COWLES FOUNDATION DISCUSSION PAPER NO. 54

Capacity Expansion and Probabilistic Growth*

by

Alan S. Manne

July 17, 1958

* Research undertaken by the Cowles Commission for Research in Economics under Contract Nonr-358(01), NR 047-066 with the Office of Naval Research.

Capacity Expansion and Probabilistic Growth

by Alan S. Manne*

1. Introduction

This study stems from an optimizing model originally suggested by Hollis Chenery for predicting investment behavior. [5] Like Chenery's paper, this one is concerned with the interplay between economies of scale and an anticipated persistent growth in demand for capacity. The generalizations discussed here are of two types: (a) the use of probabilities in place of a constant rate of growth in demand; and (b) a study of the economies and the penalties involved in accumulating backlogs of unsatisfied demand. The possibility of accumulating such backlogs raises considerable doubt with respect to Chenery's "excess capacity hypothesis".

Surprisingly, generalization (b) leads to considerably greater difficulties in analysis than (a). The use of probabilities to describe the growth process does little - if anything - to complicate matters. The probabilistic version of Chenery's model turns out to be closely related to the classical problem of gambler's ruin, and an extremely powerful tool can be borrowed from that area - the probability generating function for the duration of the game. Thanks to this generating function, the zero-backlog probabilistic model becomes no more difficult to study than the corresponding deterministic one. A direct implication is that a probabilistic growth course makes it desirable to install plant capacity of

---

* The author is deeply indebted for suggestions made by Martin Beckmann, Gerard Debreu, and Richard Rosett of the Cowles Foundation, by Ashley Wright of Standard Oil Co. (New Jersey), and by Robert Donovan, Donald MacArthur, and Gifford Symonds - all of Esso Standard Oil Co.

a somewhat larger size than would be optimal if demand were growing at a steady rate equal to the expected value of the probabilistic increments. Uncertainty, in this sense, has a stimulating effect upon the magnitude of individual investments.

Going beyond Chenery's model to include the possibility of backlogs, it turns out that there is a curious ambiguity in the effects of an increase in the variance of demand. Once the possibility of backlogs is admitted, an increase in variance can even lead to a decrease in the optimal size of individual installations.

2. The deterministic model - no backlogs in demand

In order to provide a reference point for discussion of the more difficult cases, Chenery's deterministic model itself will first be reviewed. Following this discussion will come the modifications involving (a) probabilistic growth and (b) the possibility of accumulating backlogs in demand. Chenery's model grew out of his studies of the natural gas transmission industry - a sector characterized by rapid growth and by substantial economies of scale in pipeline construction and operation. Much the same situation seems to prevail in the case of oil pipelines [8], the telephone industry [7], highway construction, electric power generation, petroleum refining, and chemicals processing [6]. Figure 1 charts the course of demand and of capacity over time under the following simplifying assumptions: (1) that demand grows linearly over time; and (2) that whenever demand catches up with the existing capacity, x units of new capacity are installed.* (The demand at t_o is denoted by D_o.)

* Chenery and Cookenboo [8] both point out that the concept of installed capacity is a slippery one - even when dealing with such a honogeneous facility as a gas or an oil pipeline. Once a line of given diameter has been laid, new pumping equipment can be added - enough to raise the ultimate installed capabilities to a level of perhaps two or three times the initial amount. From the viewpoint of our model, it seems best to regard the decision variable x as a measure of the ultimate rather than the immediate amount of pumping capacity installed. In defense of this shortcut, it should be noted that on an optimum-diameter line, all pumping station equipment - according to Cookenboo's figures - generally comprises no more than 10% of the total initial pipeline costs. [8 , pp. 65, 82, 106.]

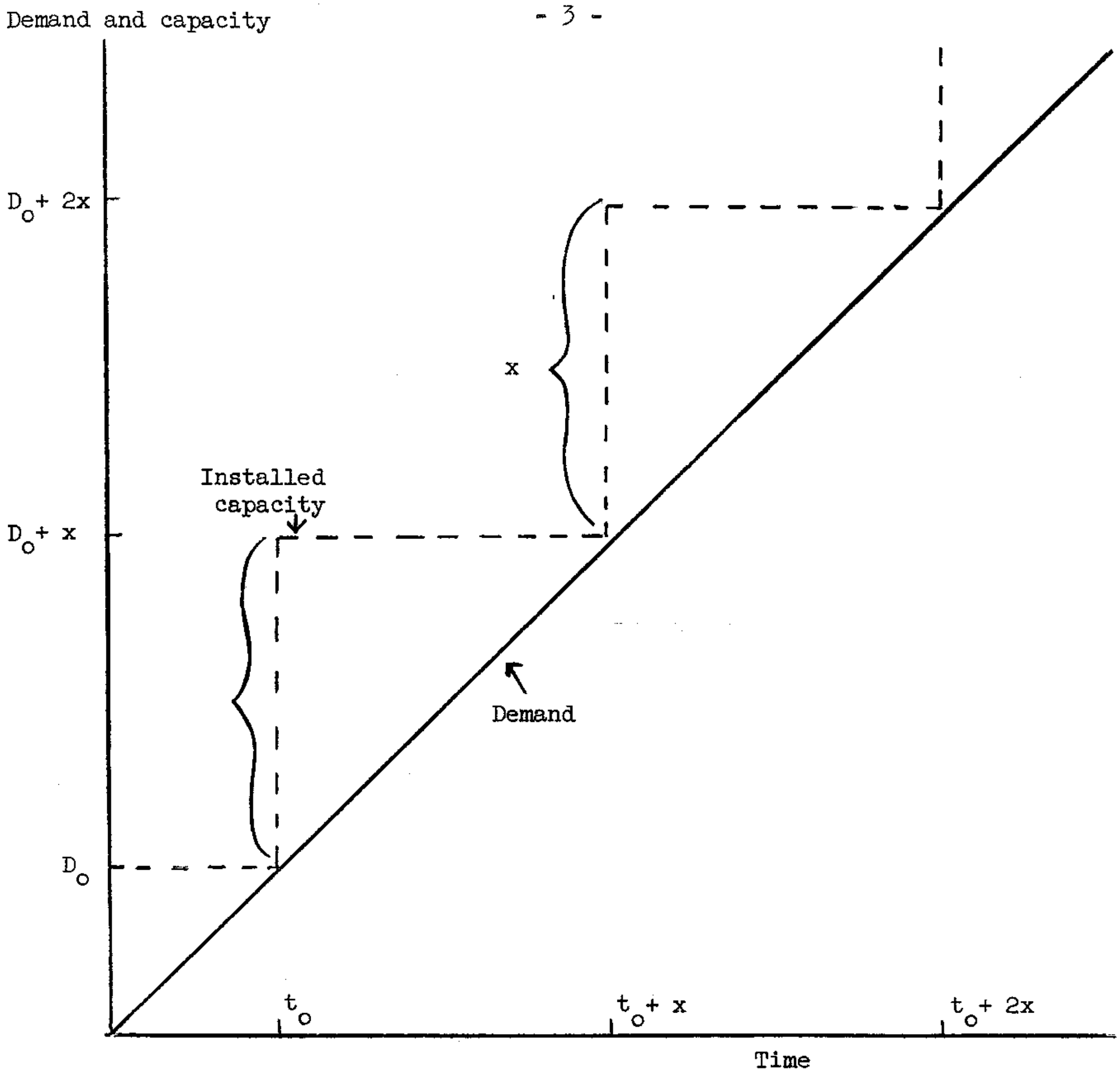


Figure 1. Growth of demand and of capacity over time.

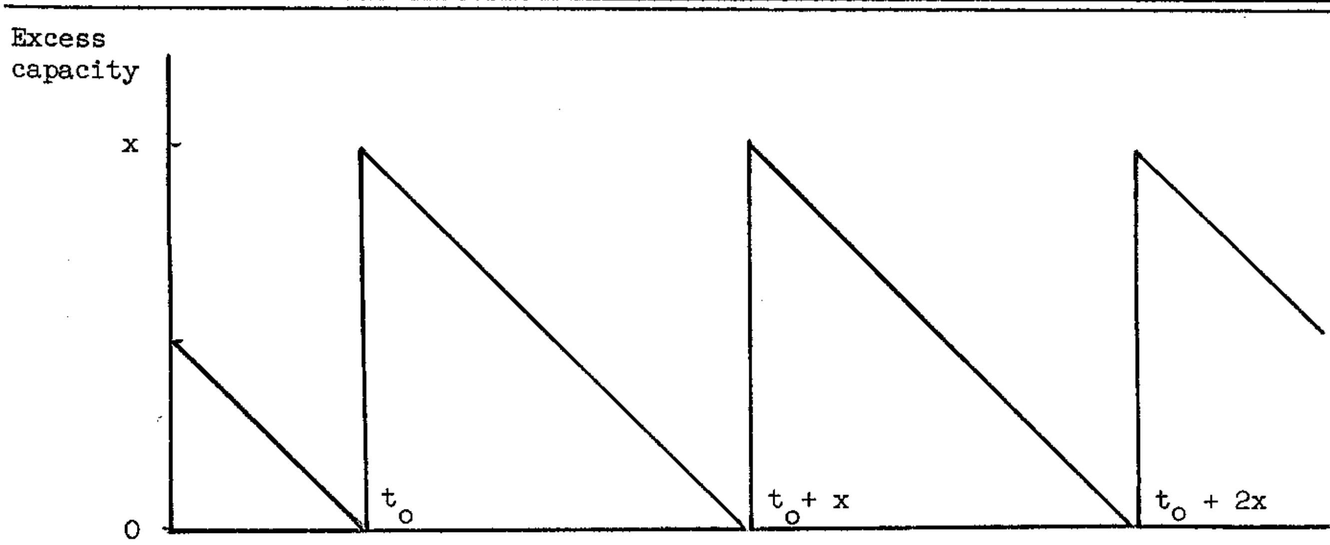


Figure 2. Evolution of excess capacity over time

Unlike Chenery, we shall assume that the planning horizon is an infinite one, and is not truncated after an arbitrary finite number of years. Excess capacity, when plotted on Figure 2, then displays a sawtooth pattern typical of the closely related Wilson-type inventory model. [1, pp. 252-255] If, for convenience, our physical unit of capacity and of demand is set equal to one year's growth in demand, this sawtooth cycle repeats itself every $\underline{x}$ years.

The installation costs that result from a single capacity increment of size $\underline{x}$ are assumed to be given by a cost function of an exponential sort:*

$$(2.1) \quad k \cdot x^{\alpha} \qquad\qquad (k > 0;\ 0 < \alpha < 1)$$

If, for example, $\alpha = 1/2$, this cost function says that a pipeline capable of handling 16 years' worth of growth in demand is only twice as expensive as one that can accommodate four years' worth. The existence of such substantial economies of scale implies the desirability of building new capacity considerably in advance of demand. But how much in advance? Here the discounting of future costs becomes crucial.

Without discounting, it would be perfectly sensible to spend a dollar now in order to save a dollar's worth of costs either next year or ten years from now, or 100 years hence. Under such circumstances, there is no limit to the size of line which it pays to build. With discounting,

---

* This cost function corresponds to Chenery's equation (3), p. 6 [5]. A square-root law ($\alpha = 1/2$) would be implied by the geometrical relationship between the cross-section area and the circumference of a circular body such as a pipe; a two-thirds law by the relationship between the volume and the surface area of a sphere.

on the other hand, this paradox can be sidestepped. Discounting reflects the very real fact that in any business enterprise there are always competing opportunities for the use of limited investment funds - alternatives with yields that are ordinarily far in excess of those anticipated on gilt-edge securities. It is fundamentally on account of this scarcity that we must consider the present value of a dollar due in $\underline{t}$ years' time as lower than that of one due at an earlier date.* In principle, of course, any monotone decreasing, non-negative formula could be employed as a present-value function. Here, however, we shall follow tradition in adopting the expression β^t for the present value of a dollar due $\underline{t}$ years in the future. $(0 < \beta < 1.)$ Hereafter, the fraction β will be known as the "discount factor".

As a time origin for subsequent calculations, it will be convenient to take any such point as t_o or $t_o + x$ or $t_o + 2x$ on Figure 2 - a point at which the previously existing excess capacity has just been wiped out. Such an event will be known hereafter as a "point of regeneration". Note that when we have reached $t_o + x$, the future looks identical with the way it appeared $\underline{x}$ units of time previously.** Then if

---

* Gifford Symonds has suggested an additional reason for the discounting of future costs - the expectation of continuing progress in pipeline technology. If the general price level remains constant, it is reasonable to suppose that in, say, 10 years' time the cost of building a line with a capacity of $\underline{x}$ units will be significantly cheaper than the cost of such a line today. The proviso about constancy of the general price level is important. If one is a believer in the inevitability of creeping inflation, one's discount factor should be higher than otherwise, and one should be quite anxious to incur debts that are fixed in money terms.

** A quote from Bertrand Russell seems rather pertinent here:

> ....our wishes can affect the future but *not* the past, the future is to some extent subject to our power, while the past is unalterably fixed. But every future will some day be past: if we see the past truly now, it must, when it was still future, have been just what we now see it to be, and what is *now* future must be just what we shall see it to be when it has become past. [11, p. 27]

we say that $C(x)$ is a function of $\underline{x}$ that represents the sum of all discounted future costs looking forward from a point of regeneration, we may write down the following recursive expression:

$$(2.2) \qquad C(x) = k \cdot x^{\alpha} + \beta^{x} C(x)$$

The first term on the right-hand side indicates the installation costs incurred directly at the beginning of the current cycle. (See equation (2.1).) The second term measures the sum of all installation costs incurred in subsequent cycles, and discounts these from the next point of regeneration back to the present one - a difference of $\underline{x}$ years. From (2.2), it follows directly that:

$$(2.3) \qquad \frac{C(x)}{k} = \frac{x^{\alpha}}{1 - \beta^{x}}$$

Differentiating $\log C(x)$ with respect to $\underline{x}$, and setting the result equal to zero:

$$\frac{d \log C(x)}{dx} = \frac{\alpha}{x} + \frac{(\log \beta)\, \beta^{x}}{1 - \beta^{x}} = 0$$

or $$(2.4) \qquad \alpha = \frac{\hat{x}}{\beta^{-\hat{x}} - 1} (- \log \beta)$$

where $\underline{\hat{x}}$ denotes the optimal size of installation.

The reader can verify for himself that (2.4) is not only a necessary condition, but also a sufficient one to ensure the determination of a unique minimum-cost solution. With this equation, the optimal capacity increment $\hat{x}$ may be determined for any combination of the two parameters α and β - a cross-plot being provided in Figure 3. From this figure, if one were interested in the economies-of-scale effect, he would observe, say, that when $\alpha = 2/3$ and $\beta = .85$, the optimal value $\hat{x}$ is approximately 5 years' worth of demand. With the discount factor β unchanged, but with the economies-of-scale factor at a level of 1/2, $\hat{x}$ rises to almost 8 years.

This deterministic model is one that lends itself readily to sensitivity-testing. To find out how the optimal level $\hat{x}$ is affected by changes in β, it is enough to transpose terms in equation (2.4) and equate the differentials. This operation yields:

$$(2.5) \quad \frac{d\hat{x}}{d\beta} = \frac{\hat{x}}{-\beta \log \beta} > 0$$

The derivative $\frac{d\hat{x}}{d\beta}$ is clearly positive for positive values of $\hat{x}$ and fractional values of β. The higher the discount factor (i.e., the lower the implicit cost of capital), the greater will become the optimal size of each installation.

3. The probabilistic model - no backlogs

With this background, we are in a position to discuss the case of probabilistic growth - still ruling out the possibility of deliberate backlogs in demand. Just as in the deterministic model, the mean annual rate of growth is taken to be the unit of physical measurement. The total expected growth over μ years therefore remains μ units. In

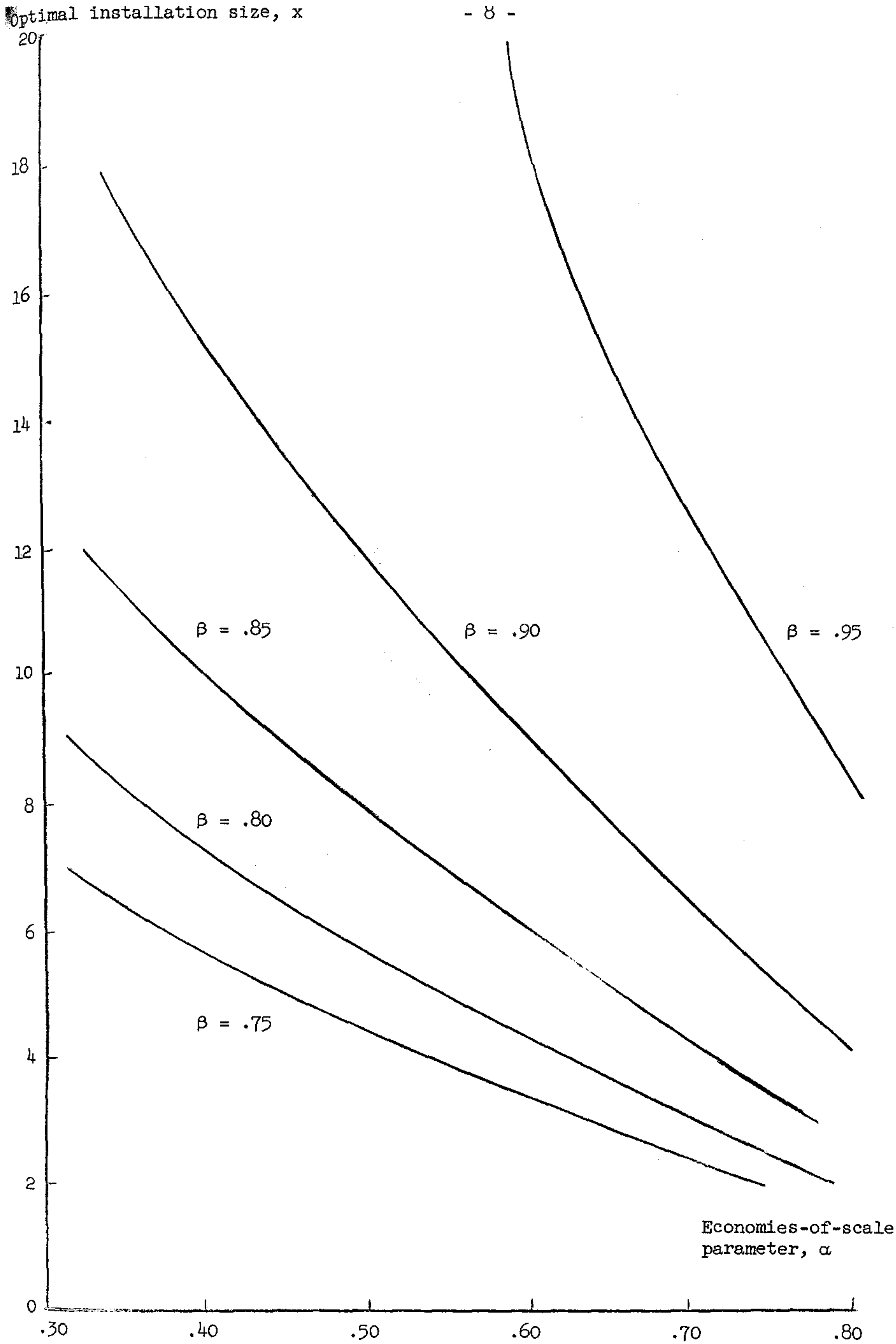


Figure 3. Optimal installation size: zero backlog assumption.

this case, however, we view the actual growth taking place in μ years not as a single-valued outcome, but rather as a random variable - the outcome of n independent Bernouilli trials. At each of these trials, the probability of a unit increase in demand is said to be p, and the probability of a unit decrease, q. Thus if one were considering a one-year period with just two points at which total demand could change, μ and n would be respectively one and two. To keep the expected amount of growth during this time equal to μ, it is necessary to set $p = \frac{1}{2} \cdot \left[1 + \frac{\mu}{n}\right] = \frac{3}{4}$. The total growth will be a binomially distributed random variable that can take on any one of three values ranging from minus two to plus two units:*

total one-year change in demand	-2	0	+2
respective probabilities	1/16	6/16	9/16

$$\text{expected change in demand} = n(p-q) = \mu = 1 \text{ unit}$$

$$\text{variance of change in demand} = \text{var} = 4\,npq = n - \frac{\mu^2}{n} = \frac{3}{2} \text{ units}$$

The basic scheme, then, is one in which time is broken down into artificial sub-periods - each $\frac{\mu}{n}$ years in length. Within each such time interval, demand either increases or else decreases by one unit. The respective probability for each of these outcomes is $\frac{1}{2}\left[1 + \frac{\mu}{n}\right]$ and $\frac{1}{2}\left[1 - \frac{\mu}{n}\right]$. Through this device, we are enabled to represent any stochastic growth process in which the individual μ-year increments follow the binomial law of distribution. Furthermore, the two

---

* There is no real necessity for working with the mean annual rate of growth as the unit of change at each Bernouilli trial. If an investigator wished to make his model appear less "lumpy", he could just as well have chosen the mean monthly rate or the mean daily rate as his unit of change.

parameters μ and n may be adjusted so as to provide for any desired ratio between the variance and the mean growth.

Suppose now that some specific values have been assigned to μ and n. From these, we determine directly the duration of the individual sub-periods (μ/n years) and the probability, p of a unit increment in demand during any one period. Let $u_{x,t}$ represent the probability with which t such units of time will elapse before the point at which demand first exceeds the initial level by x units. In gambler's ruin terminology, this is the probability with which t "trials" are needed in order for a gambler to go broke - a gambler whose initial capital is x and who is playing against an adversary with infinite wealth. At each stage of such a game, the gambler would lose one unit of capital with a probability of p, and gain one unit with a probability of $q = 1-p$. [10 , pp. 311-21.] The following relationship may therefore be written:

$$(3.1) \quad u_{x,t+1} = p\, u_{x-1,t} + q\, u_{x+1,t} \qquad (x \geq 1\ ;\ 0 \leq t \leq \infty)$$

Next we make use of a dummy variable s. $(0 < s < 1.)$ Multiplying both sides of (3.1) by s^{t+1} and summing over all t :

$$(3.2) \quad \sum_{t=0}^{\infty} s^{t+1} u_{x,t+1} = ps \sum_{t=0}^{\infty} s^{t} u_{x-1,t} + qs \sum_{t=0}^{\infty} s^{t} u_{x+1,t} \qquad (x \geq 1)$$

The generating function $U_x(s)$ of the probability sequence $u_{x,t}$ is defined by:

$$(3.3) \quad U_x(s) = \sum_{t=0}^{\infty} s^{t} u_{x,t}$$

For $x \geq 1$, $u_{x,0} = 0$. As a result, the left-hand side of (3.2) equals $U_x(s)$, and the entire equation may be rewritten in terms of the generating functions:

$$(3.4) \quad U_x(s) = ps\, U_{x-1}(s) + qs\, U_{x+1}(s) \qquad (x \geq 1)$$

Now (3.4) is a second-order linear difference equation. Its characteristic equation has two roots:

$$(3.5) \quad \lambda_1 = \frac{1 + \sqrt{1 - 4\,pqs^2}}{2\,qs}, \qquad \lambda_2 = \frac{1 - \sqrt{1 - 4\,pqs^2}}{2\,qs}$$

The general solution is consequently of the form:

$$(3.6) \quad U_x(s) = A(s)\,\lambda_1^x + B(s)\lambda_2^x$$

where $A(s)$ and $B(s)$ are constants whose values depend upon $\underline{s}$ and also upon the boundary conditions for $U_x(s)$. Now these boundary conditions are twofold: first, that $U_x(s)$ not exceed unity, and second, that $U_0(s) = 1$. Since $\lambda_1 > 1$ and $0 < \lambda_2 < 1$, the upper bound upon $U_x(s)$ can be ensured only by setting the constant $A(s)$ equal to zero. And to have $U_0(s) = 1$, the constant $B(s)$ must be unity. With these simplifications, the generating function (3.6) becomes:

$$(3.7) \quad U_x(s) = \lambda_2^x$$

Note that λ_2 is a function of $\underline{p}$ and $\underline{s}$ alone, and that it is independent of the quantity $\underline{x}$.

The weary reader will be relieved to learn that we are at last ready to make an interpretation of the dummy variable s and of the generating function $U_x(s)$ in terms of our capacity expansion model. The mysterious dummy variable s is to be regarded as nothing but the discount factor for a time period μ/n years in length. ($s = \beta^{\mu/n}$.) Then if we allow $u_{x,t}$ to represent the probability with which t such units of time will elapse between two successive points of regeneration - points between which the total demand grows by an amount x - the probabilistic analogue of (2.2) may be written:

$$(3.8) \qquad C(x) = kx^{\alpha} + \sum_{t=0}^{\infty} s^{t} u_{x,t} C(x)$$

Just as in the earlier deterministic case, the first term on the right-hand side equals the present cost of installing a facility of capacity x. The second term indicates the probability with which the next point of regeneration will occur in t units of time, discounts the corresponding cost back to the present, and sums up over all t. As in the earlier case, the function $C(x)$ gives the expected present value of all costs incurred over the indefinite future - as measured from a point of regeneration. From (3.8) and from the definition (3.3):

$$\frac{C(x)}{k} = \frac{x^{\alpha}}{1 - U_x(s)}$$

And by (3.7), this becomes:

$$(3.9) \qquad \frac{C(x)}{k} = \frac{x^{\alpha}}{1 - \lambda_2^{x}}$$

According to (3.9), then, the probabilistic model is not one bit more difficult to analyze than the deterministic one. All that has to be done is to regard the quantity λ_2 as the adjusted annual discount factor, and to insert this in place of β in equation (2.4) or else in Figure 3. From this also, it is easy to show that the greater the variance of the growth in demand, the greater will be: (1) the minimal level of discounted costs $C(x)$ and (2) the greater will be the optimal size of the capacity increments, $\hat{\underline{x}}$.

<u>Proof</u>

Hold $\underline{\mu}$ constant, and increase the value of $\underline{n}$. $(n \geq \mu.)$ With $\underline{\mu}$ constant, the expected amount of growth over $\underline{\mu}$ years remains $\underline{\mu}$. However, as the integer $\underline{n}$ increases, so does the variance. The relationship between $\underline{n}$ and the variance is one-to-one and monotone increasing. This may be verified from the fact that:

$$\text{var}(\mu, n) = 4\ npq = n - \frac{\mu^2}{n}$$

$$\text{var}\ (\mu, n+1) - \text{var}(\mu, n) = 1 + \frac{\mu^2}{n(n+1)} > 0$$

Next, we have to show that for constant $\underline{\mu}$, λ_2 is a monotone increasing function of $\underline{n}$ - i.e., a monotone decreasing function of the ratio μ/n. No attempt will be made to prove this rigorously. Instead, the reader is referred to Figure 4 - a plot of λ_2 versus the ratio $\frac{\mu}{n}$.* Finally, we observe that if the variance and λ_2 are both monotone increasing functions of the positive integer $\underline{n}$, then they are monotone increasing functions of one another. In other words, the greater the variance, the greater the value of λ_2 . Referring back to (3.9), this proves directly assertion (1) - the greater the variance, the greater will be the minimal level of discounted costs.

In order to prove assertion (2), we simply return to the sensitivity analysis at the end of the preceding section. According to (2.5), the optimal size of installation increases with the discount factor $\underline{\beta}$. In our probabilistic model, we have already shown that λ_2 may be viewed as nothing but an "adjusted" discount factor. Hence assertion (2): the greater the variance, the greater will be λ_2 and also the optimal value, $\hat{\underline{x}}$. This completes the proof.

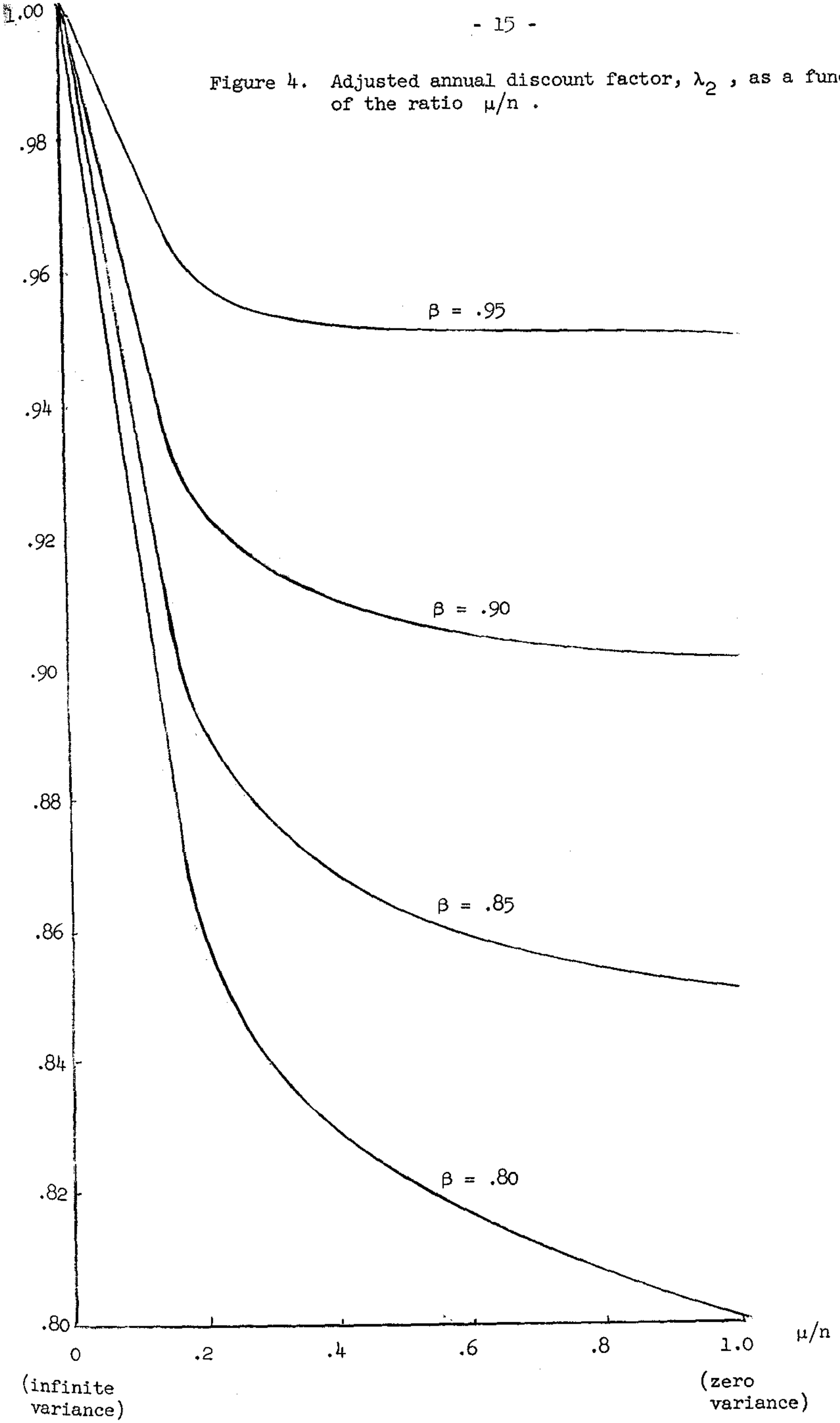
---

* When $\frac{\mu}{n} = 1$, we have the case of complete certainty - i.e., zero variance. Both the numerator and the denominator of the expression for λ_2 vanish when $\frac{\mu}{n} = 1$. It is easy, however, to show that as $\frac{\mu}{n}$ approaches 1 , the expression for λ_2 approaches the value of β .

$$\lambda_2 = \frac{1 - \left[1 - 4\, pqs^2\right]^{1/2}}{2\, qs} = \frac{1 - \left[1 - (1 - \mu^2/n^2)\, \beta^{2\mu/n}\right]^{1/2}}{(1 - \mu/n)\, \beta^{\mu/n}}$$

Differentiating the numerator of this expression with respect to the ratio μ/n, we obtain $-\beta^2$ when $\mu/n = 1$. Differentiating the denominator, we obtain $-\underline{\beta}$. The ratio of these two derivatives is $\underline{\beta}$, the limiting value of λ_2 - a result that completely accords with our intuition for the case of zero variance.

Figure 4. Adjusted annual discount factor, λ_2 , as a function of the ratio μ/n .



To illustrate these results, Table 1 provides a few calculations for several alternative values of the ratio μ/n . In each of the calculations presented in this table, the expected growth of demand over a μ-year period is, of course, identical - namely μ units. (With μ constant, an increase in n corresponds to a decrease in the ratio μ/n.) Note then that as n increases, so does the variance of demand, the optimal size of installation, and the minimum value of expected discounted costs.

Other things equal, our model indicates that the riskier the growth in demand, the larger ought to be the amount invested in each installation. To some, this result will seem to fly in the face of common sense. However, to those familiar with models of inventory stockage under conditions of probabilistic demand [e.g., 1, pp. 256-259], this should come as no paradox. In both the capacity model and the inventory case, the greater the risk of running out of capacity or out of inventory in a specified period of time, the greater the amount which it pays to invest in order to avert this contingency.

Table 1. Variance of demand and optimal installed capacities (μ = constant, α = .50, and β = .85)

μ/n	.2	.4	.6	.8	1.0
$p = 1/2\ (1 + \mu/n)$	.6	.7	.8	.9	1.0
$q = 1/2\ (1 - \mu/n)$	.4	.3	.2	.1	0
variance = $\sigma^2 = 4\ npq$	4.8000μ	2.1000μ	1.0667μ	.4500μ	0
$s = \beta^{\mu/n}$	.9681	.9371	.9069	.8780	.8500
adjusted annual discount factor = $\lambda_2 = \frac{1 - \left[1 - 4\ pqs^2\right]^{1/2}}{2\ qs}$	.8820	.8673	.8594	.8542	.8500
optimal installed capacity = $\hat{x}$	10.0	8.8	8.3	8.0	7.7
minimum expected discounted costs = $\frac{C(\hat{x})}{k} = \frac{\hat{x}^{\alpha}}{1-\lambda_2^{\hat{x}}}$	4.422	4.153	4.026	3.947	3.887

4. <u>The deterministic model - backlogs considered</u>

It is now time to examine the zero backlog assumption in a more critical way, and to explore the implications that result from discarding it. The zero backlog assumption seemed especially appropriate for the industry described by Chenery - natural gas transmission. In the case of this industry, it was reasonable to suppose that since the demand for the delivered product comes largely from individual homeowners, such individuals - if unable to obtain natural gas fuel at the time their home is initially constructed - would thereafter constitute a rather dubious sales prospect. For an individual homeowner, the initial outlay required for conversion from liquid fuels to natural gas could easily outweigh any benefits that he might conceivably derive from the switch.

This irreversibility phenomenon means that if a natural gas transmission line is operating at full capacity and if that capacity is kept unchanged, then demand for the delivered product will also remain constant. Chenery was probably quite right to have assumed that the gas industry's customers cannot be backlogged. Any attempt to do so would only result in their switching allegiance to an alternative fuel.

Even when customers cannot readily shift over to a competing product - as in the case of telephones, water, and electric power for residential purposes - it may still be sensible for the business enterprise to plan its investment outlays under the assumption of zero backlogs in demand. Certainly from the public relations standpoint, a utility company would be well advised to keep its capacity ahead of residential demand - even though its customers cannot easily rig their own telephone lines, dig their own wells, or generate their own power. In all these cases, the assumption of zero backlogs seems like quite a reasonable starting point. From this assumption, together with the economies of scale phenomenon, Chenery derives his "permanent" excess capacity hypothesis: ". . . excess capacity will occur even with perfect forecasting; this may be called 'optimum' overcapacity." [5, p. 27].

Despite the impressive list of sectors just noted, it would be a mistake to suppose that the assumption of zero backlog possibilities is a universally valid one. The economist who is accustomed to work with a downward sloping price-demand curve will certainly find it just as reasonable to believe that backlogs are admissible, and that they are accompanied by some kind of penalty cost to the firm. The zero backlog model then turns out to be a special case - the case in which backlog costs are infinite. Everything hinges upon the penalty cost assumption.

To a petroleum transporter, for example, these penalties are far less than infinite. If he is unable to ship crude or refined products via a pipeline, there is in almost all cases a transportation alternative available - tankship, barge, railroad tank-car, or tank truck. The penalty for failing to have enough pipeline capacity is simply the difference between the short-run marginal operating cost of the pipeline and the marginal cost of using the alternative mode of transport. No irreversibility effects seem significant here. As soon as new pipeline capacity becomes available, the oil transporter will not hesitate to switch over from the high-cost mode that is temporarily in use. The change-over costs would be negligible.

This kind of reasoning is surely not confined to petroleum pipelines alone. A steel producer might find that the penalty for being short of capacity in one section of the country would amount to nothing more than an increase in the amount of freight absorption needed to supply the region from a more distant point. Alternatively, the shortage penalty might consist of the profits foregone in being unable to bid on such marginal business as a large construction project or in the export market. In none of the examples just cited would it be reasonable to assume that the steel company loses permanent customers. The penalty for being short of capabity is of a temporary nature, and is confined to the period of full-capacity operations.

In graphical terms, the analogy with Figures 1 and 2 is shown on Figures 5 and 6. Just as in the earlier model, we assume that demand grows linearly at the rate of one physical unit per year. Again, x units will denote the size of each new installation and the points t_o , t_{o+x} , t_{o+2x} , . . . still mark the points of regeneration: the points at which excess capacity has just been wiped out. The entire difference between this and the earlier case is that we allow excess capacity to become negative here - in other words, permit backlogs of demand. Once such backlogs become admissible, there is no longer any *a priori* reason to believe in the necessity of Chenery's excess capacity hypothesis. With sufficiently low penalty costs, it is even conceivable that excess capacity will, on the average, be negative.

Figures 5 and 6 have been drawn on the assumption that whenever the backlog in demand grows to y units (that is, whenever excess capacity

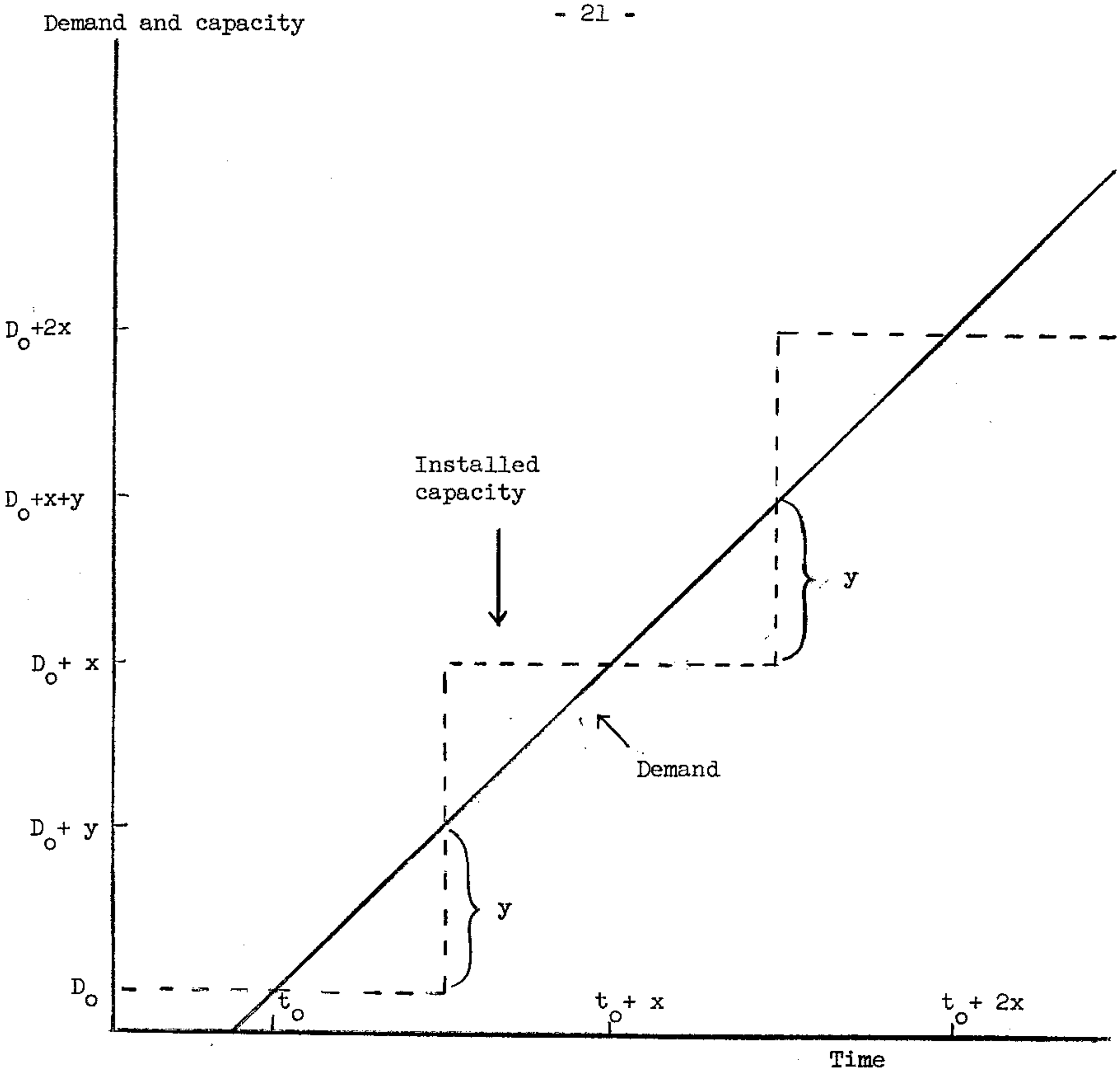


Figure 5. Growth of demand and of capacity over time.

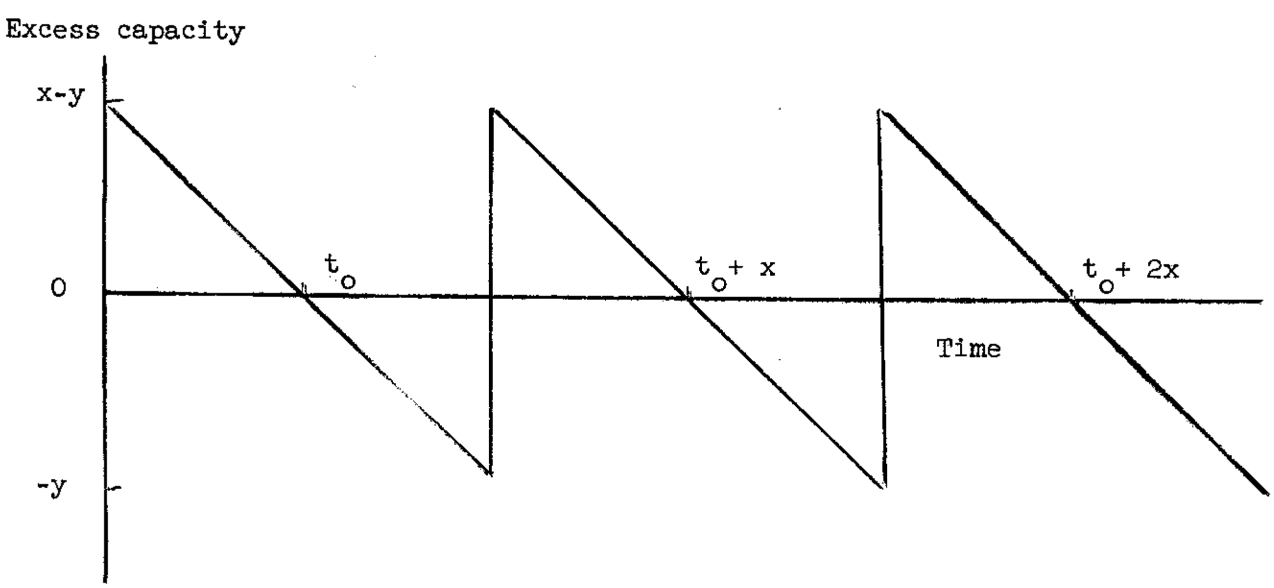


Figure 6. Evolution of excess capacity over time.

equals minus $\underline{y}$), a new facility is built - one of size $\underline{x}$. We now have two decision variables: $\underline{x}$, the size of each installation, and $\underline{y}$, the "trigger" level for backlogs in demand.* Penalty costs will be assumed strictly proportional to the dummy variable $\underline{z}$, hereafter employed to denote the size of the backlog.

Looking forward into the future from a point of regeneration, total discounted costs are a function of both $\underline{x}$ and $\underline{y}$. If we denote these discounted costs by $C(x,y)$, the expression that corresponds to (2.2) is as follows:

$$C(x,y) = \sum_{z=1}^{y} (cz)\beta^{z} + (kx^{\alpha})\beta^{y} + \beta^{x} C(x,y) \tag{4.1}$$

where $\underline{c}$ represents the penalty costs per unit of backlog.

It is easy to see that when demand is growing steadily at the rate of one unit per year, a backlog of size $\underline{z}$ occurs exactly $\underline{z}$ years after a point of regeneration. The first term on the right-hand side of (4.1) therefore measures the discounted sum of all penalty costs incurred during the course of a single construction cycle. The second term measures the installation costs, and discounts them $\underline{y}$ years back to the beginning of the cycle. Finally, the last term indicates the present value of all costs incurred in subsequent cycles, and discounts this value over a period of $\underline{x}$ years. From (4.1), we readily obtain:

$$C(x,y) = \frac{1}{1-\beta^{x}} \left[c \sum_{z=1}^{y} z\beta^{z} + kx^{\alpha}\beta^{y} \right]$$

---

* Any reader will note the striking similarity between this and the Ss theory of optimal inventory policy. One important difference tends to be concealed in the deterministic form of the two models. A replenishment lag is characteristic of the inventory studies, [1, 2]. In the interests of simplicity, however, the corresponding feature - a construction lag - is ignored in the present paper.

Dividing through by k in order to eliminate one parameter, and christening the ratio c/k with the name γ, we finally have the cost expression to be minimized:

$$(4.2) \qquad \frac{C(x,y)}{k} = \frac{1}{1-\beta^{x}} \left[\gamma \sum_{z=1}^{y} z\beta^{z} + x^{\alpha}\beta^{y} \right]$$

Expression (4.2) involves three parameters: α, the economies-of-scale factor; β, the discount factor; and γ, the penalty factor. Minimization of (4.2) with respect to both x and y could conceivably have been accomplished by calculus methods as in the earlier one-variable case, but this approach seemed rather clumsy.* Instead, refuge was taken in numerical methods. An electronic computer** evaluated $C(x,y)$ for a large number of combinations of x and y, and reported the minimum for each specified set of values of α, β, and γ. The results of three such calculations are shown in Table 2. As in the case of Table 1, the parameters α and β were set at .50 and at .85, respectively. One word of caution about the construction of this table: the decision variables x and y were restricted to integer values.

---

* If we are willing to make the approximation that $\log \beta = \beta - 1$, and then set the two partial derivatives of $C(x,y)$ equal to zero, we obtain two simultaneous nonlinear equations for the optimal values $\hat{x}$ and $\hat{y}$:

$$\beta^{\hat{y}} = [(-\log \beta)\ \beta^{\hat{x}}\ C(\hat{x},\hat{y})] \div \alpha\ \hat{x}^{\alpha-1} \qquad (4.3a)$$

and $$\hat{y} \doteq \frac{(1-\beta)}{\gamma\ \beta}\ \hat{x}^{\alpha} \qquad (4.3b)$$

From (4.3b) it is clear that the optimal y vanishes only when γ, the unit penalty cost, becomes infinite.

** The machine was the I.B.M. 650 located in the Yale University Computing Center. For help in programming and in running the machine, I am indebted both to M. Davis, Director of the Center, and to D. Ciosek.

Table 2. Shortage penalty costs, optimal backlog levels, and installed capacities. (α = .50 and β = .85)

shortage penalty costs = γ	∞	.20	.10
optimal installed capacity = $\hat{x}$	8	10	13
optimal backlog level = $\hat{y}$	0	2	6
minimum discounted costs = $\frac{C(\hat{x},\hat{y})}{k}$ (see (4.2).)	3.888	3.416	2.765

Table 2 and similar calculations strongly suggest one conjecture - a conjecture which I have not, however, attempted to demonstrate rigorously - that a decrease in the penalty cost factor γ will always lead to an increase in the optimal levels, $\hat{\underline{x}}$ and $\hat{\underline{y}}$. This conjecture is supported by the general appearance of equations (4.3a) and (4.3b).

Both to an economist and to an operations researcher, it is likely that the shape of the cost function $C(x,y)$ is of even greater interest than the optimal values $\hat{\underline{x}}$ and $\hat{\underline{y}}$ themselves. Figure 7 contains a plot for the set of parameters corresponding to the third column of Table 2 ($\alpha = .50$, $\beta = .85$, and $\gamma = .10$). The optimal point, as indicated by both that table and this figure, leads to a cost of 2.765. The figure also gives an indication of how insensitive these costs are to a fairly wide set of values assigned to the decision variables $\underline{x}$ and $\underline{y}$.

An $\underline{x}$-value as high as 19.0 and a $\underline{y}$-value as high as 8.8 will only increase costs to a level of 2.850 - a matter of about 3%. From the viewpoint of the operations researcher and the business forecaster, this insensitivity is fortunate indeed. Even a substantial error in forecasting will not lead to an egregiously bad choice for the installation size or for the "trigger" level of the backlog.

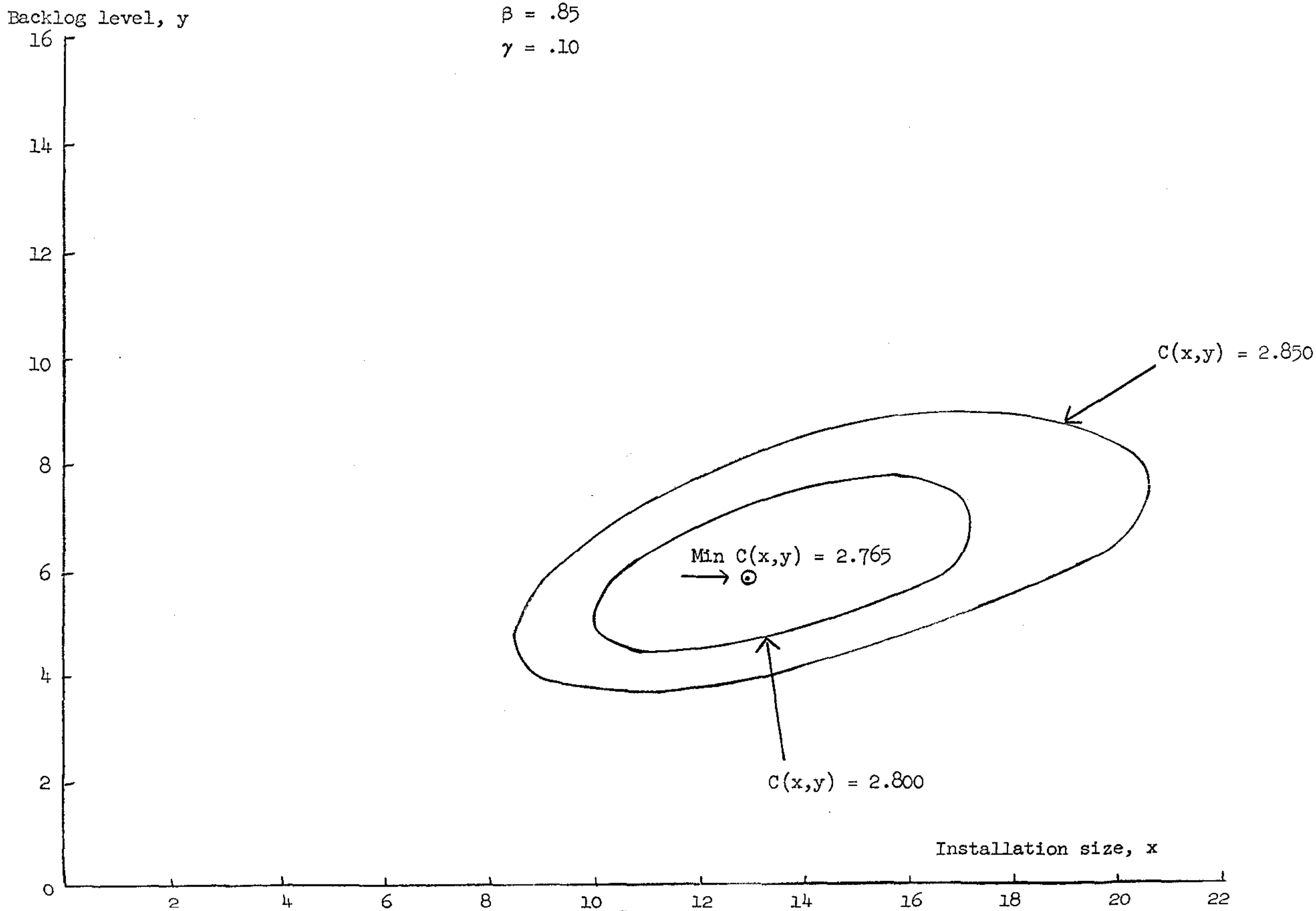
What is fortunate from the viewpoint of the businessman may be disastrous, however, from the viewpoint of an economist trying to forecast investment choices on the basis of an optimizing model. Even if the economist happens to hit upon the same values for α , β , and γ that are in the mind of the businessman, the latter will suffer no great penalty for deviating from the optimal path predicted by the economist for his behavior.

Figure 7. Discounted cost function, C(x,y)

$\alpha = .50$

$\beta = .85$

$\gamma = .10$



5. The probabilistic model - backlogs considered

The final stage of this investigation will consist of fitting together the two kinds of generalizations of Chenery's model: (a) probabilistic growth, and (b) backlogs in demand. Just as in the zero backlog case, we now assume that the mean annual rate of growth is the physical unit of measurement; that the actual growth taking place in μ years is not a single-valued outcome, but rather a random variable - the outcome of n independent Bernouilli trials; and that at each of these trials, the probability of a unit increment in demand is p and the probability of a unit decrease, q. The individual time units for our probabilistic model are μ/n years in length, and the discount factor for each such period will be denoted by s, $(s = \beta^{\mu/n})$.

As before, we shall let $u_{x,t}$ represent the probability with which t time units have elapsed at the time when total demand first exceeds the initial level of x units. Similarly, $u_{y,t}$ will denote the probability with which t time units are needed before the first occasion on which demand has increased beyond the initial level by y units. We already know the generating functions for these two probability distributions:

$$(5.1) \qquad U_x(s) = \sum_{t=0}^{\infty} s^t u_{x,t} = \lambda_2^x$$

$$(5.2) \qquad U_y(s) = \sum_{t=0}^{\infty} s^t u_{y,t} = \lambda_2^y$$

(Refer back to (3.3), (3.5), and (3.7).)

Now in order to deal with the backlog question, we shall have to introduce one additional piece of notation: $u_{y,t}(y-z)$. This symbol will denote the probability with which the backlog equals $\underline{z}$ at $\underline{t}$ time units after a point of regeneration. Why does the decision variable $\underline{y}$ enter into the definition of this probability? Because the process of building up a backlog will come to an end as soon as demand has increased by $\underline{y}$ units - that is to say, by an amount large enough to trigger off the construction of a new facility.

In random walk language, $u_{y,t}(y-z)$ is the probability with which a particle, starting $\underline{y}$ units above the origin, will on the t-$\underline{th}$ step be $\underline{z}$ units beneath its initial position, without having previously touched the absorbing barrier at the origin. Feller has already provided us with the generating function for this probability distribution [10, problem 16, p. 336]:

$$(5.3) \qquad U_y(s;\ y-z) = \sum_{t=0}^{\infty} s^t u_{y,t}(y-z) = \left[V_o\ \lambda_2^z - \left(\frac{\lambda_2}{\lambda_1}\right)^y\ \lambda_1^z\right]$$

$$(y > z > 0)$$

where λ_1 and λ_2 are as determined earlier by (3.5) and where the parameter V_o is given by:

$$(5.4) \qquad V_o = [1-4\ pqs^2]^{-1/2}$$

The cost equation for our new model may be written down by direct analogy with the deterministic one (4.1):

$$(5.5) \quad C(x,y) = \sum_{z=1}^{y-1} (cz)U_y(s;y-z) + (cy)\lambda_2^y + (kx^\alpha)\lambda_2^y + \lambda_2^x \cdot C(x,y)$$

Total expected discounted costs, $C(x,y)$, will - as in the preceding cases - be measured from a point of regeneration, a point at which excess capacity equals zero. Now the first term on the right-hand side of (5.5) measures the expected discounted sum of all backlog penalties incurred after this point of regeneration and prior to the point at which the backlog reaches the critical level, y. (See Figure 6. Also equation (5.3).) The penalty cost summation extends over all possible backlog levels other than y itself: $z = 1, 2, . . , y-1$. Note that it is quite easy for the backlog to become negative at any time after a point of regeneration. Our cost expression simply says that whenever this happens (that is, whenever demand drops off enough to create some excess capacity), no additional outlays are incurred beyond those that were previously committed.

The second term on the right-hand side is also connected with penalty costs - those that occur just once each cycle at the point when the backlog reaches the triggering level, y. The appropriate generating function in this case is as given by (5.2). In random walk language, this is the generating function of first-passage times through a point y units below the initial position.

The third term on the right-hand side of (5.5) is the one having to do with construction costs during a single cycle. Just as with the second term, these costs are all incurred at the time of reaching the level y, and so the appropriate generating function is again (5.2).

Finally, the fourth term (that measuring the discounted sum of all costs incurred in subsequent cycles) refers to a cost that is dated as of the beginning of the following cycle. This cycle will begin whenever the total demand first increases by $\underline{x}$ units over the current level - i.e., whenever the $\underline{x}$ units of new capacity are, for the first time, fully utilized. In the fourth term therefore the appropriate generating function is (5.1).

For purposes of numerical analysis, the cost function (5.5) may be rewritten:

$$(5.6) \quad \frac{C(x,y)}{k} = \frac{1}{1-\lambda_2^x} \left\{ \gamma\, V_o \sum_{z=1}^{y-1} z\lambda_2^z - \gamma\, V_o \left(\frac{\lambda_2}{\lambda_1}\right)^y \sum_{z=1}^{y-1} z\, \lambda_1^z + \gamma y \lambda_2^y + x^\alpha\, \lambda_2^y \right\}$$

where γ again measures the ratio c/k, as in the deterministic calculations of the preceding section.

The numerical analysis of (5.6) is only slightly more complex than that of (4.2).* There are still just two decision variables, $\underline{x}$ and $\underline{y}$, and three economic parameters α, β and γ . The only additional feature is that in the present case we must also take account of the generating function parameters λ_1, λ_2, and V_o . As can be seen from the row headings of Table 3, these last-mentioned parameters all depend upon the ratio μ/n . (The symbol $\underline{\mu}$ represents the expected outcome of the $\underline{n}$ independent Bernouilli trials.) Just as in Table 1, we now examine the effects of increasing the variance while holding the expected increment in demand, $\underline{\mu}$, constant. Also held constant in Table 3 are the parameters α, β, and γ . Both α and β are set at the same levels as were used for the earlier calculations.

Table 3 would be of little interest if it merely confirmed for the variable backlog case what we already knew about the zero backlog model: that an increase in variance is inevitably accompanied by an increase in $\hat{x}$ the optimal installation size.

---

* In fact, the same I.B.M. 650 program written to solve (5.6) also handled (4.2).

Table 3. Variance of demand, optimal backlog levels, and installed capacities

(μ = constant, α = .50, β = .85, and γ = .10)

μ/n	.2	.4	.6	.8	1.0
$p = 1/2\ (1 + \mu/n)$	.6	.7	.8	.9	1.0
$q = 1/2\ (1 - \mu/n)$	.4	.3	.2	.1	0
variance = $\sigma^2 = 4\ npq$	4.8000μ	2.1000μ	1.0667μ	$.4500\mu$	0
$s = \beta^{\mu/n}$	.9681	.9371	.9069	.8780	.8500
(5.4) $V_o = \left[1\text{-}4\ pqs^2\right]^{-1/2}$	3.1556	1.9520	1.4531	1.1765	1.0000
(3.5) $\lambda_1 = \dfrac{1 + \left[1\text{-}4\ pqs^2\right]^{1/2}}{2\ qs}$	1.7003	2.6895	4.6533	10.5353	∞
(3.5) $\lambda_2 = \dfrac{1 - \left[1\text{-}4\ pqs^2\right]^{1/2}}{2\ qs}$	.8820	.8673	.8594	.8542	.8500
optimal installed capacity = $\hat{x}$	11	11	12	13	13
optimal backlog level = $\hat{y}$	2	3	4	5	6
minimum expected discounted costs = $\dfrac{C(\hat{x},\hat{y})}{k}$ (see (5.6).)	3.833	3.426	3.181	2.971	2.765

Table 3 provides an immediate counter-example to this conjecture. For $\mu/n = 1$, the variance is zero and the optimal installation size is one that equals 13 times the expected annual increment in demand. When the variance increases to 4.8000μ, $\hat{x}$ drops down to 11. An increase in variance - at least over a limited range - actually leads to a decrease in the optimal size of each individual facility.

No attempt will be made to establish in an ex post manner the intuitive plausibility of these new results. All that needs to be pointed out is the rather trite moral that theorems about a one-variable problem do not necessarily carry over to a two-dimensional analogue.

6. Summary and generalizations

For the benefit of the reader who has persevered through the sensitivity analyses that went with each of the four models analyzed here, it seems only merciful to recapitulate the chief results schematically:

Backlog assumption	zero		non-zero	
Demand growth assumption	deterministic	probabilistic	deterministic	probabilistic
Cost equation	(2.3)	(3.9)	(4.2)	(5.6)
$\frac{d\hat{x}}{d\beta}$	> 0 See (2.5), (3.5), and Fig. 4	> 0	?	?
$\frac{\Delta x}{\Delta\sigma^2}$	not applicable	> 0	not applicable	$\gtreqless 0$

In each of the four cases, the regeneration point technique - when coupled with the appropriate generating functions - made it a rather simple matter to relate total costs to the decision variables in a closed analytical form. There was no need to invoke the functional equation concept with which Bellman's work on dynamic programming has made us so familiar [4]. This is not in the least to underrate the importance of the functional equation approach - only to point out that other tools may also be quite useful in dynamic optimization.

Among the problems that seem closely related to the one discussed here would be not only cases of inventory stockage [1,2], but also models of equipment replacement [3, 12], and of forestry economics.* Each of these is characterized by a cyclical spurt of build-up activity (either ordering some inventory, or installing a new piece of equipment, or cutting down and reseeding a piece of forest land) - a spurt which is followed by a fairly long gestation period (rundown of inventory, accumulation of operating inferiority, or the growth of new trees) until the beginning of the next cycle. There is good reason to believe that generating functions could prove useful in the analysis of all such stationary cyclical activities as these.

As intriguing as the research prospects in allied areas might be, the fact remains that we are far from having exhausted the subject of capacity expansion in this paper. It is easy to think of certain

---

* The forestry problem was suggested to me in an unpublished paper by R. F. Keniston of Oregon State College. A similar case was, of course, investigated a good many years ago by Knut Wicksell [13, pp. 172-184].

generalizations that would be quite straightforward: It would be a fairly simple matter, for example, to replace the construction cost function (2.1) with a different one - just as long as the new function also exhibited economies of scale. (Without economies of scale, the whole rationale breaks down for the bunching together of investment activity.) Another fairly simple generalization would consist of replacing the proportional backlog penalty cost with an arbitrary non-linear function of the backlog size, $\underline{z}$. The generating function (5.3) would still be valid. The only difficulties would arise in the numerical optimization with respect to $\underline{x}$ and $\underline{y}$. It might no longer remain true that a local optimum was a global one.

Even if one wished to alter the probability structure underlying the demand growth pattern, there are certain alterations with which no great difficulty would be experienced. One might wish to assume, for example, that at each Bernouilli trial demand either increases by one unit with a probability of $\underline{p}$, or else stays constant with a probability of $\underline{q}$. (This assumption excludes the possibility of any decline in demand.) We then obtain a first-order difference equation for the generating function of first-passage times, and find it even easier to study than the case discussed here.

Another fairly straightforward generalization would be to the case of continuous rather than discrete time. One would then want to analyze a growth process in which, during a time interval of length dt , the probability of a one-unit demand increase was p dt , and the probability of a one-unit decrease was q dt. $(p + q < 1)$ This kind of process would call for the use of Laplace transforms in place of generating functions, but would otherwise be quite similar to the problem discussed here.

Now to enumerate several of the more intractable generalizations: Suppose that we wished to deal with a case in which, at each Bernouilli trial, demand either increased by 2 units with a probability of $\underline{p}$, or else decreased by 1 unit with a probability of $\underline{q}$. We would then be in the same difficulties as those which Dvoretsky, Kiefer, and Wolfowitz raised in connection with the Ss inventory policy. [9, pp. 189-196.] No matter what value we assign to the "trigger" level of excess demand, there is no assurance at all that the random-walk process will pass through this point on each cycle. Unless demand is constrained to move from one point to an immediately adjacent one, we cannot be sure that the construction process will be triggered off at the same backlog level during each successive cycle. Our problem would then become the much more difficult one of calculating the optimal installation size as a function of the backlog level.

One final kind of generalization that is appealing from the viewpoint of economic realism, but which leads to analytical difficulties is the following: Suppose that we wish to replace an arithmetic growth process with a geometric one. Instead of assuming that the unit increment remains constant over the indefinite future, we would want to say that this increment is an increasing function of time. Again we would run into trouble. Even if we then defined the trigger value and installation size as functions of time, we would still run into the Dvoretzky, Kiefer, Wolfowitz objection. No matter what trigger value function were chosen, there would still be no assurance that the random-walk process would pass through a trigger point on each successive cycle. The optimal installation size would, under these circumstances, have to be defined as a function of both the backlog size and also of the calendar date. A linear probabilistic growth pattern seems like a far easier thing to analyze than the geometric case.

References

1. Arrow, K.J., T. Harris, and J. Marschak, "Optimal Inventory Policy," Econometrica, July 1951.

2. Beckmann, M.J., and R.F. Muth, "An Inventory Policy for a Case of Lagged Delivery," Management Science, January 1956.

3. Bellman, R., "Equipment Replacement Policy," Journal of the Society for Industrial and Applied Mathematics, September 1955.

4. ________ Dynamic Programming, Princeton, 1957.

5. Chenery, H.B., "Overcapacity and the Acceleration Principle," Econometrica, January 1952.

6. Chilton, C.H., " 'Six-Tenths Factor' Applies to Complete Plant Costs," Chemical Engineering, April 1950, pp. 112-114.

7. Clark, P., "The Telephone Industry: A Study in Private Investment," Studies in the Structure of the American Economy, W. W. Leontief ed., New York, 1953, ch. 7.

8. Cookenboo, L., Jr., "Costs of Operating Crude Oil Pipe Lines," Rice Institute Pamphlet, April 1954.

9. Dvoretzky, A., J. Kiefer, and J. Wolfowitz, "The Inventory Problem: I. Case of Known Distributions of Demand," Econometrica, April 1952.

10. Feller, W., An Introduction to Probability Theory and its Applications, Vol. I, 2nd ed., New York, 1957.

11. Russell, B., Mysticism and Logic, Penguin ed., 1953.

12. Terborgh, G. W., Dynamic Equipment Policy, New York, 1949.

13. Wicksell, K., Lectures on Political Economy, Vol. I, London, 1934.