CHAPTER 2
LITERATURE REVIEW

This research builds bridge between many disparate concepts and methodologies to examine the productive efficiency problem in a new light. The purpose of this chapter is to discuss the concepts that will be combined and serve as the foundation for the methodological contribution of this research. It also will discuss the fundamental definitions and concepts that will be used throughout this research.

Section 2.1 provides the key definitions and the basis for selecting the dynamic modeling approach. Section 2.2 provides a brief review of SD and a high level overview of SD optimization. This high level overview of SD optimization is designed to provide the reader with a historical and conceptual perspective that will be expanded in chapter 3 where the dynamic productive efficiency models are developed. Section 2.3 addresses the theory of productive efficiency. These are the concepts that will be expanded upon in Chapter 3, and serve as the core of the productive efficiency expansion. While this chapter represents a seemingly hodgepodge array of concepts, it represents the point of departure from which this research originates.

2.1 Key Definitions and Model Selection

2.1.1 Key Definitions

Blanchard and Fabrycky (1990, p. 1) define a *system* as "an assemblage or combination of elements or parts forming a complex or unitary whole." The objective of the system is to convert selected inputs to the system, to outputs of the system.

"The term *dynamic* refers to the phenomena that produces time-changing patterns, the characteristics of the pattern at one time being interrelated with those at other times" (Luenberger, 1979). Samuelson (1947, p. 314) defines a system as being dynamical "if its behavior over time is determined by functional equations in which 'variables at different points of time' are involved in an 'essential' way." Samuelson (1947) classifies dynamical systems in two distinct ways: (1) dynamic and historical; and (2) dynamic and causal. During the course of this research, a third type of dynamical system (dynamic, causal, and closed) became apparent. Since the dynamic, causal, and closed

systems are a natural continuation of Samuelson's dynamical systems, a discussion of all three will be deferred until Chapter 3.

The term *complexity* may be defined in many different ways. For the purposes of this research, complexity will be defined as two or more components joined together in such a way that it is difficult to separate them. Complexity can be further characterized as combinatorial or dynamic. Combinatorial complexity is found in systems in which the complexity lies in finding the best solution out (i.e. the best configuration of elements or components) of an astronomical numbers of possible combinations (Sterman, 2000). This type of problem is generally solved with a linear programming approach and usually ignores the time factor.

Dynamic complexity is found in systems where the interaction of the various elements over time is complex. Systems, which exhibit dynamic complexity, do not necessarily exhibit combinatorial complexity and vice-versa. Dynamic complexity problems are usually solved by some dynamic methodology (e.g. continuous time models (Forrester, 1961 and 1968; Richardson and Pugh, 1981; Sterman, 2000), discrete event models (Law and Kelton, 2000)). Sterman (2000) suggest ten reasons why dynamic complexity arises in systems.

1. Systems are dynamic – Systems in a real world environment are subject to evolution as time passes.
2. System elements are tightly coupled - Elements within a system interact with other elements within the system and elements within the environment.
3. Non-linear – Real world systems are seldom linear from. While elements in close proximity to each other may adequately be defined by linear relationships, the effects of these elements are often non-linear when examined downstream in the system after a number of factors interact to yield the current state.
4. Self-organizing – The dynamics of the system often arise spontaneously from the system's internal structure. Small disturbances to the system may be insignificant at the time and point of entry, but may yield larger consequences to the system if amplified or combined with other factors, within the system as time passes.

5. Systems are governed by feedback loops – Elements are tightly coupled and eventually feedback into themselves. Decisions lead to actions, which change the results of a system. Information results from these changes, which leads to further decisions. This process only evolves over time, thus is dynamic.
6. History dependent - As a system evolves over time, decisions are made, and the system becomes path dependent. Many of the actions that were taken are irreversible. Hence returning to the previous decision point is almost impossible in most cases.
7. Adaptive- If returning to the decision point is possible without consequences, the same decision most likely would not be made again as the passage of time would have provided data which allows the system to learn from previous experience.
8. Counterintuitive – Effects of a disturbance (i.e. a change to the system usually initiated by a decision) within a system are often unrealized until the effect is separated by space and time. However, attention is often drawn to elements in close proximity with respect to space and time. As a result, attention is often focused on the symptom rather than the underlying cause.
9. Policy resistant – Since considerable space and time often separate unanticipated system consequences, the original policy that caused the consequence is often unrealized. This often makes system policies difficult to understand.
10. Characterized by trade-offs – Time delays with a system often disguise the ultimate outcome of a system. Policy changes in the short term may prove to be problematic, while in the long term may prove to be the best option.

This research addresses combinatorial and dynamic complexity within the same system structure.

The term *causality* refers to the cause-and-effect relationships that exist between components within a system. This principle is critical to system analysis because it identifies the root causes of system performance (good or bad) and allows key components to be studied and adjusted for future performance enhancements. Causal

chains must be linked together to form disciplined *feedback* or *control loops* (Roberts, *et. al*., 1983). Forrester (1961, p.14) states that a " feedback system exists whenever the environment leads to a decision that results in action which affects the environment and thereby influences future decisions." The forming of the feedback loops will help to identify system boundaries. Undisciplined causal chains will make it virtually impossible to identify system boundaries and control the behavior of the components within the system (and hence system behavior). These concepts are crucial to understanding the interdependencies that exist within any system.

As an example of how interdependencies grow tighter, consider a real world problem from the scientific arena (Figure 2-1). We quickly learn that the amount of green house gasses (chlorofluorocarbons, carbon dioxide, methane, and nitrous oxide) in the atmosphere is causing global warming. One of the side effects of global warming is a noticeable temperature increase in the world's oceans. As ocean temperature increases, conditions become ripe for the spawning of increased hurricane activity as evidenced by the extremely active hurricane seasons in the Atlantic Ocean during the past decade. These storms, more massive than ever, are causing large amounts of property damage (especially along coastal regions) each year, resulting in huge expenditures to rebuild lost or damaged homes and businesses. As the seaside communities rebuild, new, and often innovative, attractions (e.g. modern rides and games along a boardwalk) are built, which attracts more visitors, thus helping to pay for their losses and to prosper overall. As the seaside communities become more attractive, they become more popular vacation destinations. As more people travel to the seaside resorts by fossil fuel burning modes of transportation, they are adding to the greenhouse gasses in the atmosphere.

While this example is simplistic, it does illustrate the concepts of causal relationships and how they can form a causal chain. It further shows how that causal chains form closed feedback loops. Lastly, the causal relationships define the boundary of the system. It is important to note that in this example the system boundaries exceed what many would consider factors within the system. This illustrates the point that distance and time often separate some factors that influence the system.

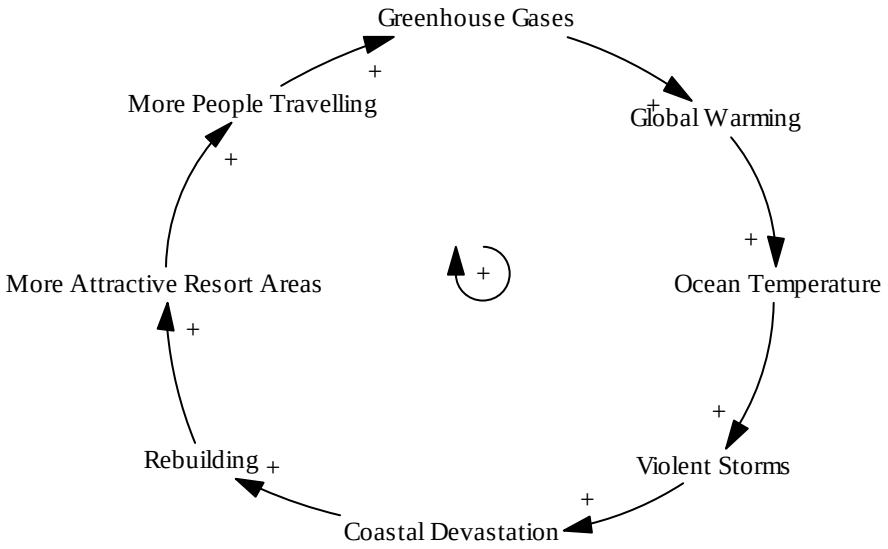


Figure 2-1. A Simplistic View of a Complex Problem.

Continuing with the example, how many vacationers being evacuated from a coastal region due to an impending hurricane would believe that they contributed to the problem? Probably most would deny the claim, but the fact remains that the majority of the vacationers had to travel to work (most in fossil fuel burning transports) to earn the money to have the beach vacation. Additionally, all of the vacationers had to travel to their vacation destination, thus contributing even more to global warming.

This example illustrates the one of the many strengths of approaching a problem from a system dynamics perspective. Humans can easily understand causal relationships that are linked closely in space and time. However, as space expands and time passes, humans have a difficult time transforming several causal linkages into a feedback structure (even a single feedback structure). Given the fact that most systems are comprised of a series of feedback structures, the utility of this modeling approach becomes evident.

At this juncture, our understanding of causal and deterministic relationships must be discussed. A system is *deterministic* when certain actions will predict the outcome of those actions with certainty. The deterministic property of a system is unrealistic in real world events because it assumes ideal conditions exist. For example, combining two parts of hydrogen to two parts of oxygen will produce water if and only if the conditions are ideal (Dantzig, 1998). While causality represents the behavior between two variables, it does not consider whether those variables are deterministic or stochastic. For a system

to be deterministic, the values of the variables and the causality between variables must be known. For a system to be stochastic, the range of possible values and causality between variables must be known.

Another common source of confusion is the comparison between causality and statistical correlation. Correlation between variables reflects the historical behavior between variables within a system, but does not represent the structure of the system. For example, there may be a strong positive correlation between one washing their car and rainfall shortly thereafter. However, it is intuitively obvious that the two events are mutually exclusive. Since statistical correlation represents the past history of a system, and does not represent the system structure. Therefore, variables that were highly correlated in the past may not be correlated in the future due to changes (internal or external) within the system's structure. This is not always the case with causality. Since causality represents relationships within a system, conditions are always subject to change. Previously dormant feedback loops may become active (through factors such as an unstable system achieving stability, or changes in system policy) which may change once believed deep-seated relationships (Sterman, 2000).

Often the boundaries of the system may not be clear. The difficulty in system analysis is to define these boundaries so that the system can be studied and controlled. To study a system, boundaries must be established so that all of the elements or components are enclosed within the system, essentially forming a *closed system*. The closed system concept is essential because the internal structure of the system must be isolated to study the behavior that one is seeking to control, without the interaction with the environment (Forrester, 1968).

One of the purposes of modeling systems is to evaluate policies that govern systems. A *policy* within a system is a general statement of how the available information is used to generate an outcome or decision. Forrester (1968) identifies four concepts found within any policy statement:

1. A goal;
2. An observed condition of the system;
3. A method to express any discrepancy between the goal and the observed condition;
4. Guidelines of which actions to take based on the discrepancy.

2.1.2 Dynamic Model Characteristics

In today's fast paced global market environment, managers seek definitive solutions to difficult organizational problems. Many organizations base their decisions on scientific and mathematical methods (e.g. linear programming). However, Forrester (as cited in Vennix, 1996) found that some of these techniques do not adequately solve the broad strategic management problems, but focus instead on finding an optimal solution.

The problem of optimizing a system that has not achieved a steady state is complicated. Many organizations attempt to optimize their system through trial and error experiments, but this approach is often inefficient and costly. Therefore, models have become commonplace at representing this complicated phenomenon. As an alternative to trial and error experimentation, this research involves embedding an optimization routine within a system dynamics model to examine the productive efficiency problem. The decision to follow this modeling strategy was based on a systematic evaluation of models in general.

Figure 2-2 adapted from Forrester (1961); Dantzig (1998)) depicts the classification of models. This diagram will be used to justify selecting the system dynamics modeling approach.

Models can either be *physical* or *abstract*. Since physical models of complex systems are difficult and expensive to replicate, abstract models are commonly used. An abstract model is one in which symbols or mathematical representations are used to describe the system instead of applying physical devices or components (Forrester, 1961).

Abstract models can be categorized as being either *static* or *dynamic*. Static models simply take a snapshot of the system at a given time and optimize the system based on that data from that specific time period. For example, an organization may seek an operational level that will maximize profit or minimize cost. The static optimization model can be represented mathematically by (Kamien and Schwartz, 1981):

$$Z = \max F(x)$$

subject to: (2-1)

$$x \geq 0$$

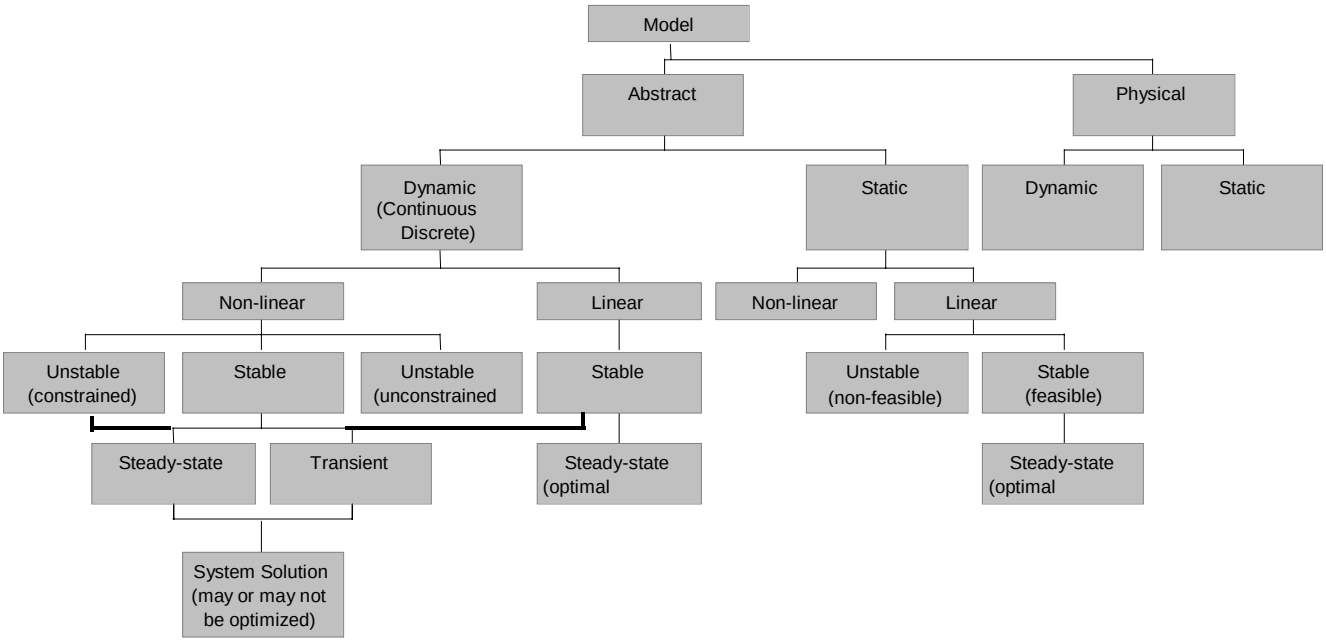


Figure 2-2. Model Classifications.

Dynamic models allow for the system to vary over a time horizon, and are useful for depicting how policies effect decision-making over time.

Models are further categorized as being *linear* or *non-linear*. While linear models offer an approximate solution to a problem, they make the assumption that the organizational systems are adequately represented by linear analysis (Forrester, 1961 and 1987). Because of the definitive and relatively easy answers provided by the linear models, most organizations are willing to base their decisions on this solution. However, most factors in an organization admittedly behave in a non-linear fashion. Because many organizations base their decisions on approximate linear solutions, many systems operate in a sub-optimized state.

Models are further sub-divided as being *stable* or *unstable*. In a static model, (linear model) a stable system is one which yields a feasible solution. Dynamic system stability is characterized by the system returning to its initial condition after a disturbance is introduced. A stable dynamical system may exhibit oscillation characteristics after the introduction of a disturbance, but the oscillations will decline with the passage of time until the system returns to its initial condition (Forrester, 1961 and 1987).

In contrast, unstable dynamical systems start at a steady state (most commonly at rest), then have an initial disturbance introduced. This disturbance is amplified leading to growth. In a non-linear system, the growth will continue until a point when growth is constrained (Forrester, 1961). For example, a newly implemented system can be categorized as being unstable and non-linear until it reaches it maximum operating capacity at which point it may be classified as stable.

Dynamic models are further sub-divided into *steady-state* or *transient* models. A steady-state model is one which has repetitive behavior from one time period to the next (Forrester, 1961). A steady-state model would represent a mature production process that yields consistent output over time.

A transient behavior describes changes where the characteristics of the system vary with time (Forrester, 1961). Transient systems are systems that represent growth. For example, a newly implemented system would represent a transient system until it reaches the point where full operating capacity is realized, at which time it would be reclassified as a steady-state system.

Other model attribute that must be considered when selecting a model to represent a system are whether the system is *open* or *closed*, *discrete* or *continuous*, and *deterministic* or *probabilistic* (i.e. stochastic).

"An open system is one characterized by outputs that respond to inputs, but where the outputs are isolated from and have no influence on the inputs" (Forrester, 1968, p. 1-5). Past actions of an open system do not influence or control future actions of the system. An open system is not aware of its own performance.

A closed system is one that functions without connection to exogenous variables (Forrester, 1961). Such a system receives no inputs from, nor sends any outputs to, the environment. Outputs from closed systems are influenced by inputs within the system. As a result, the structure allows for past actions to be studied and future actions to be controlled (Forrester, 1968).

All real world systems are open systems since they have some form of interaction with their environment. However, when studying systems, boundaries encompassing the components or variables necessary to generate the problem behavior need to be established. Once the boundaries are set, the focus of the problem is inward, feedback

loops and causal relationships can be established, and components or variables can be changed to optimize performance (Richardson and Pugh, 1981). Defining system boundaries is a source of controversy in most models. The argument can always be made that other exogenous variable should have been included and evaluated in the model. A good model will have only the components or variables necessary to solve the problem at hand.

Dynamic models are classified as discrete time or continuous time. Discrete time models consist of an ordered sequence of points, such as monthly, daily, or yearly (Luenberger, 1979). An example of a dynamic discrete time application is deciding the operational level of an organization to maximize profit or minimize costs with respect to this period and other periods. The dynamic discrete time optimization model can be defined as (Kamien and Schwartz, 1981):

$$Z = \max \sum_{t-1}^{T} F(t, y_t , y_{(t-1)}) \qquad (2-2)$$

subject to:

$$y_t \geq 0, t = 1,...,T$$

where:

y_t is output at time *t*.

Continuous time models treats time as a continuous variable as a continuum of real numbers (Luenberger, 1979). In mathematical terms, dynamic continuous time optimization models are represented by differential equations of the form (Kamien and Schwartz, 1981):

subject to:

$$Z = \max \int_0^T F(t, y_t, u_t)dt \qquad (2-3)$$

where:

$$y_0 \cap y_t \geq 0,$$

u_t is output rate at time *t*

The last model classification to evaluate is *deterministic* or *probabilistic*. As discussed previously, deterministic means that if certain actions are taken, or variables acted upon, the outcome can be predicted with certainty. Probabilistic models derive solutions to problems based on uncertainty. Thus the outcome of a given action may depend on a random or chance event (Dantzig, 1998).

2.2 System Dynamics

System dynamics modeling is a methodology that solves problems within time varying (or dynamic), non-linear, closed boundary systems. In its initial application, this methodology was used to study behavior of industrial systems where the short-term dynamics of production rates and inventory levels were analyzed (Forrester, 1961). During the past forty years system dynamics models have been used to solve problems in many diverse disciplines including sociology, biology, economics, and engineering. The strength of system dynamics modeling is that it allows system policies to be evaluated by studying the structure of the system through a series of causal relationships.

System dynamics models represent the flow of information through a system based on the policies that govern those systems. The fundamental concept that governs this methodology is that systems can be represented by a series of level and rate variables, embedded within a feedback structure (Forrester, 1961). Level variables represent accumulations within a system, or describe the state of the system. Rate variables flow from one area of the system to another and control the changes to the levels (Drew, 1994). System policies control the rate variables.

The fundamental premise of system dynamics is that system behavior is a consequence of factors endogenous to the system structure (Richardson and Pugh, 1981). This premise is based on the belief that decision–makers should focus on systemic problems within their purview. If the problem is not within their purview, then they should not expend energy trying to control the problem, as it is uncontrollable from their perspective. Actions taken to correct systemic problems not only include the physical aspects of the system, but also the policies that govern the decision-making within the system (Roberts, 1978).

The system structure relies on a series of cause and effect relationships that determine the underlying flows within a system. These underlying flows are used to bring the system elements together in a holistic manner instead of treating each element independently (Roberts, 1978). Forrester (1961) identified six flows within any system. They are: materials, money, people, capital equipment, orders, and information. System dynamics modeling will be discussed in a generic sense in section 2.2.1. The eight distinct behaviors associated with dynamical systems will be discussed in section 2.2.2

While system dynamics modeling is a powerful tool for predicting and evaluating systems to select the "right" system structure and policies, it falls short of optimizing the system. To optimize the system, an optimization heuristic[1] must be added to the system dynamics framework (Keloharju, 1983). By including optimization within system dynamics, it is possible not only to have the power to evaluate system behavior, but also to select policies that will ensure that the system is operating at its optimum. System dynamics optimization concept will be discussed further in section 2.2.3.

2.2.1 System Dynamics Overview

Volumes have been written about system dynamics modeling. This dissertation will not attempt to replicate the information in those volumes, but instead refer the interested reader to the references listed in the bibliography. However, a brief description of key system dynamics elements and their mathematical representation will be described in this section.

System dynamics models typically have two attributes in common: (1) they involve quantities that change over time; and (2) the systems have control or feedback loops. Thus actions taken in one time period influence or effect actions taken in subsequent time periods (Richardson, and Pugh, 1981). A feedback system is present whenever decision-makers will later experience the consequences of their actions (Roberts, 1978). The results of the decision may be readily apparent very quickly for a system with a short (with respect to time and space) feedback structure, or may be manifested with a long time delay for a system with many elements in the control loop.

[1] Inserting an optimization heuristic into a system dynamics framework is known as *Relativity Dynamics* in some academic circles.

This delay is due to the characteristics of closed loops and feedback processes (Roberts, 1978).

System dynamics models study the feedback structures within a system. These feedback structures can be broken down into a hierarchy of feedback elements. The different levels of this hierarchy are variables, linkage, feedback loop, and feedback system (Roberts, 1978).

A *variable* is defined as a quantity that changes over time (Roberts, 1978). Typical variables found in a system dynamics model include *levels*, *rates*, and *auxiliaries*.

The rate variables determine how fast a system is changing. A rate equation recognizes the goal towards which the system strives, compares the goal to the current system condition (level variables) and makes adjustments to correct the discrepancy (Forrester, 1961 and 1968). The relationship expressed in 2-4 is the mathematical statement for a rate variable (Vensim, 1998):

$$Rate(t) = f(levels(t), auxillaries(t), data(t), constants) \tag{2-4}$$

where:

f is an arbitrary, non-linear, time varying, vector function

Levels (also known as stocks, state variables, or integrations) are the accumulations of inflows and outflows within a system over time. Levels represent the current state of the system. Examples of these accumulations include inventory levels, number of employees, and bank balances (Forrester, 1961 and 1968; Richardson and Pugh, 1981). Levels integrate the results of actions (rates) within a system. These variables cannot be changed instantaneously. Levels are represented by the integral equation (Forrester, 1961 and 1968; Richardson and Pugh, 1981; Sterman 2000):

$$Level(t) = Level(t_0) + \int_{t_0}^{t} (Inflow - Outflow)dt \tag{2-5}$$

where:

Inflow represents the value of the quantity that has flowed into the level

Outflow represents the value of the quantity that has flowed out of the level

Level (t_0) is the initial value of the levels within the system and are governed by the relationship (Vensim, 1998):

$$Level\ (t_0) = f(level\,(t_0)\,,\ auxillary\,(t_0)\,,\ data\,(t_0)\,,\ constants) \qquad (2\text{-}6)$$

Equation 2-5 represents the accumulation of level variables within a system. Equivalently, levels can be described by their net rate of change. This relationship can be defined by the differential equation (Sterman, 2000):

$$\frac{d}{dt} level = Inflow(t) - Outflow(t) \qquad (2-7)$$

While equation 2-5 is the mathematical expression to represent the accumulations within a system, most system dynamics problems are solved using computer software, thus the expression must be redefined using *Euler Integration*. Euler Integration is the default simulation method for all system dynamics software packages. Equation 2-5 is expressed in Euler Integration terms by (Forrester, 1961; Sterman, 2000):

$$LEVEL.K = LEVEL.J + (DT)(INFLOW.JK - OUTFLOW.JK) \qquad (2\text{-}8)$$

Equation 2-8 is shown with post-scripts that are commonly used in system dynamics software packages. However, while Vensim does not use this time-script (post-script) convention, it will be used throughout this section for clarity. Figure 2-3 (adapted from Richardson and Pugh, 1981) illustrates the meaning of the time-script.

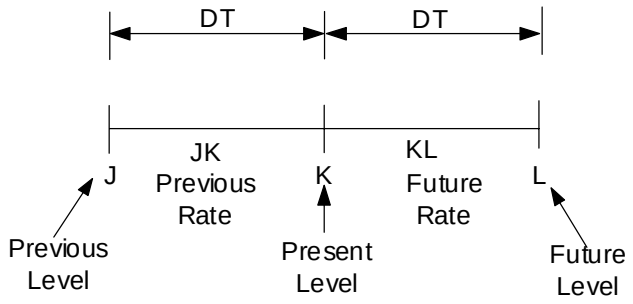


Figure 2-3. System Dynamics Time-script Convention .

There are subtle differences between "dt" in equation 2-5 and "DT" in equation 2-8. In calculus, "dt" represents a continuous time function. In Euler integration, "DT" represents the continuous advance of time broken down into small equal intervals. The intervals of "DT" must be sufficiently small so that it can approximate a constant rate of flow over the interval. This means conditions at the beginning of the interval will remain constant throughout the entire interval (Forrester, 1961). The smaller the length of "DT", the greater the accuracy of the model. When "DT" is reduced to 0, the expression will approximate a continuous-time differential equation as shown in the following relationship (Sterman, 2000):

$$\lim_{dt \to 0} \frac{Level_{t+dt} - Level_t}{dt} = \frac{d}{dt} Level = (Input_t - Output_t) \qquad (2-9)$$

To illustrate the concept of rates and levels consider the accumulation of liquid in a tank (i.e. the hydraulic metaphor). The liquid flowing into the tank represents the inflow rate, and the liquid flowing out of the tank represents the outflow rate. The accumulation of liquid in the tank at any given time represents the level. Figure 2-4 (Richardson and Pugh, 1981; Sterman, 2000) depicts this hydraulic example represented five different (but equivalent) ways.

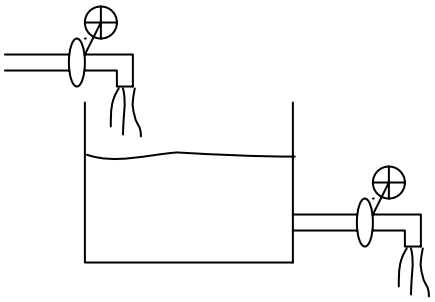
A rule of thumb is helpful to distinguish between level and rate variables. Since the rates are action variables and the levels are accumulations of past actions, rate variables become zero when action in the system stops, and levels would continue to exist at their current accumulation (Forrester, 1968). In the hydraulic metaphor, if all action was stopped, the inflow and outflow of liquid would become zero, and the level would show the current amount of the liquid.

Any feedback loop can adequately be represented by rates and levels alone (Forrester, 1968). To enhance the information in the feedback loop, auxiliary equations are used. Richardson and Pugh (1981, p.81) define an auxiliary equation as "a computation representing information in a feedback system." They are used to aid in the formulation of rate equation and to assist with system decisions. Auxiliary equations can be expressed as (Vensim, 1998):

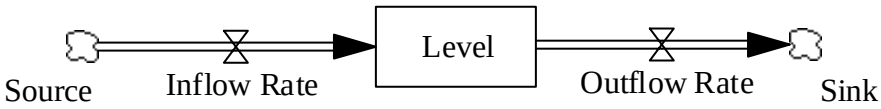
Auxiliary (t) = f(levels(t), auxiliary (t), data(t), constant) (2-10)

In the hydraulic metaphor, an auxiliary may assist the rate variable in determining how fast the fluid flows into (or out of) the tank during any given time period (DT).

Physical Description:



Flow Diagram:



Integral Equation:

$$Level(t) = Level(t_0) + \int_{t_0}^{t} (Inflow - Outflow)dt$$

Euler Integration Equation:

$$LEVEL.K = LEVEL.J + (DT)(INFLOW.JK - OUTFLOW.JK)$$

Differential Equation:

$$\frac{d}{dt} level = Inflow(t) - Outflow(t)$$

Figure 2-4. The Hydraulic Metaphor.

The computing sequence is important in system dynamics modeling. Figure 2-5 (adapted from Forrester, 1961) depicts that state of a system at time K. The first computation at any time K is the value of the levels. This computation is performed using equations 2-5, 2-7 or 2-8. Once the level equations are solved, the auxiliary variable is solved for time period K, using the information from the mathematical relationship shown in equation 2-10. The mathematical relationship in equation 2-4 shows the information that the rate variables for time period KL are calculated from.

A *linkage* is a cause-and-effect relationship between two variables (Roberts, 1978). Two variables are linked together by an arrow connecting the causal variable to the effect variable. The linkages can either be positive or negative. A plus sign indicates that there is a direct variation between the two variables (i.e. both variables tend to move in the same direction). A negative sign indicates that the variables have an inverse relationship (i.e. the variables move in different directions) (Richardson and Pugh, 1981).

Examples of positive and negative linkages are shown in Figure 2-6.

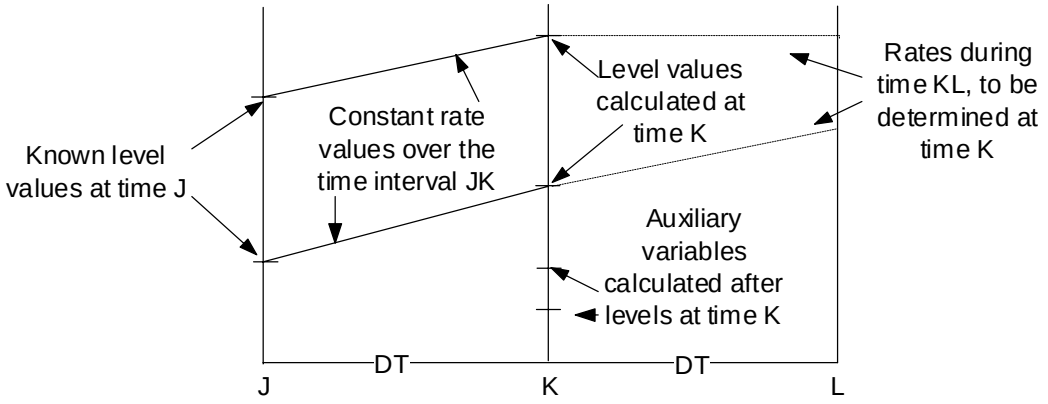


Figure 2-5. Calculations at Time K.

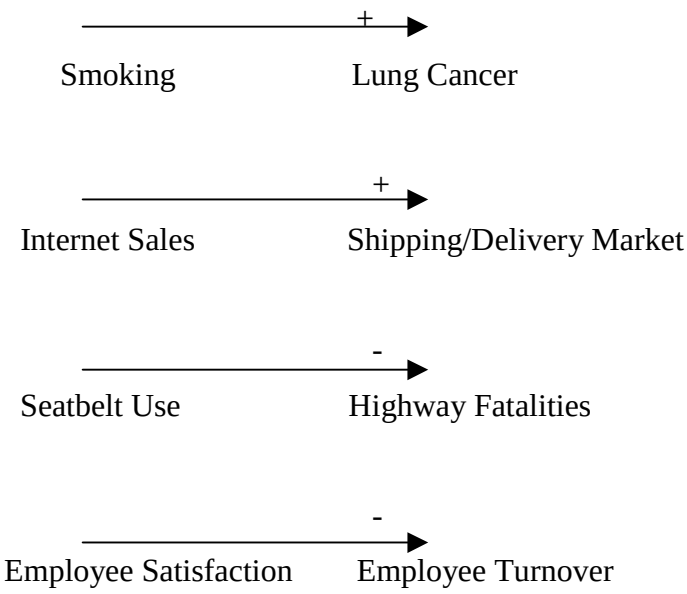


Figure 2-6. Examples of Positive and Negative Linkages.

A *feedback loop* (also known as a causal loop diagrams, or directed graphs) is two or more linkages connected so that one can begin with any variable, and follow the loop through the diagram and back to the original variable. Feedback loops form the basic structures of system dynamics problems, and contain rate and level variables at a minimum. Every decision, and all actions within the system, occur within the feedback loop (Forrester, 1968). As with individual linkages, feedback loops can be categorized as being positive or negative. As a rule of thumb, a feedback loop is positive if it contains an even number of negative linkages and, a feed back loop is negative if it contains an odd number of negative linkages (Richardson and Pugh, 1981).

To see this concept mathematically, consider that the feed back loop on the left side of Figure 2-7 (adapted from Sterman, 2000). To determine the polarity, break the loop at any point, as shown on the right hand side of Figure 2-7. Now an open loop exists, and mathematics from control theory is used to determine the polarity. Equation 2-11 (Sterman, 2000) shows that the polarity of the open loop is determined by a series of partial derivatives.

$$\text{Polarity of loop} = \partial x1(e)/\partial x1(b) \qquad (2\text{-}11)$$

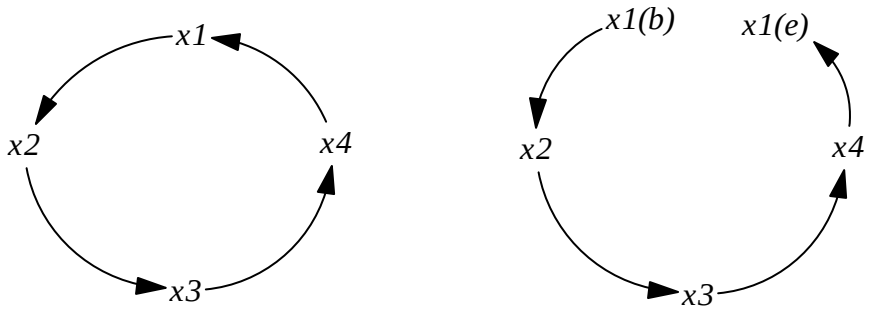


Figure 2-7. Calculating Loop Polarity.

where:

$$\partial x1(b)/\partial x1(e) = [(\partial x1(e)/\partial x4)\partial x4/\partial x3)(\partial x3/\partial x2)(\partial x2/\partial x1(b))] \quad (2\text{-}12)$$

Since the polarity of the loop is a product of all of the partial derivatives of that loop, it can easily be seen from equation 2-12 that an even number of negative signs would result in a positive loop polarity, and an odd number would result in an negative loop polarity Sterman, 2000).

Positive feedback loops are known as self-reinforcing feedback loops (Senge, 1990). Self-reinforcing systems will eventually self-destruct if not acted upon. For example, Figure 2-1 shows how automobile emissions contribute to greenhouse gases and eventually lead to coastal devastation. If closed conditions can be assumed, and this system is undisturbed, coastal devastation will increase to a point where property losses will devastate the economies of many seashore municipalities, and some states. One intervention to stop the self-reinforcing cycle is to develop programs that will reduce the greenhouse gases.

Negative feedback loops are known as self-balancing feedback loops (Senge, 1990). Self-balancing feedback loops tend to stabilize the system. For example the thermostat in a home is a self-reinforcing feedback system. When the room temperature drops to the desired setting, the thermostat sends a signal to the furnace to turn on. After the furnace heats the room to the desired temperature, the thermostat sends a signal to cease heating.

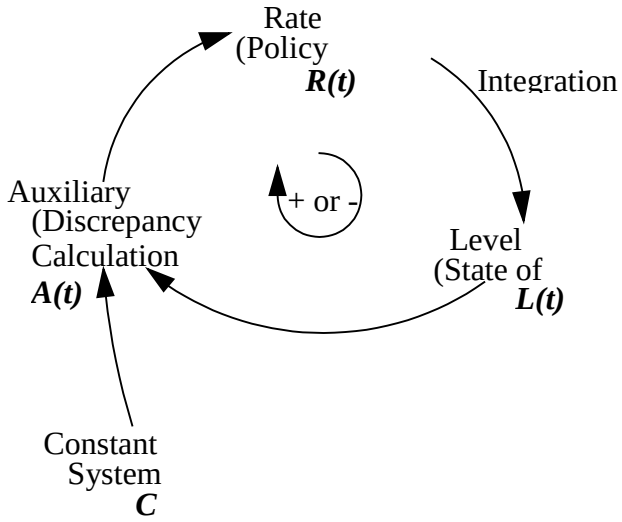


Figure 2-8. How Variables Interact with Feedback Loops.

To illustrate how rate, level and auxiliary variables interact together within a feedback loop, consider Figure 2-8. The rate variable uses the rules established by the policy to determine the flow within the system. Integration is performed to determine the value of the level variable based on the rate variable. An auxiliary variable compares the value of the level variable to the desired value, and sends the discrepancy to the rate variable. The rate variable uses this information to determine the flow for the next system iteration.

A *feedback system* is a series of two or more feedback loops. One feedback loop has the potential of affecting the entire feedback system. Typical organizational and industrial problems are generally described by a feedback system. System dynamics models derive the most information from areas where multiple feedback loops converge (Roberts, 1978).

2.2.2 Behavior of Dynamical Systems

System behavior can be categorized by eight distinct behavior patterns: static equilibrium, exponential growth, goal seeking, exponential decay, oscillation, S-shaped growth, S-shaped growth with overshoot, and overshoot and collapse (Sterman, 2000). Appendix A depicts each of these distinct behaviors and their related structures.

The most basic system behavior is equilibrium. When a system is in equilibrium, the net rate of change of the system is equal to zero. This can be achieved in two ways: First, static equilibrium is defined when there is no flow of inputs x_t into or outputs y_t [2] from the system at time period *t*. Second, when a system has achieved a state of dynamic equilibrium, the net flow of inputs equals the net flow of outputs, thus the state of the system is unchanged but the system is not idle.

A system exhibits exponential growth (Figure 2-9) when the larger the state of the system, the larger the system growth, leading to an even larger system state (Sterman, 2000). This behavior is governed by a single positive (or self-reinforcing) feedback loop[3]. Therefore the growth of the system remains unchecked such that $\lim_{t\to\infty} y_t = \infty$, where y_t is the system output at time *t*.

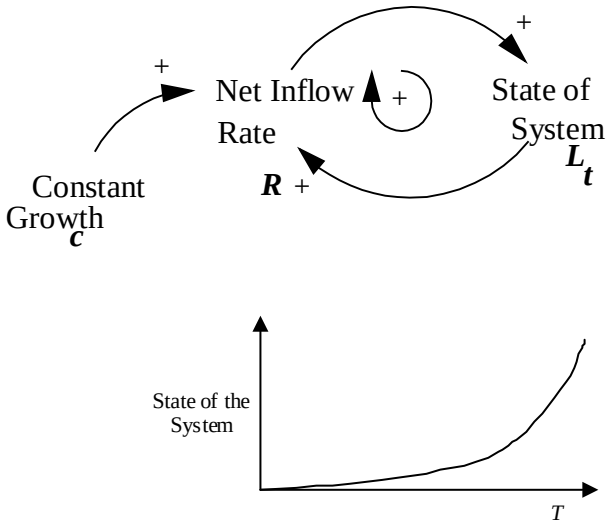


Figure 2-9. Exponential Growth Structure and Behavior.

[2] x_t *is* the input vector defined over the input space $\Re_+^N$ and y_t is the output vector in the output space $\Re_+^M$ during time period *t*.

[3] "Positive feedback loops generate growth, amplify deviations, and reinforce change" (Sterman , 2000, p. 111).

Goal seeking behavior strives to bring the state of the system in line with the systems goal. This is accomplished by a single negative (or goal seeking) feedback loop which counteracts any disturbance that moves the system away from its desired goal (Figure 2-10) (Sterman, 2000). As the system approaches its desired goal $\hat{L}$, the inputs x_{t-d} are transformed into outputs such that $\lim_{t\to\infty} y_t = \hat{y}$, where $\hat{y}$ the system requirement. In most cases, the system's desired goal $\hat{L}$ is equal to the system's requirement $\hat{y}$.

A special single negative feedback structure is the source of the exponential decay behavior (Figure 2-11). Exponential decay occurs when the relationship between the net input x_{t-d} and the system discrepancy Δ is linear (Sterman, 2000). The system discrepancy is defined as the difference between the desired state and the current state of the system, $\Delta = L_t - \hat{L}$. As the discrepancy Δ decreases, so does the net inflow rate x_{t-d}. As the discrepancy $\Delta \to 0$, and the inflow rate $x_{t-d} \to 0$, then the $\lim_{t\to\infty} y_t = 0$.

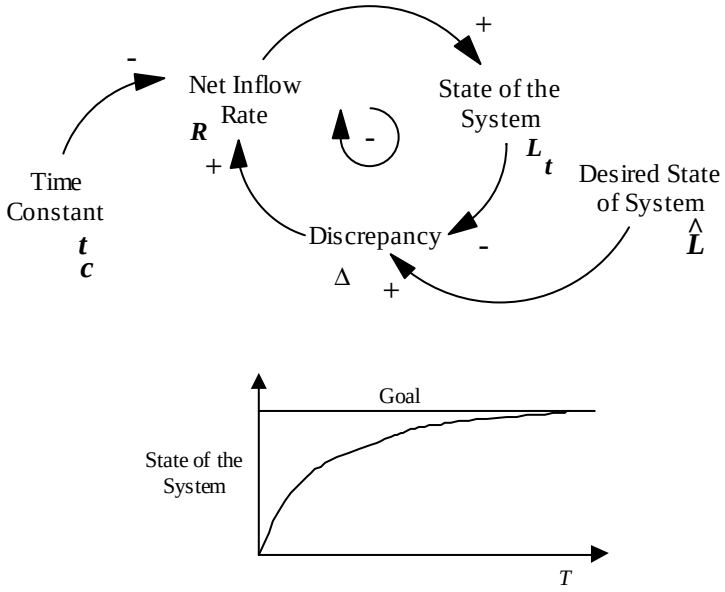


Figure 2-10. Goal Seeking Structure and Behavior.

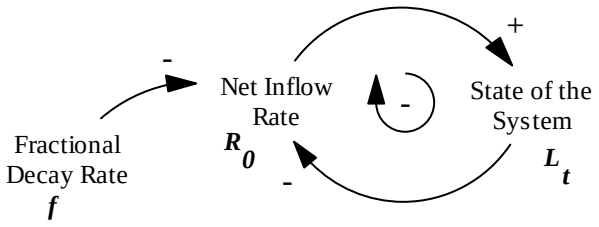


Figure 2-11. Exponential Decay Structure.

Oscillatory behavior is perhaps the most common behavior found in dynamic systems. This behavior can assume many forms, including damped oscillation, chaos or expanding oscillation. Oscillation is caused by a single negative feedback loop (Figure 2-12) that has a significant delay within at least one of its causal linkages. The negative feedback loop structure compares the state of the system to the goal, and makes adjustments accordingly. However, given the delay, the system continues to take corrective action even after the system achieves its goal. Thus the system is constantly overshooting and undershooting the desired system state (Sterman, 2000).

While positive and negative feedback loops (including those with delays) provide a firm foundation from which to evaluate system behavior, realistic production systems usually involve structures with multiple feedback loops. These multiple loops systems produce behaviors that are the composite of several system behaviors.

The S-shaped growth behavior exhibits exponential growth at first, but then exhibits goal-seeking behavior as the system approaches its equilibrium. This behavior is governed by two feedback loops: a positive feedback loop which leads to the exponential growth behavior; and a negative feedback loop which leads to the goal-seeking behavior (Figure 2-13). For this behavior to exist, two conditions must be satisfied: (1) the system production capacity must be fixed; and (2) the system must not contain any significant time delays (Sterman, 2000). The feedback loop that is dominant at time *t*, determines which behavior the system is currently exhibiting.

To further explain the concept of S-shaped growth, assume that a system has a fixed production capacity. As the production process begins, the resources available appear to be infinite, thus exponential behavior is experienced. As the system approaches its production capacity, system resources become scarce. As a result the system

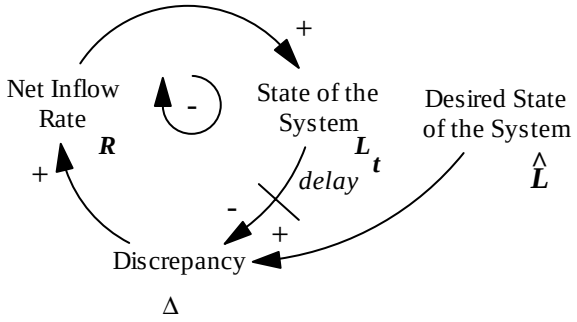


Figure 2-12. Oscillation Structure.

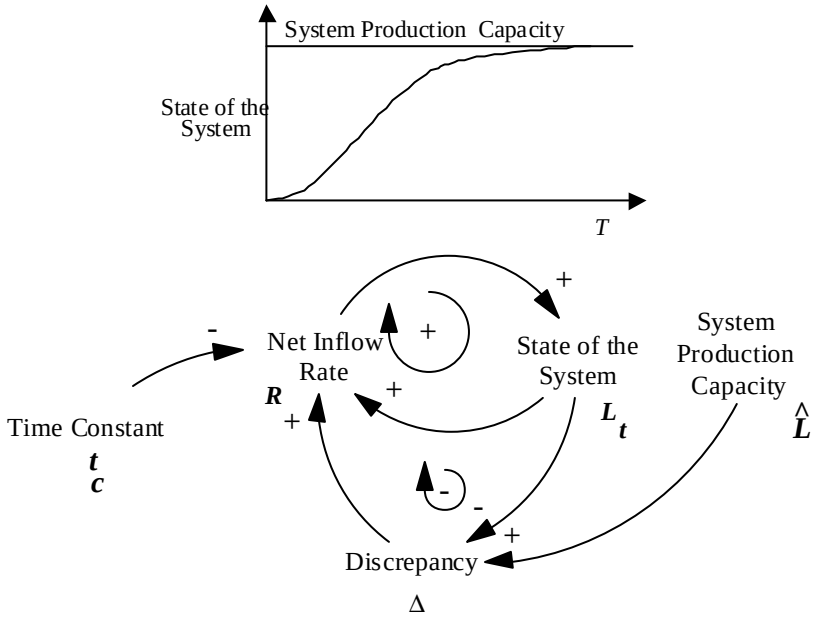


Figure 2-13. S-shaped Growth .

continues to increase towards the production capacity, but at a slower growth rate. Thus, goal-seeking behavior is experienced (Sterman, 2000).

S-shaped growth characterizes a perfect system behavior whose single positive and negative feedback loops are combined into a single system structure. However, real production systems are not perfect as demonstrated by the oscillatory behavior discussion. When a negative feedback loop with a significant delay is coupled with a positive feedback loop, S-shaped growth with overshoot occurs (Figure 2-14) (Sterman, 2000).

S-shaped growth with overshoot behavior initially behaves like exponential growth due to the dominant positive feedback structure. As the system approaches its production capacity, the system exhibits oscillatory behavior around that capacity. This oscillation occurs because of the presence of significant delays embedded in the negative feedback loop. Thus the system continues to take corrective action even after the system achieves its production capacity (Sterman, 2000).

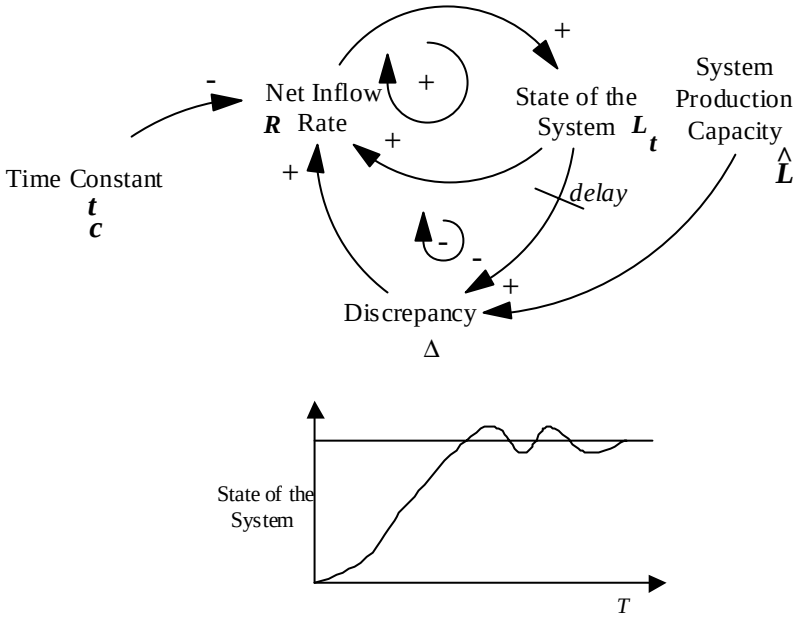


Figure 2-14. S-shaped Growth with Overshoot Structure and Behavior.

A structure that contains a positive feedback loop coupled with two negative feedback loops produces the S-shaped growth with overshoot and collapse behavior (Figure 2-15). Unlike the S-shaped growth behavior, the production capacity is not fixed, but is consumed as the ability of the system to support the system requirements erodes (e.g. the erosion of a non-renewable resource within the system) (Sterman, 2000). When the positive feedback loop is dominant, the system exhibits exponential growth. As the system matures, the discrepancy Δ begins to fall and the negative feedback loops gain in strength. When $\Delta \rightarrow 0$, the inputs x_{t-d} are transformed into outputs such that $\lim_{t \rightarrow t_m} y_t = \hat{y}$, where $0 < t_m < \infty$, but the system does not achieve dynamic equilibrium. When $\lim_{t \rightarrow t_m} y_t = \hat{y}$, production is at its maximum, but the production capacity drops because resources are being consumed to sustain the system requirement. If the production capacity is not regenerated (e.g. a non-renewable resource), the state of the system declines until static equilibrium is achieved (i.e. $\lim_{t \rightarrow \infty} y_t = 0$) (Sterman, 2000).

To clarify the S-shaped growth with overshoot and collapse behavior, consider a simple example. During a heat wave the requirement for electricity becomes so great that the electrical generators are often pushed to their operating capacity. When this occurs, if action is not taken early enough, the ability to supply electricity to customers decreases as the production capacity of the system erodes as system resources are being consumed.

2.2.3 The Systems Dynamics Optimization Concept

Heretofore in this discussion of SD modeling, the basis for the techniques and variables used in model development has been introduced. All SD models consider a real world problem, select variables that represent real world elements, and define a time step that is appropriate for analyzing the system. Once the model has been developed, the model parameters and structure are usually manually adjusted until the model achieves the desired objective (Wolstenholme, 1990).

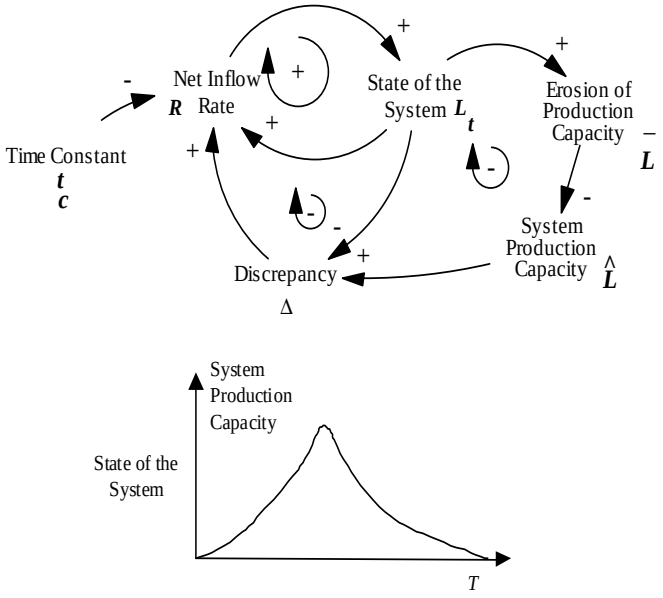


Figure 2-15. S-shaped Growth with Overshoot and Collapse Structure and Behavior.

During the early 1980's, a major new approach to SD modeling emerged. This approach is simulation by optimization. This concept is based on replacing manual model revisions to achieve system optimization with a heuristic optimization (i.e. hill-climbing search) algorithm that will automatically determine the optimal solution for the system (Wolstenholme, 1990). Figure 2-16 (Wolstenholme, 1990) illustrates the differences between traditional SD and optimized SD.

To perform SD optimization, three fundamental steps are required:

1. An objective function must be defined that represents the desired model behavior;
2. Parameters, which represent constraints within the model, with their feasible range of values, must be defined;
3. The number of iterations which the model must complete must be defined (Wolstenholme, 1990).

A level variable is generally the basis of the objective function. However, these level variables are usually only added to the model for analysis purposes, and have no physical meaning to the real world system (Coyle, 1996). Coyle (1985) suggest developing a performance index to measure efficiency. This performance index has no meaning to the system, but can represent a usable measure for the decision-maker. For

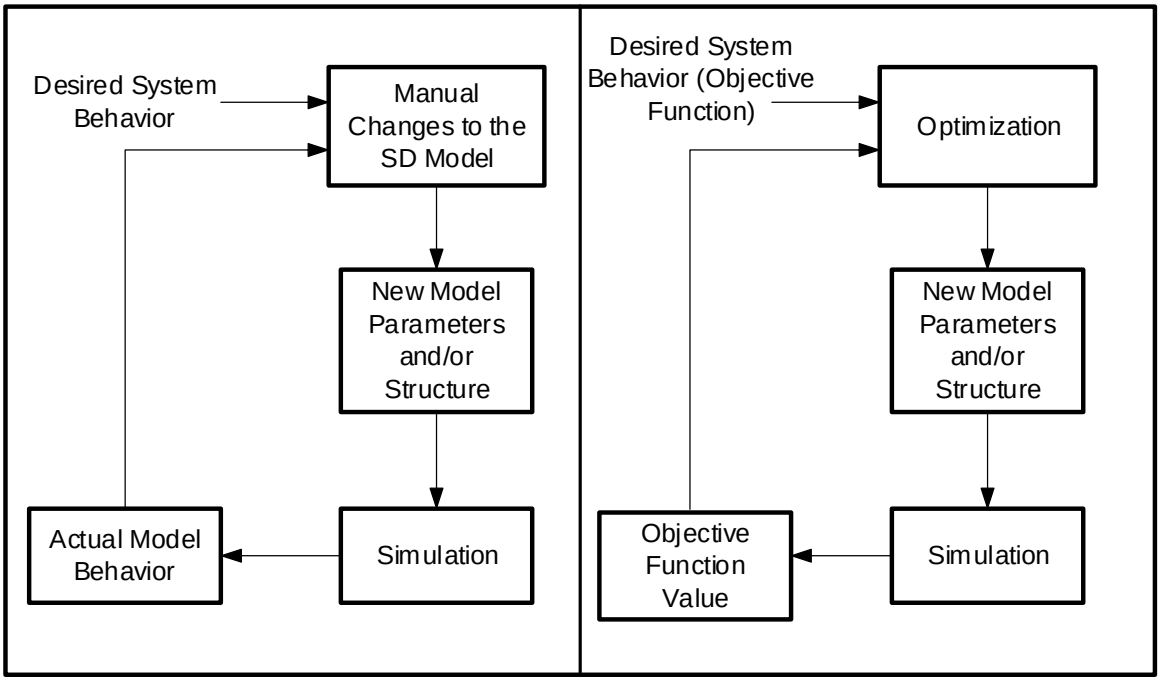


Figure 2-16. The Differences between Traditional and Optimized SD.

example, in a productive efficiency study the objective function centers on an efficiency based measure of effectiveness. While this efficiency score does have significance to the decision-maker, it has no tangible meaning to the operation of the system.

Another reason for using a surrogate level variable is because minimizing or maximizing can have unexpected consequences on the system as a whole (Coyle, 1996). For example, consider a manufacturing plant that wants to minimize inventory. If inventory is truly minimized, inventory will become zero. Zero inventory will be detrimental to other factors in the system such as timeliness of delivery.

The constraints in the model are variables that have some special significance to the system. These variables are the variables that will be changed to optimize the objective function. The range of feasible values for each constraint must be carefully considered because an unrealistic constraint could provide an erroneous solution to the objective function (Wolstenholme, 1990). Any rate or auxiliary variable can be represented as a constraint.

Unlike linear programming where the optimal solution is obvious when determined, the optimal value of the SD model is not. The SD modeler must define the

number of iterations that the model will achieve before stopping. At the end of those iterations, the model will yield the optimal value of the objective function found until that model iteration. However this does not guarantee that the solution found is optimal and feasible. The only way to determine the feasible and optimal solution is by setting the number of iterations significantly high so that the model will run through enough iterations to find the desired solution. Wolstenholme (1990) states that a good rule of thumb is 100 iterations for a medium sized model.

System dynamics optimization is achieved through a hill-climbing algorithm. The fundamental idea of the hill-climbing search algorithm is to systematically vary the variables of the objective function in order to find its minimum or maximum value (Fletcher and Powell, 1963; Vanderplaats, 1984). The heuristic operates in the sense that if a move from one point to another point in an n-dimensional space moves in the desired direction of the objective function, then the next move should be in that same direction (Kivijärvi and Tuominen, 1986).

Coyle (1996) uses the analogy of a blind man trying to find his way to the top of a mountain to describe the SD optimization heuristic. The man's strategy is to feel the shape of the ground around him. When the blind man detects the direction that the slope of the hill is steepest he moves in that direction. This move corresponds to one iteration in the model. The blind man repeats this process until he has no energy left to continue. The highest point that he found before his energy was depleted represents the maximum height that he found on the mountain. However, the height may not represent the highest peak on the mountain (Coyle, 1996). This represents achieving the maximum number of iterations in the model. Once these iterations have been achieved, there is no guarantee that the maximum peaks have been found. Thus the reason to test the model with many different iterations to achieve the best results.

Another way to describe the hill-climbing search algorithm is pictorially. Consider the two-dimensional picture of a three-dimensional diagram, in Figure 2-17 (Coyle, 1996). In this diagram, the man progresses from his initial point (at the graph's origin) and progresses to points *A* and *B*. At point *B* he selects to follow path 1, because the terrain in that direction has the steepest gradient, and proceeds to point *C*. If the man has any remaining energy, he will return to point *B* and explore the other options. This

time the man selects option 2 that leads him to point *D*. Since point *D* is the highest of the two options, the man claims success in finding the highest point.

In this simple example, point *C* represents a local optima, and point *D* represents the optimum solution. However, point *D* would never have been achievable if the man didn't have the energy to continue. The man's energy represents the number of iterations in the model. The path that the man took to reach the top of the mountain represents the dynamic production frontier. Identifying and deriving information from this frontier is the heart of the modeling effort for this research. This research uses the plane identified by the optimization heuristic to identify efficient operating conditions and to establish performance goals for the selected variables within the system. The derivation of the dynamic production plane will be discussed in Chapter 3.

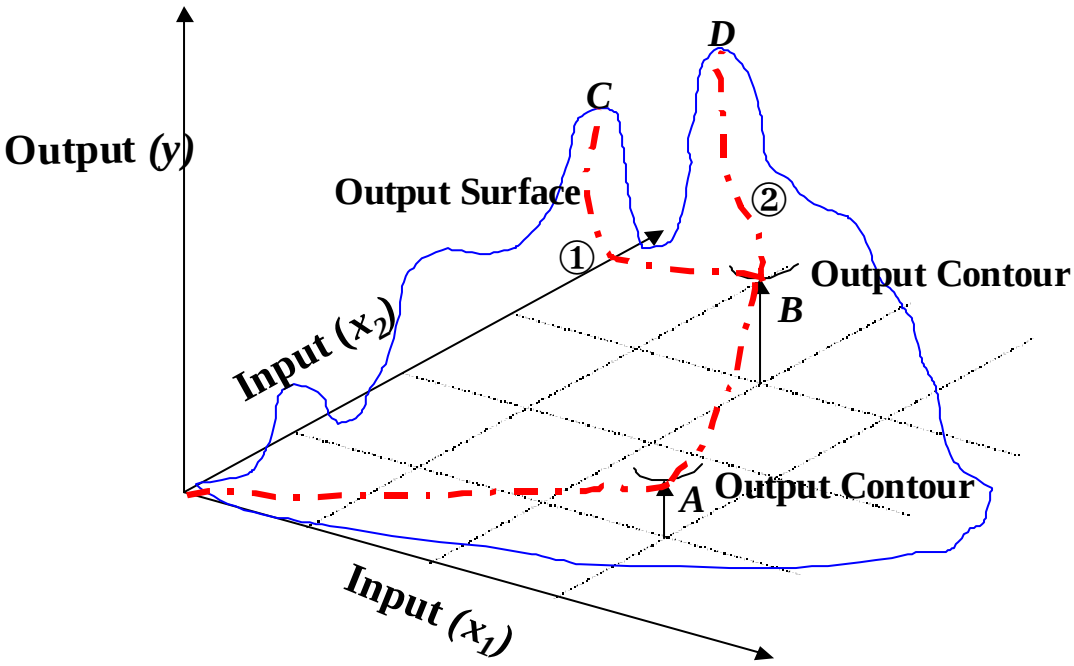


Figure 2-17. An Illustration of Hill-climbing Optimization.

2.2.4 The Generic Hill-climbing Optimization Structure

The generic hill-climbing optimization structure (Figure 2-18) (Sterman, 2000) is comprised of two feedback loops – one positive and one negative. The negative feedback loop serves to close the gap between the desired state of the system L^* and the state of the system L. The positive feedback loop adjusts the goal of the system. The desired state of the system L^* is coupled with the state of the system L. External pressures influence the desired state of the system. These external pressures represent the gradient or slope of the hill, and indicate which direction the system should move to achieve its optimal value (Sterman, 2000).

The hill-climbing optimization process is governed by three fundamental equations that are solved sequentially during each iteration of the model (Sterman, 2000). The starting point for any SD model is with the level variable (Richardon and Pugh , 1981) Thus the first equation in the model is:

$$L = L_{t_0} + \int_0^t R dt \qquad (2-13)$$

Where L_{t_0} is the initial state of the system.

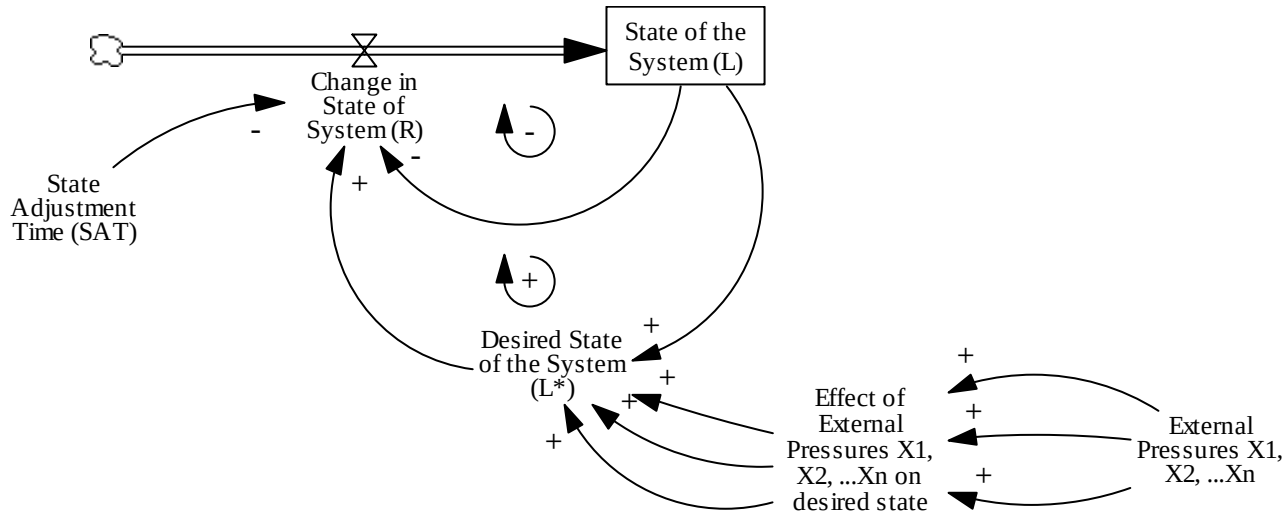


Figure 2-18. The Generic Hill-Climbing Optimization Structure.

The desired state of the system is acted upon by factors that are external to the dual loops of the optimization structure, but internal to the SD model. The desired state of the system is defined as (Sterman, 2000):

$$L^* = f(L, \text{effect of } X1 \text{ on } L^*, \ldots, \text{effect of } Xn \text{ on } L^*) \tag{2-14}$$

where:

Xi is an external pressure on the system.[4]

$$\text{Effect of } Xi \text{ on } L^* = f(Xi/Xi^*) \tag{2-15}$$

where:

*Xi** is the base value.

The change in the system *R* is the rate variable that determines how fast the state of the system *L* changes, and is defined as (Sterman, 2000):

$$R = \frac{L^* - L}{SAT} \tag{2-16}$$

where:

SAT is the state adjustment time.

When the net effect of the pressures on the desired state of the system cause $L^* > L$, *L* will increase exponentially. When $L^* < L$, *L* will decrease exponentially. When $L^* = L$, the system has achieved equilibrium and the optimal value(s) are achieved.

2.3 The Theory of Productive Efficiency

The theory of productive efficiency has been an interesting and well researched subject throughout the years. The purpose of this research is to add another chapter to the body of literature by introducing a methodological approach that combines the system dynamics paradigm with the measurement of productive efficiency. The impetus for this research effort is much like that of previous researchers. Farrell (1957, p. 11), in his seminal paper, stated it best:

[4] The external pressure governs the effect on the system. For example, the effect of stock values is governed by many factors including but not limited to political environment, traders trust in the economy, and influence of foreign markets.

> "A number of attempts have been made to solve this problem, but, although they usually produced careful measurements of some or all of inputs and outputs of the industry, they failed to combine theses measurements into any satisfactory measure of efficiency. This failure was partly due to a pure neglect of the theoretical side of the problem. Indeed, for a long time it was considered adequate to measure the average productivity of labour, and to use this as a measure of efficiency. This is a patently unsatisfactory measure, as it ignores all inputs save labour, but it was so widely used by economic statisticians that is now enjoying an extensive popular vogue, which may indeed have unfortunate effects on economic policy. More recently, attempts have been made to construct 'indices of efficiency', in which a weighted average of inputs is compared with output. These attempts have naturally run into all the usual index problems."

This section provides a rudimentary overview of some of the concepts and methodologies that have emerged over time, and provides insight into how a point of departure was derived for this research. While some of the concepts expressed here provided impetus for this research, they proved not to contribute as much as originally envisioned. However, they will be addressed with a broad overview. Other concepts proved to be more critical, and were used to develop the methodology this research is introducing. These concepts will be discussed casually here, and in more detailed in the Chapter 3 where they are expanded to the dynamic realm.

2.3.1 The Production Axioms

The production axioms are a set of properties that explain and govern the activities that occur within a production environment. These axioms are representative of any condition found where inputs are converted into outputs. Thus only a subset of these axioms may apply to any given circumstance. However, these axioms provide a starting point from which assumptions about production systems can be made.

The production technology can be represented in a variety of ways (i.e. output correspondence, input correspondence, and graph theory) (Färe and Grosskopf (1996)). Without the loss of generality, this dissertation addresses the production axioms from the output correspondence perspective.

The production technology axioms govern the transformation of inputs into outputs. Specifically, production technology uses a set of inputs:

$$x = (x_1, x_2, ..., x_n) \in \Re_+^N \qquad (2-17)$$

to produce a set of outputs.

$$y = (y_1, y_2, ..., y_M) \in \Re_+^M \qquad (2-18)$$

where x is the input vector defined in the input space $\Re_+^N$, and y is the output vector defined in the output space $\Re_+^M$.

The static production technology is represented by the output correspondence as (Färe and Grosskopf (1996)):

$$P : \Re_+^N \rightarrow P(x) = \{y : x \in S(y)\} \qquad (2-19)$$

where $P(x)$ is the output set. The output set contains all output vectors y that are produced by the input vector x, given the transformation process P. $S(y)$ is the input set. This set contains all input vectors x that can produce output vectors y, given the transformation process S (Färe and Lovell (1978); Färe and Grosskopf (1996)).

Since the basic concepts of the static production axioms are the same as the dynamic production axioms, a comprehensive discussion of them is deferred until Chapter 3. However a casual discussion follows.

Axiom A.1(a) simply states that it is always possible for a firm to produce no outputs (Färe and Primont, 1995; Färe and Grosskopf, 1996).

$$A.1(a) \qquad 0 \in P(x), \forall x \in \Re_+^N \qquad (2-20)$$

Axiom A.1(b) states that it is not possible to produce outputs without inputs (Färe and Primont, 1995; Färe and Grosskopf, 1996).

$$A.1(b) \qquad y \notin P(x = 0), \text{ i } y > 0 \qquad (2-21)$$

The weak input disposability axiom (A.2(a)) states if all inputs are increased proportionality, then output will not increase (Färe and Primont, 1995; Färe and Grosskopf, 1996). Conversely, if inputs are not increased proportionally, then outputs may decrease. By increasing the amount of x by a factor λ, then the production function may increase the output by that factor providing that $\lambda>1$. It is critical that $\lambda>1$, because if $\lambda<1$, than the output from the production function will decrease. For example if an employee who is fully trained to perform operations on a new system leaves the project, and a less experienced person is hired in their place, the output from the system will not be equivalent to its previous level. Therefore a loss of productivity will be realized.

$$A.2(a) \qquad \text{If } y \in P(x) \wedge \lambda \geq 1 \Rightarrow y \in P(\lambda x) \qquad (2-22)$$

The strong input disposability axiom (A.2(b)) states if inputs are increased, whether proportional or not, output will not decrease (Färe and Primont, 1995; Färe and Grosskopf, 1996).

A typical example of this axiom is when one element of a production line containing a bottleneck is improved. In this case output may or may not increase depending on if a bottleneck or non-bottleneck resource is increased. If A.2(b) is true than A.2(a) is also true. However, A.2(a) being true does not imply that A.2(b) is true.

$$A.2(b) \qquad \text{If } y \in P(\breve{x}) \wedge x \geq \breve{x} \Rightarrow y \in P(x) \qquad (2-23)$$

Axiom A.3(a) is known as the weak output disposability axiom (Färe and Primont, 1995; Färe and Grosskopf, 1996). It allows for a reduction of outputs without decreasing inputs. This axiom is especially useful when producing both desirable and undesirable outputs.

$$A.3(a) \qquad \text{If } y \in P(x) \wedge 0 \leq \varphi \leq 1 \Rightarrow \varphi y \in P(x) \qquad (2-24)$$

The strong output disposability axiom (A.3(b)) allows outputs to be disposed of without incurring any additional cost (Färe and Primont, 1995; Färe and Grosskopf, 1996). This axiom does not hold true for the reduction of undesirable outputs. For example, a production may produce desirable products and undesirable (rework)

products. This axiom implies that disposable of these undesirable products has not associated cost. Axiom A.3(b) implies that A.3(a) is true. However, the converse is not valid.

$$A.3(b) \qquad \text{If } y \in P(x) \wedge \breve{y} \leq y \Rightarrow \breve{y} \in P(x) \qquad (2-25)$$

Axiom A.4 states that a firm cannot produce an infinite number of outputs from a finite number of inputs (Färe and Primont, 1995; Färe and Grosskopf, 1996).

$$A.4 \qquad \forall x \in \Re_+^N, P(x) \text{ is a bounded} \qquad (2-26)$$

The mathematical condition of closedness is addressed in A.5. Assume the output y is a series of vectors $y_j = (y_1, y_2, y_3, ...y_m)$ such that $\lim_{j\to\infty} y_j = y$. If every sequence of output vectors y_j can be produced from inputs x_i, then x can produce y (Takayama, 1985; Färe and Primont, 1995).

$$A.5 \qquad \forall x \in \Re_+^N, P(x) \text{ is a closed} \qquad (2-27)$$

The convexity axiom (A.6) assumes that the production function can be represented by a series of graphs for differing levels of production (Färe and Primont, 1995, and Färe and Grosskopf, 1996). It can be concluded from this axiom that the production function is not only bounded but is also closed.

$$A.6 \qquad \forall x \in S(y) \in \Re_+^N if\ 0 \leq \lambda \leq 1 \Rightarrow \lambda x + (1-\lambda)\breve{x} \in S(y) \qquad (2-28)$$

2.3.2 Productive Performance Measurement

Färe and Lovell (1978), define the *production function* as a scalar output that specifies the maximum output obtainable from an input vector. Thus the production function axioms describe a technical relationship between the inputs to the production process and the outputs from the production process.

The technical relationship actually describes the degree to which a unit is technically efficient. A *technically efficient* unit reflects the firm's ability to produce the

maximum output for a given input. If this condition is met, the unit is said to be operating on its production frontier. A *technically inefficient* unit will be operating below the production frontier, thus not optimally using its inputs to produce outputs (Farrell, 1957; Coelli, *et. al.* 1998).

As technical efficiency characterizes the physical efficiency of transforming inputs into outputs, *allocative efficiency* characterizes the economic or price efficiency associated with transforming inputs into outputs (Farrell, 1957).

Figure 2-19 dichotomizes productive efficiency into its two sub-components: technical and allocative (Farrell, 1957). The figure examines points α and β during a single time period (static sense), with two inputs x_1 and x_2, and one output y, by comparing them to the unit isoquant (line yy')and isocost (line cc') lines. This approach allows for the analysis of productive efficiency by comparing the measured values with respect to the isoquant and isocost lines. If a Decision-Making Unit (DMU) lies on the isoquant, it is deemed technically efficient. If is lies on the isocost line, it is deemed allocatively efficient. And if it lies at the intersection of the isoquant and isocost lines, the DMU is deemed both technically and allocatively efficient. The area above the isoquant is the inefficient region, and the area below the isoquant is the infeasible region.

With the assistance of Figure 2-19, the mathematical relationships for allocative efficiency, technical efficiency, and overall productive efficiency can be shown. Technical efficiency *TE* is mathematically defined as (Farrell, 1957):

$$TE = \frac{0\alpha'}{0\alpha} \qquad (2-29)$$

Allocative efficiency *AE* is defined as (Farrell, 1957):

$$AE = \frac{0\alpha''}{0\alpha'} \qquad (2-30)$$

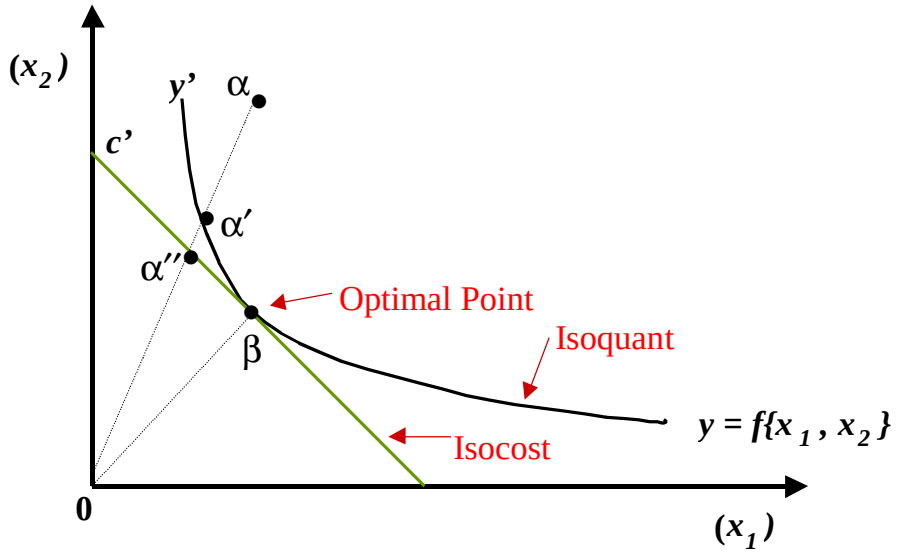


Figure 2-19. The Concept of Technical and Allocative Efficiency.

Overall productive efficiency *OPE* is defined as (Farrell, 1957):

$$OPE = \left(\frac{0\alpha''}{0\alpha'}\right)\left(\frac{0\alpha'}{0\alpha}\right) = (AE)(TE) \qquad (2-31)$$

In Figure 2-19, point β lies at the intersection of the isoquant and isocost lines, thus is deemed both technically and allocatively efficient with respect to point α. Point α is initially located to the northeast of the isoquant, thus is deemed technically inefficient. By incorporating process improvements, a weighting factor less than 1 is applied to point α, the inputs are used more efficiently such that the DMU is able to produce the same amount of outputs with less inputs. This action plots point α' on the isoquant. Incorporating process improvements based on cost, a different weighting factor , also less than 1 is applied, and point α is adjusted further. This action now places point α'' on the isocost line. Subsequently the system can now be deemed to technical efficient or allocative efficient but not both. For point α to be both technically and allocatively efficient, the processes must be changed to resemble the process used at point β.

A few caveats to Figure 2-19. First, the assumption is being made that the production function is being operated under constant returns to scale. Second, the

production frontier is known. In general production frontier is not generally known and must be estimated (Coelli, *et. al.* 1998).

2.3.3 Data Envelopment Analysis

During the early months of this research it was thought that Data Envelopment Analysis (DEA) and Dynamic Data Envelopment Analysis (DDEA) could be incorporated into the system dynamics framework. That has proven not to be the case, for reasons described in Chapter 5. However, since DEA and DDEA have been used for the analysis of productive efficiency for many years, and provided a portion of the initial motivation for this research. A short discussion of DEA and DDEA is warranted.

Data Envelopment Analysis (DEA) is a mathematical application for comparing the measures of relative efficiency in firms, organizations, or systems, in which there are multiple inputs and/or outputs, and there is no way of aggregating the inputs or outputs into a single measure of relative efficiency (Charnes, *et. al.* 1978). The objective of DEA is to optimize each decision-making unit. This approach is in contrast to regression analysis in that the statistical procedure seeks to find the mean value of each decision making unit (DMU), or the central tendency of the system overall.

Each decision-making unit governed by the simple ratio (Boussofiane, *et. al.* 1991):

$$Efficiency = \frac{output}{input} \qquad (2-32)$$

Organization usually have more than one input and output, therefore equation 2-32 needs to be expanded to (Boussofiane, et. al., 1991):

$$Efficiency = \frac{Weighted\ Sum\ of\ Outputs}{Weighted\ Sum\ of\ Inputs} \qquad (2-33)$$

Equation 2-33 is also difficult for most firms to use on a practical basis because it requires that the weights be defined. Since weights of the inputs and outputs are generally unknown, equation 2-33 needs to be further refined. Charnes *et. al*. (1978)

developed a model based on linear programming that considers the unknown weights and solves the model for its maximum possible values.

$$Max\ e_k = \frac{\sum_{r=1}^{M} u_r y_{r0}}{\sum_{i=1}^{N} v_i x_{i0}} \qquad (2\text{-}34)$$

subject to:

$$\frac{\sum_{r=1}^{M} u_r y_{rj}}{\sum_{i=1}^{N} v_i x_{ij}} \le 1; \qquad j = 1,2,...,n$$

$$u_r, v_i \ge 0; \qquad r = 1,2,...,m; \qquad i = 1,2,...,m$$

where:

e_k is the efficiency measure

u_r is the weight given to output *r*

y_{rk} is the amount of output *r* from unit *k*

v_i is the weight given to input *i*

x_{ik} is the amount of input *i* to unit *k*

n is the number of units

m is the number of outputs

n is the number of inputs.

The logic for solving equation 2-34 is that the system is maximized to attain the maximum possible value of each DMU. Since each DMU is being given the maximum possible efficiency rating (as opposed to an efficiency rating based on a measure of central tendency), if it is judged inefficient, then the DMU truly inefficient (Charnes *et. al.*, 1994).

Equation 2-34 assumes that each of the DMUs being evaluated is evaluated for a common and single time period. This approach is impractical for evaluating processes that occur over multiple time periods, and use outputs from one time period as inputs to

another time period. During the last decade, Färe and Grosskopf (1996) defined and developed the dynamic data envelopment analysis (DDEA) methodology. This methodology adds the element of time to the DEA model. This is accomplished by extending the static DEA model into an infinite sequence of static equations (Färe and Grosskopf, 1996). While this methodology does evaluate organizational performance over time, it is important to understand that this methodology is static at each discrete point in time, and provides a linear solution.

Figure 2-20 (adapted from Färe and Grosskopf, 1996) illustrates the concept of DDEA. Each production cycle (*P*) has three types of inputs: fixed (*x(f)*), variable (*x(v)*), and inputs that were intermediate outputs to the last production cycle (*y(i)*). The outputs for each production cycle are the final outputs (*y(f)*) which represent the products that go to the customer, and intermediate outputs that are used as inputs to subsequent production cycles (*y(i)*).

As an example of how DDEA is applied, consider a manufacturing plant that has many outputs. Some of the outputs can be sent directly to market without further processing after the first going though the first production process at time *t*. This output is representative of *y(f)*. Other products will have to go to other process for finishing at a later time *t+1*. These outputs can be represented by *y(i)*.

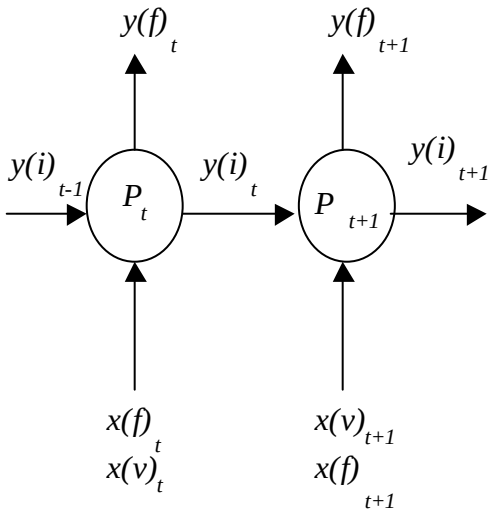


Figure 2-20. The DDEA Basic Structure.