ISE

Industrial and
Systems Engineering

Impact of Improved Forecasting on Operations

S. David Wu
Lehigh University

Mehmet O. Atan
Lehigh University

Report: 10T-005

Impact of Improved Forecasting on Operations

Mehmet O. Atan[*], S. David Wu[†]

Abstract

In this report, we aim to establish high-level tradeoffs between forecast accuracy and operational costs. We model a high-tech factory to investigate the optimal operational strategies that should be employed under diverse scenarios of capacity and demand realizations. We derive the relationship between forecast variance and operational costs in closed form and show that expected operational costs increase with the increase of forecast variance. Findings suggest that variance reduction techniques should be used in forecasting to obtain operational cost savings; especially when cost parameters for supply overage and underage are higher.

1 Introduction

The volatility observed in the high-tech markets such as semiconductors often leads to various operational difficulties. In existence of long lead-times and frequently changing demand signals, it is challenging to maintain a prompt product supply. Avoiding shortages during ramp-up and maturity stages of a product's lifecycle is crucial to sustain a desired rate of adoption amongst potential customers. Considering the expensive capacity costs, a satisfactory return on investment can only be obtained by high utilization of facilities for a long period of time throughout the lifecycle and by exploiting market sales to its full potential. Then again, over-production to prevent backorders is also risky since (1) excess inventories diminish profits, and (2) re-allocation of this expensive capacity considering other critical products would have been beneficial. Consequently, technological forecasting is critical to mitigate the risks instigated in these volatile markets.

To maintain competitiveness in high-tech markets, companies should adapt to fast advancing technology, and continuously deliver state-of-the-art products. Unfortunately, these efforts translate to substantial research and development costs. In addition, manufacturing resources are extremely expensive to build, and become quickly obsolete as new technology generations are introduced. Therefore, efficient planning and execution of operations is key to sustaining the profitability.

[*]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, Pennsylvania, 18015, mehmet@lehigh.edu

[†]P.C. Rossin College of Engineering and Applied Science, Lehigh University, Bethlehem, Pennsylvania, 18015, david.wu@lehigh.edu

One of the important operational decisions of a high-tech manufacturer is the supply rate. Inaccurate supply planning usually yields to both loss of revenues and significant increases in operational costs. Demand forecast is the primary tool that a supply planner employs to determine the optimal supply rate and accordingly build adequate production resources. However, there is always some uncertainty inherent in these forecasts. Therefore, instead of using point forecasts, ranges of demand scenarios can be considered to portray the uncertainty in a better way. Apparently, the narrower the range of demand scenarios is, the easier the supply planner's job should be. In the extreme case, if demand was deterministic, supply planning would be trivial. Consequently, smaller forecast variance is expected to lead to more efficient planning of supply.

Fortunately, to reduce the forecast variance, researchers have introduced many approaches including but not limited to use of leading indicators, forecast combinations, etc. In this study, we show that, independent of the approach that is taken to achieve, a reduction in forecast variance yields to direct operational cost savings. The impact of lower forecast variance is amplified when cost parameters are greater. We presume that the benefit can only notably increase if we move from a single product factory to a multi-product, multi-generation manufacturing model.

In the next section, we review the relevant high-tech forecasting and manufacturing literature. In Section 3, we describe the particular high-tech manufacturing environment we deal with, model the factory, and discuss optimal supply policies. Next, in Section 4, we demonstrate the impact of forecast reduction on operational costs, which is followed by concluding remarks in Section 5.

2 Literature Review

Rapid innovation processes, expensive resources, and market competition are important characteristics of high-tech industry. Manufacturers in this industry suffer from difficult capacity management problems triggered by volatile customer demand, products with short lifecycles and frequent product/technology transitions. Therefore, the capability of forecasting future market potential and rate of adoption has made technology diffusion models a popular research topic.

Demand lifecycle for a technological product characteristically follows an S-shaped profile. After market introduction, rate of product adoption goes through stages of ramp-up, maturity and ramp-down, which should be sustained by uninterrupted supply for maximum sales. To characterize product demand lifecycles, a significant portion of the forecasting literature employ technology diffusion models. Meade and Islam (1998) and Kumar and Kumar (1992) provide extensive surveys of diffusion models that have been proposed by researchers to be used for technological forecasting. These models differ in the percentage of adoption achieved when the peak diffusion rate is reached as well as the steepness of growth and decline of the diffusion rate.

The most well-known, widely used and extended diffusion model is introduced by Bass (1969). Bass model hypothesizes that potential adopters of a new technology consists of two groups. Innovators are influenced by mass media and boost the initial demand, while imitators count on word-of-mouth and drive the later demand. Innovation and imitation effects are represented by two separate parameters that together determine the shape of the model. Bass model was first tested on consumer durable goods, and provided accurate predictions on timing and magnitude of peak sales. Since then, this model has been revised to consider additional features of technology diffusion such as pricing and advertisement, and used for diffusion forecasting in various markets including but not limited to retail, education, pharmaceutics, high-tech and agriculture (Mahajan et al., 2001).

Although individual diffusion models have been successful in predicting demand for a wide variety of products in many industries mentioned above, they fail to provide accurate estimations for high-tech demand of the recent years. Significant volatility in high-tech markets' demand has encouraged researchers to develop more complex forecasting models, which focus on reducing the forecast variability through use of advanced demand signals called leading indicators and combination of multiple models' intelligence. Wu et al. (2006) use several diffusion models and demand patterns of leading indicator products to generate accurate forecasts in a custom semiconductor manufacturing setting. Aytac and Wu (2008) present a theoretical demand characterization framework that employs a Bayesian updating procedure to systematically reduce forecast variation. In Wu et al. (2010), we use various dynamic market information as leading indicator of future microprocessor demand in a semiconductor manufacturer. The implementation results in reduced forecasting errors and forecasting effort, as well as significant cost savings.

To illustrate the impact of improved forecast accuracy (through reduced forecast variance), we model a capacitated production system and investigate the relationship between operational costs and supply policies that rely on the demand forecast. Since building the necessary manufacturing resources takes significant time, high-tech companies have to rely on inventories for unaccounted demand overages. Furthermore, underutilization of production capacity is not an option until well into the maturity of the demand because the resources are too expensive to build. Consequently, the planner has to balance the risks of having too little and too much manufacturing capacity.

Diffusion of a new technology under supply constraints has been an interesting research topic. This literature consider the possibility that a manufacturer may not be able to satisfy customer demand on time, especially if the product becomes popular very quickly as in the case of video game consoles or high-tech mobile phones. Jain et al. (1991) consider supply constraints when the first time the telephone technology was introduced. They assume that customers who do not get the product wait in a queue until they are satisfied. Authors modify the Bass model to predict how the diffusion pattern changes in existence of supply constraints. More recently, Ho et al. (2002) present a make-to-stock model that a firm can build up inventory before market introduction to satisfy customer demand. They decide on optimal capacity size, timing and inventory build-up using a modified Bass model. This study shows that although delaying market introduction time may be optimal, it is never optimal to delay satisfying the demand. Nevertheless, in a similar study, Kumar and Swaminathan (2003) claim that fulfilling the demand to the maximal possible amount is not always optimal. They introduce a heuristic, which delays demand filling to build up an inventory that prevents lost sales once selling is resumed.

In addition to demand, managing supply is also challenging in high-tech manufacturing such as semiconductor industry. High-tech production systems usually operate at high utilization for economic justification of costly resources. Multiple products competing for resources and a reentrant material flow that cycles through several bottleneck operations complicates the manufacturing environment significantly. Consequently, realistic and detailed operational planning requires a combination of complex simulation and mathematical programming models (Hung and Leachman, 1996).

Although we recognize the complexity of the high-tech manufacturing, our aim is not making detailed operational decisions. We want to demonstrate the impact of efficient demand management on high-tech supply. For this purpose we use a simpler yet sophisticated enough factory model. High-tech manufacturing is fundamentally a queuing system that involves very high resource utilization. Therefore, assumption of constant, exogenous production lead times would not be a realistic representation considering the fact that lead times increase nonlinearly with the increase in utilization (Hopp and Spearman, 2001). We incorporate load-dependent lead times into the model to obtain an accurate congestion effect. Graves (1986) was first to present clearing functions that relate the expected output of a manufacturing system to the expected work-in-progress inventory level over a given period. Several different clearing functions have been proposed in literature. A simulation study can be useful to determine the clearing function that best captures the specific system dynamics (Orcun et al., 2006). To make detailed operational decisions, actual parameter values of clearing functions are essential and can be estimated through simulation (Asmundsson et al., 2006). In our model we use nonlinear saturating clearing functions, which were introduced by Karmarkar (1989) and Srinivasan et al. (1988).

In the following sections we will briefly introduce the demand and supply models that we employ to investigate the impact of improved forecasting on operational costs.

3 Factory/Operations Analysis

3.1 The Factory Model

High-tech manufacturing such as semiconductors has very complex operational dynamics that is notoriously challenging to model. There is an extensive literature focusing on realistic and detailed modeling of such environment so that it is possible to make detailed operational decisions (c.f.,(Asmundsson et al., 2006), (Orcun et al., 2006)). Such model often involves complex stochastic analysis and/or discrete simulation that attempt to capture the interconnection of key bottleneck operations and reentrant flow of material, which is further complicated by the interaction of multiple products competing for resources. The purpose of our factory/operational analysis is not intended for operational decision making. Rather, we are interested in higher level trade-offs that are relevant from a supply-demand analysis perspective, e.g., how does improvement in forecast accuracy affects operational efficiency? Can we quantify such impact in some way? In the following, we will describe a factory model that is streamlined to capture the essence of trade-off we need for analysis at this level, which is much simpler than a factory model used for detailed operational planning. In particular, we are interested in deriving clean theoretical insights that can be offered from straightforward closed-form relationships.

We describe a generalized factory model, which can be, but not limited to a semiconductor fabrication plant (fab). In this factory, overall capacity is a long-range planning activity that considers expected technology advancement over the product lifecycle and is consistent with the company's technology release policy. First, we will assume that factory capacity is given and represents the maximum possible output of the factory.

The factory is modeled to consider the congestion caused by the workload, which significantly affects production lead times. A clearing function defines the relationship between the WIP level and expected output rate for a factory. In our case, this relationship is nonlinear, and given by a

Concave Saturating Clearing Function (CSCF), as suggested by Karmarkar (1989):

$$S_t = \frac{C \cdot W_t}{W_t + K},$$

where S_t is the output level in period t, while W_t denotes the WIP level during the same period. K is the curvature parameter and C is the capacity, which are constant for the factory. This equation captures the fact that the output level increases with WIP but at a decreasing rate since higher utilization causes congestion in the factory (Asmundsson et al., 2006; Orcun et al., 2006). Typical relationship between WIP, output rate, and capacity suggested by a CSCF is given in Figure 1.

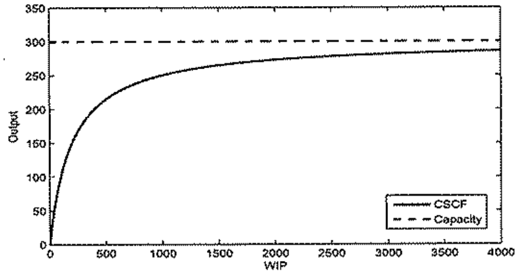


Figure 1: Output rate vs WIP

In Figure 2, we illustrate the interaction between the flow of material (thick arrows) and the flow of information (thin arrows) in high-tech demand/supply model. The material flow begins with raw material (e.g., wafer starts) releasing into the factory based on a certain inventory/production policy, which produce output at a rate characterized by the CSCF; the factory output provides the "supply" (S_t) into the finished goods inventory, which in turn satisfies customer demands. The information flow indicates that demand forecast, together with WIP and finished goods inventory levels, provide the basis for the inventory/production policy that release raw materials into the factory. The figure captures the basic relationship between the *demand signal* (demand forecast derived from customer needs that trigger the production), the *supply* (output from the factory as result of production), and the *demand* (the finished goods inventory delivered to the customer to satisfy demand). This conceptual view helps us to link the demand/forecast analysis with the factory/operations analysis. Note that the terms supply and demand are defined from the customer's perspective; when considering the factory perspective, it is more convenient to think of *supply* as factory *output*. We will make this point clear throughout the analysis.

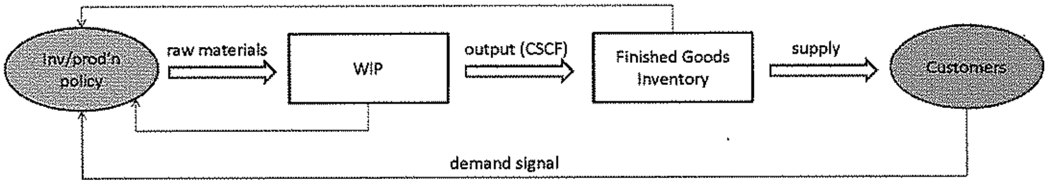


Figure 2: Demand/Production/Inventory Model

3.2 Matching Supply and Demand

To streamline our analysis, we will first establish a few conditions, focusing on the scenarios most relevant to the industry setting under investigation. We will focus our attention on a single product, since this provides the most straightforward connection between the factory and the demand/forecast analysis. Once the production starts, the following operational policies are employed: (1) Running the factory at full capacity, and (2) keeping the output rate constant during ramp-up and maturity. These are policies common in the semiconductor industry, because maintaining a high utilization is important to obtain a satisfying return on investment, while maintaining a steady output rate only requires a steady flow of wafer starts into the fab, making it easier to manage the nonlinear relationship between WIP and output. Also, avoiding backorders is crucial during the phases of ramp-up and maturity. Given the above policy, the factory will operate at the maximum output level (i.e., maximum supply S for the customer).

In the following, we investigate strategies that match the supply levels S with respect to customer's demand diffusion curve. Note that the analysis is applicable to any diffusion model, including combined diffusion curves. The Bass model is used as an example throughout the section to demonstrate that closed form solutions can be obtained.

Given the forecast for the diffusion of the new product ahead of its planned market introduction time, an optimal supply policy can be adopted while taking the supply capacity and preproduction option into account. The following analysis investigates the optimal supply policies under various scenarios that are classified based on whether (1) diffusion (sales) of the product is constrained by the supply, and (2) market introduction time is flexible.

3.2.1 Supply Constrained Diffusion:

When maximum supply rate is less than the maximum rate of demand, the diffusion process is supply constrained (Figure 3). In this scenario, peak demand rates that are faced during the maturity phase of the diffusion process can only be met on time by the use of inventory buffers.

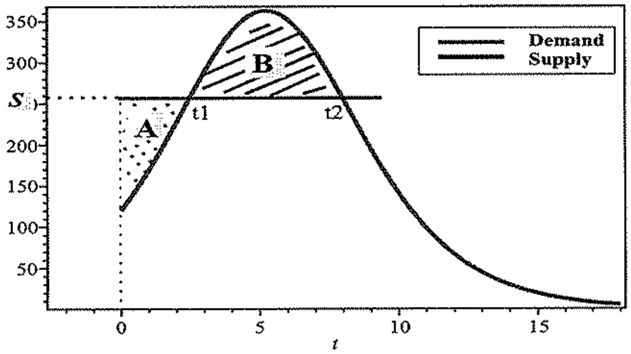


Figure 3: Supply Constrained Diffusion

As shown in the figure, supply and demand curves intersect twice, at time points t_1 and t_2. We assume that t_2 denotes the end of maturity phase, and we focus our attention to diffusion until this point. Let cumulative adoption by time t is denoted by a technology diffusion model $F(t)$. Then, we can find closed form expressions for t_1 and t_2 in terms of diffusion parameters by solving the equation:

$$S = M \cdot f(t),$$

where $f(t)$ is the rate of demand diffusion at time t ($f(t) = dF(t)/dt$) and M is market potential. For example, if demand follows the Bass model, then:

$$t_1, t_2 = \frac{-\ln\left(\frac{1}{2}\,\frac{M(p(p+q))^2 - 2\,Spq \mp \sqrt{M^2(p(p+q))^4 - 4\,M(p(p+q))^2 Spq}}{Sq^2}\right)}{(p+q)}$$

The analysis of optimal supply strategy depends on whether the extra production into buffers during (0,t_1) is sufficient to prevent backorders later in maturity. We compare the production into inventory (denoted by area A) and the excess of demand over supply (denoted by area B) to distinguish between the possible cases, where they are calculated respectively as:

$$A = S \cdot t_1 - M \cdot F(t_1), \quad and$$
$$B = M \cdot (F(t_2) - F(t_1)) - S \cdot (t_2 - t_1)$$

Case 1: A>B $(M \cdot F(t_2) < S \cdot t_2)$

When maximum output rate is realized starting with the planned introduction time ($t_p = 0$), the inventory buffer established during ramp-up is sufficient to prevent backorders during maturity. We recognize that maintaining the maximum output rate results in unnecessary inventory holding costs, while decreasing the rate would be against production policy, which enforces maximum utilization of costly fabrication facilities. Although production capacity cannot be expanded, we assume that it is possible to decrease it before the production starts by allocating some of the capacity to be utilized for other purposes. In this way, inventory holding costs are reduced as well as efficient use of production capacity -which was clearly overestimated- is provided. We assume that capacity cost is incurred long before start of the production, thus, there is no savings in terms of production costs when the capacity is decreased. Consequently, the optimal supply policy is implied by the magnitude of the capacity cut that minimizes inventory holding costs. Notice that according to CSCF, when capacity is reduced by a certain ratio, output rate also decreases by the same ratio if WIP and material release into factory are not altered.

Proposition 1. *The optimal supply strategy is to cut the production capacity at the beginning of diffusion so that $A = B$.*

Proof. Without a capacity cut, production is maintained at maximum rate (S) until decreasing demand rate hits S on t_2. At this time, there are $(A - B)$ units left in finished goods inventory. In Figure 4, t_e denotes the time point up to which the demand can be satisfied without backorders by use of leftover inventory while producing at a lower rate, $\bar{S}$ (via reducing capacity).

Note that inventory level is zero at t_e, and there is a unique pair $\bar{S}$, t_e that satisfies this condition.

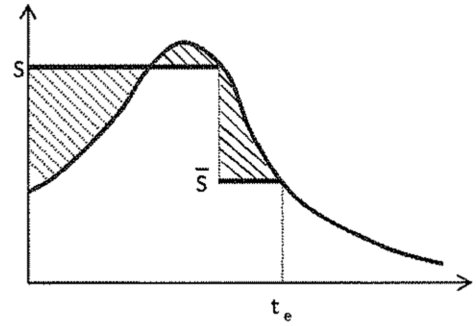


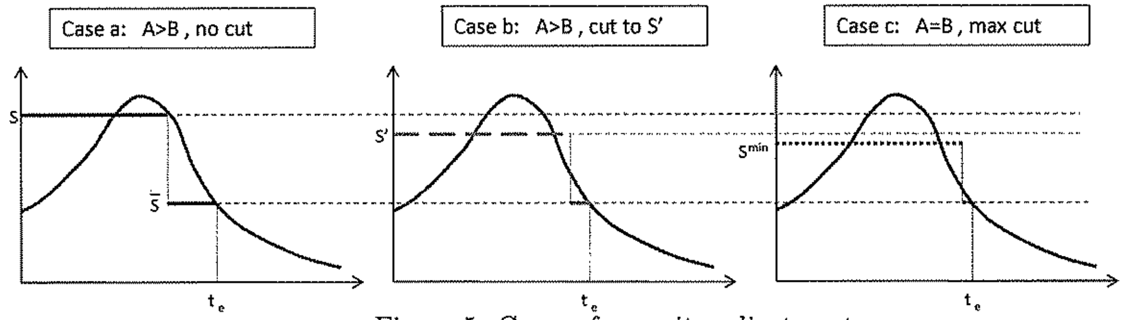
Figure 4: Case 1: $A > B$

To find $\bar{S}, t_e$, numerically solve:

$$M \cdot F(t_e) = S \cdot t_2 + (t_e - t_2) \cdot \bar{S}$$

$$M \cdot f(t_e) = \bar{S}$$

The maximum capacity that can be cut at the beginning of the diffusion is constrained by the "no backorders" policy, and provided by the relationship $A = B$ $(M \cdot F(t_2^{S^{min}}) = S^{min} \cdot t_2^{S^{min}})$. Notice that t_2 changes when the rate of supply is different. Besides these two extreme strategies that employ either maximum or no rate change, we should also consider other scenarios that differ in magnitude of the rate reduction. To achieve a fair comparison of strategies, we consider that the supply policies are run for the same period of time, until t_e, with zero leftover inventory.


Case a: A>B , no cut

Case b: A>B , cut to S'

Case c: A=B , max cut

Figure 5: Cases of capacity adjustment

We illustrate these cases that differ in the amount of initial capacity cut in Figure 5. Case a represents no cut, Case c refers to the maximum cut (down to S^{min}), and Case b illustrates a cut between the two extremes ($S^{min} < S' < S$). In all three cases, the capacity is reduced to output $\bar{\bar{S}}$ before reaching t_e. The timing of this reduction is such that there are no inventories left at t_e. There are also no backorders in any of the cases.

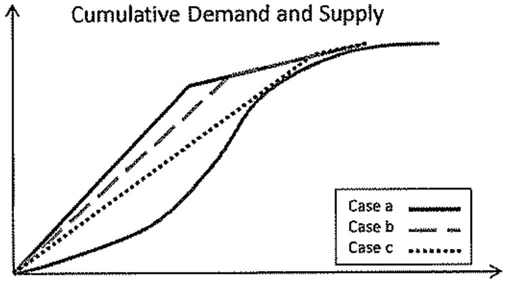


Figure 6: Cumulative supply corresponding to all scenarios

Capacity costs are incurred when they are built, so they are not relevant at this point. As illustrated in the cumulative chart (Figure 6), the inventory holding costs are the lowest when $A = B$ in Case c, since it is the closest a supply curve can be to the demand profile without having backorders. Consequently, the optimal strategy is to reduce the initial capacity as much as possible without causing backorders. We can explicitly describe the required amount of capacity cut (Solve $M \cdot F(t_2) = S \cdot t_2$), and numerical solutions are easy to obtain.

□

<u>Case 2: B>A</u> $(M \cdot F(t_2) > S \cdot t_2)$

When sales start on the planned market introduction time ($t_p = 0$), the excess production (A) in (0,t_1) is not sufficient to fill excess demand (B) during maturity. To prevent backorders, production should start earlier than planned since capacity expansion is not possible (Figure 7).

Proposition 2. *The optimal supply policy is to start production* $t_p = \frac{M \cdot F(t_2) - S \cdot t_2}{S}$ *periods before the start of diffusion.*

Proof. Latest production start time without backorders satisfies $A' = B$, where A' is the production

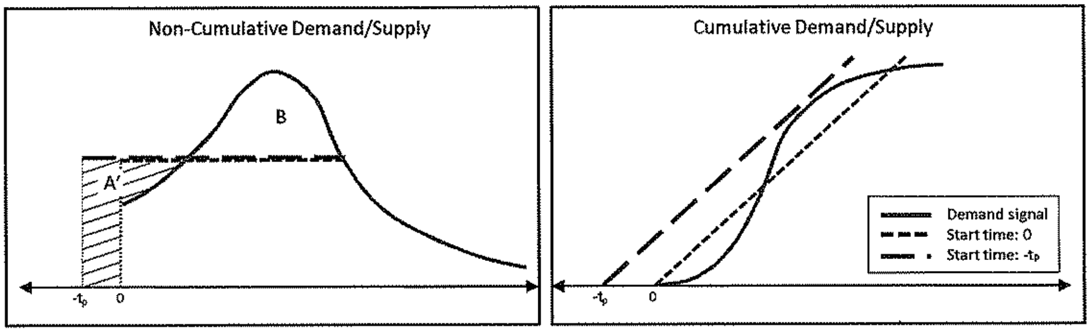


Figure 7: Case 2: Preproduction

into inventory in $(-t_p, t_1)$:

$$(t_p + t_2) \cdot S = M \cdot F(t_2)$$

After rearranging the terms, we get $t_p = \frac{M \cdot F(t_2) - S \cdot t_2}{S}$. Note that, if production starts any earlier than $-t_p$, unnecessary inventory holding costs would have incurred.

It is also possible to satisfy demand on time if production starts even earlier, outputting at a lower supply rate. However, it is not the optimal strategy since inventory costs would be higher. We investigate this scenario in Proposition 3, while considering the limited preproduction case.

□

Next, we consider some special cases that can be faced under this scenario.

i. Special Case 1: Limited preproduction ($t_p \leq x$)

Under this scenario, preproduction cannot start earlier than a certain time point $-x$. Remember that preproduction is necessary to meet demand on time.

Special Case 1.1: $A' > B$ $(M \cdot F(t_2) < S \cdot (x + t_2))$

If preproduction starts earlier than x periods before the start of diffusion while running at the initial supply rate S, there will be inventories left at the end of maturity, since $A' > B$ (Figure 8a). It is possible to prevent inventories at t_2 while avoiding backorders by reducing the production capacity, by delaying the start of production, or by using a combination of the two.

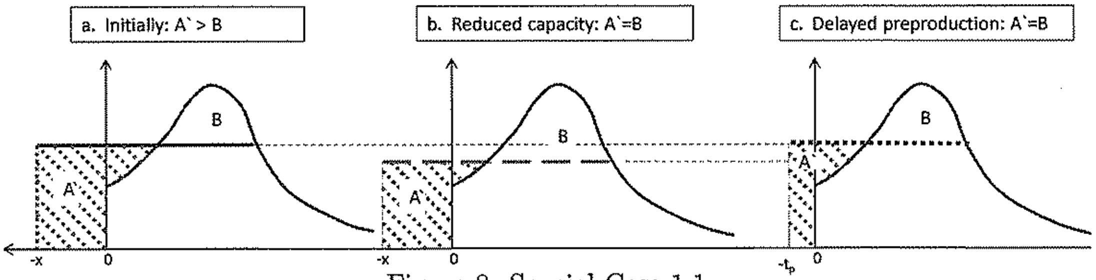


Figure 8: Special Case 1.1

Proposition 3. *Keeping the initial supply level constant and delaying start of production until $A' = B$ is the optimal strategy.*

Proof. Figure 8b and Figure 8c illustrate pure strategies of reduced capacity and delaying production, respectively, until $A' = B$. Figure 9 clearly illustrates using cumulative supply demand curves that delaying production provides lower inventories. Note that a strategy, which uses a combination of production delay and capacity decrease corresponds to a cumulative supply curve in between the curves for the two pure strategies, since the production start time and slope of the cumulative supply line would be between the values of those for the pure strategies.

□

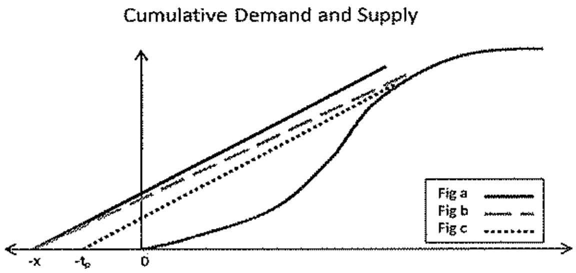


Figure 9: Special Case 1.1 - Cumulative chart

Special Case 1.2: *$B > A'$* $(M \cdot F(t_2) > S \cdot (x + t_2))$

In this case, it is impossible to avoid backorders even if the full capacity production starts x periods earlier than start of the diffusion, since $A' < B$ (Figure 10)

To fill the backorders caused by the difference $B - A'$, production continues at full capacity even after t_2. Since the backorder cost is significant, filling backorders as quickly as possible is the optimal policy. Factory outputs at full capacity until either all backorder during ramp-up is met $(A' + C = B)$

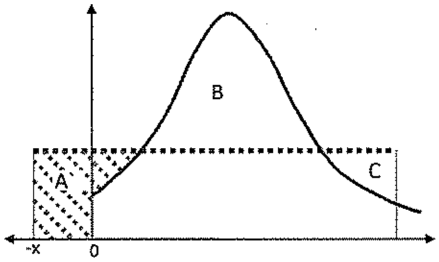


Figure 10: Special Case 1.2

or WIP is equal to the remaining demand potential plus backorders. If the latter is reached first, material release into factory should be stopped.

ii. Special Case 2: Initially constrained diffusion ($S < f(0)$)

Under this scenario, supply rate is less than the demand rate throughout ramp-up and maturity. Therefore, preproduction is required to meet demand on time. The optimal strategy is equivalent to that in Case 2. If preproduction cannot start early enough to build sufficient inventory to meet demand on time, the optimal strategy is trivial and similar to the Special Case 1.2, in which the supply tries to catch up with the backorders.

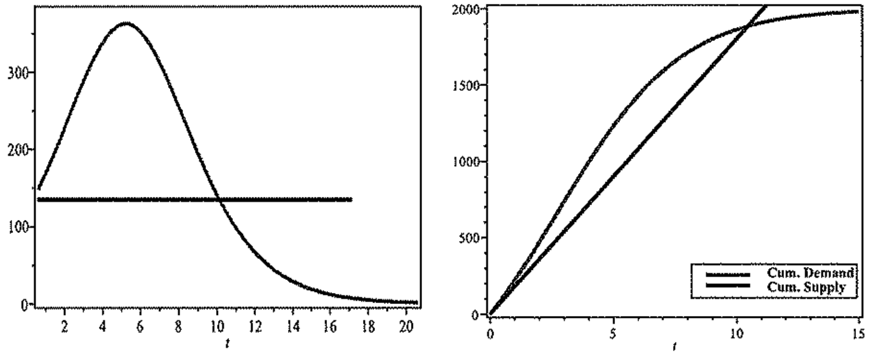


Figure 11: Special Case 2

Next, we briefly discuss the case of unconstrained diffusion.

3.2.2 Unconstrained Diffusion:

In this case, demand realization is extremely lower than the rate that capacity is planned for. Best option is to allocate some of this expensive capacity for other purposes and continue as a constrained diffusion case. Problem is similar to supply constrained Case 2, in which the optimal strategy that minimizes inventory holding costs is to decrease the supply level at time zero to obtain $A = B$.

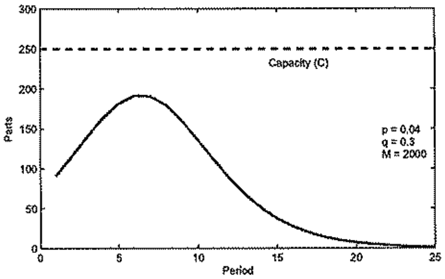


Figure 12: Unconstrained Diffusion

After analyzing the optimal supply strategies during ramp-up and maturity, we now look into the end of life phase and corresponding optimal ramp-down strategies.

3.2.3 Ramp-Down Strategy

During the end of life phase, backorders and inventories are not considered significant costs. Therefore, a more dynamic, short-term planning activity that is based on a period-to-period demand forecasts is employed. As the demand rate decreases, the supply level can be manipulated by decreasing the WIP level. Under this strategy, we can determine the amount of required wafer starts in closed form using the following steps:

1. In period t, required supply level is equal to the demand in that period.

$$S_t = M \cdot (F(t) - F(t-1))$$

2. WIP level that would provide the required output can be determined by CSCF:

$$W_t = \frac{K \cdot S_t}{C - S_t}$$

3. Then, required amount of releases (R) into factory is found by using the balance equation:

$$R_{t-1} = W_t - W_{t-1} + S_{t-1}$$

To summarize, the release into factory in period $t-1$ is given by:

$$R_{t-1} = \frac{K \cdot M \cdot (F(t) - F(t-1))}{C - M \cdot (F(t) - F(t-1))} - W_{t-1} + \frac{C \cdot W_{t-1}}{W_{t-1} + K}$$

where W_{t-1} is WIP level at the beginning of period $t-1$ and known to the planner.

For instance, if Bass model is used to characterize the demand process, W_t is derived to be the following term:

$$\frac{-Mp\left(-qe^{-(p+q)(t-1)} + e^{-(p+q)t}p + qe^{-(p+q)t} - e^{-(p+q)(t-1)}p\right)K}{\left(p+qe^{-(p+q)t}\right)\left(p+qe^{-(p+q)(t-1)}\right)\left(C + \frac{Mp\left(-qe^{-(p+q)(t-1)}+e^{-(p+q)t}p+qe^{-(p+q)t}-e^{-(p+q)(t-1)}p\right)}{\left(p+qe^{-(p+q)t}\right)\left(p+qe^{-(p+q)(t-1)}\right)}\right)}$$

When remaining market potential is equal to WIP level at any time, the material release into factory stops, and WIP is output according to CSCF until all demand is satisfied in order to prevent leftover items at the end of diffusion. The optimal supply strategies that we presented will be useful in the next section, which demonstrates the impact of better forecasts on operational efficiency.

4 Impact of Forecasting on Operations

In this section, we aim to illustrate that reduction of forecast variance would yield to decrease of operational costs in the factory. Recall that we assume an operating policy that avoids backorders during ramp-up and maturity while producing at full capacity in a steady rate. Therefore, for any given demand diffusion profile, there is a corresponding optimal supply level that satisfies this policy.

Although point estimates are used widely in practice, accuracy of such forecasts is rarely perfect when obtained using diffusion models with inherent uncertainty. The uncertainty that originates from the nonlinearity of model fitting and the errors in parameter estimation is expected to be passed on to the projection of the lifecycle model. Therefore, representing forecasts in terms of ranges instead of a single point has been useful for planning purposes.

The uncertainty in the estimate of a future realization of the random variable is described by a *prediction interval.* Meade and Islam (1995) provide a detailed discussion about determining the prediction intervals, and introduce several methods to estimate the forecast error such as bootstrapping, explicit density, and approximated variance approaches. At time T, a $100\,(1-\alpha)\,\%$ prediction interval for τ-periods-ahead demand, $X(T+\tau)$, provided by diffusion model k is given as:

$$\hat{X}_k(T+\tau|\Theta(T)) \pm k_{\alpha/2} \cdot \sigma_k,$$

where $k_{\alpha/2}$ is the random variable that describes the forecast error. The prediction intervals for the forecast obtained by a diffusion model are illustrated in Figure 13.

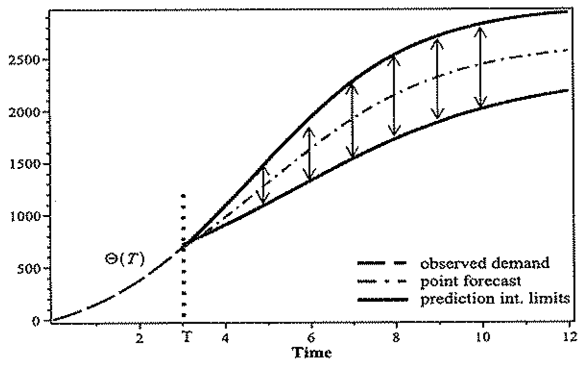


Figure 13: Prediction intervals

Since there is an uncertainty associated with future demand realizations (Figure 13), the forecast can take any shape between the prediction limits rather than having a single set of values. The supply level should account for all possible demand realizations between these limits.

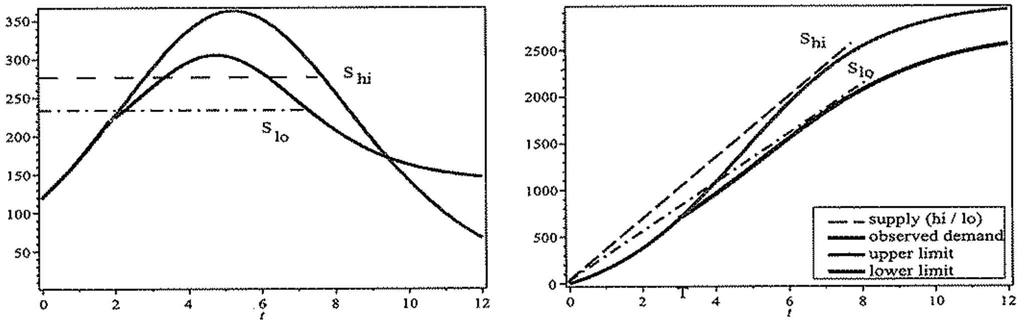


Figure 14: Demand realization and supply plan

In Figure 14, we illustrate the range of optimal supply levels corresponding to possible realizations of demand diffusion between the prediction limits. The lowest and highest possible supply levels are denoted by S_{lo} and S_{hi}, respectively. The planner's problem is to minimize the operational costs by determining the supply level Y, given the range of possible demand diffusion realizations.

We will use cumulative demand and supply figures to reduce the computational difficulties caused by noncumulative transcendental diffusion functions. Notice that for each demand scenario (a particular diffusion profile), there is a corresponding implied optimal supply rate, denoted by S^*. We assume that S^* is Uniformly distributed in (S_{lo}, S_{hi}), between the lowest and highest possible optimal supply levels, according to the prediction interval.

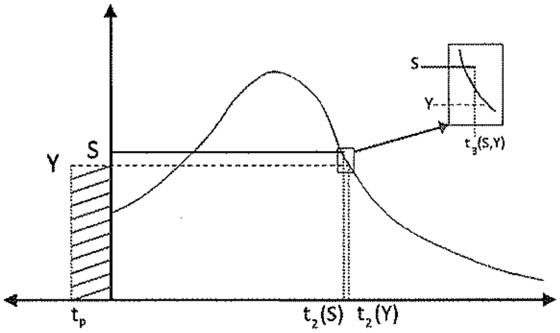


Figure 15: Supply shortage

Consequently, the planner should build a supply rate Y between (S_{lo}, S_{hi}). For any diffusion realization that requires a supply level lower than Y $(S < Y)$, the planner incurs an overage cost. This can be considered as the cost of underutilization of expensive equipment or opportunity cost. On the other hand, if diffusion realization requires a higher supply level than Y $(S > Y)$, part of the customer demand is backordered or lost; amount of which is equal to the preproduction quantity that would have avoided backorders (shaded area in Figure 15). The underage is penalized per unit demand not satisfied on time, exact quantity of which can be computed using the following formula:

$$Y \cdot (t_p + t_2(Y)) = S \cdot t_3(S, Y) + Y \cdot (t_2(Y) - t_3(S, Y))$$

where $t_3(S, Y)$ is the time point for optimal supply case that rate S should be reduced to rate Y for the inventory to be zero at $t_2(Y)$. Then, the backorder quantity is calculated as:

$$Y \cdot t_p = (S - Y) \cdot t_3(S, Y)$$

Next, we show that tighter prediction intervals that are provided by lower forecast variance lead to less operational costs.

Proposition 4. *Expected operational costs decrease as forecast variance decreases (i.e., the prediction intervals get tighter, $S_{hi} - S_{lo}$ decreases).*

Proof. The expected operational costs can be written as follows:

$$C_o \cdot \int_{S_{lo}}^{Y} (Y-S) \cdot \frac{1}{S_{hi}-S_{lo}} dS + C_u \cdot \int_{Y}^{S_{hi}} (S-Y) \cdot t_3(S,Y) \cdot \frac{1}{S_{hi}-S_{lo}} dS,$$

where C_o and C_u are overage and underage costs, respectively. The first term represents the cost of capacity that goes unused. The second term is the cost of backorders in case of a supply shortage. Although $t_3(S,Y)$ is variable with respect to the value of Y, the change in its value is insignificant inside the prediction intervals when the diffusion curve is steep. Therefore, approximation of this term as a constant should not have any implications on our analysis. Let $\bar{C}_u = C_u \cdot t_3(S,Y)$. Then, the expected total cost is:

$$\frac{C_o}{S_{hi}-S_{lo}} \cdot \left(Y^2/2 - S_{lo}Y + S_{lo}^2/2\right) + \frac{\bar{C}_u}{S_{hi}-S_{lo}} \cdot \left(S_{hi}^2/2 - S_{hi}Y + Y^2/2\right) \tag{1}$$

The optimal supply rate is found after differentiating the above term with respect to Y, than solving it for zero:

$$0 = (Y - S_{lo}) \frac{C_o}{S_{hi}-S_{lo}} + (Y - S_{hi}) \frac{\bar{C}_u}{S_{hi}-S_{lo}}$$

$$\Rightarrow Y^* = \frac{S_{lo}C_o + S_{hi}\bar{C}_u}{C_o + \bar{C}_u}$$

Plugging Y^* in equation 1, we get the expected total cost:

$$Cost = \frac{C_o \cdot \bar{C}_u \left(S_{hi} - S_{lo}\right)}{2 \cdot \left(C_o + \bar{C}_u\right)}$$

□

We may conclude that the expected total operating cost increases proportionally to the increase in the difference $S_{hi} - S_{lo}$. This difference decreases as forecast variance decreases and prediction intervals get tighter. To sum up, a decrease in prediction intervals via methods that reduce forecast variance translates into a direct decrease in expected operational costs. More importantly, as the costs C_o and $\bar{C}_u$ increase, the impact of forecast variance to operating cost will sharply increase.

5 Conclusions

In this chapter, we modeled the factory of a high-tech manufacturer to gain theoretical insights on the relationship between supply and demand. In particular, we investigated the optimal high-level operational strategies that are employed under diverse scenarios of capacity and demand realizations. Later, we used these strategies to illustrate the impact of forecasting activities on operational costs.

We derive the relationship between forecast variance and operational costs in closed form. We show that expected operational costs increase with the increase of forecast variance. In addition, the impact of forecast variance is amplified when supply overage and underage cost parameters are higher. Consequently, the variance reduction methods should be used in forecasting process since they lead to operational cost savings.

In our analysis, we considered a single product factory and simple operational strategies to achieve high-level tradeoffs between forecast accuracy and operational costs. However, in reality, high-tech manufacturing involves multiple products being manufactured using common resources, as well as simultaneous ramp-up and ramp-down of different product generations. In such environment, operational impact of one product's forecast affects operational decisions regarding other products that share the production resources. Understanding the relationship between systemwide operational costs and individual products' forecasts' accuracy is an intriguing research direction.

Although it is a significantly difficult task, considering a multi product factory model should lead to more realistic supply strategies with short term volatilities. Notice that in this case, the clearing function deals with a WIP consisting of multiple product types, each at a different point in their demand lifecycles. In addition, manufacturing lead time of a certain product type depends also on the amount of other product types in process. Investigating the demand-supply tradeoffs in such a more comprehensive setting should be a challenging yet promising research direction.

References

Asmundsson, J., R. L. Rardin, R. Uzsoy. 2006. Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. *IEEE Transactions on Semiconductor Manufacturing* **19**(1) 95–111.

Aytac, B., S. D. Wu. 2008. Characterization of demand for short-lifecycle technology products. Tech. rep., ISE Dept. Lehigh University, Bethlehem, PA.

Bass, F. M. 1969. A new product growth model for consumer durables. *Management Science* **15**(5) 215–227.

Graves, S. C. 1986. A tactical planning model for a job shop. *Operations Research* **34**(4) 522–533.

Ho, T. H., S. Savin, C. Terwiesch. 2002. Managing Demand and Sales Dynamics in New Product Diffusion Under Supply Constraint. *Management Science* **48**(2) 187–206.

Hopp, W. J., M. L. Spearman. 2001. *Factory Physics*. Irwin.

Hung, Y.-F., R. C. Leachman. 1996. A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. *IEEE Transactions on Semiconductor Manufacturing* **9**(2) 257–269.

Jain, D., V. Mahajan, E. Muller. 1991. Innovation diffusion in the presence of supply restrictions. *Marketing Science* **10**(1) 83–90.

Karmarkar, U. S. 1989. Capacity loading and release planning with work-in-progress (WIP) and leadtimes. *Journal of Manufacturing and Operations Management* **2** 105–123.

Kumar, S., J. M. Swaminathan. 2003. Diffusion of Innovations under Supply Constraints. *Operations Research* **51**(6) 866–879.

Kumar, U., V. Kumar. 1992. Technological innovation diffusion: The proliferation of substitution models and easing the user's dilemma. *IEEE T. Eng. Manage.* **39**(2) 158–168.

Mahajan, V., E. Muller, F. M. Bass. 2001. New-product diffusion models: A review and directions for research. *Journal of Marketing* **54** 1–26.

Meade, N., T. Islam. 1995. Prediction intervals for growth curve forecasts. *Journal of Forecasting* **14** 413–430.

Meade, N., T. Islam. 1998. Technological forecasting – Model selection, model stability, and combining models. *Management Science* **44**(8) 1115–1130.

Orcun, S., R. Uzsoy, K. Kempf. 2006. Using system dynamics simulations to compare capacity models for production planning. *WSC '06: Proceedings of the 38th conference on Winter simulation*. Winter Simulation Conference, 1855–1862.

Srinivasan, A., M. Carey, T. E. Morton. 1988. Resource pricing and aggregate scheduling in manufacturing systems. GSIA working papers, Carnegie Mellon University, Tepper School of Business.

Wu, S. D., B. Aytac, R. T. Berger, C. A. Armbruster. 2006. Managing Short-Lifecycle Technology Products for Agere Systems. *Interfaces* **36**(3) 234–247.

Wu, S. D., K. G. Kempf, M. O. Atan, B. Aytac, A. Mishra, S. A. Shirodkar. 2010. Extending bass for improved forecasting at Intel. To appear in *Interfaces*.