# Operational Efficiency

**Chia-Yen Lee[1] Andrew L. Johnson[2]**

[1]Institute of Manufacturing Information and Systems, National Cheng Kung University, Tainan City 701, Taiwan
[2]Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77840, USA

[1]E-mail: cylee1980@neo.tamu.edu
[2]E-mail: ajohnson@tamu.edu

## 1. Introduction

The fields of engineering and management associate *efficiency* with how well a relevant action is performed, i.e. ‘‘doing things right’’, and *effectiveness* with selecting the best action, i.e. ‘‘doing the right thing’’. Thus, a firm is *effective* if identifies appropriate strategic goals, and *efficient* if it achieves them with minimal resources. This chapter focuses on *operational efficiency,* or the ability to deliver products and services cost effectively without sacrificing quality. In this chapter we investigate a firm’s operational efficiency with both queueing models and productivity and efficiency analysis methods that identify maximum productivity and measure efficiency as a ratio of observed productivity to maximum productivity. The maximum productivity levels serves as a benchmark for desired perform. The methods for analysis will vary depending on the level of analysis. For example, at the micro-level, we measure operational efficiency at points (machine, workstation, laborer) on the shop floor, whereas the macro-level might be at the firm, industry, or nation level. We begin by evaluating performance at the operational level, and then apply productivity and efficiency analysis to aggregate performance at higher levels.

The analysis of productivity and efficiency is associated with production economics which focuses on assessment and uses an aggregate description of technology to answer questions such as (Hackman, 2008):

- How efficient is the firm in utilizing its input to produce its outputs?
- Is the firm using the right mix of inputs or producing the right mix of outputs given prevailing prices?
- How will the firm respond to a price hike in a critical input?
- How efficient is the firm in scaling its operations?
- Has the firm improved its productive capability over time?
- How does the firm compare to its competitors?

Figure 1 shows the three levels of production and operational planning and defines productivity and efficiency analysis (PEA) role. The *strategic level* includes long-term planning issues such as make-or-buy decisions. The *tactical level* describe midterm actions that are done perhaps on a weekly or month basis, while the *operational level* emphasizes daily scheduling and shop floor control. PEA supports tactical-level decisions and is part of mid-term production planning. PEA provides performance

benchmarking and production guidance. It can also provide ex post analysis to quantify efficiency for complex production processes that use multiple inputs to generate multiple outputs, or ex ante analysis to suggest guidelines for resource allocation.

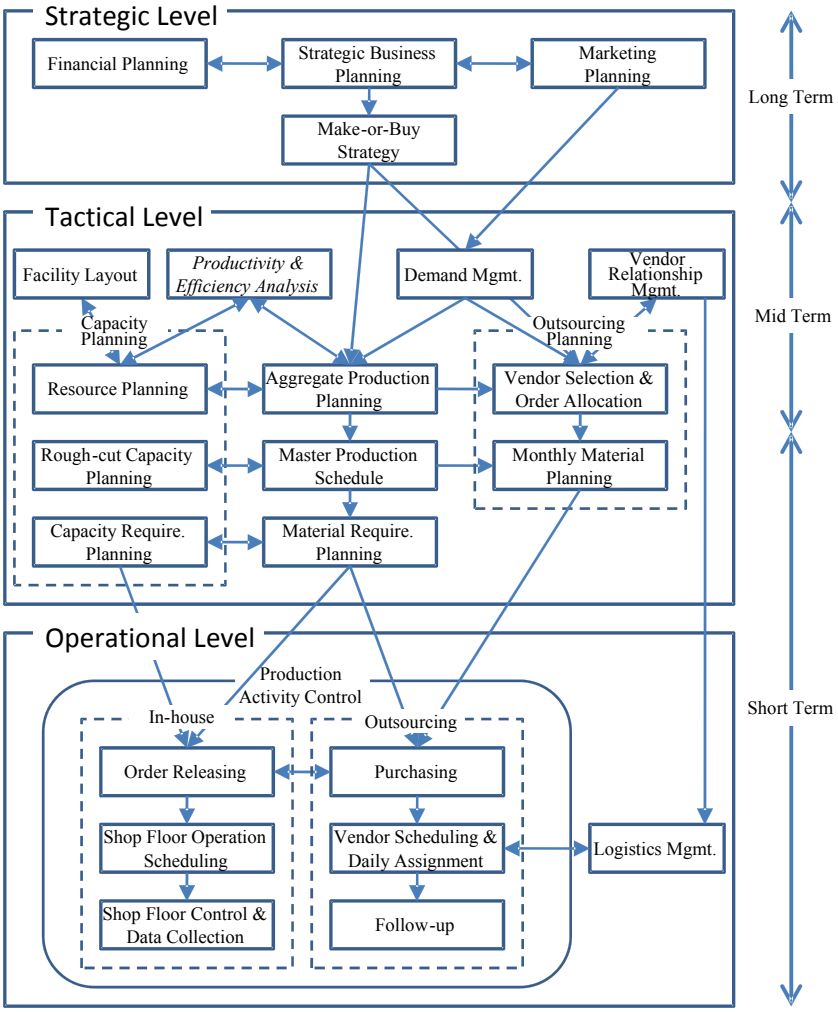



Figure 1 A General description of the analysis levels in production and operations planning

*Absolute Operational Efficiency*

*Ideal benchmarks* to measure efficiency are usually developed in a design laboratory under perfect operating conditions. However, it is not easy to identify the sources of efficiency loss between ideal performance and the best observed performance. For instance, in a manufacturing process operating in perfect conditions, one machine's ideal throughput is 100 units per hour, yet the actual throughput is 80 units per hour due to operator's skill, scheduling, etc. We can estimate an absolute operational efficiency (AOE) as:

$$AOE = \frac{actual\ throughput}{ideal\ throughput} = \frac{80}{100} = 0.8$$

Note ideal benchmarks can be observed at the machine or process level, but are almost never observed at the firm level. Thus alternative metrics are beneficial in the cases when ideal benchmarks are not observable.

*Relative Operational Efficiency*

Relative operational efficiency (ROE) is the ratio of actual throughput compared to best observed throughput. *Relative benchmarks* are often used to measure efficiency because similar comparable machine, process, firm, etc. are often easily identifiable. We estimate ROE by identifying the best observed performance in a data set of multiple operations performing the same task, for instance, a data set of multiple machines performing the same manufacturing process. We find that the best observed throughput is 90 units per hour, but machine A produces 80 units per hour. We can estimate the relative operational efficiency (ROE) of machine A as:

$$ROE = \frac{actual\ throughput}{best\ observed\ throughput} = \frac{80}{90} = 0.88$$

*Best observed throughput* is often determined by using historical performance data under the assumption, if all conditions are unchanged, actual throughput should be equal to/or close to the historically best performance.

In the real world, a firm's resources are always limited. When a firm would like to provide a product or service, it must consume input resources to generate the output level. In this setting, operational efficiency is determined by the outputs produced as well as the input resources or costs consumed. Thus, we can define productivity and efficiency as:

$$productivity = \frac{output}{input}$$

$$efficiency = \frac{productivity}{productivity\ of\ best\ pratice}$$

In other words, productivity is the ratio of output level to the input level and efficiency is the ratio of the current productivity level to the best practice productivity level. *Best practice* is defined as the largest productivity achievable.

The relationship between the output level produced as input levels change is the *production function.* Figure 2 shows an S-shaped production function with a single input and a single output. We say that Firm A is technically inefficient because, given the same input level, Firm B is able to produce more output than Firm A. We can also say that Firm B is efficient because, holding the input level fixed, it produces the highest possible output level. The concept of production function is explained in section 2.2.

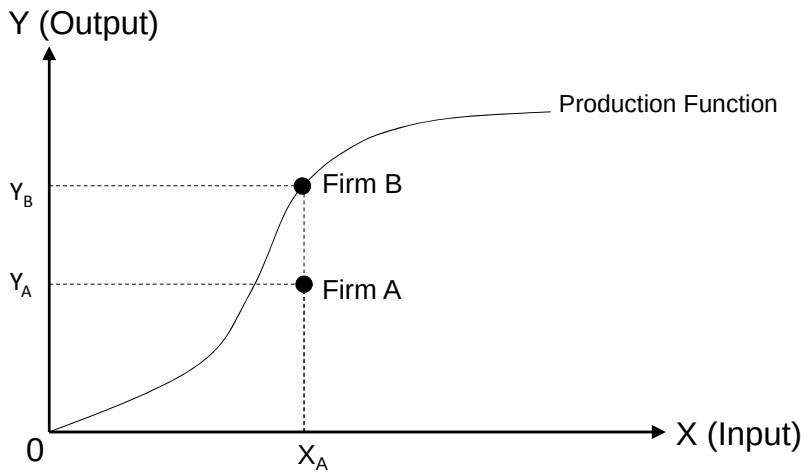



Figure 2 Production function and efficiency estimation

## 2. Efficiency Evaluation and Performance Indices

This section describes efficiency evaluation and related performance indices. Section 2.1 discusses how to evaluate efficiency by queueing theory in the shop-floor level. Section 2.2 discusses the use of a production function characterization of aggregate performance at the system or firm-level as the production process becomes larger and includes workers with uncertain behavior and longer time horizons. Section 2.3 introduces the three approaches, stochastic frontier analysis (SFA), data envelopment analysis (DEA), and Stochastic semi-Nonparametric Envelopment of Data (StoNED), to estimate technical (operational) efficiency by using the observed inputs and outputs levels of a set of firms to estimate a production function.

### *2.1 Shop-Floor Performance and Queueing Theory*

At the shop-floor level, queueing models provide a method for evaluating machine performance. In the model below, we use the notation M/M/1 to describe the inter-arrive process, the service process for a single-server queueing system. The first M indicates customer arrivals follow a Poisson (Markovian) Process and the inter-arrival time is exponential distribution. The second M indicates the service time follows an exponential distribution. The 1 indicates there is a single server.

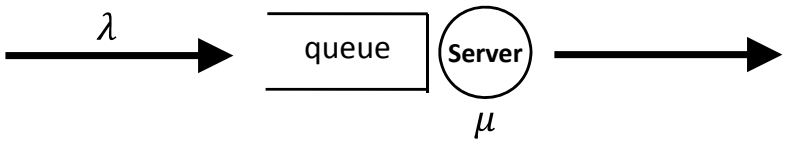



Figure 3 M/M/1 queue

We use two parameters to describe the M/M/1 queueing system. Let $\lambda$ be the arrival rate and $\mu$ be the service rate. For example, if $\lambda = 2.5$ customers per hour, it means on average 2.5 individuals arrive every hour. Thus, $1/\lambda$ is the mean inter-arrival time and $1/\mu$ is the mean service time. Figure 4 shows the Markov state-transition diagram.

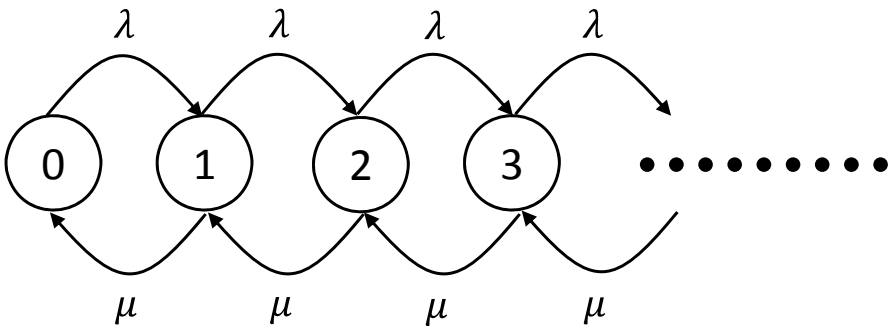



Figure 4 Markov-state transition diagram of M/M/1 queueing system

The condition $\lambda < \mu$ is necessary for the system to be stable, i.e. for the queue to be finite in length. $\rho = \lambda/\mu$ is the probability the server is busy and $p_0$ is the probability the server is idle. $p_i$ is the probability of the server with $i$ customers. We use the following set of algebraic equations to analyze the queue's performance.

In the beginning, we want to know the stable probability $p_0$. To characterize a queueing system with the state transition between 0 and 1 a rate-balance equation between arrival rate and service rate can be shown as

$$\lambda p_0 = \mu p_1 \rightarrow p_1 = \left(\frac{\lambda}{\mu}\right) p_0 = \rho p_0.$$

Intuitively, an empty system needs 1 arrival to become state 1; a system with 1 customer needs 1 departure to become state 0. This idea is the foundation of the rate-balance equation.

Similarly, we can derive the rate-balance equation for state 1 associated with state 0 and state 2.

$$(\lambda + \mu)p_1 = \lambda p_0 + \mu p_2 \rightarrow p_2 = (1 + \rho)p_1 - \rho p_0 = \rho^2 p_0$$

We can also derive a general formula, $p_n = \rho^n p_0$, for the probability that there are *n* customers in the system ($p_n$).

We obtain $p_0$ since the sum of all probability $p_n$ for $n = 1, \dots, \infty$ must be equal to 1:

$\sum_{n=0}^{\infty} p_n = p_0 \sum_{n=1}^{\infty} \rho^n = \frac{p_0}{1-\rho} = 1 \rightarrow p_0 = 1 - \rho.$

Thus, we derive the steady-state probability:

- $P[server\ idle] = p_0 = 1 - \rho$
- $P[server\ busy] = 1 - p_0 = \rho = \lambda/\mu$ (also called the "utilization")
- $P[n\ customers\ in\ the\ system] = p_n = \rho^n(1 - \rho)$

To derive the probability of $n$ or more customers in the system:

$\sum_{m=n}^{\infty} p_m = (1-\rho) \sum_{m=n}^{\infty} \rho^m = (1-\rho) \sum_{k=0}^{\infty} \rho^{n+k} = \rho^n (1-\rho) \sum_{k=0}^{\infty} \rho^k = \rho^n (1-\rho) \frac{1}{1-\rho} = \rho^n$

- $P[n\ or\ more\ customers\ in\ the\ system] = \rho^n$
- $P[less\ than\ n\ customers\ in\ the\ system] = 1 - \rho^n$

So far the probability distribution of steady state is derived for a single-server queueing system. We can construct two indices to evaluate the queueing system's performance by asking:

- What is the expected number of customers in the system/ in the queue?
- What is the expected time of a customer staying in the system/ in the queue?

Let *L* be the expected number of customers in the system.

$$L = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = \sum_{n=0}^{\infty} n(\rho^n - \rho^{n+1}) = 1(\rho^1 - \rho^2) + 2(\rho^2 - \rho^3) + 3(\rho^3 - \rho^4) + \cdots$$

$$= \rho + \rho^2 + \rho^3 + \rho^4 + \cdots = \rho(1 + \rho + \rho^2 + \rho^3 + \cdots) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

The expected number of customers in the queue, $L_q$, can be derived similarly. Note that we assume *customer being served is not in the queue,* so $n$ customers in the systems means the queue length is $n - 1$.

$L_q = \sum_{n=1}^{\infty} (n-1)p_n = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n = L - (1 - p_0) = \frac{\rho}{1-\rho} - \big(1 - (1-\rho)\big) = \frac{\rho}{1-\rho} - \rho$

$L_q = L - \rho = L\rho$

Let *W* be the expected time spent in the system by a customer. Intuitively, it is equal to the expected number of customers in the system divided by arrival rate $\lambda$. The equation is:

$$L = \lambda W$$

This equation, or Little's Law, defines the relationship between *L* and *W*. Similarly, $L_q = \lambda W_q$, where $W_q$ denotes the expected time spent in the queue by a customer:

$W = \frac{L}{\lambda} = \frac{L\rho}{\lambda\rho} = \frac{L_q/\lambda}{\rho} = \frac{W_q}{\rho} \rightarrow W_q = \rho W$

$$W = \frac{L}{\lambda} = \frac{L_q + \rho}{\lambda} = \frac{\mu L_q + \lambda}{\lambda \mu} = \frac{L_q}{\lambda} + \frac{1}{\mu} = W_q + \frac{1}{\mu}$$

The relationship between $W$ and $W_q$ results because the expected time spent in the system is equal to the expected time spent in the queue plus the mean service time.

Above Little's Law is defined for a general queueing system. In a manufacturing system, Little's Law is interpreted as the relationship among work-in-process (WIP), throughput (TH), and cycle time (CT):

$$WIP = TH \times CT$$

WIP is the number of unfinished units in the production system, TH is the number of finished products manufactured per unit of time and CT is the amount of time the units remain in the production system. Given a fixed WIP, an inverse relationship characterizes TH and CT, i.e. an increase in TH will decrease CT.

Little's Law is useful because it applies to a wide variety of production systems. Given a fixed TH, WIP and CT will maintain an almost linear relationship until the capacity limit is approached, but if WIP continue to increase, CT will deteriorate rapidly. Figure 5, an example of a workstation, shows that when utilization approaches 100%, the increase of arrival rate $\lambda$ will deteriorate WIP or CT. Thus, $\lambda > \mu$ implies the workstation is no longer stable.

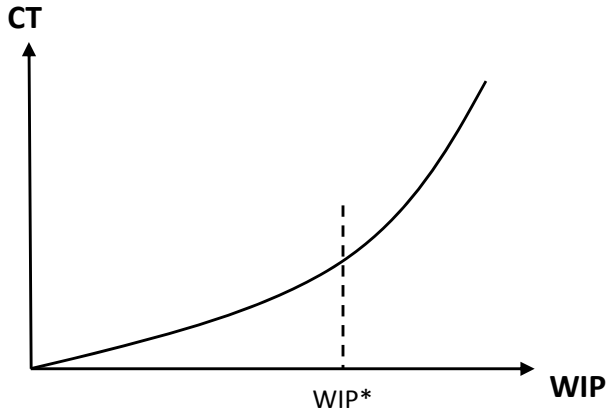



Figure 5 CT deterioration

The typical performance metrics for queueing systems are utilization and throughput. We calculate utilization as:

$$Utilization = \frac{\lambda}{\mu} = \frac{actual\ throughput}{theoretical\ (ideal)\ throughput}$$

Given CT and the level of WIP, we use Little's Law to calculate the M/M/1 system's productivity by dividing TH by 1. More complicated network analyses are possible with multiple processors linked in a network, Gautam (2012). Queueing theory can be used to calculate throughput, productivity can be

estimated by dividing throughput by the number of processors. However, all processors may not be identical and throughput will clearly be impacted by the underlying network structure. Further the human component of operating machines adds additional complications and uncertainty that are difficult to capture in queueing models. Thus production functions are useful for estimating complex systems or firm level performance.

### *2.2 Production Function*

A production function $f(\mathbb{x})$ is the *maximum outputs* that can be achieved using input vector $\mathbb{x} = (x_1, \ldots, x_N)$ (Hackman, 2008). *Outputs* are units a firm generates and *inputs* are the factors of production, or the commodities used in production. In economics, there are at least five types of factors of production: capital, labor, land, energy, and raw materials. We can analyze the performance of a firm's production system in using either the long-run production function or the short-run production function. In the short run the factors can be divided into fixed factors and variable factors. Fixed factors are the factors that cannot be changed in the short run such as building and land, and variable factors are the factors that can be changed in the short run such as temporary worker. In the long run all of the production factors are variable.

Theoretically, four properties characterize a production function (Chambers, 1988; Coelli, *et al.*, 2005):

- Nonnegativity: The production output is a finite, non-negative, real number.
- Weak Essentiality: The production output cannot be generated without the use of at least one input.
- Monotonicity: Additional units of an input will not decrease output; also called *nondecreasing* in $\mathbb{x}$.
- Concavity: Any linear combination of the vectors $\mathbb{x}^0$ and $\mathbb{x}^1$ will produce an output that is no less than the same linear combination of $f(\mathbb{x}^0)$ and $f(\mathbb{x}^1)$. That is, $f(\lambda\mathbb{x}^0 + (1-\lambda)\mathbb{x}^1) \geq \lambda f(\mathbb{x}^0) + (1-\lambda)f(\mathbb{x}^1)$. This property implies the "law of diminishing marginal returns".

These properties can be relaxed to model specific production behaviors. For example, monotonicity is relaxed to model input congestion (Färe *et al.*, 1985; 1994) [1] and concavity is relaxed to characterize an S-shaped production function (Frisch, 1964 and Henderson and Quandt, 1980).

#### *2.2.1 Short-Run Production Function*

Because of the fixed factors in the short run, the production function is characterized by monotonicly increasing levels and diminishing returns, that is, increasing one variable factor of production will increase output levels at a decreasing rate while holding all others constant. The fixed factors limit the growth of the output. This is also called *the law of diminishing marginal returns* (*product*).

[1] Input congestion indicates that the output level may decrease even though we increase more input due to a difficulty of management and organization.

Three concepts of production characterize a short-run production function:

- Total product ($TP$): the total amount of output generated from the production system, $TP = y = f(x)$.
- Average product ($AP$): the average amount of output per unit input, $AP = \frac{f(x)}{x}$.
- Marginal product ($MP$): the marginal change while adding one more unit of input, $MP = \frac{df(x)}{dx}$.

Figure 6 illustrates a single-input and single-output production function when all other factors are fixed. As the firm increases its input levels, the output levels also increase. The firm reaches point A, an inflection point, i.e. where the maximal marginal product is achieved. As inputs continue to increase, the single-input and single-output production function shows diminishing marginal product as it reaches the *Most Productive Scale Size* (MPSS). MPSS is the point on the production function that maximizes the average product (or productivity). Finally, input and output levels continue to increase until point B, beyond which input congestion occurs due to the fixed factors and negative marginal product.

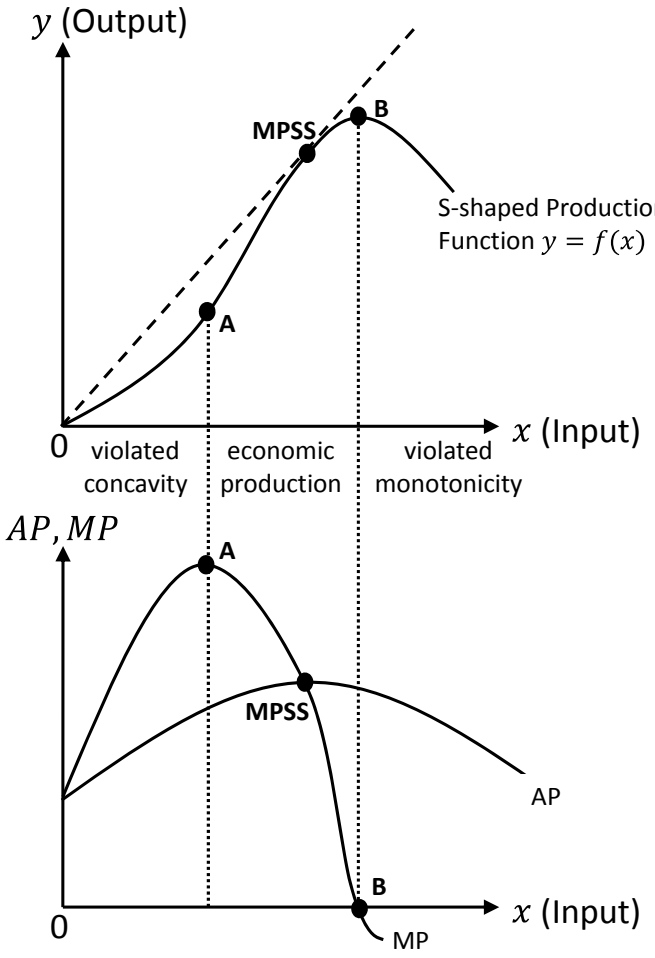



Figure 6 Single-input and single-output production function

*2.2.2 Long-Run Production Function*

All of the factors of production are variable in the long-run. Consider production using multiple-inputs. It is common practice to plot the relationship between two of the variables while holding all other fixed. Figure 7 shows the relationship between the inputs $x_n$ and $x_m$ while holding the output fixed at the value $y^0$ and holding all other inputs fixed. The resulting curve is the *input isoquant*, that give all combinations of $x_n$ and $x_m$ capable of producing the same output level $y^0$. It is convex towards the origin if it satisfies all properties of the production function. For different output levels $y^2 > y^1 > y^0$, these isoquants form non-intersecting functions. The slopes of the isoquants are the *marginal rate of technical substitution* ($MRTS$) which measures the rate of using $x_n$ to substitute $x_m$ while holding the output level constant:

$$MRTS_{nm} = -\frac{\partial x_m(x_1, \dots, x_{m-1}, x_{m+1}, \dots, x_N)}{\partial x_n} = \frac{MP_n}{MP_m}$$

$$MP_n\, \partial x_n + MP_m \partial x_m = 0$$

where $x_m(x_1, \dots, x_{m-1}, x_{m+1}, \dots, x_N)$ is an implicit function indicating how much $x_m$ is needed to produce the same output level given fixed levels of $x_1, \dots, x_{m-1}, x_{m+1}, \dots, x_N$. Thus, the rate of substitution of input *m* for input *n* along the isoquant is equal to the ratio of the marginal productivity of *n* relative to the marginal productivity of *m*. To remove the unit of measurement, the *direct elasticity of substitution* ($DES$) is the percentage change in the input ratio relative to the percentage change in the *MRTS,* and quantifies the *curvature* of the isoquant:

$$DES_{nm} = \frac{d(x_m/x_n)}{d(MP_n/MP_m)} \times \frac{MP_n/MP_m}{x_m/x_n}$$

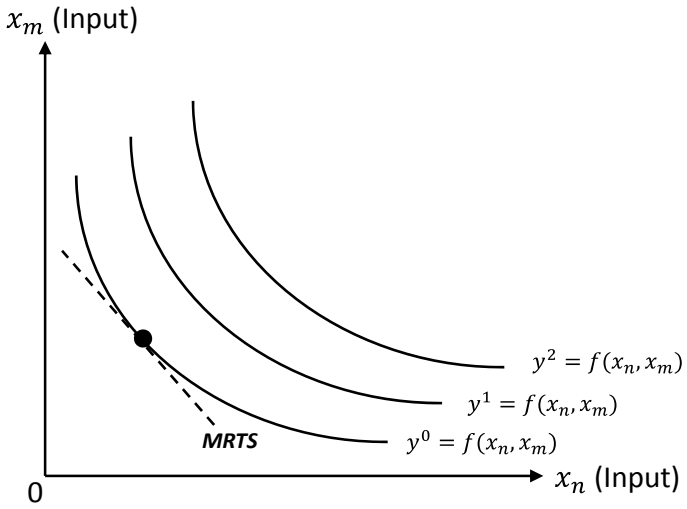



Figure 7 Input isoquants

Next, we describe three typical production functions for a two-input case.

*Leontief production function*

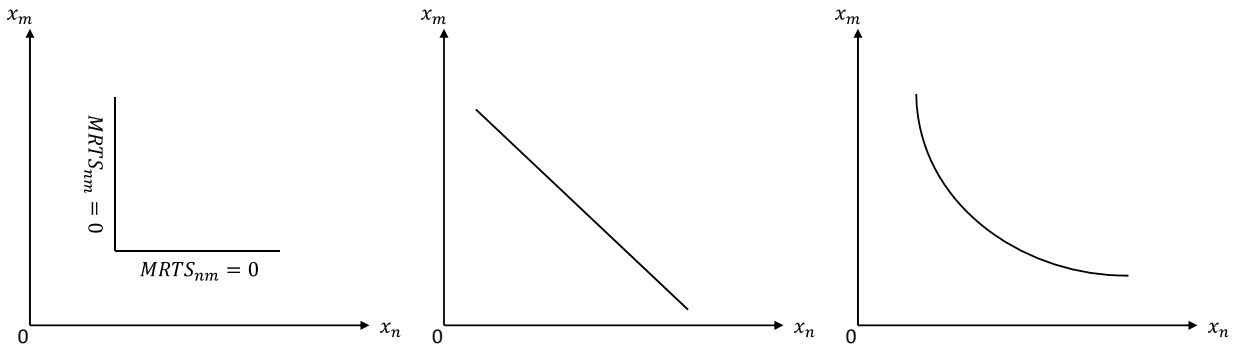
*Leontief production functions* or fixed proportions functions describe production that occurs in fixed proportions, for example, cars that require wheels ($x_n$) and bodies ($x_m$). The mathematical form is $y = \min\{\beta_n x_n, \beta_m x_m\}$ and $\beta_n, \beta_m > 0$; figure 7(a) shows how the horizontal part of the isoquant indicates that an increase in $x_n$ does not contribute to the output ($y$), and $MP_n = 0$ and $MRTS_{nm} = 0$, and that the vertical part of the isoquant indicates that an increase in $x_m$ does not contribute to the output ($y$), and $MP_m = 0$ and $MRTS_{nm} = \infty$. $MRTS_{nm}$ is not defined at the corner. Therefore, a Leontief production function is used to model production where there is no substitution between $x_n$ and $x_m$, i.e. $DES_{nm} = 0$.

*Linear production function*

A linear production function assumes that inputs are substituted at a constant rate regardless of the level of either input or output. The mathematical form is $y = \beta_n x_n + \beta_m x_m$ and $\beta_n, \beta_m > 0$; figure 7(b) shows that the production function implies a constant rate of substitution, $MRTS_{nm} = \frac{MP_n}{MP_m} = \frac{\beta_n}{\beta_m}$, and also imposes perfect substitution between $x_n$ and $x_m$, i.e. $DES_{nm} = \infty$.

*Cobb-Douglas production function*

A Cobb-Douglas production function assumes that inputs are substitutable. However, consistent with the law of diminishing marginal productivity, additional inputs are needed to maintain the same output level as the mix of inputs becomes more skewed. The mathematical form is $y = \alpha x_n^{\beta_n} x_m^{\beta_m}$ and $\alpha, \beta_n, \beta_m > 0$; Figure 7 shows that the production function is a smooth curve and convex towards the origin, and that $MRTS_{nm} = \frac{MP_n}{MP_m} = \frac{\alpha\beta_n x_n^{\beta_n - 1} x_m^{\beta_m}}{\alpha\beta_m x_n^{\beta_n} x_m^{\beta_m - 1}} = \frac{\beta_n x_m}{\beta_m x_n}$ decreases with respect to $x_n$. Thus, substitution exists in this production function and $0 < DES_{nm} < \infty$.




(a) (b) (c)

Figure 8 Production function for a two-input case

*Properties of production function*

Figure 8 shows that the production functions are convex towards the origin, because the absolute value of the slope of the isoquant decreases while increasing $x_n$, thus, $MRTS_{nm}$ also decreases. This is called the *law of diminishing marginal rate of technical substitution*. The mathematical representation is $\frac{\partial}{\partial x_n} MRTS_{nm} < 0$.

In addition, if a proportionate increase in all inputs results in a less than proportionate increase in output, we say that the production function exhibits *decreasing returns to scale* (DRS). Alternatively, if increasing all inputs results in the same proportional increase in output, we say that it exhibits *constant returns to scale* (CRS). Finally, if the increase of all inputs results in a more than a proportionate increase in output, we say that the production function exhibits *increasing returns to scale* (IRS). Table 2 shows a mathematical illustration of these three properties where $\lambda > 1$.

Table 2 Returns to scale

| Return to Scale | Mathematical Formulation |
|---|---|
| Decreasing Returns to Scale (DRS) | $f(\lambda \mathbb{x}) < \lambda f(\mathbb{x})$ |
| Constant Returns to Scale (CRS) | $f(\lambda \mathbb{x}) = \lambda f(\mathbb{x})$ |
| Increasing Returns to Scale (IRS) | $f(\lambda \mathbb{x}) > \lambda f(\mathbb{x})$ |

There are many reasons why firms may exhibit different returns to scale. For example, a firm may exhibit IRS if hiring more personnel allows specialization of labor, but the firm may eventually exhibit DRS if the firm becomes so large that management is no longer able to control operations. Firms that can replicate all aspects of their operations exhibit CRS. Operating at decreasing returns to scale would indicate decentralization or downsizing might be appropriate whereas operating at increasing returns to scale would indicate mergers, acquisitions, or other changes in organizational structure might be appropriate.

*2.3 Firm-Level Performance and Efficiency Estimation*

We construct the production function to define a benchmark to measure how efficiently production processes use inputs to generate outputs. Given the same level of input resources, inefficiency is indicated by lower levels of output. In a competitive market, if a firm is far from the production function and operates inefficiently, it needs to increase its productivity to avoid the going out of business.

Production theory provides a useful framework to estimate the production function and efficiency levels of a firm in three ways: 1) using parametric functional forms in regression based methods, e.g., SFA (Aigner et al., 1977; Meeusen and van den Broeck, 1977), 2) using nonparametric linear programming methods, e.g., DEA (Charnes et al., 1978; Banker et al., 1984) or 3) integrating regression and programming methods, e.g., StoNED (Kuosmanen and Kortelainen, 2011; Kuosmanen and Johnson, 2010). In this section we describe how to use the three methods to estimate efficiency based on cross-sectional data for $K$ firms.

*2.3.1 Stochastic Frontier Analysis*

Aigner and Chu (1968) use the logarithmic form of the Cobb-Douglas production function to estimate a deterministic frontier:

$$\ln y_k = \boldsymbol{x}_k'\boldsymbol{\beta} - u_k$$

where $k = 1, \dots, K$ and $y_k$ indicates the single output of the firm $k$; $\boldsymbol{x}_k$ is a $I \times 1$ vector with the elements of logarithm inputs; $\boldsymbol{\beta}$ is a vector of unknown parameters; and $u_k$ is a nonnegative random variable associated with technical inefficiency. Several methods can be used to estimate the parameter $\boldsymbol{\beta}$, such as maximum likelihood estimation (MLE) or ordinary least squares (OLS) (Richmond, 1974). However, the Aigner and Chu method neglects statistical noise and assumes that all deviations from the frontier are a result of technical inefficiency. Therefore Aigner et al. (1977) and Meeusen and van den Broeck (1977) proposed the stochastic frontier production function and introduced the random variable representing statistical noise as:.

$$\ln y_k = \boldsymbol{x}_k'\boldsymbol{\beta} + v_k - u_k$$

where $v_k$ models the statistical noise using a symmetric random error. The function is bounded from above due to the stochastic variable $\exp(\boldsymbol{x}_k'\boldsymbol{\beta} + v_k)$. To illustrate, we use a Cobb-Douglas stochastic frontier model with single input variable:

$$\ln y_k = \beta_0 + \beta_1 \ln x_k + v_k - u_k$$

$$y_k = \exp(\beta_0 + \beta_1 \ln x_k) \times \exp(v_k) \times \exp(-u_k)$$

In this functional form, $\exp(\beta_0 + \beta_1 \ln x_k)$ is the deterministic component, $\exp(v_k)$ is the statistical noise, and $\exp(-u_k)$ is the inefficiency component. Figure 9 illustrates the deterministic frontier $y_k = \exp(\beta_0 + \beta_1 \ln x_k)$, the noise effect and the inefficiency effect of firm A and firm B. Firm A has a negative random noise component, whereas firm B has a positive noise random noise component. The *observed output* level $y_k = \exp(\beta_0 + \beta_1 \ln x_k + v_k - u_k)$ and the *frontier output* level (i.e. without the inefficiency effect) is $y_k^* = \exp(\beta_0 + \beta_1 \ln x_k + v_k)$. The observed output of firm B lies below the deterministic part of the frontier, because the sum of the noise and inefficiency is negative.

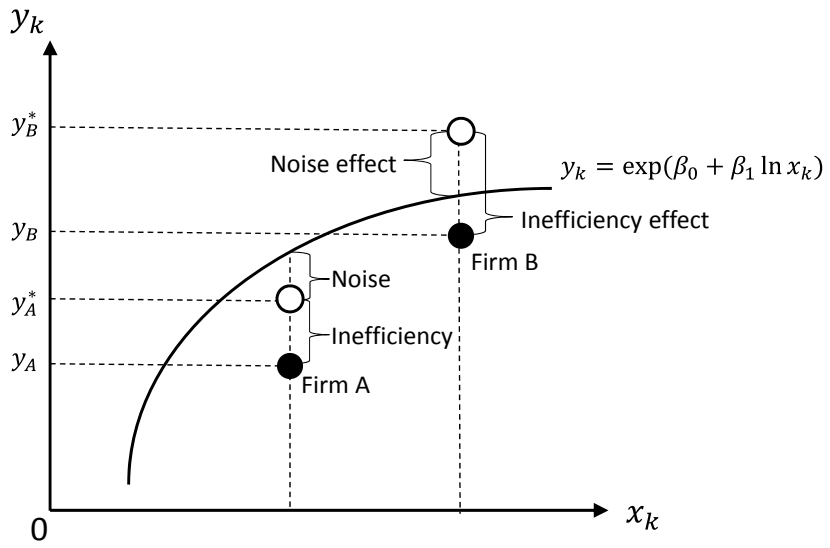



Figure 9 An example of the stochastic frontier analysis estimate of the production function

Since SFA estimates the inefficiency effects, we can define the output-oriented measure of technical efficiency ($TE$) by using the observed output over the frontier output:

$$TE_k = \frac{y_k}{\exp(\boldsymbol{x}_k'\boldsymbol{\beta} + v_k)} = \frac{\exp(\boldsymbol{x}_k'\boldsymbol{\beta} + v_k - u_k)}{\exp(\boldsymbol{x}_k'\boldsymbol{\beta} + v_k)} = \exp(-u_k)$$

This $TE_k$ estimate shows the measure of observed output of firm $k$ relative to the frontier output of an efficient firm given the same input vector. This benchmarking with best practice provides the estimation of technical inefficiency.

We need to estimate the parameter vector $\boldsymbol{\beta}$ before calculating $TE$. Note that the model is complicated by the two random terms, $v_i$ and $u_i$, where $v_i$ is usually a symmetric error and $u_i$ is a nonnegative term. The parameter $\boldsymbol{\beta}$ is estimated under the following assumptions:

- $E(v_k u_l) = 0, \forall k, l$: uncorrelated
- $E(v_k) = 0$: zero mean
- $E(v_k^2) = \sigma_v^2$: homoskedastic
- $E(v_k v_l) = 0, \forall k \neq l$: uncorrelated
- $E(u_k^2) = \text{constant}$: homoskedastic
- $E(u_k u_l) = 0, \forall k \neq l$: uncorrelated

Further, $v_k$ and $u_k$ are uncorrelated with the explanatory variables $\boldsymbol{x}_k$. Note that $E(u_k) \neq 0$ since $u_k \geq 0$.

To estimate $\boldsymbol{\beta}$, Aigner et al. (1977) assume $v_k \sim N(0, \sigma_v^2)$ and $u_k \sim N^+(0, \sigma_u^2)$, where $v_k$ follows the independently and identically distributed ($iid$) normal distribution with zero mean and variance $\sigma_v^2$, and $u_k$ follows the $iid$ half-normal distribution which is a truncated normal distribution with zero mean and variance $\sigma_u^2$. This is called the "half-normal model" in SFA. Under these assumptions, the OLS estimator will provide consistent estimators of slope in $\boldsymbol{\beta}$ but a downward-biased intercept coefficient since

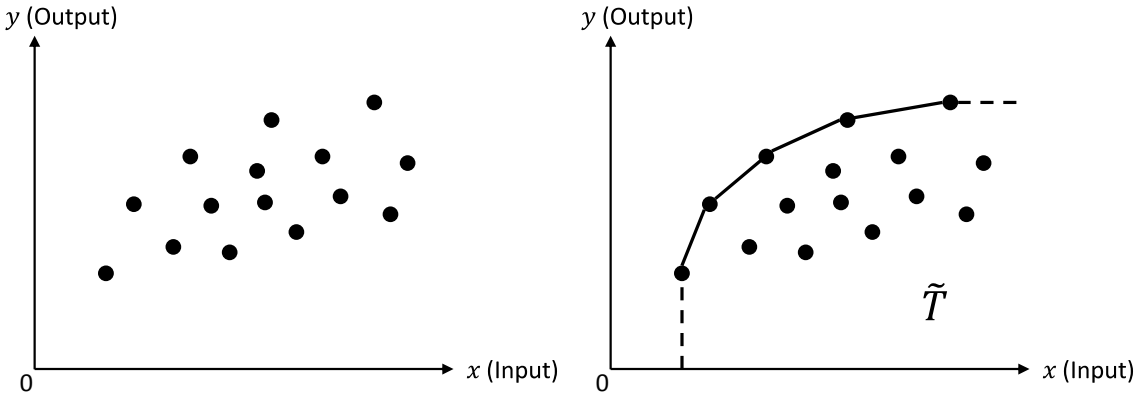
$E(u_k) \neq 0$. Therefore, we use the maximum likelihood estimator (MLE) on the log-likelihood function with $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\xi^2 = \sigma_u^2/\sigma_v^2$:

$$\ln L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma, \lambda) = -\frac{K}{2}\ln\left(\frac{\pi\sigma^2}{2}\right) + \sum_{k=1}^{K}\ln\Phi\left(-\frac{\varepsilon_k\xi}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\varepsilon_k^2$$

where $\boldsymbol{y}$ is a vector of log-outputs, $v_k - u_k = \ln y_k - \boldsymbol{x}_k'\boldsymbol{\beta}$ defines a composite error term $\varepsilon_k$ i and $\Phi$ is a cumulative distribution function of the stand normal random variable. Finally, we use the iterative optimization procedure to estimate the coefficient $\boldsymbol{\beta}$ (Judge et al., 1985).

### *2.3.2 Data Envelopment Analysis*

DEA is an optimization-based approach that imposes the axiomatic assumptions of monotonicity and convexity and the minimum extrapolation principle (MEP) (Banker et al., 1984). MEP identifies the smallest set that satisfies the imposed production assumptions and envelops all the data. Thus, DEA estimates a piecewise linear production function based on the observed data points.




(a) (b)

Figure 10 DEA frontier with 15 observations

Figure 10(a) illustrates 15 production observations and Figure 10(b) illustrates the DEA frontier. The dashed line segment of the DEA frontier represents the strong disposability hull (SDH). That is, the firm on the SDH can decrease the input level without reducing the output level or decrease the output level without changing the input level. We measure the slack in inputs or outputs along the dashed line segments distinguishing Farrell efficiency measure (Debreu, 1951; Farrell, 1957) and Koopmans efficiency measure (Koopmans, 1951). The Farrell measure defines technical efficiency as the maximum radial reduction in all inputs consistent with equivalent production of output. The Koopmans measure states that it is impossible for a firm to increase any output without simultaneously reducing another output (or increasing any input). Note that after all inputs have been radially reduced, additional slack may still exist in some but not all inputs. Thus, a Farrell efficient firm may not be Koopmans efficient.

In this section, we focus on the widely used Farrell measure. First, we introduce the linear programming technique to estimate the production function and production possibility set. Let $x \in R_+^I$ denote the inputs

and $y \in R_+^J$ denote the outputs of the production system. We define the production possibility set as $T \equiv \{(x,y): x \text{ can produce } y\}$. $X_{ik}$ is the $i^{th}$ input resource, $Y_{jk}$ is the amount of the $j^{th}$ production output, and $\lambda_k$ is the multiplier for the $k^{th}$ firm. The following model defines the feasible region of the production possibility set $\tilde{T}$. This is called the Variable Return to Scale (VRS) DEA model (Banker et al., 1984), because decreasing marginal product is observed along the frontier:

$$\tilde{T} = \{(x,y): \sum_k \lambda_k Y_{jk} \geq Y_j, \forall j; \sum_k \lambda_k X_{ik} \leq X_i, \forall i; \sum_k \lambda_k = 1; \lambda_k \geq 0, \forall k\}$$

We use the DEA estimator to measure the efficiency. We describe the input-oriented technical efficiency ($ITE$) as measured using the distance function $D_x(x,y) = \inf\{\theta | (\theta x, y) \in \tilde{T}\}$.

*Input-oriented DEA efficiency model*

$$\min_\theta \{\theta | \sum_k \lambda_k Y_{jk} \geq Y_j, \forall j; \sum_k \lambda_k X_{ik} \leq \theta X_i, \forall i; \sum_k \lambda_k = 1; \lambda_k \geq 0, \forall k\}$$

*Output-oriented DEA efficiency model*

$$\max_\omega \{\omega | \sum_k \lambda_k Y_{jk} \geq Y_j \omega, \forall j; \sum_k \lambda_k X_{ik} \leq X_i, \forall i; \sum_k \lambda_k = 1; \lambda_k \geq 0, \forall k\}$$

We calculate $\theta = 1/\omega$ from the output-oriented DEA efficiency model $i$ to get an output-oriented technical efficiency ($OTE$), $\theta$. $\theta = 1$ implies an efficient firm and $\theta < 1$ implies an inefficient firm.

Figure 11 illustrates the input-oriented efficiency measure. Three firms, A, B and C are located in an input space constructed by holding the output level constant at $y = \bar{y}$. The solid line is the piecewise linear efficient frontier estimated by DEA. Firms B and C are located on the frontier, but firm A is on the interior of the estimated PPS, $\tilde{T}$. Using the Farrell measure to estimate the technical efficiency shows that the inputs of firm A can be reduced radially. Point D is the intersection of the line segments $\overline{OA}$ and $\overline{BC}$. In fact, point D is a convex combination of firms B and C. We estimate Firm A's technical efficiency as:

$$TE_A = \theta = D_x(x_A, y_A) = \frac{\overline{OD}}{\overline{OA}}$$

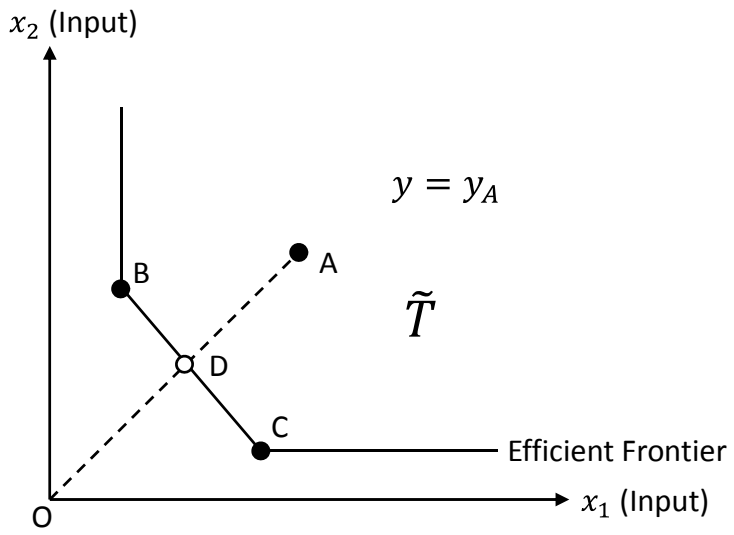



Figure 11 Efficiency estimation relative to a DEA Input Isoquant

*2.3.3 Stochastic semi-Nonparametric Envelopment of Data*

The benefits of both SFA and DEA can be achieved using the nonparametric regression approach, StoNED. The first stage of StoNED uses Convex Nonparametric Least Squares (CNLS) proposed by Hildreth (1954) and extended by Hanson and Pledger (1976) to estimate a function satisfying continuity, monotonicity and global concavity – the standard regularity conditions for a production function. To include both random noise and technical inefficiency, Kuosmanen and Kortelainen (2011) combine the CNLS piecewise linear production function with the composite disturbance term concept from SFA.

Let $\boldsymbol{x}_k \in R_+^I$ be an input vector, $y_k \in R_+$ be an output and $f$ be an unknown frontier production function satisfying continuity, monotonicity and concavity. The regression model is:

$$y_k = f(\boldsymbol{x}_k) + \varepsilon_k \qquad \forall k = 1, \dots, K$$

where $\varepsilon_k$ is a disturbance term with $E(\varepsilon_k) = 0\ \forall k$, $\mathrm{Var}(\varepsilon_k) = \sigma^2 < \infty\ \forall i$ and $\mathrm{Cov}(\varepsilon_k \varepsilon_j) = 0\ \forall k \neq j$. We formulate the CNLS problem as the quadratic program:

$$\min_{\alpha, \boldsymbol{\beta}, \varepsilon} \sum_k \varepsilon_k^2$$

s.t. $\quad \varepsilon_k = y_k - (\alpha_k + \boldsymbol{x}_k' \boldsymbol{\beta}_k) \qquad \forall k = 1, \dots, K$

$$\alpha_k + \boldsymbol{x}_k' \boldsymbol{\beta} \leq \alpha_h + \boldsymbol{x}_k' \boldsymbol{\beta}_h \qquad \forall h, k = 1, \dots, K$$

$$\boldsymbol{\beta}_k \geq 0 \qquad \forall k = 1, \dots, K$$

where $\alpha_k$ and $\boldsymbol{\beta}_k$ are the coefficients characterizing the hyperplanes of the frontier production function $f$. Note that $\alpha_k$ and $\boldsymbol{\beta}_k$ are specific to each firm $k$. The objective function minimizes the sum of squared disturbance terms. The equality constraint defines the disturbance term as the difference between an

observed output and an estimated output. The inequality constraints comprise a system of Afriat inequalities (Afriat, 1972), imposing the underlying frontier production function to be continuous and concave. The last constraints enforce monotonicity. Unlike DEA, CNLS uses all of the data points to estimate a production function, making it more robust to outliers.

The CNLS estimator of the production function, $\hat{f}(\boldsymbol{x})$, is generally not unique, but the fitted output values at observed inputs, $\hat{f}(\boldsymbol{x}_k)$, are unique (Kuosmanen, 2008). In fact, given the fitted output values, it is possible to derive the tightest lower bound of the frontier production function as the explicit lower bound representor function:

$$\hat{f}_{\min}(\boldsymbol{x}) = \min_{\alpha,\boldsymbol{\beta}} \{\alpha + \boldsymbol{x}_k'\boldsymbol{\beta} | \alpha + \boldsymbol{x}_k'\boldsymbol{\beta} \geq \hat{y}_k \quad \forall k = 1, \dots, K\}$$

where $\hat{y}_k = \hat{f}(\boldsymbol{x}_k)$ is the fitted output value. Since the tightest lower bound $\hat{f}_{\min}$ is a piecewise linear function satisfying continuity, monotonicity and concavity, we can use it as the unique CNLS estimator of the frontier production function $f$.

StoNED uses a similar approach to SFA for modeling inefficiency and noise terms. Consider the composite disturbance term:

$$\varepsilon_k = v_k - u_k \qquad \forall k = 1, \dots, K$$

where the same properties for $v_k$ and $u_k$ are assumed to as in the SFA section.

The composite disturbance term violates the Gauss-Markov property that $E(\varepsilon_k) = E(-u_k) = -\mu < 0$; therefore, we modify the composite disturbance term as:

$$y_k = [f(\boldsymbol{x}_k) - \mu] + [\varepsilon_k + \mu] = g(\boldsymbol{x}_k) + \vartheta_k \qquad \forall k = 1, \dots, K$$

where $\vartheta_k = \varepsilon_k + \mu$ is a modified composite disturbance with $E(\vartheta_k) = E(\varepsilon_k + \mu) = 0$ and $g(\boldsymbol{x}_k) = f(\boldsymbol{x}_k) - \mu$ is an average production function. Since $g$ inherits the continuity, monotonicity and concavity, the CNLS method can find the estimator of the average production function $g$. We formulate the composite disturbance CNLS problem as:

$$\min_{\alpha,\boldsymbol{\beta},\vartheta} \sum_k \vartheta_k^2$$

s.t. $\quad \vartheta_k = y_k - (\alpha_k + \boldsymbol{x}_k'\boldsymbol{\beta}_k) \qquad \forall k = 1, \dots, K$

$$\alpha_k + \boldsymbol{x}_k'\boldsymbol{\beta}_k \leq \alpha_h + \boldsymbol{x}_k'\boldsymbol{\beta}_h \qquad \forall k, h = 1, \dots, K$$

$$\boldsymbol{\beta}_k \geq 0 \qquad \forall k = 1, \dots, K.$$

where $\alpha_k$ and $\boldsymbol{\beta}_k$ are the coefficients that characterize the hyperplanes of the average frontier production function $g$. Note that the composite disturbance CNLS problem only differs from the CNLS problem the sum of squared modified composite disturbances is minimized.

To illustrate the StoNED estimator, 100 observations of a single-input single-output Cobb-Douglas production function are generated, $y = x^{0.6} + v - u$. The observations, $x$, were randomly sampled from a Uniform [1,10] distribution, $v$ was drawn from a normal distribution with standard deviation of 0.5, and $u$ was drawn from a half-normal distribution with standard deviation of 0.7. Figure 12 shows the obtained StoNED estimator.

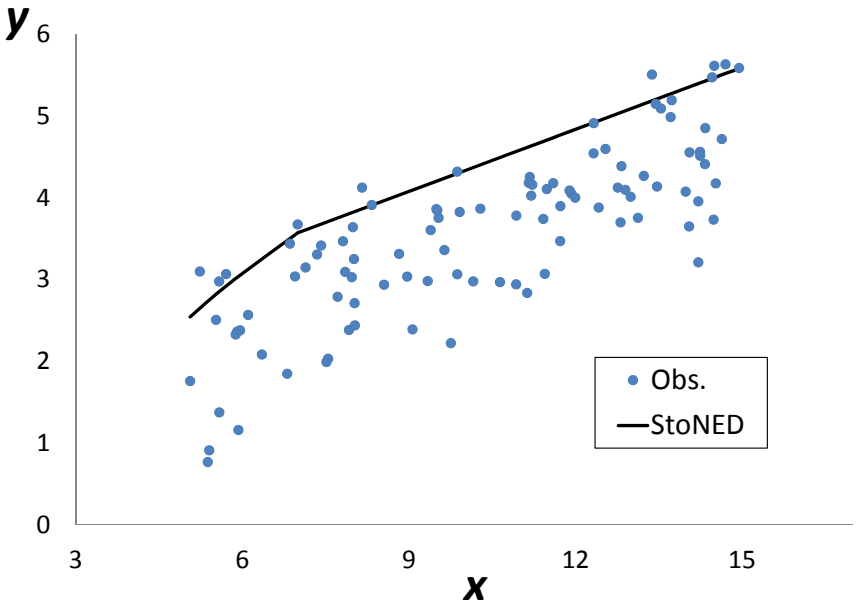



Figure 12 The StoNED frontier with 100 observations

The second stage of StoNED uses the modified composite residuals, $\hat{\vartheta}_k \; \forall k$, to separate the technical inefficiencies and random noises by applying the method of moments (Aigner et al., 1977; Kuosmanen and Kortelainen, 2011). Assuming that technical inefficiency has a half normal distribution, $u_k \sim |N(0, \sigma_u^2)|$, and that random noise has a normal distribution, $v_k \sim N(0, \sigma_v^2)$, the estimated standard deviation of technical inefficiency and random noise is:

$$\hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{\left(\frac{2}{\pi}\right)\left(1 - \frac{4}{\pi}\right)}}$$

$$\hat{\sigma}_v = \sqrt{\hat{M}_2 - \left(\frac{\pi - 2}{\pi}\right)\hat{\sigma}_u^2}$$

where $\hat{M}_2 = \frac{1}{n}\sum_k \left(\hat{\vartheta}_k - \hat{E}(\vartheta_k)\right)^2$ and $\hat{M}_3 = \frac{1}{n}\sum_k \left(\hat{\vartheta}_k - \hat{E}(\vartheta_k)\right)^3$ are the second and third sample central moments of the modified composite residuals. Moreover, $\hat{M}_3$ should be negative so that $\hat{\sigma}_u$ is positive. Intuitively, the composite residuals should have negative skewness reflecting the presence of the technical inefficiency. We calculate the expected technical inefficiency by:

$$\hat{\mu} = \hat{\sigma}_u\sqrt{2/\pi}.$$

Given $(\hat{\alpha}_k, \widehat{\boldsymbol{\beta}}_k)$ from the CNLS problem, we write the unique StoNED estimator of the frontier production function as:

$$\hat{f}_{\min}(\boldsymbol{x}) = \min_{\alpha,\boldsymbol{\beta}}\{\alpha + \boldsymbol{x}_k'\boldsymbol{\beta} | \alpha + \boldsymbol{x}_k'\boldsymbol{\beta} \geq \hat{y}_k \quad \forall k = 1, \dots, K\} + \hat{\mu}$$

where $\hat{y}_k = \min_{h\epsilon\{1,\dots,n\}}\{\hat{\alpha}_h + \boldsymbol{x}_k'\widehat{\boldsymbol{\beta}}_h\}$. We obtain the unique CNLS estimator of the average frontier production function, $\hat{g}_{\min}$, by using the tightest lower bound representor function with the fitted output values, $\hat{y}_k$. Recall that $\hat{y}_k$ is calculated from the representor function and $(\hat{\alpha}_k, \widehat{\boldsymbol{\beta}}_k)$. Therefore, we obtain the frontier production function by additively shifting the unique CNLS estimator of the average frontier production function upwards by the expected value of technical inefficiency.

Given $\hat{\sigma}_u$ and $\hat{\sigma}_v$, the method introduced in Jondrow et al. (1982) can estimate firm-specific inefficiency. Specifically:

$$\hat{E}(u_k|\hat{\varepsilon}_k) = -\frac{\hat{\varepsilon}_k\hat{\sigma}_u{}^2}{\hat{\sigma}_u{}^2 + \hat{\sigma}_v{}^2} + \frac{\hat{\sigma}_u{}^2\hat{\sigma}_v{}^2}{\hat{\sigma}_u{}^2 + \hat{\sigma}_v{}^2}\left[\frac{\phi(\hat{\varepsilon}_k/\hat{\sigma}_v{}^2)}{1 - \Phi(\hat{\varepsilon}_k/\hat{\sigma}_v{}^2)}\right]$$

where $\hat{\varepsilon}_k = \hat{\vartheta}_k - \hat{\mu}$, $\phi$ is the standard normal density function and $\Phi$ is the standard normal cumulative distribution.

## 3. Efficiency Improvement

Section 2 provided models to estimate the system performance and efficiency. This section provides some methodologies for driving productivity. Section 3.1 introduces overall equipment effectiveness (OEE) and Section 3.2 describes lean manufacturing.

### 3.1 *Overall Equipment Effectiveness*

OEE is a time-based metric to assess productivity and efficiency, particularly for the semiconductor manufacturing industry (Ames et al., 1995; SEMI, 2000, 2001; de Ron and Rooda, 2005). The traditional single index metrics of productivity, throughput, and utilization do not allow easy identification of root cause for reduced productivity. The OEE definition describes six standard equipment states:

- *Nonscheduled state*: Equipment is not scheduled to be used in production, such as unworked shifts, weekends, or holidays (including startup and shutdown).
- *Unscheduled down state*: Equipment is not in a condition to perform its intended function due to unplanned downtime events, e.g., maintenance delay, repair, change of consumables or chemicals, and out-of-spec input.
- *Scheduled down state*: Equipment is not available to perform its intended function due to planned downtime events, e.g., production test, preventive maintenance, and setup.

- *Engineering state*: Equipment is in a condition to perform its intended function but is operated to conduct engineering experiments, e.g., process engineering, equipment engineering, and software engineering.
- *Standby state*: Equipment is in a condition to perform its intended function but is not operated; the standby state includes no operator available (including breaks, lunches, and meetings), no items available (including no items due to lack of available support equipment), and no support tools.
- *Productive state*: Equipment is performing its intended functions, e.g., regular production (including loading and unloading of units), work for third parties, rework, and engineering runs done in conjunction with production units.

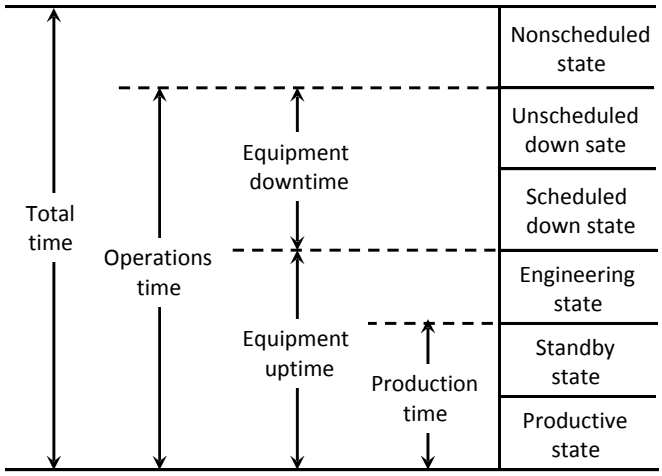



Figure 13 OEE and equipment states

We define OEE as:

$$OEE = \frac{theoretical\ production\ time\ for\ effective\ units}{total\ time}$$

We decompose OEE into the following subcomponents: availability efficiency (AE), operational efficiency (OE), rate efficiency (RE), and quality efficiency (QE) (de Ron and Rooda, 2005):

$$OEE = AE \cdot (OE \cdot RE) \cdot QE = Availability \cdot Performance \cdot Quality$$

where

- $Availability = AE = \frac{equipment\ uptime}{total\ time}$
- $Performance = OE \cdot RE$
- $OE = \frac{production\ time}{equipment\ uptime}$
- $RE = \frac{theoretical\ production\ time for\ actual\ units}{production\ time}$

- $Quality = QE = \frac{theoretical\ production\ time\ for\ effective\ units}{theoretical\ production\ time\ for\ actual\ units}$

The availability captures the difference between machine breakdown and processing. Performance characterizes the production time and throughput. The quality is described by the yield metric which is typically driven by scrap, rework, defects, and reject types. In other words, OEE is a metric to estimate the efficiency of theoretical production time for effective units. In particular, the theoretical production time means the production time without efficiency losses. In addition, two popular indices can be integrated into the OEE framework: *mean time between failure (MTBF),* or the average time a machine operates before it fails, and *mean time to repair (MTTR),* or the average time required to repair a failed component and return the machine to operation:

$$AE = \frac{equipment\ uptime}{total\ time} = \frac{equipment\ uptime}{equipment\ uptime \times \frac{(MTTR + MTBF)}{MTBF}} = \frac{MTBF}{MTTR + MTBF}$$

OEE has two practical benefits. First, we can use its subcomponents to identify bottlenecks and improve productivity. In general, machines with high utilization are typically the bottlenecks. Because bottlenecks can shift depending on the product mix, it is important for engineers to identify and release bottlenecks quickly to maintain high throughput levels. Note that the utilization is a necessary condition for bottleneck identification, but does not mean that all high-utilization machines are bottlenecks. If the processing time of each product is the same and the variation in the production line is low, a machine may have high utilization without affecting throughput. Second, we can use OEE to separate machine's status into regular operating conditions and down. The availability level quantifies the time used for production. A lower throughput is sometimes the result of low availability rather than poor performance. Thus, OEE decomposition helps with machine diagnosis and productivity improvement.

### 3.2 *Lean Thinking and Manufacturing*

*Lean manufacturing* has its roots in the manufacturing processes developed by Henry Ford in the 1920s. The Ford Motor Company increased its revenue during the post-World War I depression by developing assembly line methods and eliminating activities that were either unnecessary or did not add value to the cars produced. Toyota coined the name and the concept of lean manufacturing in its production system in the 1980s, and also developed additional supporting methods and concepts such as the Just-in-Time (JIT) system (Ohno 1988a, 1988b). We call a production system "lean" if it produces the required output levels with minimal buffering costs.

In fact, the only time a machine adds value is when it processes a part. Figure 14 provides a Gantt chart to visualize processing time, transportation time and wait time. Note that loading products into tools is handling, not processing, and thus a non-value adding activity. Most of the processing time of a product is waiting and non-value-adding activities. Smith (1998) proposed a manufacturing performance index called manufacturing cycle efficiency:

$$\text{Manufacturing Cycle Efficiency} = \frac{\text{Value} - \text{adding time}}{\text{Total cycle time}}$$

He pointed out that this index is often less than 1 percent in practice, meaning that firms usually waste resources performing non-value-adding activities.

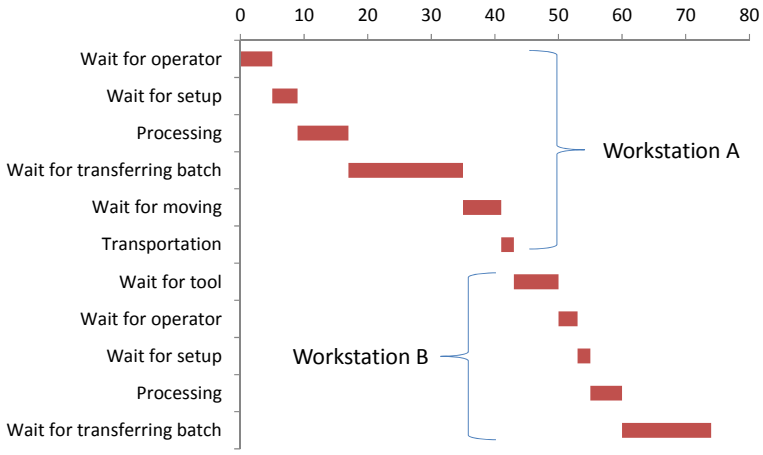



Figure 14 Gantt chart of product transition

The basic philosophy of lean manufacturing is to eliminate the waste by *buy(ing) only enough material to fit the immediate needs of the production plan considering the transportation resource*.

Below, we describe the three main principles of lean manufacturing:

1. Waste elimination
2. Continuous flow
3. Pull production system

*Implementation and Benefits*

As the term implies, waste elimination reduces all forms of waste in the manufacturing process. Continuous flow smoothes and balances the production flow. Pull production system, or "make-to-order production" allows a firm to produce units only when it receives an order. There are four steps to implementing lean manufacturing:

1. Eliminate waste: 7 types of waste are identified and need to be eliminated.
2. Use buffers: build up, adjust and swap buffers to manage for variability.
3. Continuous improvement: a commitment to productivity improvement
4. Reduce variability: identify and reduce internal and external causes

A firm can allocate resources dynamically and switch buffers to manage for internal or external variability. Internal variability results from uncertain processing times, setups, machine breakdown, yield loss, rework, engineering change orders, etc., and external variability results from demand fluctuation, customer change orders, supplier uncertain delivery, etc. Lean manufacturing uses three buffers: inventory, capacity, and time. Inventory hedges against uncertain demand. Capacity is somewhat flexible

due to hiring/layoffs of temporary workers, adjusting overtime, or outsourcing some activities. Time coordinates supply chain or manufacturing activities.

The benefits of lean manufacturing include:

- Productivity improvement
- Total manufacturing time saved
- Less scrap
- Lower inventory
- Quality improvement
- Plant space saved
- Better labor utilization
- Lower production cost, higher profits and wages
- Shorter cycle time: make-to-order vs make-to-stock
- Safety of operations

### 3.2.1 Waste Elimination

Womack and Jones (2003) describe seven types of "muda", or waste, in a production system:

1. Transportation: move products or materials that are not being processed between workstations, or between supplier and customer.
2. Inventory: hold excess inventory of raw materials, WIP, or finished units.
3. Motion: worker or tools move more than necessary, such as picking, sorting, stacking or storing parts.
4. Waiting: wait for upcoming tools, materials, parts or for the next production step.
5. Overproduction: generate excess products beyond the demand level.
6. Overprocessing: working more than is necessary because of poor tool or product design
7. Defects: cost of poor quality such as rework, scrap, inspection and repair.

In general, all seven types of waste described above belong to the category of non-value-adding activities. Table 3 lists some of the tools Toyota developed to eliminate waste.

Table 3 Tools developed by Toyota to eliminate waste

| Tool | Description |
|---|---|
| Flexible manufacturing | A flexible production system allows quick response to change, in particular, change in product mix. Machine flexibility allows the operator to change the configuration to produce different product types. Routing flexibility allows multiple machines to perform the same function on a product. |
| Standardize work | Standardize regular operations according to the benchmarking of best practice; post at workstations. |
| 5S | • Seiri (Sort), or "Tidiness": Throw away unrelated materials; only leave necessary items at workstation.<br>• Seiton (Set-in-order), or "Orderliness": Put everything in its place for quick pick-up and storage.<br>• Seiso (Shine), or "Cleanliness": Clean up the workplace.<br>• Seiketsu (Standardize): Hold the gains and maintain cleanliness.<br>• Shitsuke (Sustain), or "Discipline": Commitment to practice 5S for ongoing improvement. |
| Automation (Jidoka) | A supervisory function uses automation instruments to detect abnormalities and identify root causes; if an error arises, the production line shuts down immediately; to prevent the defective product and overproduction. |
| Others | Continuous improvement (Kaizen); Error proofing (Poka-yoke); Radical change (Kaikaku); Worker suggestions (Teien systems); Dynamic allocation of workers (Shojinka); etc. |

3.2.2 Continuous Flow

Continuous flow, or the series of continuous and smooth processes, is the second principle. Each production step performs only the jobs necessary for the next step. Workstations do not hold unnecessary WIP and materials that block incoming and downstream flows. Table 4 lists some tools to achieve continuous flow.

Table 4 Tools for continuous flow

| Tool | Description |
|---|---|
| Single-Minute Exchange of Die, SMED (Shingo) | Rapid changeover and setup time reduction in converting current manufacturing process to manufacture the next product; improves production flow and reduces lot sizes. |
| Andon | Uses signboard or visual signals to indicate the location of the alert for abnormality detection. |
| Takt time | Identifies the allowable time for process steps; calculated by taking available production time over customer demand; used to reduce the gap between current CT and the minimum possible time. |

| | |
|---|---|
| Line balancing | Organize tasks into groups, with each group of tasks being performed at a single workstation; each workstation has identical loading and CT; No workstation is overburdened, no one waits, and the variation is smoothed at each workstation. |
| Nagara (smooth production flow) | Shortens the lead time between manufacturing processes and reduces WIP inventories to adjust for fluctuations in demand; batch size reduction is a way to reduce inventory and smooth production flow. |
| Others | Cross-train workers to manage for inherent variability, etc. |

*Single Minute Exchange of Die* (*SMED*)

Single Minute Exchange of Die (SMED), or "Shingo", can significantly reduce setup time and improve productivity. Long setup time leads to a small number of setups, larger batch sizes, larger WIP inventories and poor process flow. SMED divides the setup time into internal and external activities. An internal activity is one that can only be done when the machine is stopped such as multi-chambers adjustments; an external activity is anything that can be performed before or after the setup without stopping the machine, such as pre-heating of raw material. To achieve a quick setup and changeover of dies, SMED recommends reducing internal setup time or converting internal activities to external activities.

*Production Line Balancing*

Line balancing is a typical problem of the assembly system design in industrial engineering (Nof et al., 1997). To compensate for demand fluctuations, the goal is to organize tasks into different groups with each group taking the same amount of time.

The line balancing problem is an NP-hard problem (Garey and Johnson, 1979); thus, heuristic methods are usually applied to provide good solutions. Helgeson and Birnie (1961) proposed a heuristic method called the *ranked positional weight technique*. This heuristic is a task-oriented technique considering the combination of precedence relationships and task processing time. Three steps are applied in this algorithm.

1. Calculate the positional weight (PW) of each task using the processing time (PT) of the task plus the processing time of all tasks having this task as a predecessor.
2. Rank tasks in descending order in terms of PW.
3. Assigns tasks to workstations sequentially in the ranked order, given the precedence relationships and CT constraint.

Figure 15 shows 8 tasks with their PT (unit: minute) and the precedence relationships. If the CT is 10 minutes for each workstation, we calculate the minimal number of workstations according to the sum of the 8 task times over the CT, that is, 38/10=3.8 and round up to 4. However, this minimum number does not consider the precedence constraints. Thus, we use the ranked PW technique for line balancing as shown in Table 5. We find that the required number of workstation is 5 and the total idle time is 12, both of which tend to increase at downstream stations.

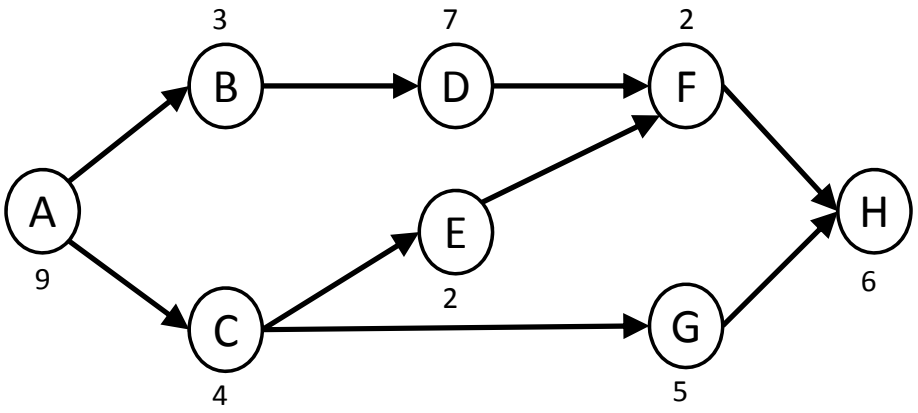



Figure 15 Precedence relationships and processing time

Table 5 Ranked positional weight technique

| | PT | PW | Order | Station |
|---|---|---|---|---|
| A | 9 | 40 | 1 | 1 |
| B | 3 | 18 | 3 | 2 |
| C | 4 | 19 | 2 | 2 |
| D | 7 | 15 | 4 | 3 |
| E | 2 | 10 | 6 | 2 |
| F | 2 | 8 | 7 | 3 |
| G | 5 | 11 | 5 | 4 |
| H | 6 | 6 | 8 | 5 |

The smoothing can be done by product type or by volume; both are quite efficient and can bring substantial efficiencies and savings. Note that a smoothed and continuous flow can be reviewed from a firm's internal production or its supply chain. The benefits include:

- Enhance flexibility by reducing batch size to accommodate to changes in product mix or demand fluctuation.
- Reduce material, WIP and inventory levels since there is no severe over or under production.
- No bottlenecks because of similar burden for each workstation.
- Enhance loyalty and commitment to the firm, i.e. a stable workforce without temporary labor.
- Shorten changeover and setup times to reduce machine idleness.

### 3.2.3 Pull Production System

*Push systems* release work without consideration of system status and hence do not have an inherent limitation on WIP. The work is released based on a schedule of demand and controlled release rates,

typically referred to as a due-date-driven production system. A *pull system* developed by Toyota releases work based on the status of the systems and has an inherent WIP limitation. The system authorizes work releases based on system status and controls WIP level. It is an order-driven production system (Hopp and Spearman, 2004).

There are two techniques in the customer-pull production system: just-in-time and kanban. Just-in-time attempts to reduce inventory, holding costs, and WIP using small lot size or even single unit processing. A "kanban" is a signboard or visual for realizing just-in-time and often leads to significant quality improvement. The advantages of using pull production system include:

- Reduce WIP and CT: limit releases into the production line.
- Improve quality: short queues allow errors to be identified quickly and shut down the production line to correct the problems.
- Reduce cost: switch the control from release rate to WIP level and reduce WIP progressively.
- Logistical benefits: less congestion, easier control, and WIP cap control.
- Kanban provides for efficient lot tracking and predetermines WIP level by the number of kanban.

In fact, based on Little's Law, $WIP = CT \times TH$, given the same rate of throughput, reducing the WIP level will lead to a reduction in CT. Thus, a pull production system reduces CT by controlling the WIP level. For further study of the pull system, see Ohno (1998a; 1998b), Liker (2004), and Nahmias (2009).

## 4. Conclusion

Operational efficiency can be measured and improved using the approaches described in this chapter. Today, many manufacturing firms define a metric for efficiency and concentrate on operational improvement activities to increase it. The specific approaches developed to identify best practice performance or to determine if a particular activity adds value are often product or industry specific. However, the evolution of new – and global – industries will require more sophisticated efficiency analysis techniques and metrics.

## References

Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review* 13 (3): 568-598.

Aigner, D. J. and S. F. Chu (1968). On estimating the industry production function. *American Economic Review* 58: 826-839.

Aigner, D., C. A. K. Lovell, and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6: 21-37.

Banker, R. D., A. Charnes, and W. W. Cooper (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30(9): 1078-1092.

Chambers, R. G. (1988). *Applied Production Analysis: A Dual Approach*. New York: Cambridge University Press.

Charnes, A., W. W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units, *European Journal of Operational Research* 2(6): 429-444.

Coelli, T. J., D. S. Prasada Rao, C. J. O'Donnell, and G. E. Battese (2005). *An Introduction to Efficiency and Productivity Analysis*. (2nd ed.). New York: Springer.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica* 19: 273-292.

Färe, R. S., S. Grosskopf, and C. A. K. Lovell (1985). Technical Efficiency of Philippine Agriculture. *Applied Economics* 17: 205-214.

Färe, R. S., S. Grosskopf, and C. A. K. Lovell (1994). *Production Frontiers*. Cambridge: Cambridge University Press.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, Series A, 120(3), 253-281.

Frisch, R. (1964). *Theory of Production*. Chicago: Rand McNally & Company.

Garey, M. R. and D. S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Gautam, N. (2012). Analysis of Queues: Methods and Applications. Boca Raton: CRC Press (Taylor and Francis).

Gililland, J., J. Konopka, K. Barber, R. Schnabl, and V.A. Ames (1995). Semiconductor manufacturing productivity: Overall equipment effectiveness (OEE) guidelines. Technology transfer 950327443 A-GEN, Revision 1.0. Sematech.

Hackman, S. T. (2008). *Production Economics: Integrating the Microeconomic and Engineering Perspectives*. Heidelberg: Springer-Verlag.

Hanson, D. L. and G. Pledger (1976). Consistency in concave regression. *Annals of Statistics* 4(6): 1038-1050.

Helgeson, W. P. and D. P. Birnie (1961). Assembly line balancing using the ranked positional weight technique. *Journal of Industrial Engineering* 12: 394-398.

Henderson, J. M., and R. E. Quandt (1980). *Microeconomic Theory: A Mathematical Approach*. (3rd ed.). New York: McGraw-Hill.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association* 49(267): 598-619.

Hopp, W. J. and M. L. Spearman (2004). To pull or not to pull: What is the question? *Manufacturing and Service Operations Management*, 6(2): 133-148.

Jondrow, J., C. A. K. Lovell, I. S. Materov, and P. Schmidt (1982). On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. *Journal of Econometrics* 19(2-3): 233-238.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T.-C. Lee (1985). *Introduction to the Theory and Practice of Econometrics*. New York: John Wiley & Sons, Inc.

Koopmans, T. (1951). An analysis of production as an efficient combination of activities. In: Koopmans, T.C. (Ed.), Activity Analysis of Production and Allocation. Cowles Commission for Research in Economics, Monograph No. 13. New York: John Wiley & Sons, Inc.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econometrics Journal* 11: 308-325.

Kuosmanen, T. and A. L. Johnson (2010). Data envelopment analysis as nonparametric least squares regression. *Operations Research* 58(1): 149-160.

Kuosmanen, T. and M. Kortelainen (2012). Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis*, in press.

Liker, J. K. (2004). *The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer*. New York: McGraw-Hill.

Meeusen, W. and J. van den Broeck (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18(2): 435-444.

Nahmias, S. (2009). *Production and Operations Analysis*. (6th ed.). New York: McGraw-Hill.

Nof, S. Y., W. E. Wilhelm, and H.-J. Warnecke (1997). *Industrial Assembly*. Chapman & Hall.

Ohno, T. (1988a). *Toyota Production System: Beyond Large-Scale Production*. Productivity Press Inc.

Ohno, T. (1988b). *Just-In-Time for Today and Tomorrow*. Productivity Press Inc.

Richmond, J. (1974). Estimating the efficiency of production. *International Economic Review* 15: 515-521.

de Ron, A. J. and J. E. Rooda (2005). Equipment effectiveness: OEE revisited. *IEEE Transactions on Semiconductor Manufacturing*, 18(1): 190-196.

Smith, W. (1998). *Time Out: Using Visible Pull Systems to Drive Process Improvements*. New York: John Wiley & Sons.

Standard for Definition and Measurement of Equipment Productivity, Semiconductor Equipment and Material International (SEMI) E79-0200, 2000.

Standard for Definition and Measurement of Equipment Reliability, Availability, and Maintainability, SEMI E10-0701, 2001.

Womack, J. P. and D. T. Jones (2003). *Lean thinking: Banish waste and create wealth in your corporation*. (2nd ed.), New York: Free Press Simon & Schuster.