we describe the production smoothing models and give their analysis. We first present a model for a one-stage production system. We then give two extensions to this model to treat a two-stage production system. To compare the models and evaluate their performance, we report in section 3 on a computational study based on data gathered from the electronic equipment manufacturer. In particular, we simulate the models over a wide range of parameter settings to assess their accuracy and to generate insight into their behavior.

DEVELOPMENT AND ANALYSIS OF PRODUCTION SMOOTHING MODELS

We develop here a set of production smoothing models that can be used to look at the tradeoffs between production smoothing and inventory requirements. We assume that production smoothing is done at an aggregate level for families of products that have similar processing requirements. For now, we will make the following simplifications:

we ignore lot-sizing considerations;

we assume no uncertainty in the build time for the products;

we assume no constraint on production capacity;

other simplifications will be introduced as needed.

The models should permit focus on the production smoothing/inventory considerations in the light of forecast uncertainty, demand variability, and yield uncertainty. We will use a simulation to examine the consequences of the simplifying assumptions.

We model the production activity in a very gross way. We are not concerned with detailed scheduling issues, but rather with the planning issues. We assume that the production system can be decomposed into a series of stages where we may maintain an inventory between successive stages and in finished goods (see Figure 1). Then we assume a known and fixed lead time for each stage: all production started at a stage at time t finishes that stage at time $t+\ell$, for ℓ being the lead time of the stage. We are concerned both with setting the aggregate production level for each stage, and with deciding how to disaggregate it into a production level for individual items. Furthermore, we want to understand how much inventory is needed as safety stock, and where it is needed. We desire to keep the aggregate production levels as smooth as possible at each stage, and to keep the safety stocks as low as possible, while providing satisfactory customer service. We have developed a production smoothing model to illuminate these tradeoffs.

To present the production smoothing models, we first describe the aggregate forecast process that drives the production smoothing models. We then consider production smoothing for one production stage. We use the production smoothing model for one stage as the building block for developing two distinct models for smoothing two production stages.

Forecast Process

A key input to the production smoothing process is the forecast of aggregate demand. We assume that this forecast is by month for up to the next 12 months, and that it is revised monthly. To model the process of forecast revision, we define $F_t(s)$ as the forecast made at month t of aggregate demand in month s (s>t). Then, we define $\delta_t(s)$ by

(1) $F_t(s) = F_{t-1}(s) + \delta_t(s)$

so that $\delta_t(s)$ denotes the change at time t in the aggregate forecast for time s. We assume that $\delta_t(s)$ is a random variable with the following characteristics:

$$E\ [\delta_t(s)] = 0\ ,$$

$$\mathrm{Var}\ [\delta_t(s)] = \sigma^2(s-t) \quad ,$$

and $\delta_t(s)$ and $\delta_{t'}(s)$ are independent for all $t \neq t'$.

We denote the aggregate demand in month t as $F_t(t)$: the 'forecast' of demand in t made at time t. From (1), we can express the aggregate demand as

$$(2) \quad F_t(t) = F_{t-k}(t) + \sum_{i=1}^{k} \delta_{t-k+i}(t)$$

where $F_{t-k}(t)$ is the initial forecast for demand in t made k months ago. The k-month forecast error is given by

$$F_t(t) - F_{t-k}(t) = \sum_{i=1}^{k} \delta_{t-k+i}(t) \quad ,$$

and has zero mean and variance equal to $\sum_{i=1}^{k} \sigma^2(k-i) = \sum_{j=0}^{k-1} \sigma^2(j)$. Thus, the forecast not only is unbiased, but also improves over time.

We assume that the demand process, $D(t) = F_t(t)$, is stationary with $E\ [D(t)] = \bar{D}$, and $\mathrm{Var}\ [D(t)] = \sigma_D^2$ for all t. (We will use D(t) and $F_t(t)$ interchangeably to denote aggregate demand in month t.) From (2) we now can compute the variance of $F_{t-k}(t)$:

$$\mathrm{Var}\ [F_{t-k}(t)] = \sigma_D^2 - \sum_{j=0}^{k-1} \sigma^2(j) \quad .$$

Thus, the variance of the initial forecast depends upon the forecast horizon length k. The longer is the forecast horizon, the smaller is the forecast variance; presumably, for a longer forecast horizon, there is less information available on aggregate demand and there is more likelihood to use $\bar{D}$ as the forecast.

Production Smoothing for One Stage

We now develop a production smoothing model for one production stage. We assume that the production stage produces one family and produces it to stock, i.e., a finished goods inventory. We use a production smoothing model to set an aggregate production rate (e.g. aggregate starts per month) for the family for each month. To set the actual production starts we must disaggregate the aggregate production plan according to the net requirements for individual items.

To smooth production, we must maintain an inventory stock since the aggregate production rate, being a smoothed average of the aggregate demand rate, will deviate from the aggregate demand rate over short time intervals. To determine the aggregate inventory requirements we need to analyze the behavior of the production smoothing model. In particular, we expect that the more we smooth production, the greater will be the stock requirements. But since customer service is determined by the stocks for individual items, we also need to understand how the aggregate plan is to be disaggregated.

The production smoothing model is given by

$$(3)\quad P(t+\ell) = \frac{1}{n+1}\,[F_t(t+1,t+\ell+n) - P(t+1,t+\ell-1) - \hat{I}(t) + SS]$$

where t denotes the current time period, and

$P(t+i)$ = the planned aggregate production started at time $t+i-\ell$ to be completed by time $t+i$;

ℓ = the lead time for production;

rule in terms of its control parameters, n and SS, and will try to show that it is both an interesting and reasonable rule to consider. In this respect, it is very similar to the study by Cruickshanks et al (1984). First, it is simple and does permit analysis of its behavior. In particular, we will see that under suitable assumptions we obtain the disaggregation implications from this rule. X Second, there is evidence that linear rules of this form are optimal or near optimal for not only quadratic cost functions but also for more general cost functions (Schneeweiss 1971, 1974). Third, while the rule may not be the optimal form, we nevertheless can find the optimal parameter choice for this form.

We need to be careful to express all variables in (3) in common units. In particular, the units need be aggregate measures of production capacity. For instance, for the manufacture of integrated circuits it might be natural to express planned production in wafer starts. However, the actual inventory is likely to be known in terms of chips for individual items. Thus, we need translate, using yield factors, this actual inventory into the equivalent wafer starts for the items and then aggregate over all items in the family. Similarly, for the assembly of electronic equipment, we might express planned production in terms of required assembly hours; for a parts fabrication facility, the units might be machine hours required on the bottleneck facility.

To analyze (3) we use the inventory balance equation,

$$(4) \quad \hat{I}(t) = \hat{I}(t-1) + \hat{P}(t) - D(t)$$

where $\hat{P}(t)$ is the actual family production completed at time t and D(t) is the family demand at time t. We assume that

$F_t(t+1, t+\ell+n) = F_t(t+1) + \ldots + F_t(t+\ell+n)$;

$P\ (t+1, t+\ell-1) = P\ (t+1) + \ldots + P\ (t+\ell-1)$;

$\hat{I}(t)$ = aggregate inventory on-hand at start of time period t;

SS = target safety stock level;

n = window length.

Thus, in period t we set $P(t+\ell)$, the planned production to be completed ℓ periods later. We assume that the lead time is deterministic, but that there is uncertainty in the production yield so that actual production completed at time $t+\ell$ will deviate from $P(t+\ell)$. In (3), the term within the brackets represents the forecast over the time interval $t+1, \ldots t+\ell+n$ minus the cumulative production planned for completion by time $t+\ell-1$ and minus an inventory adjustment (the difference between the actual and target inventory). We term the quantity within the brackets to be the net requirements over the time interval $t+\ell$, $t+\ell+1$, ... $t+\ell+n$. Thus, we set $P(t+\ell)$ to be the average of the net requirements over this time window $t+\ell$, ... $t+\ell+n$ where n is an integer decision variable equal to the length of the time window. The value of n will determine the level of production smoothing. The safety stock target, SS, is also a decision variable that is set to provide acceptable customer service on all items in the family.

We note that this production smoothing model is a linear rule. In this respect it is similar to the linear-decision-rule model developed by Holt et al. (1955,1956). However, our decision rule differs from that of Holt et al. in that it is not a consequence of minimizing some cost function. Rather, we just pose this decision

(5) $\hat{P}(t) = P(t) - \varepsilon(t)$

where $\varepsilon(t)$ is a random variable that reflects the yield uncertainty. We assume that $\varepsilon(t)$ is independent and identically distributed over time, has zero mean and has variance σ_P^2 .

Now we use (3) to consider the difference $P(t+\ell) - P(t+\ell-1)$:

$$(6)\quad P(t+\ell) - P(t+\ell-1) = \frac{1}{n+1}\{F_t(t+\ell+n) + \delta_t(t+1,t+\ell+n-1) - F_{t-1}(t) - P(t+\ell-1) + P(t) - \hat{I}(t) + \hat{I}(t-1)\}$$

where $\delta_t(t+1,t+\ell+n-1) = \delta_t(t+1) + \ldots + \delta_t(t+\ell+n-1)$.

We use (4), (5) and the substitution $F_t(t) = D(t)$ to simplify (6) to

$$P(t+\ell) - P(t+\ell-1) = \frac{1}{n+1}\{F_t(t+\ell+n) + \delta_t(t,t+\ell+n-1) - P(t+\ell-1) + \varepsilon(t)\} ,$$

from which we obtain

$$(7)\quad P(t+\ell) = \frac{1}{n+1}\{F_t(t+\ell+n) + \delta_t(t,t+\ell+n-1) + \varepsilon(t)\} + \frac{n}{n+1}\{P(t+\ell-1)\} .$$

Thus, we see that the production smoothing model (3) is equivalent to (7), which is a simple smoothing equation. Planned production for time period $t+\ell$ is a weighted average of the planned production for the previous time period $t+\ell-1$, and the initial forecast for $t+\ell+n$, modified by the change in the cumulative forecast over the interval $(t,t+\ell+n-1)$ and by the realized yield deviation for the production completed in t. The window length n determines the weights so that $1/(n+1)$ is the smoothing parameter.

The significance of (7) is not only its interpretation as a simple smoothing equation, but also its usefulness for characterizing the long-term behavior of this production smoothing model. From (7), we can show that in steady-state

(8) $$E[P(t+\ell)] = \bar{D}$$

and

(9) $$\mathrm{Var}[P(t+\ell)] = (\sigma_D^2 + \sigma_P^2) \,/\, (2n + 1) .$$

We defer the details of this derivation to the appendix. We can also use (7) to characterize the production 'step', $[P(t+\ell) - P(t+\ell-1)]$, as well as study other cases in which the demand process is not stationary (e.g., the demand process has a trend component or a seasonal component).

From (8) and (9) we see that the window length n does not affect the expected aggregate production rate, but does control the variability of the aggregate production rate. As we increase n, the production variance decreases and the aggregate production becomes smoother. Thus, the choice of n dictates the degree of production smoothing given by (3).

Aggregate Inventory Behavior

In addition to the aggregate production level, we also want to understand the behavior of the aggregate inventory level. This is necessary to see the effect of the production smoothing on customer service for individual items. To do this, we focus on $I(t+\ell)$: the planned aggregate inventory level at time $t+\ell$. At time t, $I(t+\ell)$ is defined as the current inventory at time t, plus planned production to be completed over $t+1, \dots t+\ell$, minus the cumulative demand forecast through $t+\ell$:

(10) $$I(t+\ell) = \hat{I}(t) + P(t+1,t+\ell) - F_t(t+1,t+\ell).$$

By using (3) to substitute for $P(t+1,t+\ell-1)$ in (10), we obtain

(11) $$I(t+\ell) = SS + F_t(t+\ell+1,t+\ell+n) - nP(t+\ell) .$$

From (11), we see that

(12) $$E[I(t+\ell)] = SS .$$

In addition, we can show its variance (after substantial algebra) to be

(13) $$\mathrm{Var}[I(t+\ell)] = \frac{n^2}{2n+1} \cdot (\sigma_D^2 + \sigma_P^2) + \sum_{j=0}^{n-1} [1 - 2(\frac{n}{n+1})^{j+1}] [\sigma_D^2 - \sum_{i=0}^{\ell+j} \sigma^2(i)] .$$

This characterization of the planned aggregate inventory level will be useful in determining the customer service level for a specification of the target safety stock SS and of the smoothing parameter n. While the expectation of planned inventory level equals the target safety stock, its variance depends on the window length n. From (13) we argue that the variance of the planned inventory level essentially increases linearly with the window length n, since the second term in (13) is dominated by the first (linear) term.

Disaggregation and Customer Service

To determine the proper choice for the target safety stock SS, we need to know how the aggregate inventory is spread over the individual items or equivalently, how the aggregate production level is disaggregated into item production. We must set the production starts for each item so that the planned production level for the family is achieved; that is,

$$\sum_k P(k,t+\ell) = P(t+\ell)$$

where $P(t+\ell)$ is given by (3) and $P(k,t+\ell)$ is the production level for item k started at time t for completion by time $t+\ell$. This implies that the planned inventory level for item k at time $t+\ell$, call it $I(k,t+\ell)$, must satisfy

$$(14) \qquad \sum_k I(k,t+\ell) = I(t+\ell)$$

where $I(t+\ell)$ is given by (10).

To characterize the service level for item k in period $t+\ell$ for a particular disaggregation, we can express its planned inventory as

$$I(k,t+\ell) = \hat{I}(k,t) + \sum_{i=1}^{\ell} [P(k,t+i) - F_t(k,t+i)]$$

where $\hat{I}(k,t)$ is current inventory and $F_t(k,t+i)$ is the current demand forecast for month t+i . The actual inventory at time $t+\ell$ will be

$$\hat{I}(k,t+\ell) = \hat{I}(k,t) + \sum_{i=1}^{\ell} [\hat{P}(k,t+i) - D(k,t+i)]$$

where for time t+i, $\hat{P}(k,t+i)$ is actual production and $D(k,t+i)$ is demand, and where negative inventory denotes a backorder. We assume that the difference $I(k,t+\ell) - \hat{I}(k,t+\ell)$, which reflects the yield uncertainty and cumulative forecast error for k over the lead time of ℓ time periods, is a normally-distributed random variable with zero mean and a variance σ_k^2. Consequently, we can use $z_k = I(k,t+\ell) / \sigma_k$ to indicate the service level for k at time $t+\ell$ from a particular disaggregation (14). z_k denotes the number of standard deviations of safety stock protection provided by $I(k,t+\ell)$.

Now suppose we set $P(k,t+\ell)$, or equivalently $I(k,t+\ell)$, so that the likelihood of stockout in period $t+\ell$ is the same for all items. Thus, we want $z_k = z$ for all items k, or equivalently

(15) $I(k,t+\ell) = z\ \sigma_k$

for all k. By substituting (15) into (14), we find that

$$z = I(t+\ell) \ / \sum_k \sigma_k \quad .$$

Thus, we can use (12) and (13) to characterize the service level (as specified by z) for a choice of SS and n. We see that z is a random variable, and thus the service level will vary from month to month. The expected service level depends only on the safety stock target SS. But the amount of variability in the service level depends on the smoothing parameter n; the more that we smooth production, the greater will be the variability in service from month to month.

We have assumed here that it will always be possible to disaggregate production to satisfy (15). This is not guaranteed. Indeed, if current inventories are 'out-of-balance', the simultaneous satisfaction of (14) and (15) may imply negative production ($P(k,t+\ell) < 0$) for items with large current inventories. Obviously, this is not possible. However, we expect that the assumption of 'balanced' inventories will be appropriate for the family of items that we produce to stock; we will test this assumption with the simulation exercise. We note that a similar assumption of 'balanced' inventories is made by Eppen and Schrage (1981) and by Federgruen and Zipkin (1984) in their studies on centralized ordering policies for multilocation distribution systems.

Two-Stage Models

In this section we extend the one-stage model to a production operation with two stages, e.g. fabrication and assembly. For such an operation we may have not only a finished goods inventory, but also an intermediate inventory between the two stages (see Figure 1). The intermediate inventory permits one to decouple, at least partially, the production of the upstream stage (parts fabrication) from that for the downstream stage (assembly). Thus, the production smoothing model needs to set aggregate production rates for both stages. In addition, we will need to characterize how each stage disaggregates the aggregate production plan.

We mention three approaches for production smoothing for two stages. The first approach is to view the two stages as one stage without an intermediate inventory, and then to use the previous model; we will not discuss this approach any further, but will include it in the simulation study of the various models. The second and third approach use the intermediate inventory to permit some decoupling of the stages. They differ in terms of how the upstream stage smooths its production. In the second model, production smoothing by the upstream stage relies on information about the intermediate inventory, and on forecasts of the production level for the downstream stage. In the third model, production smoothing by the upstream stage is based on the echelon inventory.

The notation will be very similar to that for the one-stage model. We let ℓ_1 and ℓ_2 denote the lead times for the upstream and downstream stage, respectively, and define $\ell=\ell_1+\ell_2$. Then at time t we want to find for the upstream stage $Q(t+\ell_1)$, planned production that starts in t to be completed by $t+\ell_1$, and for the downstream stage $P(t+\ell_2)$, planned production that starts in t to be completed by $t+\ell_2$. We use $J(t+\ell_1)$ to denote the planned intermediate inventory, and $I(t+\ell_2)$ to denote the planned finished-goods inventory, at the times $t+\ell_1$ and $t+\ell_2$, respectively.

We present these models for the simplest two-stage production environment. We assume that there is a one-to-one mapping between items produced in the upstream stage and items produced by the downstream stage. That is, each item from the upstream stage receives further processing at the downstream stage and results in a unique end item. Furthermore, we assume that the production units are defined so that one unit of downstream production starts requires as input one unit of completed upstream production. Although we present the models for this setting, we can extend it to more complex environments, such as when the downstream stage performs assembly of sets of components produced by the upstream stage.

To extend the one-stage model to a two-stage system requires the specification and analysis of a production smoothing rule for each stage. We desire to do this in a way that permits one to see how the planning at one stage impacts the other stage, and how the intermediate inventory can be used to decouple the stages. We discuss first the downstream stage, since it will be simpler.

Downstream Stage

The production smoothing model for the downstream stage is identical to that given for a one-stage system. Namely, we set the planned production level at time t for completion by time $t+\ell_2$ by the rule

$$(16) \qquad P(t+\ell_2) = \frac{1}{n+1}\{F_t(t+1,t+\ell_2+n) - P(t+1,t+\ell_2-1) - \hat{I}(t) + SS_2\} \quad ,$$

where n is the window length and SS_2 is the finished-goods safety-stock target. Thus, as in (3) we set the planned production to equal the average of the net requirements over the time window $t+\ell_2$, ... $t+\ell_2+n$. We assume that sufficient component inventory is available to permit the execution of (16). As a consequence, the analyses from the previous section of the planned inventory apply directly to $P(t+\ell_2)$ given by (16) and to $I(t+\ell_2)$ as implied by (16).

The extension of (3) to the upstream stage is less clear. This is because the demand on the upstream stage is not independent, but is set by the production by the downstream stage: $P(t+\ell_2)$ is the demand on the intermediate inventory at time t. Consequently, to extend (3) we need an appropriate forecast of the planned production for the downstream stage over the smoothing window for the upstream stage. We suggest below two approaches.

Upstream Stage: Production to Intermediate Inventory

The first approach is to smooth the upstream production using an explicit forecast of the downstream production over the smoothing window. The production smoothing rule is

$$(17) \qquad Q(t+\ell_1) = \frac{1}{m+1}\{P_t(t+\ell_2+1, t+\ell+m) - Q(t+1, t+\ell_1-1) - \hat{J}(t) + SS_1\}$$

where m is the window length, $P_t(t+\ell_2+1, t+\ell+m) = P_t(t+\ell_2+1) + \ldots + P_t(t+\ell+m)$ and $P_t(t+\ell_2+i)$ is a forecast made at time t of $P(t+\ell_2+i)$ for $i=1, \ldots \ell_1+m$. $\hat{J}(t)$ denotes the actual intermediate inventory at time t, net of the downstream production started at time t; that is,

$$(18) \qquad \hat{J}(t) = \hat{J}(t-1) + \hat{Q}(t) - P(t+\ell_2) \quad ,$$

and SS_1 is the target safety stock for the intermediate inventory. We note here that $P(t+\ell_2+i)$ will be the demand on the upstream stage at time t+i. Hence, $P_t(t+\ell_2+1, t+\ell+m)$ denotes the demand forecast for the upstream stage over the time interval $t+1, \ldots t+\ell_1+m$.

Using (18) we can rewrite (17) as a simple smoothing equation that is analogous to (7):

$$(19) \qquad Q(t+\ell_1) = \frac{1}{m+1}\{P_t(t+\ell+m) + [P_t(t+\ell_2, t+\ell+m-1) - P_{t-1}(t+\ell_2, t+\ell+m-1)] + \varepsilon_1(t)\} + \frac{m}{m+1}\{Q(t+\ell_1-1)\} \quad ,$$

where $\varepsilon_1(t) = Q(t) - \hat{Q}(t)$. Thus, planned production for the upstream stage is a weighted average of the production level from the previous period and the initial forecast of cumulative net requirements in time $t+\ell+m$.

To specify $P_t(t+\ell_2+i)$, we note that (16) is equivalent to a simple smoothing equation

$$(20) \qquad P(t+\ell_2) = \frac{1}{n+1}\{F_t(t+\ell_2+n) + \delta_t(t,t+\ell_2+n-1) + \varepsilon_2(t)\} + \frac{n}{n+1}\{P(t+\ell_2-1)\}$$

where $\delta_t(\; , \;)$ is the forecast revision as defined before, and $\varepsilon_2(t) = P(t) - \hat{P}(t)$. Now the forecast of $P(t+\ell_2+1)$ made at time t should be

$$(21) \qquad P_t(t+\ell_2+1) = (\frac{1}{n+1}) F_t(t+\ell_2+n+1) + (\frac{n}{n+1}) P(t+\ell_2) \quad ,$$

since the expected forecast revision is zero and the expected yield variation is zero. Similarly we can use (21) to forecast $P(t+\ell_2+i)$ for $i=2, \ldots \ell_1+m$:

$$(22) \qquad P_t(t+\ell_2+2) = (\frac{1}{n+1}) F_t(t+\ell_2+n+2) + (\frac{n}{n+1}) P_t(t+\ell_2+1)$$

$$\vdots$$

$$P_t(t+\ell+m) = (\frac{1}{n+1}) F_t(t+\ell+m+n) + (\frac{n}{n+1}) P_t(t+\ell+m-1) \quad .$$

In spite of the fact that the smoothing equations for the upstream stage [(17) and (19)] have the same form as those for a single-stage system [(3) and (7)], we have not analyzed fully the random variable for planned production $Q(t+\ell_1)$. This is because the planned upstream production $Q(t+\ell_1)$ is a smoothed average of the downstream planned production, which itself is a smoothed average of customer demand. We have not been able to 'decode' this double smoothing in a way to find the variance of $Q(t+\ell_1)$, or any other measure of production smoothing. (However, it is easy to show that $E[Q(t+\ell_1] = \bar{D}$.) We will need to use simulation to see exactly the behavior of this smoothing model. Despite this lack of an analytic result, we do expect that the behavior of this smoothing model for the upstream stage will closely parallel that predicted for a single-stage system. In particular, due to the double smoothing, we expect that the variability of the upstream production will depend upon the sum of the window lengths, m+n.

The planned intermediate inventory at time $t+\ell_1$ is given by

$$J(t+\ell_1) = \hat{J}(t) + Q(t+1,t+\ell_1) - P_t(t+\ell_2+1,t+\ell),$$

which can be rewritten as

$$J(t+\ell_1) = SS_1 + P_t(t+\ell+1,t+\ell+m) - mQ(t+\ell_1)$$

by substituting from (17) for $Q(t+1,t+\ell_1-1)$. Again, although it is clear that $E[J(t+\ell_1)] = SS_1$, we have not been able to characterize analytically the variability in the planned intermediate inventory. Nevertheless, we need the planned intermediate inventory to specify how the upstream aggregate planned production $Q(t+\ell_1)$ is disaggregated into planned production for individual items, e.g. $Q(k,t+\ell_1)$ for item k.

For this model we disaggregate the upstream production to try to equalize the service levels from the intermediate inventories across all items. To do this, we set $Q(k,t+\ell_1)$ such that $\sum\limits_k Q(k,t+\ell_1) = Q(t+\ell_1)$, and such that the planned intermediate inventory for item k, $J(k,t+\ell_1)$, provides the same level of protection against stockout for all k. Define $\tilde{\sigma}_{k1}^2$ to be the variance of the deviation between the planned intermediate inventory and the actual intermediate inventory at time $t+\ell_1$:

$$\tilde{\sigma}^2_{k1} = \mathrm{Var}[J(k,t+\ell_1) - \hat{J}(k,t+\ell_1)]$$

$$= \mathrm{Var}[\sum_{i=1}^{\ell_1} \{Q(k,t+i) - \hat{Q}(k,t+i)$$

$$- P_t(k,t+\ell_2+i) + P(k,t+\ell_2+i)\}]$$

Then, for given $\tilde{\sigma}^2_{k1}$ we disaggregate $Q(t+\ell_1)$ so that the planned intermediate inventory for k is given by

$$J(k,t+\ell_1) = z\ \tilde{\sigma}_{k1},$$

where z is

$$z = J(t+\ell_1)\ /\ \sum_k \tilde{\sigma}_{k1}.$$

Unfortunately, though, we are not able to determine analytically $\hat{\sigma}^2_{k1}$, since we cannot determine the forecast error for downstream production, $\mathrm{Var}[P_t(k,t+\ell_2+i) - P(k,t+\ell_2+i)]$. Hence, we would either need to estimate it empirically or substitute the known forecast error for demand, $\mathrm{Var}[F_t(k,t+\ell_2+i) - D(k,t+\ell_2+i)]$, for the forecast error for downstream production.

Upstream Stage: Production to Echelon Inventory

The second approach for smoothing the upstream production is based on the notion of echelon inventory. The echelon inventory for a production stage equals all inventory, including work-in-process, that is downstream of the stage. In a two-stage system, the echelon inventory for the upstream stage is the intermediate inventory, plus the work-in-process within the downstream stage, plus the finished-goods inventory. Each period this echelon inventory is increased by the amount of production completed by the upstream stage, and is decreased by the amount of customer demand. We set the planned production level for the upstream stage to be the average of the net requirements on the echelon inventory over the interval $t+\ell, t+\ell+1, \ldots t+\ell+m$, where m is the window length:

$$(23)\qquad Q(t+\ell_1) = \frac{1}{m+1} \{F_t(t+1,t+\ell+m) - P(t+1,t+\ell_2) - Q(t+1,t+\ell_1-1) - \hat{I}(t) + SS_2 - \hat{J}(t) + SS_1\} \quad .$$

To explain the bracketed term, note that $(\hat{I}(t) - SS_2) + P(t+1,t+\ell_2) + (\hat{J}(t) - SS_1)$ is the current echelon inventory (exclusive of safety stocks), and $Q(t+1,t+\ell_1-1)$ is the planned input to the echelon inventory over $(t+1,t+\ell_1-1)$. The sum of the current echelon inventory and the planned input over $(t+1,t+\ell_1-1)$ should cover customer demand over $(t+1,t+\ell_1+\ell_2-1)$ since ℓ_2 is the lead time for the downstream stage. Production at the upstream stage that is started at t will be input into echelon inventory at time $t+\ell_1$, but will not be available for customer demand until $t+\ell_1+\ell_2 = t+\ell$. Now, since $F_t(t+1,t+\ell+m) = F_t(t+1,t+\ell-1) + F_t(t+\ell,t+\ell+m)$, we see that the bracketed term in (23) corresponds to the net requirements on the echelon inventory over $(t+\ell,t+\ell+m)$. As we have done previously, we set planned production to be the average of the net requirements on the appropriate inventory over the chosen smoothing window.

The analysis of (23) is identical to that for (3) for the one-stage system where we replace n by m, $P(t+\ell)$ by $Q(t+\ell_1)$, SS by (SS_1+SS_2) and $\hat{I}(t)$ by $(\hat{I}(t) + P(t+1,t+\ell_2) + \hat{J}(t))$. In particular, we can rewrite (23) as

$$Q(t+\ell_1) = \frac{1}{m+1}\ \{F_t(t+\ell+m) + \delta_t(t,t+\ell+m-1)$$
$$+ \varepsilon_1(t) + \varepsilon_2(t)\} + \frac{m}{m+1}\ \{Q(t+\ell_1-1)\}\quad ,$$

from which we can obtain

$$E[Q(t+\ell_1)] = \bar{D}$$

and

$$\mathrm{Var}[Q(t+\ell_1)] = (\sigma_D^2 + \sigma_P^2 + \sigma_Q^2)\ /\ (2m+1)$$

where $\sigma_P^2 = \mathrm{Var}[\varepsilon_2(t) = P(t) - \hat{P}(t)]$ and $\sigma_Q^2 = \mathrm{Var}[\varepsilon_1(t) = Q(t) - \hat{Q}(t)]$. Thus, as expected, the aggregate upstream production rate equals, on average, the demand rate, while its variance is inversely proportional to the window length for smoothing.

While the characterization of the planned intermediate inventory $J(t+\ell_1)$ is not immediate for this model, it is for the planned echelon inventory at time $t+\ell_1$, $I(t+\ell_1) + P(t+\ell_1+1,t+\ell) + J(t+\ell_1)$. In fact, the analysis for the planned echelon inventory parallels that for the planned inventory for the one-stage system. (10) - (13). To see this, we write the counterpart to (10) to define the planned echelon inventory:

$$(24)\quad I(t+\ell_1) + P(t+\ell_1+1,t+\ell) + J(t+\ell_1) = \hat{I}(t) + P(t+1,t+\ell_2) + \hat{J}(t)$$
$$+ Q(t+1,t+\ell_1) - F_t(t+1,t+\ell_1)$$

Then, we use (23) to substitute for $Q(t+1,t+\ell_1-1)$ in (24) to obtain

$$I(t+\ell_1) + P(t+\ell_1+1,t+\ell) + J(t+\ell_1) = SS_1 + SS_2 + F_t(t+\ell_1+1,t+\ell-m)$$
$$- m\ Q(t+\ell_1)\quad .$$

From this, we find that the mean echelon inventory equals $SS_1 + SS_2$

+ $\ell_2\bar{D}$, while its variance is given by

$$(25) \qquad \frac{m^2}{2m+1} \cdot (\sigma_D^2 + \sigma_P^2 + \sigma_Q^2)$$

$$+ \sum_{j=0}^{\ell_2+m-1} \left[1 - 2\left(\frac{m}{m+1}\right)^{j+1}\right] \left[\sigma_D^2 - \sum_{i=0}^{\ell_1+j} \sigma^2(i)\right] \quad .$$

To disaggregate $Q(t+\ell_1)$ we need to find the planned production for each item k, $Q(k,t+\ell_1)$ such that

$$\sum_k Q(k,t+\ell_1) = Q(t+\ell_1) \quad .$$

We can do this in a similar manner to that for the one-stage system. Namely, we set $Q(k,t+\ell_1)$ so that the planned echelon safety stock for item k, defined as $I(k,t+\ell_1) + P(k,t+\ell_1+1,t+\ell) + J(k,t+\ell_1) - \ell_1\bar{D}(k)$, equals $z\sigma_{k1}$ where z is given by

$$z = \frac{I(t+\ell_1) + P(t+\ell_1+1,t+\ell) + J(t+\ell_1) - \ell_1\bar{D}}{\sum_k \sigma_{k1}} \quad ,$$

and σ_{k1}^2 equals the variance of

$$\sum_{i=1}^{\ell_1} [Q(k,t+i) - \hat{Q}(k,t+i)] + [F_t(k,t+i) - D(k,t+i)] \quad .$$

In words, σ_{k1}^2 is the variance of the deviation between the planned echelon inventory for time $t+\ell_1$ and the actual echelon inventory for time $t+\ell_1$ for item k. Thus, this disaggregation equalizes the planned echelon safety stocks for all items, where we have normalized by the standard deviation of the cumulative upstream yield uncertainty and forecast error over the lead time ℓ_1. The rationale for this disaggregation scheme is to try to spread the planned echelon safety stock (i.e. either in the intermediate or finished-goods inventory), 'evenly' over the items. By 'evenly', we intend for each item's echelon safety stock to provide the same level of protection against uncertainty both from the forecast errors and from yield uncertainty in the upstream stage over its lead time ℓ_1.

As before, the measure of the level of protection from the disaggregation, z, is a random variable. Its mean is given by $(SS_1+SS_2) / \sum \sigma_{k1}$, while its variance is directly proportional to (25).

It will be convenient to term the first (produce to intermediate inventory) approach as the decoupled model and to term the second (produce to echelon inventory) as the nested model. In the first case, the intermediate inventory decouples the stages. The upstream stage produces to this inventory based on a forecast of downstream usage, while the downstream stage draws its raw materials from this inventory. In the second case, both the upstream and downstream stages produce to the same forecast, namely the demand forecast. But the stages are 'nested' in that the upstream stage produces to the echelon inventory which contains the downstream stage and its inventory.

COMPUTATIONAL STUDY

To examine the proposed production smoothing models. we conducted a computational study based on data gathered on the set of items produced in one manufacturing facility. The purpose of the computational study is threefold: to understand how much inventory is needed in a make-to-stock system and where that inventory should be positioned; to compare the effectiveness of the different production smoothing models for a two-stage system; and to determine, via a comparison to a simulation, the accuracy of the approximate analyses of these smoothing models. We first describe the test data, then the simulation program that we have developed, and finally the computational results.

Test Scenario

To test the production smoothing models we abstracted a test scenario based on data gathered from a manufacturing operation. For the results given here, we have disguised this data to protect the identity of the manufacturing firm. However, the reported results are representative of the results obtained using the actual data.

In Table 1 we give the mean demand rate (units per month) and its standard deviation for the family of items. Of the 38 items in this family, we consider only 25 as candidates for the make-to-stock system. We exclude all items that have a start rate of less than 200 units per month, since it would be too risky to plan to stock these low-demand items.

We assume that there are two production stages and that each item requires processing from both stages. The lead time for the first stage is three months (ℓ_1=3) and is one month for the second stage (ℓ_2=1). The first stage processes each item in a lot of exactly 500 units. The second stage can process lots of any size. We will find it convenient to express the simulation results in terms of the lots for the upstream stage (500 units/lot).

For the sample of 25 items, the mean monthly demand is 81 lots per month (500 units/lot) and the standard deviation of aggregate demand σ_D is 28.5 lots per month ($\sigma_D^2 = 811$).

For the test scenario we assume that the forecast process is unbiased, and that for each item k, the variance of the forecast revision step is given by

$$\sigma^2(0) = \sigma^2(1) = .1 \ \sigma_D^2$$

$$\sigma^2(2) = .13 \ \sigma_D^2$$

$$\sigma^2(3) = .2 \ \sigma_D^2$$

$$\sigma^2(4) = \sigma^2(5) = .23 \ \sigma_D^2$$

$$\sigma^2(j) = 0 \quad \text{for } j > 5.$$

Thus, any forecast for 6 or more months ahead equals the mean monthly demand, and is not revised until there is 5 or less months to go.

We assume that there is no uncertainty in the production yield in each stage. There are two reasons for this. First, we had no data from which to estimate the parameters of the yield uncertainty. Second, we had observed in an earlier study of a comparable production operation that the uncertainty in the forecast errors dominated that in the yield realization. Hence, we hope that by including only the uncertainty in the forecast errors we will capture the primary behavior of the smoothing models. Nevertheless, given information on yield uncertainty, we could easily incorporate this into our study.

We can find σ_k^2 , the variance of the cumulative forecast error for item k over a given lead time ℓ, from the above specification of the variance of the forecast revision steps and from the knowledge of σ_D^2 for item k (Table 1). For instance, if $\ell=3$ then $\sigma_k^2 = \sigma^2(0) + (\sigma^2(0) + \sigma^2(1)) + (\sigma^2(0) + \sigma^2(1) + \sigma^2(2)) = .63\sigma_D^2$. For later reference, we note that the sum over the 25 make-to-stock items of the standard deviations of the cumulative forecast errors, $\sum_k \sigma_k$, is 25 lots when $\ell=1$ month, 63 lots when $\ell=3$ months, and 86 lots when $\ell=4$ months.

Simulation Program

We have written in PL-1 a program to simulate the production planning process for a two-stage production system. The simulation is a discrete-time simulation where the time unit is one month. Each month the simulation generates customer demand and a demand forecast for the next twelve months for each item. The end inventories for the items are increased by completed downstream production and are depleted by the demand amount. Backorders are created when insufficient inventories exist. The simulation then applies a specified production smoothing model [e.g. (3)] and its disaggregation to determine the production starts for the downstream stage. The intermediate inventories for the items are increased by completed upstream production and are depleted by usage from the production starts by the downstream stage. Note that the intermediate inventory cannot have backorders; rather, if insufficient inventory exists, then the downstream production starts need to be reduced. Finally, the simulation then applies a specified production smoothing model [e.g. (17)] and its disaggregation to determine the production starts for the upstream stage. This process is then repeated each month for the run length of the simulation. Note that, in effect, the simulation acts as if the events, customer demand, production starts, and completed production, occur at the start (or end) of every month.

For our tests the simulation is run for 1000 months where the first 40 months are for initialization and various statistics are collected over the remaining 960 months. The simulation relies on a common demand and forecast time series in order to increase the comparability of simulation runs for different smoothing models. Thus, any differences in performance between two simulation runs will be due to the smoothing models and not due to any differences in the realization of the demand or forecast process. To generate the demand and forecast time series, the forecast revision process obtains $F_t(s)$ as a lognormal random variable with mean $F_{t-1}(s)$ and with variance $\sigma^2(s-t)$ [see (1)].

In the previous section we develop the disaggregation procedure with no restrictions on the sign of the production outcome. Indeed, this procedure may suggest the impossible, namely negative production for an item with an excessively high inventory. The simulation does not permit negative production. Rather, if the disaggregation results in this outcome for a particular item, then the production starts for that item are set to zero and the disaggregation procedure is repeated for the remaining items.

Similarly, the disaggregation procedure assumes that sufficient raw material is available to make the desired production starts. For the upstream stage, the simulation retains this assumption. However, for the downstream stage, the simulation will not start more production of a particular item than is available in its intermediate inventory (i.e. we do not permit backordering on the intermediate inventory). Rather, if the desired production for an item exceeds the available raw material, we set its production equal to the raw material level and then repeat the disaggregation procedure for the remaining items (after we reduce the planned aggregate production by the amount preset for the excluded item).

For the simulation we impose a lot size of 500 units for all items for the upstream stage. Thus, for each item the monthly production starts must be a multiple of 500 units. To accomplish this we have to modify the disaggregation procedure to reflect this restriction. The modification is to compute the desired number of units to start, divide by 500 to convert to lots, and then round to the nearest integer. We assume no such restriction for the downstream stage, although we could impose a fixed lot size, if appropriate.

The simulation also has the capability to limit the aggregate production for each stage. For instance, the upstream stage might not be capable of starting more than 90 lots per month. In this case, we would modify the production smoothing model to set production starts equal to the minimum of the desired start rate from the model and the capacity limit, say 90 lots. However, in the computational work that we report, we do not use this capability, but assume that there are no limits to the production at each stage.

The disaggregation of upstream production for the decoupled (produce to intermediate inventory) approach requires an item forecast of downstream production. We generate this forecast via an item-level version of (21) - (22), with one modification. We replace the actual production starts for item k, $P(k, t+\ell_2)$, in (21) by the amount that would be started if there were no shortages in the intermediate inventory. This modification is necessary because stockouts in the intermediate inventory perturb the actual production starts from their desired level; hence, the actual production starts may not be an accurate reflection of future production by the downstream stage.

Computational Results

In this section we present and discuss our computational work. We do this in two parts. First, we consider the application of a one-stage model and study its behavior on the test scenario. Second, we consider the two versions of the two-stage model and compare them against each other and versus the one-stage model for the test scenario.

To apply the one-stage model to the test scenario, we assume a make-to-stock system with a finished-goods inventory but with no intermediate inventory. In effect, we combine the two stages in the test scenario into one stage with a production lead time $\ell = 4$ months. Each month we use (3) to set the aggregate production start rate, which is disaggregated based on the finished-good inventories via (14)-(15).

We used the simulation to compare the "actual" behavior (as found from the simulation) of the one-stage model with the predicted behavior from our analysis of the one-stage model, e.g. (9), (13). We report our results in Table 2. We have run the simulation for window lengths n=0,1,2,...6 and for safety stock targets of SS = 80, 120 and 150 lots (recall that $\sum\sigma_K$ = 86 lots for ℓ = 4 months). To show the consequences from production smoothing, we report the standard deviations of aggregate production P(t) and of the actual aggregate inventory position $\hat{I}(t)$. As a measure of service we report the fill rate, which equals the fraction of demand that is satisfied by inventory without any delay.

From the simulation results in this table we can see how the two decision parameters, the window length and the safety stock target, affect the reported performance measures. Namely, the window length essentially determines the measures of production smoothing [the standard deviations of P(t) and $\hat{I}(t)$], while the safety stock target determines the fill rate. This is consistent with the analysis of the production smoothing model. Furthermore, the analytic results (9), (13) are reasonably accurate predictions of the actual (simulated) values.[1] This indicates that the approximations made by the analysis (i.e. ignoring lot-sizing, and assuming that inventories remain balanced so that the proposed disaggregation is always feasible) do not give significant errors in this test scenario.

Finally, from Table 2 we see the implications of a make-to-stock system for this family of items. The average production starts needed for the sample of items in the test scenario is 81 lots per month. We see that maintaining a finished-goods inventory of 80 lots (about one month of demand), on average, will result in a 90% fill rate. Increasing this safety stock by 50% to 120 lots improves service to a 95% fill rate. To get to a 98% fill rate requires an investment in another 30 lots. The choice of window length for the production smoothing model dictates the extent of production smoothing. With no smoothing (n=0), the monthly production start rate, while equal to 81 lots on average, has a standard deviation of nearly 30 lots. Thus, we expect production starts to exceed 120 lots 16% of the time, and similarly to fall below 50 lots 16% of the time. Substantial production smoothing is possible by increasing the window length, but with decreasing returns. For instance, a window length of n=2, which corresponds to using a six-month cumulative forecast ($6=n+\ell$), reduces, the standard deviation of production starts by almost 60% over no smoothing. The cost from increased production smoothing is a slight degradation in fill rate, and an increased variability in the actual inventory position.

We examined the two-stage models to see what improvement is possible by inserting an intermediate inventory between the stages. To examine the smoothing behavior, we first simulated the two smoothing models for a fixed safety stock but with varying window lengths. As with the one-stage system, the smoothing behavior is effectively independent of the safety stock targets for reasonable stocking levels. Table 3 gives the results for the decoupled approach (produce to intermediate inventory) while Table 4 gives the results for the nested approach (produce to echelon inventory). In both cases the safety stock targets are SS_1 = 100 and SS_2 = 120.

Table 3 shows that for the decoupled approach the production smoothing behavior for the downstream stage is as predicted and is independent of the upstream stage. The production behavior for the upstream stage, for which we do not have an analytic prediction, reflects the effect of double smoothing: the upstream production is set by smoothing the forecast of downstream production, which itself is a smoothed average of the demand forecast. Consequently, for a fixed window length for the upstream stage there is greater smoothing as the downstream window length grows. In contrast with the one-stage system, the fill rate provided by the safety stocks is very dependent on the level of production smoothing. In particular, the fill rate declines dramatically with longer window lengths (more smoothing) for the downstream stage. This is due to the fact that the item forecasts of downstream production used by the upstream stage become less accurate with longer window lengths for the downstream stage.[2] As a consequence, the intermediate inventory has frequent stockouts, which ultimately results in poor customer service.

In Table 4 we see the comparable production smoothing behavior for the nested approach. Here, the analytic predictions of the production smoothing for both stages are reasonably accurate. Furthermore, the level of smoothing for the upstream stage is totally independent of the downstream stage. And the fill rate is virtually independent of the choice of smoothing parameters, as we saw for the one-stage system.

To explore the impact of the safety stock levels, we contrast in Table 5 the fill rates from the two approaches for a series of safety stock choices. Due to the fact that the fill rate for the decoupled approach is very sensitive to the level of production smoothing, we simulated a set of cases with substantial smoothing (m=3, n=2) and another set with limited smoothing (m=1, n=0). We chose the safety stock levels to provide insight into the proper positioning of these stocks and to allow comparison with the one-stage model (Table 2). We did not permit the downstream safety stock target to be set below 40 lots, since below that we have no hope of providing reasonable service (recall that $\sum\sigma_k$ = 25 lots when ℓ = 1). Based on the results in Table 5, we make the following observations:

a) For the nested approach (produce to echelon inventory), for a fixed total safety stock (SS_1+SS_2), the fill rate is relatively insensitive with slight improvement as more safety stock is placed downstream.

b) For the coupled approach (produce to intermediate inventory) the fill rate is very sensitive to both the amount of smoothing and the positioning of the safety stock. As seen in Table 3, service again deteriorates with increased smoothing. For a fixed total safety stock, fill rate improves as more stock is placed in the intermediate inventory as long as a minimal level (SS_2=40) is kept downstream; beyond this minimum, service will be degraded.

c) In comparing the two approaches, we see that the nested approach dominates the decoupled approach for the case with substantial smoothing. When there is limited smoothing, however, the decoupled approach is slightly better, provided the appropriate inventory positioning.

In comparing either of the approaches for the two-stage model with the one-stage model, we need more inventory with a two-stage model than with a one-stage model for a given fill rate. For instance, a safety stock of 120 lots with the one-stage model, gives a 95-96% fill rate. The nested approach for the two-stage model provides a 90%-92% fill rate for the same total safety stock; the decoupled approach can provide a 94% fill rate but with limited smoothing. However, it typically will cost less to hold stock in the intermediate inventory than in the finished-goods inventory. Thus, a two-stage model can be prefereable to the one-stage model if the holding cost for the intermediate inventory is low enough relative to the cost for the finished-goods inventory. For instance, suppose a 90% fill rate is desired. We can achieve this with the one-stage model with a finished-goods safety stock of 80 lots for window lengths n=2 and n=3. For comparable smoothing with the two-stage model, we would need the nested approach with smoothing windows m=3, n=2 and with intermediate safety stock of 80 lots and a finished-goods safety stock of 40 lots. Thus, the two-stage model would be preferable if the holding cost for 80 lots of intermediate inventory is less than that for 40 lots of finished-goods inventory.

FOOTNOTES

[1] Note that we report the standard deviation of actual inventory, rather than that of planned inventory as given by (13). Since the actual inventory is given by

$$\hat{I}(t+\ell) = I(t+\ell) + \sum_{i=1}^{\ell} \{P(t+i) - \hat{P}(t+i)\}$$

$$+ \sum_{i=1}^{\ell} \{F_t(t+i) - D(t+i)\} \quad ,$$

we can express its variance in terms of the variance of $I(t+\ell)$ given by (13).

[2] To disaggregate upstream production for the decoupled approach, we need an item forecast of downstream production. We obtained this forecast via an item-level version of (21) - (22). By using an alternate forecasting method, namely proportioning the aggregate forecast by the expected demand level, we could avoid the degradation in fill rate seen in Table 3. However, this alternate forecast method resulted in system performance that was strictly dominated by the nested model (Table 4).

APPENDIX

Derivation of (8) and (9)

By recursive substitution for P(t) in (7) we obtain

$$P(t+\ell) = \sum_{i=0}^{\infty} \left(\frac{1}{n+1}\right)\left(\frac{n}{n+1}\right)^{i} \{F_{t-1}(t-1+\ell+n) + \delta_{t-i}(t-i,\ t-i+\ell+n-1) + \varepsilon(t-i)\} \quad ,$$

where we assume that an infinite time history exists. By assumption we have that for all i

$$E[F_{t-i}(t-i+\ell+n) + \delta_{t-i}(t-i,\ t-i+\ell+n-1) + \varepsilon(t-i)] = \bar{D} \quad ,$$

and

$$\mathrm{Var}[F_{t-i}(t-i+\ell+n) + \delta_{t-i}(t-i,\ t-i+\ell+n-1) + \varepsilon(t-i)] = \sigma_D^2 + \sigma_P^2 \quad ,$$

Furthermore, we have assumed that these bracketed terms are independent across time. Thus, we obtain

$$E[P(t+\ell)] = \sum_{i=0}^{\infty} \left(\frac{1}{n+1}\right)\left(\frac{n}{n+1}\right)^{i} \bar{D}$$

$$= \bar{D} \quad ,$$

$$\mathrm{Var}[P(t+\ell)] = \sum_{i=0}^{\infty} \left(\frac{1}{n+1}\right)^{2} \left(\frac{n}{n+1}\right)^{2i} (\sigma_D^2 + \sigma_P^2)$$

$$= (\sigma_D^2 + \sigma_P^2) \ / \ (2n+1) \quad .$$

References

Cruickshanks, A. B., R. D. Drescher and S. C. Graves, "A Study of Production Smoothing in a Job Shop Environment," Management Science 30, 3 (March 1984, 168-380.

Eppen, G. and L. Schrage, "Centralized Ordering Policies in a Multiwarehouse System with Lead Times and Random Demand," in Schwarz, L. (ed.), Multi-Level Production/Inventory Control Systems: Theory and Practice, North-Holland Amsterdam, 1981.

Federgruen A. and P. Zipkin, "Approximations of Dynamic, Multilocation Production and Inventory Problems," Management Science, 30, 1, (January 1984), 69-84.

Hax, A. C., and H. C. Meal, "Hierarchical Integration of Production Planning and Scheduling," In Studies in Management Sciences, Vol. I, Logistics, M. A. Geisler (ed.). North Holland-American Elsevier, New York, 1975.

C. Holt, F. Modigliani, and H. Simon, "A Linear Decision Rule for Production and Employment Scheduling," Management Science 2, 1-30 (1955).

_______, _______, and J. Muth, "Derivation of a Linear Decision Rule for Production and Employment," Management Science 2, 159-177 (1956).

Schneeweiss, C.A., "Smoothing Production by Inventory - An Application of the Wiener Filtering Theory," Management Science 17, 7 (March 1971), 472-483.

_______ "Optimal Production Smoothing and Safety Inventory," Management Science, 20, 7 (March 1974), 1122-1130.

Silver, E. A., "A Tutorial on Production Smoothing and Work Force Balancing," Operations Research, 15, 6 (November-December 1967), 985-1010.

Figure 1: Two-Stage System

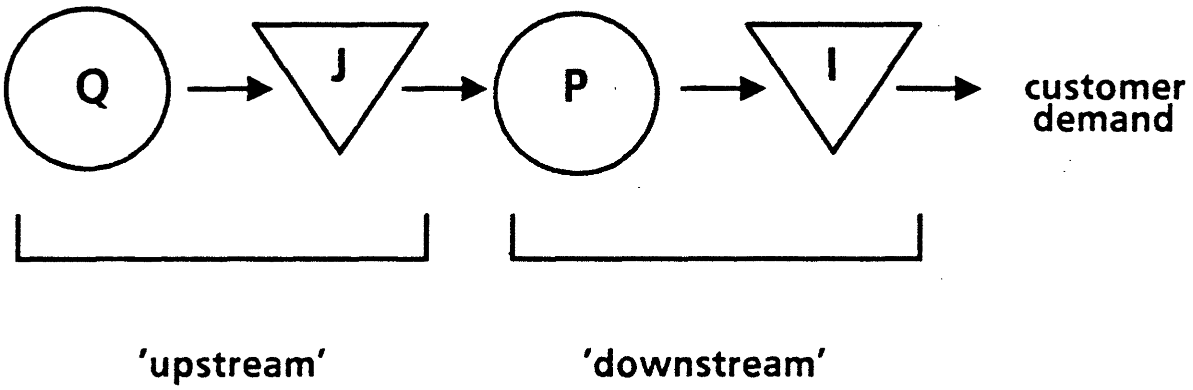

TABLE 1: DEMAND FOR TEST SAMPLE.
500 demand units equals one lot.
Starred items(*) are excluded from Make-to-Stock System.

ITEM CODE	DEMAND	
	MEAN	STD. DEV
*1	2	2
*2	1	1
*3	18	21
*4	8	16
5	651	961
6	214	362
*7	137	298
*8	11	15
9	325	238
10	237	293
11	483	569
*12	2	6
13	1315	1121
14	1953	1685
15	6729	4793
*16	1	3
17	554	722
*18	48	49
*19	20	30
20	1887	1951

TABLE 1 CONTINUED ON NEXT PAGE.

TABLE 1 (continued)

ITEM CODE	DEMAND	
	MEAN	STD. DEV
21	2193	1908
22	587	604
23	615	728
24	833	665
*25	47	151
*26	8	11
27	1206	1603
28	15691	12278
29	1142	2286
30	336	677
*31	81	242
32	246	660
33	749	890
34	817	1199
35	232	636
36	360	822
37	279	634
38	819	1336

Safety Stock

		SS = 80	SS = 120	SS = 150	Predicted Std. Deviations
Window Length	n = 0	29.8, 30.1* .92	29.8, 30.2 .96	29.1, 30.1, .98	28.5, 30.8
	n = 1	17.4, 34.0 .91	17.3, 34.3 .96	17.2, 34.1, .98	16.4, 34.9
	n = 2	13.1, 37.5 .90	13.1, 37.9 .96	13.0, 37.6, .97	12.8, 39.4
	n = 3	10.8, 40.8 .90	10.8, 41.3 .95	10.8, 41.0, .97	10.8, 43.6
	n = 4	9.3, 44.0 .89	9.4, 44.4 .95	9.3, 44.0, .97	9.5, 47.4
	n = 5	8.3, 46.8 .89	8.4, 46.9 .95	8.2, 46.7, .97	8.6, 51.0
	n = 6	7.6, 49.0 .89	7.6, 49.4 .94	7.4, 49.3, .97	7.9, 54.4
	Predicted fill rate	.90	.96	.98	

Table 2: Results from One-Stage Model

* $\boxed{\begin{matrix}x,y\\z\end{matrix}}$: x = standard deviation of P_t; y = standard deviation of $\hat{I}_t$; z = fill rate.

downstream window length upstream window length	n = 0	n = 1	n = 2	n = 3	upstream stage prediction
m = 0	29.2, 27.4,* .99	24.1, 15.9,. 98	22.0, 12.4, .93	19.9, 10.6, .88	**
m = 1	17.2, 27.4, .99	15.9, 15.9, .98	14.7, 12.4, .93	13.6, 10.5, .88	
m = 2	13.0, 27.4, .99	12.6, 15.9, .97	11.9, 12.4, .92	11.2, 10.5, .87	
m = 3	10.7, 27.2, .99	10.7, 15.9, .97	10.2, 12.4, .92	9.8, 10.5, .87	
downstream stage prediction	28.5	16.5	12.8	10.8	

Table 3: Production Standard Deviations for Two-Stage Model: Production to Intermediate Inventory ($SS_1 = 100$, $SS_2 = 120$)

* x,y, z : x = standard deviation of Q_t ; y = standard deviation of P_t ; z = customer fill rate.

** No analytic prediction is currently available.

downstream window length upstream window length	n = 0	n = 1	n = 2	n = 3	upstream stage prediction
m = 0	29.2, 27.4*	29.2, 15.9	29.2, 12.4	29.2, 10.6	28.5
m = 1	17.2, 27.4	17.2, 15.9	17.2, 12.4	17.2, 10.6	16.5
m = 2	13.1, 27.3	13.1, 15.9	13.1, 12.4	13.1, 10.6	12.8
m = 3	10.9, 27.2	10.9, 15.9	10.9, 12.4	10.9, 10.6	10.8
downstream stage prediction	28.5	16.5	12.8	10.8	

Table 4: Production Standard Deviations for Two-Stage Model: Production to Echelon Inventory ($SS_1 = 100$, $SS_2 = 120$)

*$\boxed{x,y}$: x = standard deviation of Q_t; y = standard deviation of P_t.

Note, the fill rate is .98 for all instances.

		FILL RATE (m = 3, n = 2)		FILL RATE (m = 1, n = 0)	
SS_1	SS_2	PRODUCE TO ECHELON INVENTORY	PRODUCE TO INTERMEDIATE INVENTORY	PRODUCE TO ECHELON INVENTORY	PRODUCE TO INTERMEDIATE INVENTORY
40	40	.83	.69	.84	.81
0	80	.83	.66	.85	.39
80	40	.90	.78	.91	.94
40	80	.91	.77	.92	.88
0	120	.91	.74	.92	.43
110	40	.93	.84	.94	.97
70	80	.94	.84	.95	.96
30	120	.94	.82	.95	.87
140	40	.95	.88	.96	.98
100	80	.96	.88	.97	.98
60	120	.96	.87	.97	.97

Table 5 : Comparison of Fill Rates for Two-Stage Models