ECO 7997

Major Paper

Submitted as part of the requirement

for M.A. (Economics)

Submitted

to

The Department of Economics

University of Ottawa

by

G. Jeffrey Providence

August 29, 1990

TITLE:

Natural Monopoly: Is There A Case
For Local Service

Author:

G. Jeffrey Providence

CONTENTS

	Page
Chapter 1: Introduction	1
Chapter 2: Natural Monopoly: Issues	6
Chapter 3: Theory of Natural Monopoly	11
Chapter 4: Natural Monopoly: Estimation Methods	24
Chapter 5: Natural Monopoly: The Empirical Evidence	41
Chapter 6: Conclusions	58
Figures and Charts	66
Footnotes	75
References	80

Chapter 1

Introduction

The telecommunications industry, from the turn of the century to the early 1970's, was viewed as having an industry structure which results in natural monopoly. That is, the services of the industry demanded by society could be provided at least cost by a single firm. From the 1970's, the federal regulators, the Federal Communications Commission (FCC) in the United States and the Canadian Radio-Television and Telecommunications Commission (CRTC) in Canada, adopted a more competitive posture and began permitting entry to their respective country telecommunications markets. This change in direction, by the CRTC and FCC, appears to meet with the current view that those conditions which characterized the industry as a natural monopoly are no longer significant and more specifically, they do not exist in the various market segments of the industry.

In the USA, the FCC has permitted competition in the terminal and long distance telecommunications markets while in Canada, the CRTC has allowed competition in the terminal and private line markets. With these moves to open telecommunications markets to competition have come statements that the provision of local telephone service should also be opened to competition[1]. It is an examination of these statements, in regard to the provision of local telephone service, which forms the basis of this paper. These statements are of significance, since it is now argued by several authors that little, if any, efficiency would be sacrificed by permitting competitive entry. However, in an industry as large as the telecommunications industry small changes in efficiency can have a substantial impact on consumer welfare.

Associated with a discussion on monopoly are the issues of the abuse of the monopolist's market power and the presence of economic barriers to entry[2]. The absence of barriers to entry is considered the only condition which can constrain the firm in its exercise of market power. The literature on industrial organization suggests that there are three forms of economic barriers to entry. These arise as a result of cost advantages which stem from: economies of scale; economies of scope; and product differentiation. The first two conditions where they exist, enable the existing firm to enjoy cost advantages in meeting market demand and in the production and distribution of that firm's goods and services. Associated with these cost advantages in the production and distribution of the firm's goods and services are large, up-front, capital requirements which tend to act as barriers to entry.

The third type of economic barrier to entry, that of product differentiation, also results in cost advantages to the existing firm. Although telecommunications services tend to be homogeneous in nature, differences in quality, customer service, reputation, or the ability to offer complementary services are factors which can establish preference for the existing firm.

There are two principal approaches to the study of scale economies: engineering cost studies whose approach employs detailed specifications of technology; and econometric or empirical estimation of the structure of cost and production functions which provide a broad view of the relationship among the major aggregate variables. The engineering cost study approach is thought of as being best suited for studying scale economies in specific services while the econometric approach for studying economies of scale across the industry.

The econometric literature on telecommunications highlights two kinds of tests which are used to investigate the presence of natural monopoly. Usually, the presence of natural monopoly is determined by evidence of cost sub-additivity - a necessary and sufficient condition for natural monopoly. The first of these tests is based on global cost conditions developed by Baumol, Panzar and Willig. They derived two sets of conditions for cost sub-additivity, which are: first, the existence of product-specific economies of scale and economies of scale; second, declining average cost and trans-ray convexity of the cost function. Trans-ray convexity is a measure of the cost function under certain conditions which approximately hold the overall scale of production constant. In general, these tests require information about the costs of the separate production of each type of output. As a result of the extreme information demands of this test, for which data is usually not available, methods have been devised to extrapolate cost functions to zero output levels to facilitate the process. Faced with this problem, other authors, for example Evans and Heckman (1983), have developed tests to measure for cost subadditivity which do not require any data inference outside the range of data.

The objective of this paper is to investigate the evidence to date and determine whether that portion of the industry supplying local telephone service can be characterized as a natural monopoly. If it is found that the evidence indicates significant economies of scale and scope then a natural monopoly would be the most efficient industry structure under which local telephone service can be provided. However, if the results indicate no scale economies, then competitive entry can be considered a viable alternative. This review will be conducted by investigating the empirical findings to-date on economies of scale and scope, specifically in regard to the producton of local service by Bell Canada.

The study is divided into five chapters. Chapter 2 considers the issues associated with natural monopoly and the arguments which can be used to support or deny competitve entry. Chapter 3 reviews the economic theory which underlies the concept of natural monopoly and establishes the theoretical framework which would permit the existence of a natural monopoly in the telecommunications local service market segment. Chapter 4 discusses the methods of econometric estimation currently in use for estimating scale economies, economies of scope and the extent of market power. It also discusses the problems associated with the various functional forms used in the estimation process. Chapter 5 reviews the empirical results and makes comparisons with the experience in the United States. In Chapter 6, the conclusions are presented.

Chapter 2

Natural Monopoly: Issues

A natural monopoly is said to exist when economies of scale[3] exist for a firm's production technology for a well defined market. When economies of scale are found, it generally follows that one producer - a monopolist - should be able to satisfy all market demand at the least possible cost. However, the single producer can also achieve monopoly *profits* with prices much above average costs. In contrast, the static model of pure competition, where there exists many producers, prices are usually maintained at levels just sufficient to cover average costs, including the costs of attracting any necessary capital.

Several factors, however, are important when making the association between economies of scale and natural monopoly. These include issues that relate to the significance of scale economies and the extent of the monopolist's market power. These issues have been central to the debate surrounding the policies governing the telecommunications industry both in Canada and in the United States.

Chapter 2

Extent of Scale Economies

The presence of scale economies in itself in not sufficient to label an industry a natural monopoly. The extent of these scale economies or how significant they are, play an important role. Scale economies come in different forms. Different possibilities for the existence of scale economies are shown in figures 2.2, 2.3, and 2.4. The average cost curve shown in figure 2.2 declines steadily throughout the entire range of output, whereas in figure 2.3 the average cost curve declines very slowly over the range of output. In figure 2.4 the average cost curve declines very quickly over low levels of output and remains stable over the remaining levels of output. In this case, unit cost decreases occur very slowly or not at all after a basic threshold level of output has been achieved.

The significance of these cost curve patterns lie in the fact that cost economies are likely to be minimal if most scale economies are exhausted at low levels of output, as is the case demonstrated by figure 2.3 and to a greater extent by figure 2.4. In a industry which demonstrates cost characteristics similar to those shown in figure 2.3 and 2.4, it is plausible to have more than one firm in operation. In contrast, if the industry has a cost structure as shown in figure 2.2, then entry by another firm implies a significant loss of efficiency.

The importance of scale economies in policy matters is, therefore, an empirical issue. Any pertinent estimate of scale economies depends, in large measure, on the size of the market and on how large that market is expected to become. If for example, an industry had a cost curve which resembled that of figure 2.4 and the current market was sufficient to permit one firm to exhaust most of the economies of scale, then the decision as to whether or not to establish a legal monopoly would depend greatly on the expected market outlook for the industry. If growth was expected to be steady but not large, then the argument for the creation of a legal monopoly would be strenghtened.

Extent of Market Power

A major issue which must be considered before one can relate the significance of scale economies with natural monopoly is the extent to which market power accrues to the monopolist. Market power of the natural monopolist can be gauged by the elasticity of the demand curve with respect to price. If the demand for the monopolist's output is characterized by a relatively low elasticity, figure 2.5, then the monopolist has extensive market power. Under these conditions, the potential for extracting monopoly profits from the market is substantial. The reverse may be true in the case where the demand curve is relatively more elastic[4], as in figure 2.6.

There is an interaction between the extent of market power and the extent of scale economies. If for example, an industry is characterized by large scale economies but faces a relatively elastic demand curve, its limited market power may act to discipline the monopolist in terms of its price setting behaviour. If however, the monopolist faced a relatively inelastic demand curve with moderate scale economies, monopoly rents could certainly be extracted.

The monopolist exhibiting significant scale economies and market power may find that there exist factors which limit that market power. If there exists acceptable substitutes which are not regulated, they may act as a check on the monopolist's market power. This form of interindustry competition serves to increase the elasticity of demand for the product offered by the monopolist. Even though there may be room for only one producer within a given industry, other acceptable sources of supply from without the industry may be suitable for the same purposes. In this case, a rise in price of the monopolist's product may lead consumers to purchase the substitute hence limiting the monopolist's market power.

Market power may also be limited by the existence of alternative but slightly higher cost technologies. If the monopolist who exhibits scale economies attempted to charge monopoly prices, entry by suppliers using the alternative technology would tend to hold prices at levels closer to the monopolist's cost.

There are, then, two distinct issues with regard to the provision of Bell Canada's local service. The first is that which deals with the estimation of scale economies in the cost functions which are associated with the production of Bell's local telephone service and the second, that which deals with the extent of Bell Canada's market power in the provision of that service. The former relates to the matter of the most efficient means for the production of local service and the latter to the matter of whether the monopoly producing that service should be regulated. This paper is concerned with with the first of these issues.

Theory of Natural Monopoly

A firm transforms inputs such as labour and capital into products and services. The various techniques used by that firm for transforming its inputs into some final product or service defines its state of technology. Included are the methods used for managing the firm and those for physically changing the inputs into the final product. Typically, the cost function or the production function is the tool used for stating the technology used by the firm. The cost function, the tool used for analytical purposes in this paper, describes how production costs vary with changes in production levels and input prices when the most efficient combination of inputs and production techniques are used. Where a firm produces two products using capital and labour, the cost function can be denoted by $C(Q_1, Q_2, \overline{r}, \overline{w})$, where Q_1 is the level of output one, Q_2 the level of output two, $\overline{r}$ *and* $\overline{w}$ *are parametric factor prices associated with the production of* Q_1 *and* Q_2. When all firms use the same technology, the cost function has certain properties that can define the most efficient industry structure for producing output demand. These properties, and their usefulness in determining whether an industry is a natural monopoly, are discussed below. First, the single-product techology will be considered followed by the more complicated case of the multi-product technology.

Single-Product Technology

An industry can be defined as a natural monopoly if a single firm can meet market demand more efficiently than several firms. Where all the firms use the same technology, the concept of subadditivity can be used to determine whether the industry can be defined as a natural monopoly. Suppose firm C produces Q units of output at a cost of $C(Q)$ dollars (for simplicity, the factor prices r and w are not expressed). If a is some number between zero and one, and firm A produces aQ units of output and firm B produces $(1 - a)Q$ units of output then total output is:

$$aQ + (1 - a)\ Q = Q$$

units of output. Firm A spends $C(aQ)$ dollars and firm B spends $C[(1 - a)Q]$ dollars for a total cost of $C(aQ) + C[(1 - a)Q]$ dollars. If $C(aQ) + C[(1 - a)Q]$ exceeds $C(Q)$ for all values of a, such that the single firm, firm C, can produce Q more cheaply than the two other firms, then the cost function is said to be subadditive at the level of output Q. Generally, the cost function is subadditive at output Q if the cost of a single firm producing Q is less than the cost of having two or more firms produce Q for all possible allocations of output Q across these firms.

For the single-product example, the subadditivity of the cost function can be demonstrated by examining the shape of the average cost function $C(Q)/Q = AC(Q)$. Figure 3.1 shows the shape of two possible average cost curves labeled AC along with two demand curves labeled D. The first average cost curve is shown to be constantly declining. In this case if a single firm, constrained to break-even by some regulatory body, it will produce Q_M units at an average cost of AC_M. Dividing Q_M units of output between two firms would always result in higher production costs. This is demonstrated by letting firm A produce Q_A and firm B produce Q_B, with $Q_A + Q_B = Q_M$. The average cost of firm A, AC_A, and the average cost of firm B, AC_B, exceed that of firm C, AC_M. The second average cost curve declines to Q_A and then rises. As before, a single firm, constrained by a regulator, would produce Q_M units of output at an average cost AC_M. Dividing Q_M between two firms would lead to one firm producing at an average cost higher than AC_M and the other at an average cost lower than AC_M. Jointly, these two firms would produce at an average cost higher than AC_M. This follows since the average cost AC_A is slightly lower than AC_M, but the average cost AC_B is much higher than AC_M. At the output level Q_M, the cost functions shown in the diagram are both subadditive and the industries with these cost functions are natural monopolies.

When analyzing subadditivity and natural monopoly, it is useful to make the distinction between local and global concepts. The concept of local subadditivity applies at a specific level of output, while the concept of global subadditivity applies to the entire range of feasible output. Figure 3.1 helps in making the distinction between local and global natural monopoly. Suppose demand shifts from D to D'. The first cost function is subadditive at the new output Q'_M. Verification is similar to the exercise conducted earlier. In fact, this cost curve is subadditive over the entire range of output and is therefore globally subadditive. The second cost function is not subadditive at the new output Q'_M. As shown, $Q'_M = 2Q_A$. The average cost of a single firm producing Q'_M is AC'_M. The average cost of two firms producing Q'_M is AC_A when each firm is producing half of the the total output. The two-firm, average cost AC_A is lower than the single-firm, average cost AC'_M. The second cost function is not globally subadditive. This distinction between local and global subadditivity and local and global natural monopoly is important. As demonstrated, a natural monopoly which is only local may collapse as demand increases, whereas as natural monopoly which is global remains a natural monopoly as demand increases.

The older economic theory often used the degree of scale economies to determine whether an industry was a natural monopoly. There are scale economies when a proportionate increase in output results in less than a proportionate increase in cost[5]. It is a well-known result of economic theory that scale economies exist when the average cost curve slopes down and scale diseconomies when the average cost curve slopes upward. Thus the first cost curve shown in figure 3.1 demonstrates scale economies over its entire range of output whereas the second cost curve demonstrates scale economies over a limited range of output. A cost function which has scale economies at some level of output Q is subadditive at output Q. However, a cost function which is subadditive at output Q does not necessarily have scale economies at output Q. This is demonstrated with the second cost function in figure 3.1 where at the point Q_M there exists scale diseconomies (the average cost curve slopes upwards) but is subadditive.

It is generally true that researchers tend to infer from the presence of scale economies that an industry is a natural monopoly and that, as a result, a single firm may produce more efficiently than several firms in that industry. There are others, however, who argue that this inference may be misguided based on the fact that there are several assumptions which may not hold true, in particular, the assumption that all firms use the same level of technology and so would exhibit the same costs for all levels of production.

Multiproduct Technology

For the discussions of the multi-product case, examination of the simple case of an industry that produces two products will be reviewed. In this case, no new concepts arise. Firm *A* produces aQ_1 units of the first product and bQ_2 of the second. Firm *B* produces $(1 - a)Q_1$ of the first product and $(1 - b)Q_2$ units of the second, where *a* and *b* are parameters which lie between zero and one. Firm *A* spends $C_A = C(aQ_1, bQ_2)$ dollars and firm *B* spends $C_B = C[(1 - a)Q_1, (1 - b)Q_2]$ dollars. Jointly, firm *A* and firm *B* produce Q_1 units of the first product and Q_2 units of the second product at a cost of $C_C = C(Q_1, Q_2)$ dollars. The cost function is subadditive at Q_1 and Q_2 if $C_A + C_B$ exceeds C_C for all possible values of *a* and *b* or equivalently for any possible division of the production of the two products between firms

A and *B*. In this case, a single firm can produce both products, in the amounts of Q_1 and Q_2, more cheaply than two separate firms. As a result, there is a natural monopoly over both products when the market demands Q_1 and Q_2 units of product one and product two respectively. The cost function is globally subadditive, and there is a global natural monopoly, when the cost function is subadditive at all levels of demand for both products.

In the multi-product case, several scale economy concepts exist of which the most useful is that of product-specific scale economies. Suppose a firm originally produces product one in an amount of Q_1, spending $C(Q_1, 0)$ dollars. It then begins to produce product two in an amount of Q_2, spending $C(Q_1, Q_2)$ dollars, the cost of producing both products. The average incremental cost of producing Q_2 is:

$$AIC_2 = [C(Q_1, Q_2) - C(Q_1, 0)] \;/\; Q_2$$

If the average incremental costs decline with a small increase in output of product two, product two is said to have product-specific scale economies at output Q_2. Product two is said to have global, product-specific scale economies if it has product-specific scale economies at all levels of output.

Single-product scale economies imply single-product natural monopoly. However, product-specific scale economies in both products do not necessarily imply multiproduct natural monopoly. If there are no cost complementaries from common production of both products, firm *A* can specialize in the production of product one and firm *B* in product two without any loss in efficiency. If common production of both products result in cost complimentarities over some narrow range of output and if product-specific scale economies are not extensive, several multiproduct firms can produce the products cheaper than either one multiproduct firm or two single-product monopolies. Further, a multiproduct natural monopoly may not have product-specific scale economies in either product it produces. If common production gives rise to significant cost complimentary which increases as production of both outputs increase, a multiproduct natural monopoly could exist even though there are no product-specific scale economies in either product. Product-specific scale economies by themselves provide little or no information about the existence of a natural monopoly.

The concept of economies of scope is also useful in the determination of the existence of a natural monopoly. Economies of scope are said to exist when the cost of producing product one and the cost of producing product two separately exceed the cost of producing the two products jointly. Formally, this can be expressed as follows: If $C(Q_1, 0) + C(0, Q_2)$ exceeds $C(Q_1, Q_2)$, there are economies of scope at output levels Q_1 and Q_2. There are global economies of scope when there are economies of scope at all levels of output. Local economies of scope together with local product-specific scale economies imply the existence of a local natural monopoly. Likewise for the existence of a global natural monopoly.

The concept of natural monopoly is somewhat more complex in the case of a multiproduct firm than it is for a single firm. Further, to prove the existence of a natural monopoly, much more has to be known about the cost function of the multiproduct firm than of the single-product firm. It is these reasons which make the estimation of the multiproduct cost function difficult and why researchers are tempted to use single-product cost functions as a proxy.

Baumol, Panzar and Willig (1982) have derived necessary conditions for subadditivity and sufficient conditions for subadditivity which require somewhat less information. Economies of scope is a necessary condition for subadditivity for without it, two single product monopolies could produce more cheaply than a multiproduct monopoly. Economies of scope and product-specific scale economies are sufficient conditions for subadditivity. This follows since dividing production between several firms would reduce cost complementaries from joint production and scale economies from volume production. Baumol, Panzar and Willig suggest that the necessary and sufficient conditions be tested separately. If the necessary condition is rejected, subadditivity is decisively rejected. If the sufficient conditions are accepted, subadditivity is decisively accepted. If the necessary condition is accepted and the sufficient conditions are rejected, the test for subadditivity is inconclusive. This paper will adopt these conditions for investigating the issue of the existence of natural monopoly in Bell Canada's local service sector.

Telcommunications Technology

Telecommunications is the process of transmitting messages between two or more points via an electronic medium. Traditionally, the telecommunication firms use their networks in the transmission of voice messages, however, these networks can also be used to transmit data, hardcopy and video signals. Recent developments in digital switching technology has made it possible to transmit large volume data over the same network more economically and reliably, as well as, making it possible to provide a wide range of enhanced services, such as calling waiting, last number redial, programme dialing and so on.

The telecommunications industry can be divided into three broad market segments: the terminal equipment market; the local service market; and the long distance market. (See figure 3.3). Terminal equipment can be defined as the device that is attached to the end of the telephone line and is usually located on the customer's premises. A wide variety of terminal equipment is presently available and can range from the basic black rotary telephone to a sophisticated computer terminal. Without terminal equipment, there is no method by which to transmit or receive messages.

The local service and long distance service markets, from a technology point of view, are very similar. They involve the routing and carrying of messages from one point to another. The main difference between these two markets segments lie in the geographical areas served. The local service market provides access to subscribers within a given calling area, usually associated with a city or town while the long distance market provides access to subscribers in different calling areas, usually other urban centers. The routing of messages is performed by switching equipment. Without the switch, each subscriber or customer would require an individual connection to each other subscriber in order to communicate. There are a wide variety of methods which are currently used to transmit signals from one point to another. These include copper, the traditional form of carriage in the local service market, microwave relay and satellite transmission and copper wire systems for long distance service, and more recently new technologies such as fibre optics, for local and long distance service, and cellular radio, for local service.

Looked at broadly, the natural monopoly debate and the shift in policy with regard to competitve entry can be more easily understood diagramatically. Figure 3.2 illustrates, in a highly simplified way, the traditional end-to-end model of the telephone industry organization. It suggests that the telephone system should be viewed as a seamless network connecting any telephone through local and long distance facilities to any other telephone. Figure 3.3 illustrates, again in a simplied way, that in a discrete market model there should be separate markets for terminal or customer premises equipment, local service and long distance service.

So far, competitive entry has been confined to the equipment and long distance market segments and has not yet extended to the local service market. All three markets, however, remain interconnected and interdependent. This interconnectedness and interdependence gives rise to a high degree of common costs between the local and long distance markets and raise many interesting questions from a regulatory viewpoint on competitive entry.

Chapter 4

Natural Monopoly: Estimation Methods

In econometric studies an attempt is normally made to determine whether scale economies are present by estimating the parameters describing relationships among economic measures of a company's operations, based on historical data. An econometric study, thus, proceeds from the assumption that there are fundamental relationships generating the observed historical data. Included among the fundamental relationships are basic assumptions regarding the economic motivations and behaviour of firms. Simplifying assumptions characterizing the relevant technology are also included. This last assumption is quite crucial since scale economies are commonly embedded in technology.

The most basic assumption regarding technology, or the production process, is that quantities of inputs are used to produce quantities of output. However, choosing a mathematical specification, which adequately represents the production process and that facilitates estimation by standard econometric techniques, can be quite troublesome. Nonetheless, the adequacy of the specification can be tested statistically when the parameters are estimated using historical data. Tests of specification can also be carried out by comparing forecasts of the economic variables using the estimated parameters to actual real world values.

The production process can be formalized in several ways: as a production function in which a single output is produced by one or more inputs; as a cost function in which the total cost of producing a single output varies systematically with the level of output and the price levels of inputs; as a transformation function which is a multi-output analogue to the single output production function; or as a multi-output analogue to the single output cost function.

Based on the work conducted in the 1970's on the theory of duality[6], it is widely recognized that cost functions are more suitable than production functions for estimating the characteristics of industry structure. One advantage of the cost function is that output is an exogenous variable. This makes it especially suitable for regulated utilities, whose prices are determined by a regulatory agency, hence their output levels are exogenously set. Other advantages stem from the relative ease with which the cost function can serve the multiproduct case. When investigating cost complementarity, economies of scope and product-specific scale economies, a multiproduct specification is normally required. In addition, cost functions yield direct estimates of important properties such as marginal cost, cost elasticities and elasticities of factor substitution. Major drawbacks to the use of the cost function are the difficulty associated with the measurement of the cost of capital and the lack of direct marginal product estimates.

The major methodological difficulty associated with production functions is that, if all the variables grow and their growth is inter-related, multicollinearity is very likely to be present. The presence of multicollinearity leads to statistical estimation of coefficients whose variance is large. As a result, it is not possible to judge the reliability of the estimated parameters. Since one of the parameters is a measure of scale economies, there is an inherent problem.

The cost function of choice for estimating industry structure has been the translog cost function, a second order Taylor series expansion in logarithms. Its principal advantage lies in that it imposes no *a priori* restrictions on the nature of technology, allowing the values for economies of scale, economies of scope and factor price elasticities to vary with output. For this reason, the translog cost function is usually referred to as a flexible functional form. The translog cost function and a variant referred to as the generalized translog cost function takes the following form:

$$\text{Log } C = a_0 + \sum_{i=1}^{m} a_i \text{ Log } W_i + \sum_{k=1}^{n} a_{Qk} Q^*_k + b T^*$$

$$+ 1/2\, b_T (T^*)^2 + 1/2 \sum_{i=1}^{m} \sum_{i=1}^{m} c_{ij} \text{ Log } W_i \text{ Log } W_j$$

$$+ 1/2 \sum_{k=1}^{n} \sum_{l=1}^{n} d_{kl} Q^*_k Q^*_l + \sum_{i=1}^{m} \sum_{k=1}^{n} e_{ik} Q^*_k \text{ Log } W_j$$

$$+ \sum_{i=1}^{m} b_i T^* \text{ Log } W_i + \sum_{k=1}^{n} b_{Qk} Q^*_k T^*$$

where $Q_1, \ldots, Q_k, \ldots Q_n$ denote output volumes, $W_1, \ldots, W_m$ denote factor input prices, T is an index of technological change, and C refers to the total production cost, defined as the sum of payments of m factor inputs $X_1, \ldots, X_m$, or more formally as $\Sigma_i(W_i X_i)$, where $i = 1, \ldots, m$. The Q^*_k and T^* variables represent logarithmic transformation of the translog function and Box-Cox transformation in the generalized translog cost function respectively. For example, $Q^*_k = \text{Log } Q_k$ and $T^* = \text{Log } T$ in the translog cost function.

Chapter 4

The translog cost function is not well defined in the regions near zero, since the log 0 is infinity. In order to overcome this problem, the generalized translog cost function is used[7]. The generalization of the translog cost function involves applying a Box-Cox transformation to the variables. Thus the variable Q^*_k and T^* in the equation above will take the form $Q^*_k = (Q_{kk} - 1)/\lambda_k$ and $T^* = (T^T - 1)/\lambda_T$ in the generalized translog cost function. The terms on the right hand side of these expressions represent Box-Cox transformations, where Q_k and T are as defined earlier, with λ_T and λ_K parameters to be estimated. As both λ_T and λ_K approaches zero, the transformation approaches the translog function. Thus, the translog cost function is a special limiting case of the generalized translog cost function.

Typically, the number of data observations associated with the firm are limited. In addition, if the data reflect a high degree of multicollinearity, which is often the case, reliable estimates of the cost function cannot be determined. The economic theory of production, however, provides additional information to facilitate the estimation process. This information is incorporated in the concept of symmetry[8] and the degree of homogeneity of the cost function in its input prices. With the degree of homogeneity equal to one[9], the following conditions hold:

$$\sum^{m} a_i = 1; \quad \sum^{m} c_{ij} = \sum^{m} e_{ik} = \sum^{m} b_i = 0,$$
$$(j = 1, \ldots, m;\ k = 1, \ldots, n).$$

The advantage of the translog cost function, as well as other flexible functional forms, is that a variety of assumptions regarding technology may be tested. Earlier studies often assumed separability or homotheticity without any empirical justification; assumptions which can lead to biased results. The hypothesis of separability implies that the possibility exists whereby a consistent aggregate of output can be constructed. Its rejection implies that no suitable aggregative function exists and, as a result, a multiproduct cost function is the appropiate specification.

The test for joint production is the hypothesis of non-jointness. Should the hypothesis of non-jointness be accepted, then the cost of joint production is equal to the cost of producing products separately. Non-jointness implies that economies of scope are not present. Thus, if non-jointness is rejected, economies of scope are present.

A function is said to be exhibit homotheticity if it is a positive monotonic transformation of a linear homogeneous function. Should the function be homothetic, then the economies of scale and factor price elasticities are constant globally.

Estimation Problems

There are a number of problems which are encountered in the estimation of cost functions. One of the problems which has to be dealt with when estimating translog cost functions is that the amount of information to be obtained - parameter estimates - is usually large compared to the available number of observations. This difference between the number of observations and parameters - the degrees of freedom - gives rise to estimated parameters with smaller degrees of error, the larger the degrees of freedom. If the data suffers from a high degree of multicollinearity, the data must provide many more degrees of freedom in order for the statistical method to provide reliable results.

When the data exhibits a high degree of multicollinearity and observations are limited, as is usually the case for most industry studies, the economic theory of production provides additional information which can be used to alleviate the problem of poor parametric estimation. As already discussed, these porperties are: first, the firm's demand function for capital and labour depend on the same parameters as its cost function; second, the cost function is homogeneous of degree one; and third, the cost function exhibits a mathematical property referred to as symmetry. These properties of production theory provide sufficient information which can be added to the statistical analysis to afford reliable estimation of the parameters consistent with economic theory. Incorporating this information can significantly improve the degrees of freedom. This is obtained through the spreading of the parameters over more equations and through a combination of some of these parameters.

In addition to the informational demands that specification pose, there are also other major practical difficulties which are encountered in econometric studies of the telecommunications industry. First, there is a problem with the available data. The data on labour, capital and technological change is not readily available in a format - physical units - that is most suited for econometric estimation. Instead, dollars are used as a common denominator for the inputs and different types of outputs or inputs are summed up to totals using their prices as weights. Using dollar values of total output and inputs does create other practical problems because apparent changes in levels of output and inputs can occur simply because of price changes. Where this occurs, scale economies cannot be properly measured. This problem is particularly true in the case of telecommunications local service where the flat rate structure is in general use in Canada. Under this structure, there can be changes in the usage of the switching equipment, both in terms of the number of calls and the duration of calls, without any change in the local service component of revenues. However, where dollar values are used, the problem is controlled by deflating revenues and plant by indices of price changes. The deflation process and, as a result, the estimated parameters, are only as good as the indices used.

Another important problem related to the inputs is the measure of capital used in the estimation process. Ideally, the capital inputs should measure capital services, however, these are usually not available and measures of capital stock are are generally used as proxy. To convert stocks to usage, it would be helpful to have measures of capital utilization but these are difficult to obtain for the telecommunications industry.

In the case of labour inputs, physical measures are available - namely the number of hours. The hours for employees, however, cannot be added together in any meaningful way because there are different skill levels of employees. Where labour hours are used, they should be weighted by wage rages to reflect the different skills levels that exist across the firm. Also, the labour inputs should include hours of overtime worked in addition to regular hours worked. This information on overtime hours is not available for Bell Canada and other telecommunications firms. As a result, most econometric studies on the structure of the industry suffer from at least one of these problems.

Another problem associated with the estimation procedure, is the representation of technological change. Often the techniques used usually implicitly assume that technological change has the same proportional impact on capital as it has on labour, and the same proportional impact on the productivity of equipment of all vintages. However, there is some evidence that technological change in the telephone industry may impact the productivity of capital to a larger extent than labour.

It can be seen, then, that even if there were no data or measurement problems, the econometric studies would face other technical estimation problems as discussed earlier. As a result, different interpretations of the econometric evidence concerning the existence and extent of economies of scale are likely. The presence of these interpretation problems does not preclude the use of econometric results for policy evaluation. However, it does suggest that careful evaluation of the evidence on scale economies should incorporate information from other sources in addition to the econometric results.

Tests for Production Characteristics

A wide range of production characteristics can be derived from the estimation of the translog and generalized translog cost functions. Since they form part of the econometric estimation process, these characteristics are discussed briefly below. The production characteristics can be grouped into three broad categories. First, there is a category which describes the economic behaviour of the factor inputs. This category includes the properties of factor demand and factor substitution. The demand for factor inputs is measured by the own-price elasticities and factor substitution by the cross-price elasticities. Factor substitution is indicated by positive cross-price elasticity, factor complementarity by negative cross-price elasticity, while zero cross-price elasticity indicates that the inputs are independent.

The second category of production characteristics provide information on the output elasticities of the inputs and on aggregate economies of scale. Inference on these properties are defined by the cost elasticities with respect to output, and are discussed in greater detail below.

The third category of production characteristics relate to the changes in technology. The technology elasticity of cost indicates, based on its sign, whether technology changes are cost saving. The technology elasticities of inputs indicate input biases in technological changes. Negative, zero or positive elasticities indicate that technology changes are input saving, input-neutral or input using, respectfully.

Testing for Economies of Scale and Scope

Most econometric studies which examine the structure of telecommunications costs estimate at least one of the following properties.

Chapter 4

1. Overall Economies of Scale: a methodology which measures the change in total cost due to a change in aggregate output.

2. Global Output Specific Economies of Scale: a methodology measuring the same properties as in paragragh 1. above but with respect to a specific output category. The process, however, requires the extrapolation to zero output levels of production in order to estimate the incremental costs of the relevant output being measured.

3. Local Cost Complementarity: a measure of the change in marginal cost of one output resulting from a change in the volume of a second output.

4. Global Output Specific Economies of Scale: a measure of the relative cost of producing a specific output separately and jointly with the remaining outputs. The process also requires extrapolation of the data to zero output levels.

5. Local Cost Subadditivity: a comparison of the relative costs of a single-firm to that of a two-firm production of historically observed output levels and mixes. This test was first proposed by D. Evans and J. Heckman.

Panzar and Willig (1979) and Fuss and Waverman (1977) have shown that overall scale economies can be measured by the scale elasticity, which is simply the inverse of the cost elasticity. The cost elasticity is the derivative of the cost function with regard to all output. Hence, the scale elasticity can be expressed as follows:

$$S = \left(\sum_{j=1}^{n} \xi_{CQj} \right)^{-1},$$

where S is the scale elasticity and where $\xi_{C}Qj$ is the cost-output elasticity of the j-th output. If *S* is greater than unity, local scale economies are present. If *S* is less than unity, local scale diseconomies are present, and if *S* is equal to unity, there are constant scale economies.

Economies of scope can be determined by comparing the cost of producing each output separately with the cost of producing all outputs jointly. A local measure of economies of scope at specific levels of output is referred to as cost complememtarity. The test statistic for cost complementarity, in a function that is twice differentiable, is the second order derivative of the cost function with respect to any two outputs. This is given as:

$$CC = \partial C \ / \ \partial Q_1 \ \ \partial Q_2 \ ,$$

where CC is the measure of cost complementarity, Q_1 and Q_2 are two output levels.

If CC is greater than zero, economies of scope exist. If CC is less than zero, diseconomies of scope are present and if CC equals zero, joint production yields no cost savings or increases.

Global economies of scope are said to exist if the joint production of two or more outputs is lower than producing each individual output separately. This can be expressed formerly by

$$C\ (Q_1,Q_2) \ < \ C(Q_1,0) \ + \ C(0,Q_2),$$

where $C(Q_1,0)$ is the cost , say, of producing local telephone service and $C(0,Q_2)$ is the cost of producing some other telephone service.

The tests discussed, so far, pose problems which make it difficult to test for the existence of a natural monopoly. As a result, Baumol, Panzar and Willig (1983) have suggested that for subadditivity to hold, first, economies of scope is a necessary condition. If not, the outputs would be produced by different firms. Second, economies of scope and product specific economies of scale are sufficient conditions.

One method proposed, by Panzar and Willig (1979), for measuring product specific economies of scale is to determine the ratio of the average incremental cost to the marginal cost of that service. This can be expressed as

$$(IC_k / Q_k) / (\partial C / \partial Q_k) = (IC_k / C) / CE$$

where IC_k / Q_k is the average incremental cost, $\partial C / \partial Q_k$ is the marginal cost, and CE is the cost elasticity of service k. When the ratio of average incremental cost to marginal cost is larger than one, product specific scale economies exist. In the case of the translog cost function, the determination of IC_k poses a problem since the function is not well defined about zero. In this case the generalized translog cost function is more suiable for the task.

This test, however, is not without its shortcoming since global economies of scope and product specific scale economies require that the cost of producing the individual services be known. For Bell Canada, these cost are difficult to obtain. If the data used in the studies are not observations which relate to the costs of producing a specific service, it may result in poor estimates of incremental costs, product specific economies of scale and economies of scope.

Chapter 5

Natural Monopoly: The Empirical Evidence

This chapter reviews the evidence, based on the estimation of the cost function, with regard the natural monopoly hypothesis as it relates to Bell Canada and specifically, to the provision of local service.

Historical Review

Econometric studies on the telecommunications industry in North America began emerging in the 1970's. The early studies were primarily concerned with the estimation of production functions, namely variations of the Cobb-Douglas function[10]. The emergence of the theory of duality, in production economics in the 1970's, led to the recognition that the cost function was more attractive than the production function since, as already discussed, these functions promised a wide range of estimates for the economic properties of production technology, which included direct measures of marginal cost and the cost elasticities with respect to output, factor prices and technological change.

The cost function models which have been estimated for the telecommunications industry since the late 1970's can be grouped into the following broad categories: static cost models of single and multi-output specifications; and dynamic cost models of mainly single output specifications. Of the static cost models, the early models assumed instantaneous and costless factor input adjustments in response to changes in output, factor prices and technological change. The first models estimated in this group generally used a single ouput translog cost function and included such studies, mainly on Bell Canada, as Fuss and Waverman (1977); Fuss and Waverman (1978); and Denny et. al. (1979). Single output models for the United States industry were estimated by Nadiri and Schankerman (1979) and Christensen et. al. (1980, 1981, 1983). Work on multi-output cost models quickly followed that of the single output models. The early work in multi-output cost functions were pioneered by Bernstein (1980) and Breslaw and Smith (1980) who estimated two output translog cost functions for British Columbia Telephone and Bell Canada respectively. Early works on three output, cost functions using a generalized translog cost function emerged in 1981, and consisted of studies by Fuss and Waverman (1981) and Kiss et. al. (1981, 1983).

Two-output models were estimated for Bell Canada by Smith and Corbo (1979) and by Kiss et. al. (1981, 1983). Similar studies were estimated for AT&T by Evans and Heckman

(1983, 1984), Charnes et. al. (1985), Kiss and Lefebvre (1986), Gentzoglanis and Cairns (1989), and Ngo (1990). Of these studies, only that of Kiss et. al. produced estimates of product-specific scale economies for local service, however, they all test for cost complimentarity, economies of scope and subadditivity. The primary objective of two-ouput models is an attempt to improve the estimates of the production technology by disaggregating the firm's output into its broad components. This is justified in the case of Bell Canada and the United States Bell System since the toll or long distance service and local service outputs form the two major broad categories of output and they are important from a policy making point of view.

Three-output cost models strive to acheive the same objectives as the two-output costs models, but attempt to carry the disaggregation process one step further. The categories usually considered in the three-ouput model are: local service output; monopoly toll service output; and competitive services output, which includes private line and miscellaneous toll services. Thus far, the three-output, cost models have only been used to examine Bell Canada data. Estimations of the three-output cost function have been conducted by Fuss and Waverman (1977, 1978), Denny et. al. (1979, 1981a, 1981b), Breslaw and Smith (1981), Kiss et. al.

(1981), and Fuss and Waverman (1981). Of these studies, only the papers by Fuss and Waverman evaluate product-specific scale economies for the local service sector, however, they all examine the question of cost complementarity, economies of scope and subadditivity.

Dynamic cost models were first estimated, in 1981, for both Bell Canada and the Bell System in The United States. They evolved from a need to distinguish between short-run utilization and long-run scale economies. This work included studies by Christensen et. al. (1981) and Schankerman and Nadiri (1984) who estimated single output cost functions. To overcome the limitations of the variable cost functions, Bernstein (1986) estimated a single output adjustment cost function for Bell Canada. In that paper, the assumption of long-run equilibrium was relaxed by assuming that the firm must incur adjustment costs as it invests in new capital. However, the experts conclude that the dynamic cost models have not produced satisfactory econometric estimates for the industry, however, they do open up a new area of enquiry[11].

In reviewing the studies and their evidence, consideration will first be given to the results obtained from the two-output models followed by the evidence from the three-output models.

Chapter 5

Two-Output Cost Models: The Evidence

Smith and Corbo (1979). This study of Bell Canada's technology groups the firm's output into two categories: monopoly services which consists of monopoly local and monopoly message toll or long-distance services; and competitive services which consist of other message toll services, in addition to miscellaneous services. The monopoly output was assumed to be exogenous whereas the output for competitive services was considered endogenous. The model estimated a four equation system consisting of a two-output translog cost function, the first order condition for profit maximization with respect to competitive services, that is a revenue share equation, and two cost share equations.

The estimated results indicated that the output for Bell Canada exhibited overall scale economies, measured at 1.20. The study also indicated that there was an indication of local cost complementary. In addition, increases in the volume of competitive output were found to have caused a decline in unit cost of monopoly services over most of the period studied. However, no clear indication of global economies of scope were demonstrated. This study, however, did not test for product-specific scale economies.

Kiss, Karabadjian and Lefebvre (1981, 1983).[12] This study, using data for the period 1957 - 1975, classified Bell Canada's output into local service, which also included miscellaneous services, and long-distance services, which included private line and other toll services - the competitive services. This classification differs from that of Smith and Corbo (1979). Although they tested several specifictions, their final specified model was a two-output cost model, in which the local and toll output variables were transformed by a Box-Cox transformation. The other variables were transformed logarithmically.

The study concluded, that the production process for Bell Canada exhibits robust scale economies, with the scale elasticity estimated at 1.62 at the expansion point. Aggregate scale economies, to a large measure, appeared to be a product of scale economies in the production of local service. The local output specific scale economies were estimated at 1.56 at the expansion point. The study also found that there were economies of scope but that these were not statistically significant. It was estimated that the separate production of local and toll output would have increase Bell Canada's cost by about 9% over the actual production costs and by somewhat larger amounts for most of the sample period.

Although the estimates from this study appear reasonable, Kiss and Lefebre (1987) concluded that some caution should be exercised in their use. The reasons are as follows. Reasonable results are only obtained when certain parameters are restricted to zero, specifically the output interaction term, a restriction which limits the model's capabilities to investigate local cost complementaries in the production process. Economies of scope and output specific scale economies were estimated with some difficulty and some of the annual estimates for the economies of scope are not reliable.

Gentzoglanis and Cairns (1979). This study examined for natural monopoly characteristics in both Bell Canada and Alberta Government Telephones Companies. The study made use of the two-output translog cost function with an estimation process based on Zellner's (1962) Iterated Seemingly Unrelated Regression method and makes use of the Evans and Heckman (1984) subadditivity test.

The results of the study found significant overall economies of scale for Bell Canada with an overall scale elasticity measured at 1.54. The model using the Evans and

Heckman test, for the years 1961 to 1986, consistently found evidence that local and toll costs are subadditive.

Ngo (1990). This study estimated a two-output, generalized translog cost function utilizing the Box-Cox transformations of the output and technology variable for Bell Canada data. The two outputs were assumed to be local and toll services. The local output included local, directory advertising, and miscellaneous services. Toll output included message toll and WATS services, private line and other toll services

The estimation period used in the study was the years from 1953 to 1987 with variables being normalized around their 1967 values. All models used in the study were estimated with a first-order correction for autocorrelation in order to produce satisfactory results.

The study showed that the models used in the estimation process exhibited increasing returns to scale in the early years of the sample. At the expansion point, scale elasticities were found to be significant, between 1.58 and 2.10. In regard to tests for cost subadditivity, all the models used in the study produced evidence of local cost subadditivity.

Chapter 5

The following two studies, although not direct measures of scale economies in Bell Canada, are discussed here since they represent a significant element in the debate about economies of scale and scope in the telephone industry. They arrive at completely opposite conclusions about the United States telecommunications industry.

Evans and Heckman (1983, 1984).[13] The main objective of these studies was to investigate the presence, or lack thereof, of natural monopoly features in the technology of AT&T. In their study, period of estimation 1947 - 1977, some eight models were estimated including: a translog cost function; a generalized cost function similar to that of Fuss and Waverman (1981), and Kiss et. al. (1981, 1983); a Box-Tidwell function, essentially a translog cost function where a Box-Cox transformation is used to transform all the explanatory variables. These three specifications were estimated with and without an adjustment for first order correlation. The remaining estimated functions were variations of the adjusted translog cost function but in one case technological changes were restricted to be non-output augmenting and in the other they were assumed to be neutral. Evans and Heckman conclude that the adjusted translog cost function, with an unrestricted form of technological change, best represented AT&T's production technology.

The study finds that the eight specifications produce substantial overall aggregate scale economies for AT&T. The scale elasticities for the various models tested ranged from 1.39 to 1.57. The main purpose of these studies, however, was to investigate for the presence or lack of natural monopoly characteristics in AT&T's technology. Since it is very difficult to test cost functions for global subadditivity[14], Evans and Heckman contructed a test for cost subadditivity which they refer to as *within sample subadditivity*. The test is based on the definition of subadditivity but restricted to the range of observable data. They conclude, that their findings reject the hypothesis that the AT&T's cost function is subadditive and, therefore, the hypothesis that AT&T is a natural monopoly. In addition, they rejected the hypothesis that there exists an aggregate measure for local and long-distance service and therefore, the single-output cost functions estimated by previous researchers must be rejected since the scale economy estimates derived from these single-output models provide little information about the structure of the industry[15].

Charnes, Cooper and Sueyoshi (1985). The model used for this study is based on methods of goal programming and constrained regression. They argue that the objective of the firm is to minimze the difference between observed and estimated costs subject to certain constraints on the cost function. The cost function specified is the two-output adjusted translog model of Evans and Heckman above and is estimated on their sample data, period of estimation 1947 - 1977.

The study by Charnes et. al. was a very controversial one which criticized econometric models for being based on the assumption of technical efficiency. They argued that it was not appropiate to assume technical efficiency and at the same time estimate a cost surface over the sample points of cost observations. It would be more appropiate, they suggested, to estimate a cost surface below the observed cost point, an efficient frontier, which would represent the cost of wasteless production.

The results of this study, indicate that there are aggregate scale economies for AT&T, with a scale elasticity measured at 1.39 equal to that of Evans and Heckman (1983). Kiss and Lefebvre (1987) argued, scale elasticity similarities aside, that the Charnes et. al. aggregate estimates of scale economies are flawed since the cost elasticity for message toll services, and with it the marginal cost of message toll output, is negative and the strongly trended annual scale elasticity estimates, over the latter half of the sample period, are unrealistic. Economies of scope and output specific economies of scale are not reported. However, Charnes et. al. findings contradict the Evans and Heckman conclusions on the subadditivity of AT&T's cost function as a result of changes in model specification.

Chapter 5

Three-Output Cost Models: The Evidence

Denny, Fuss and Everson (1979)[16]. This study classifies the output of Bell Canada into local service, message toll service and competitive services, with the period of estimation being 1952 - 1976. The major assumptions are that local service output is exogenous, and that message toll service and competitive service outputs are endogenous. The study also assumed that technology was output augmenting and used proxy variables to represent these assumptions. In arriving at the technology proxy variables for the different categories of service, the percentage of telephones connected to modern switching equipment was assumed to augment local service output and the percentage of telephones with direct access to long distance dialing was assumed to augment the toll service and competitive service outputs. The model consists of five estimated equations which include the three-output translog cost function, two cost share and two revenue share equations.

The results of the study indicate that there are robust aggregate scale economies for Bell Canada's production technology, with a scale elasticity measured at 1.46. Annual scale elasticities are not reported, however, the study reports a strongly trended local output elasticity of costs which causes an upward trend in the annual scale elasticities. The study does not investigate product-specific scale economies or economies of scope. The study, however, does provide some evidence on local cost complementarities between local service output and both the toll services and competitive service outputs.

Breslaw and Smith (1981). This study, using Bell Canada data for the period 1952 - 1978, is very similar to the Denny et. al. (1979) study in all respects except for the specification of technological change, the parametric restrictions, and the normalization of the variables. In this study, technology is assumed to be output augmenting. However, it appears that the variables of the model were not normalized[17]. Because of this difference a direct comparison with other studies can not be undertaken although the conclusions of the study remain valid.

The results of this study indicate that there are robust aggregate scale economies for Bell Canada, measured at 1.60. While the functional form used, the translog function, does not permit the estimation of product-specific scale economies and global economies of scope, it did allow the investigation of local cost complementary between the firm's outputs. The study indicates, that increases in the production of any of the three outputs lowered the marginal cost of both the other two outputs and found that cost complementary between local service and toll service to be statistically significant. The same result, however, was not found between local and competitive services. The results of the study also indicate a significant degree of cost subadditivity.

Kiss, Karabadjian and Lefebvre (1981). This study, making use of Bell Canada data for the period 1957 - 1975, estimated twenty-five flexible, three-output translog and generalized translog cost functions with the main objective of investigating whether there was aggregate scale economies and economies of scope in Bell Canada's production technology. By estimating this number of cost functions, they were able to evaluate a number of restrictions with respect to the transformation of the technology and output variables.

The main conclusions of this study were: that the results indicated significant aggregate economies of scale for Bell Canada, with measures ranging from 1.22 to 1.67; that economies of scope and cost complementarities were generally indicated but statistically insigificant; and that the models failed to produce statistically significant and reasonable estimates of product-specific scale economies with any consistency. However, Kiss et. al. declared the experiment a failure and concluded that the three-output model proved to be too general to provide the estimates that are most desired.

Fuss and Waverman (1981). This study was performed on Bell Canada data for the period of 1957 - 1977. Its purpose was to investigate economies of scope and output-specific scale economies for the firm's production technology. The model estimated was a generalization of the translog cost function used by Denny et. al. (1979). The output variables were transformed by a Box-Cox transformation while the price variables were transformed logarithmically. Technological changes were assumed to be output augmenting in a similar fashion to that of Denny et. al. The study assumed that local service output was exogenous and that toll service and competitive services outputs were endogenous[18]. The Fuss and Waverman model contained the same five equations as in the Denny et. al. (1979) study, in addition to, two toll service demand equations.

The results of the study found that the hypothesis of constant returns to scale for Bell Canada could not be rejected. Aggregate scale economies were measure at 0.94.

The study also finds that the production of local service is characterized by constant returns to scale, whereas, the production of toll and competitive services display significant scale economies, measured at 2.58 for toll and 2.24 for competitive services. The study also found that there are strong diseconomies of scope between Bell Canada's local service and its toll service and between toll service and competitive and local services[19]. This study and an earlier study by Fuss and Waverman (1978) are the only studies of Bell Canada that suggests constant returns to scale, both at the aggregate and local service sector, and significant diseconomies of scope and as such remain the focal point for discussion of most of the other studies on Bell Canada production technology.

Conclusions

Of the ten studies reviewed, six were two-output cost models with outputs defined as local and toll and four were three-output cost models with outputs defined as local, message toll service and miscellaneous miscellaneous toll services. Of these: one used the translog cost function; one used the Box-Tidwell transformation of the output variables; one used the translog cost function with an estimation process based on Zellner's Iterated Seemingly Unrelated Regression method; one use the Box-Cox transformation on the output and technology variables. Four of the studies, all for Bell Canada, used three-output cost models with the output being local, message toll and competitive services. The studies using two and three output specifications all test for and reject the existence of a single aggregate measure of output.

The evidence, based on the studies reviewed, tend towards the affirmation of overall economies of scale in Bell Canada's production function, with a somewhat stronger indication for economies of scale in the production of local service. The only study examining the Canadian telecommunications industry which produced results in contrast to the findings, in general, was that of Fuss and Waverman (1981).

Of the studies which examine for economies of scope and cost complememtary, none of the four three-output models, all for Bell Canada, found any significant economies of scope. Of the studies using two outputs specification, one for Bell Canada found no clear indication of economies of scope (Smith and Corbo (1979)); and one found statistically invalid evidence of economies of scope (Kiss et. al.(1981, 1983), Bell Canada). Evans and Heckman ((1986), AT&T) and Charnes, Cooper and Sueyoshi ((1986), AT&T) tested for natural monopoly criteria and reached opposite conclusions for the industry in the United States. The evidence on economies of scope, however, is insufficient to arrive at any definite conclusion. Where such evidence exists, it appears weak and sometimes conflicting. The evidence on economies of scope appears to be hampered by limitations in the data and tests which demand too much of the data. As a result, the conclusion as to whether the local telephone service industry is a natural monopoly remains somewhat ambiguous.

The contrast between the results of the Evans and Heckman (1986) study and the Charnes et. al. (1986) study, which looked exclusively at the U.S. industry, provide some information on the sensitivity of the specification of the models. These studies were performed on the same data, over the same period 1947 - 1977, and yielded dramatically different results. The main difference between the two studies are in the specification of the objective function to be maximized. Charnes et. al. by changing the specification of the objective function, reversed the results obtained by Evans and Heckman.

As the results of the studies reviewed, it becomes quite apparent that there should be some caution exercised when interpreting the estimates produced. This caution is justified given the results of Evans and Heckman and Charnes et. al. The concern stems from biases which are introduced to the various parametric estimates necessary for an examination of a telephone company's technology function. These arise from essentially two retrictions, that of limited data and appropriate methods for representing technical change. First, testing for subadditivity requires the estimation of product specific economies of scale. Because of data limitations, the firm is usually taken as the unit of operation. The use of the firm's broad aggregates of output introduces a bias to the estimates of product specific scale economies. The Baumol, Panzar and

Willig sufficiency condition for the existence of natural monopoly requires an estimate of product specific economies of scale. This estimation involves the calculation of the impact of the cost of adding a single subscriber to the local network, all other conditions remaining unchanged. The econometric evidence measures, however, the impact across all local exchanges of increasing local output somewhere. If output at all Bell Canada local exchanges increased at the same rate over the period studied, the estimated product specific economies of scale would indicate the unit cost changes from a one percent increase in local output. Since it can be safely assumed that output at Bell Canada local exchanges did not grow at the same rate, output bias would result affecting the estimates of the studies and, hence, their interpretation. Aggregation bias is present at the different levels of service as well. It is safe to assume that for other services the estimated parameters would be subject to similar biases and, thus, the same caution extended.

Second, the representation of technological change is also an important element in the econometric estimation of scale economies. Over the period reviewed by the studies, 1947 to 1977, technological change had a greater impact on some types of plant and equipment when compared to others.

Chapter 6

In the production of local service, different technologies were introduced starting with the replacement of *step-by-step* switching equipment with *crossbar* in the 1950's followed by the introduction of analogue electronic switches in the 1960's and digital electronic switches in the 1970's. The scale economies which result from these different generations of switching equipment vary significantly. Electronic switches exhibit greater economies of scale than either *crossbar* or *step-by-step* plant, whereas, digital switches exhibit far greater economies of scale than all of the former. However, technological change has been represented in different ways by different researchers. Proxies for technological change include expenditures on research and development, the number of telephones connected to modern switching equipment and the number of phones with access to direct long distance dialing. The technological impact on the local loop came later. However, with the introduction of transmission concentrators or remotes - a device capable of extending the use of a single loop - and, most recently, the virtual unlimited carrying capacity of fibre optic technology, significant scale economies are now present in the local loop. The proxies used in the studies reviewed are not all directly related to the pace of technological change in the provision of local and various other telecommunications services, therefore, introducing unknown biases to the estimates.

Third, economies of scope or cost complementarity is measured for Bell Canada's technology by the second partial derivative of the cost function with respect to local service output and with respect to message toll service output for two-output cost models. This implies that increases in message toll service lowers the marginal cost of providing local service and vice versa. At the individual switching center level, such cost complementarity can be examined. However, aggregation of output data, to the firm level, introduces bias to the estimated parameters. This occurs, since on the aggregation of output data, the correlation of message toll output with that of local service output cannot be controlled. Aggregate production data is too broad a production unit to address the question of cost complementarity between the various services provided.

As discussed earlier, the test for the existence of natural monopoly is that of subadditivity. In most of the studies reviewed, with the exception of Evans and Heckman and Charnes et. al., these tests have been applied indirectly by measuring economies of scale and scope.

Chapter 6

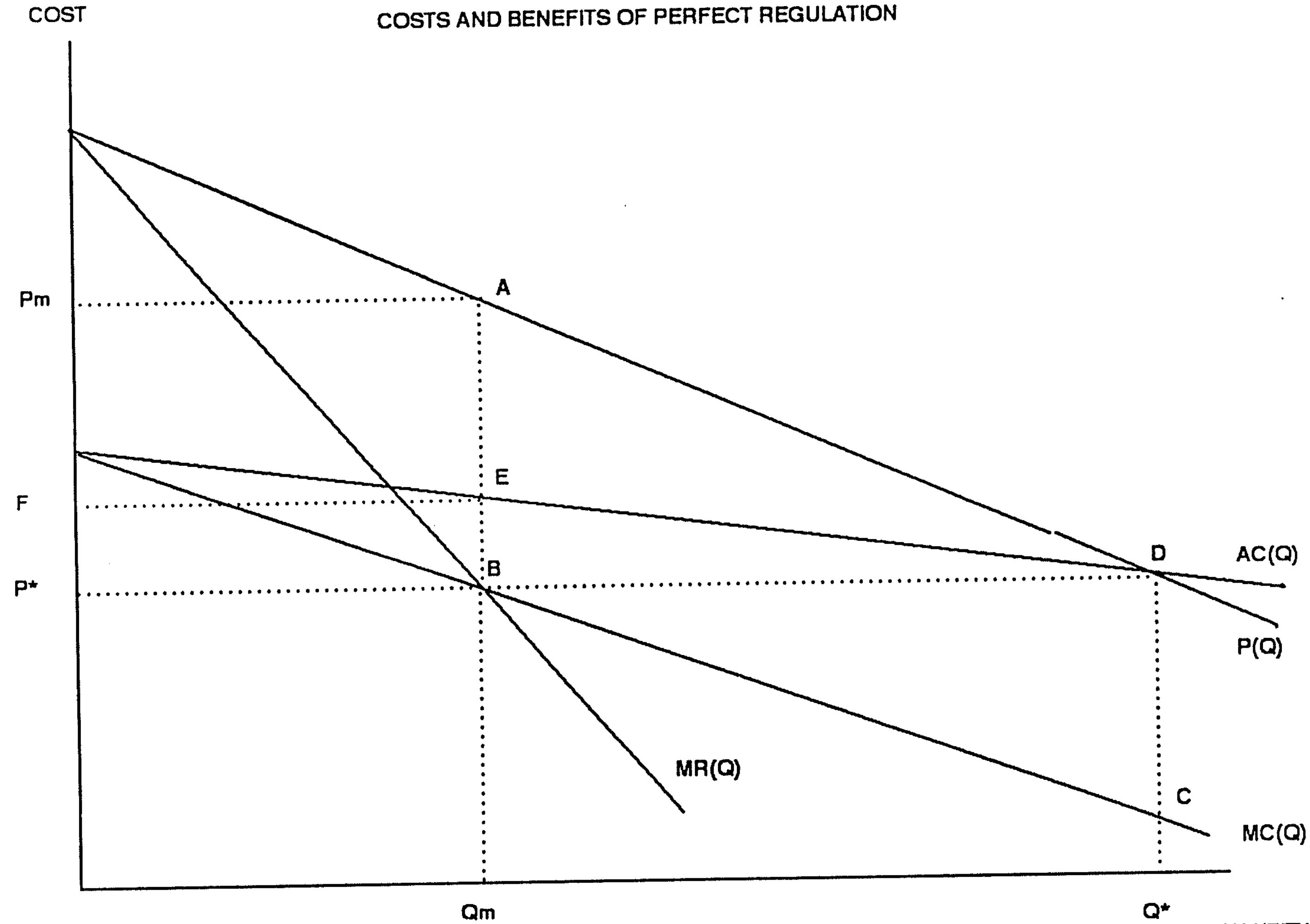
However, the tests for subadditivity pose problems in their calculation. The calculation requires hypothesizing that the firm, Bell Canada, produces zero output of either message toll or local service. Under these circumstances, where it is necessary to work beyond the boundaries of actual data, the results will tend to be unreliable. Evans and Heckman and Charnes <u>et</u>. <u>al</u>. are the only two studies reviewed that directly test for subadditivity and their results are dramatically different.

Although several concerns were raised, in regard to the interpretation of the results, this paper must conclude, based on the evidence from the studies reviewed, that a monopoly of some sort is the appropiate industry structure for most efficiently producing local telelphone service. This conclusion is supported by the positive impacts witnessed within the telecommunications industry as a result of technical change over the last few decades, and to which there appears to be wide agreement at the engineering and plant level. The evidence, however, does not support the existence of a natural monopoly.

There appears to be somewhat of a paradox in public opinion in regard to statements being made about the existing and proposed industry structure, that is, monopoly versus competition. There appears to be concensus, in large measure, about the level of scale economies and, thus, the efficiency of the industry in the production of local service. This position finds support in the technological advancements in the industry, at both the central office and local loop levels, over the last three decades. While technology has acted to reduce barriers to entry and permit several players to enter the industry, at the same time, however, technology has acted to increase the efficiency of plant and equipment used in provision of local telephone service. Thus, while the focus has been, to a large degree, centered on the issues of barriers to entry - euphemistically, reducing the height of the wall surrounding the playing field - little attention has been paid to what has been occurring on that field. As demonstrated, the analytical tools which are being used to assist in the formulation of telecommunications policy suffer from limitations. Until better tools are found, care should be exercised in regard to the forces which are being unleached on the industry since, given its size, small changes in efficiency can lead to substantial consumer losses. Dismantling the existing structure on shaky evidence may prove costly.

FIGURE 2-1

COSTS AND BENEFITS OF PERFECT REGULATION



SOURCE: Adapted from Richard Schmalensee, "Estimating the Costs and Benefits of Utility Regulation," Quarterly Review of Economics and Business 14, No. 2 (Summer 1974): 51-64

FIGURE 2-2

STRONG ECONOMIES OF SCALE

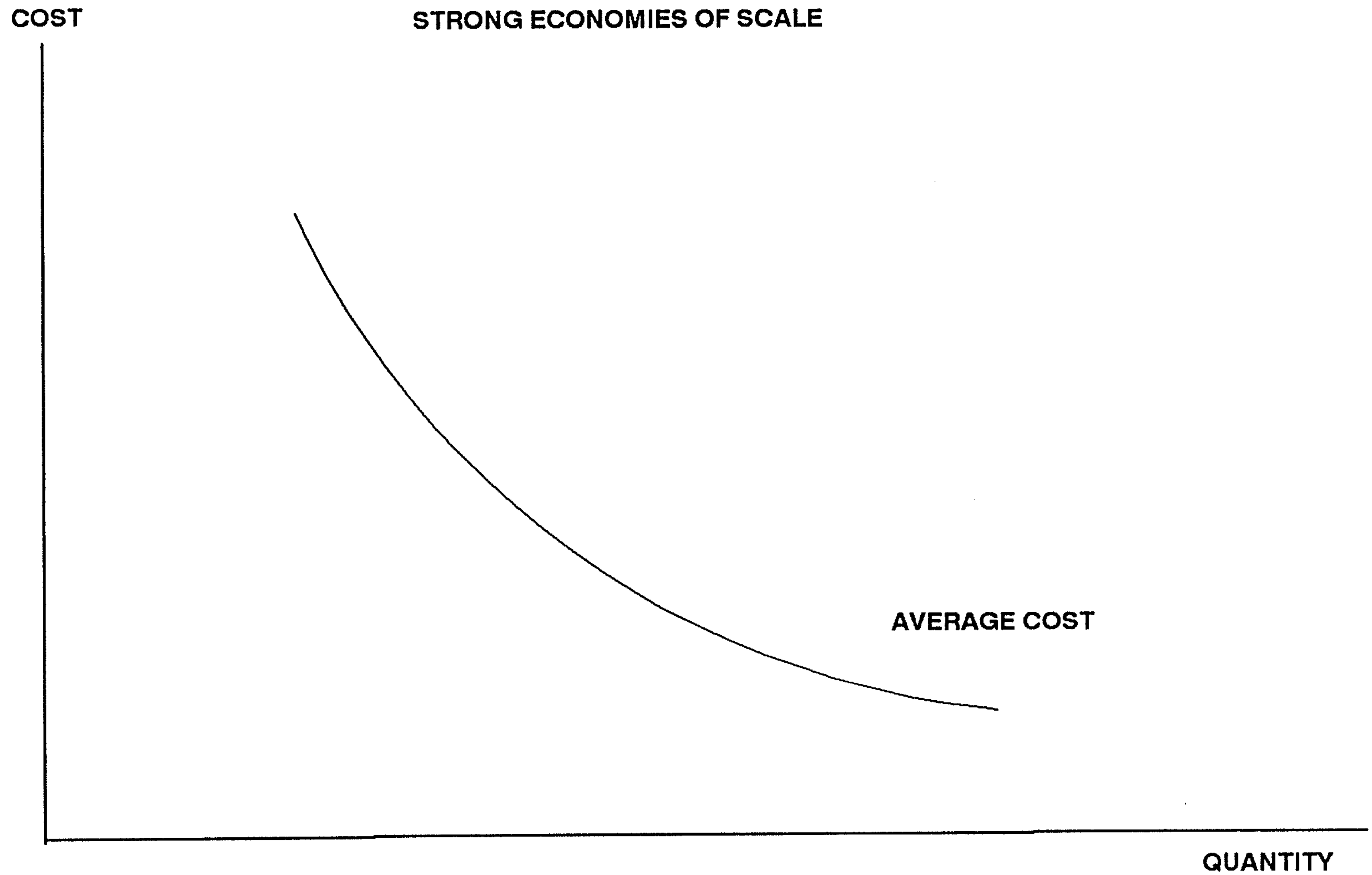


FIGURE 2-3

SLIGHT ECONOMIES OF SCALE

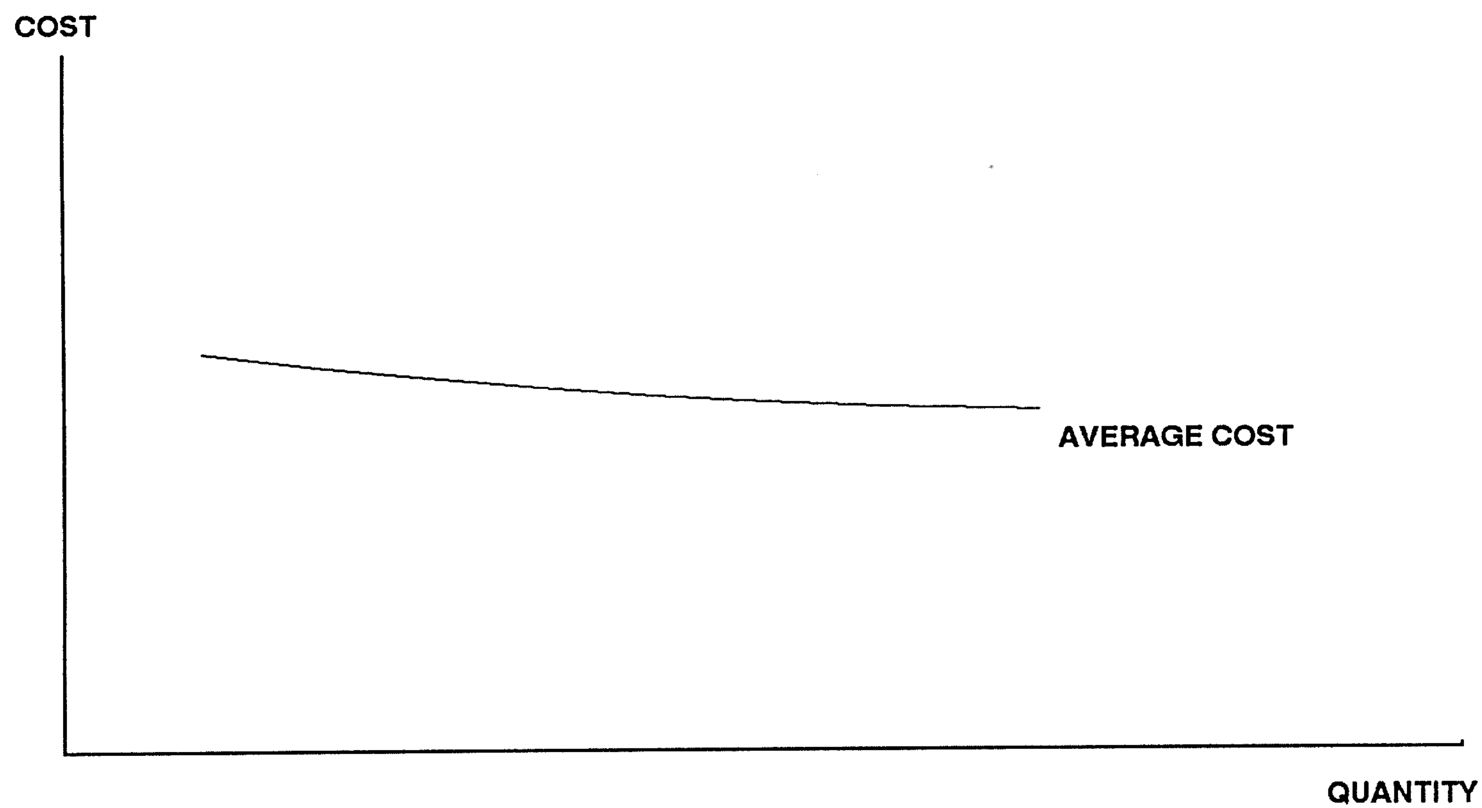


FIGURE 2-4

ECONOMIES OF SCALE EXHAUSTED AT
LESS THAN HALF THE SIZE OF THE MARKET

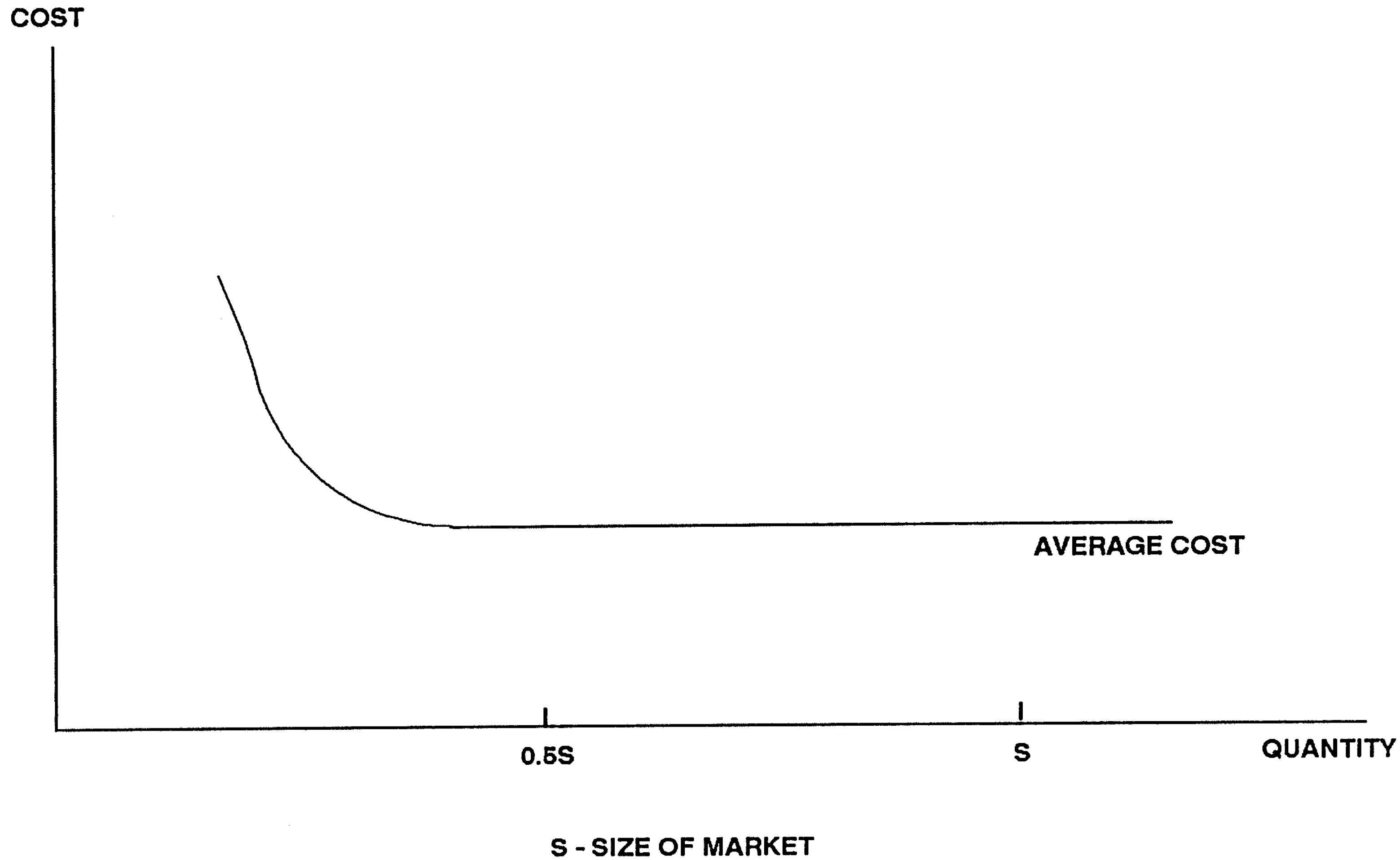


FIGURE 2-5

MONOPOLY PRICING WITH LOW ELASTICITY OF DEMAND

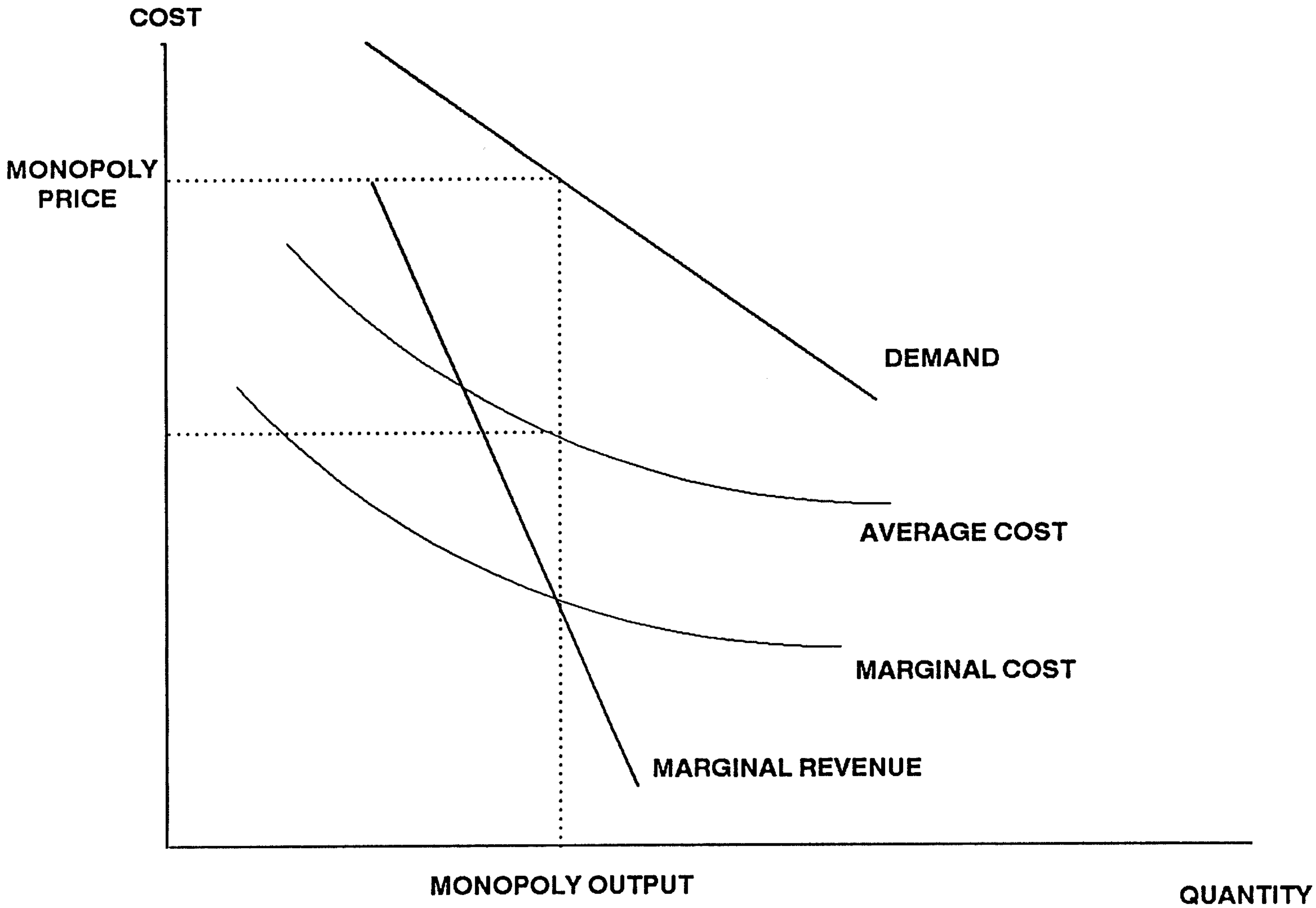


FIGURE 2-6

MONOPOLY PRICING WITH HIGH ELASTICITY OF DEMAND

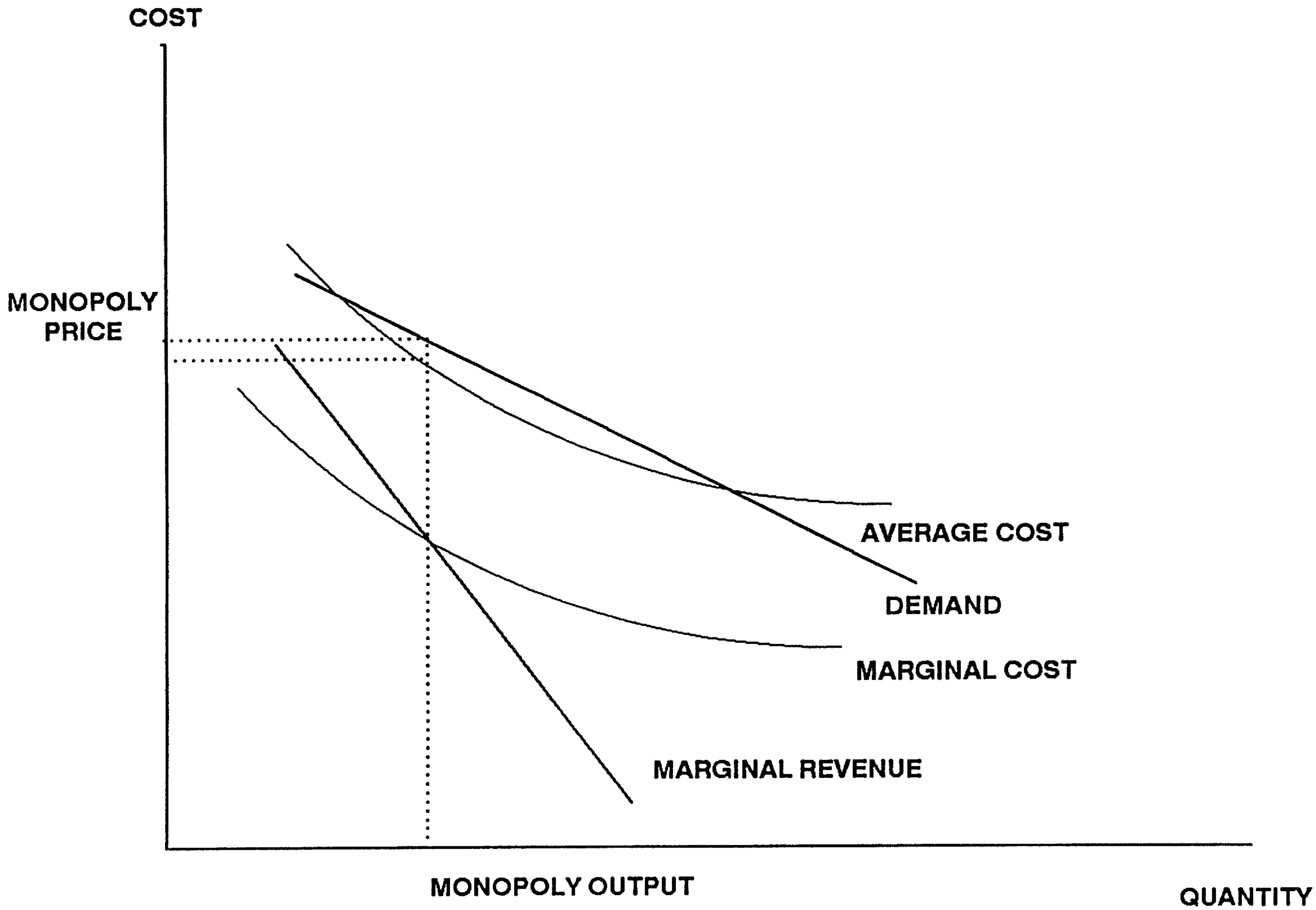
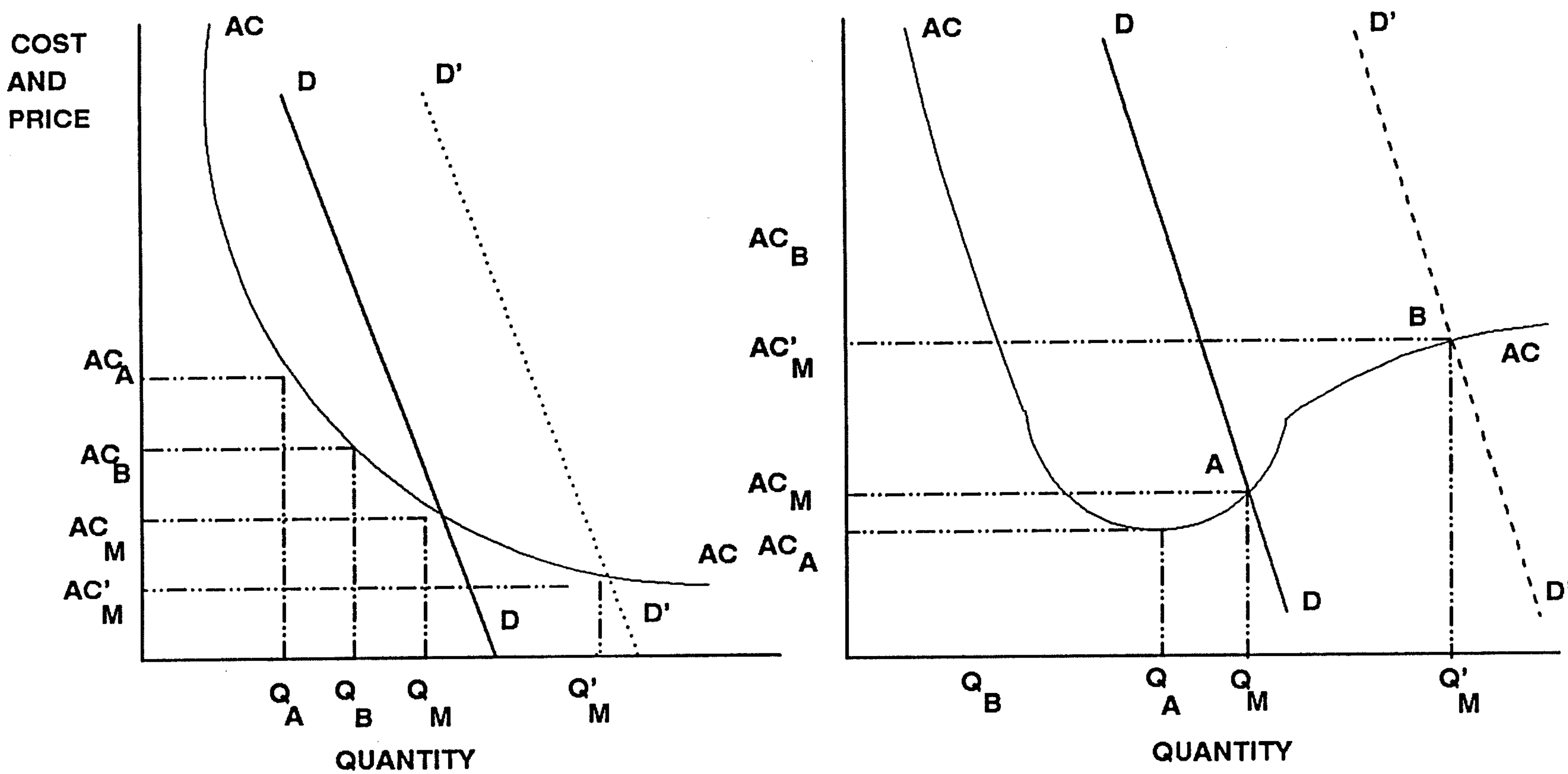


FIGURE 3-1

SUBADDITIVE COST FUNCTIONS



SOURCE: Adapted from Evans and Heckman (1983), "Natural Monopoly", In David S Evans (ed.) Breaking Up Bell - Essays on Industrial Organization and Regulation, Amsterdam, North-Holland.

FIGURE 3-2

END-TO-END MODEL OF TELEPHONE INDUSTRY

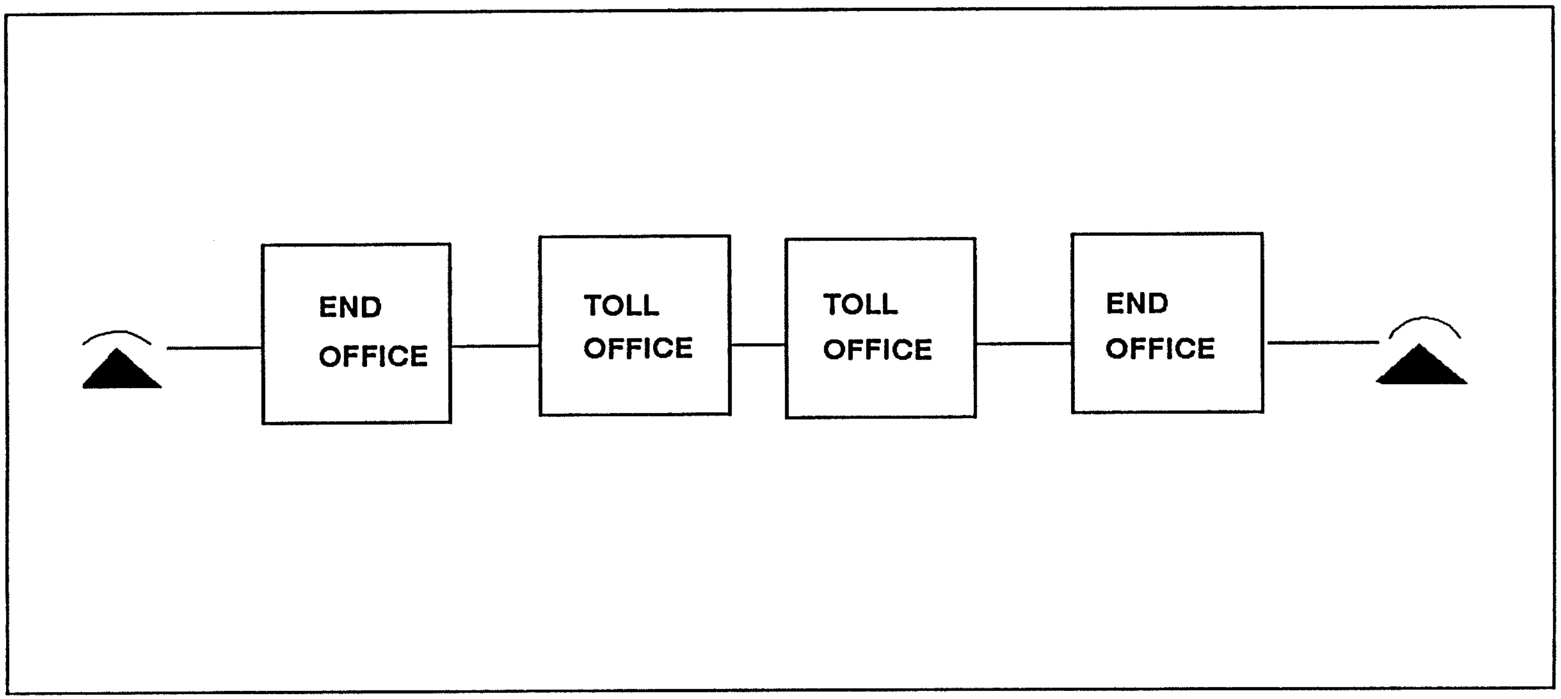
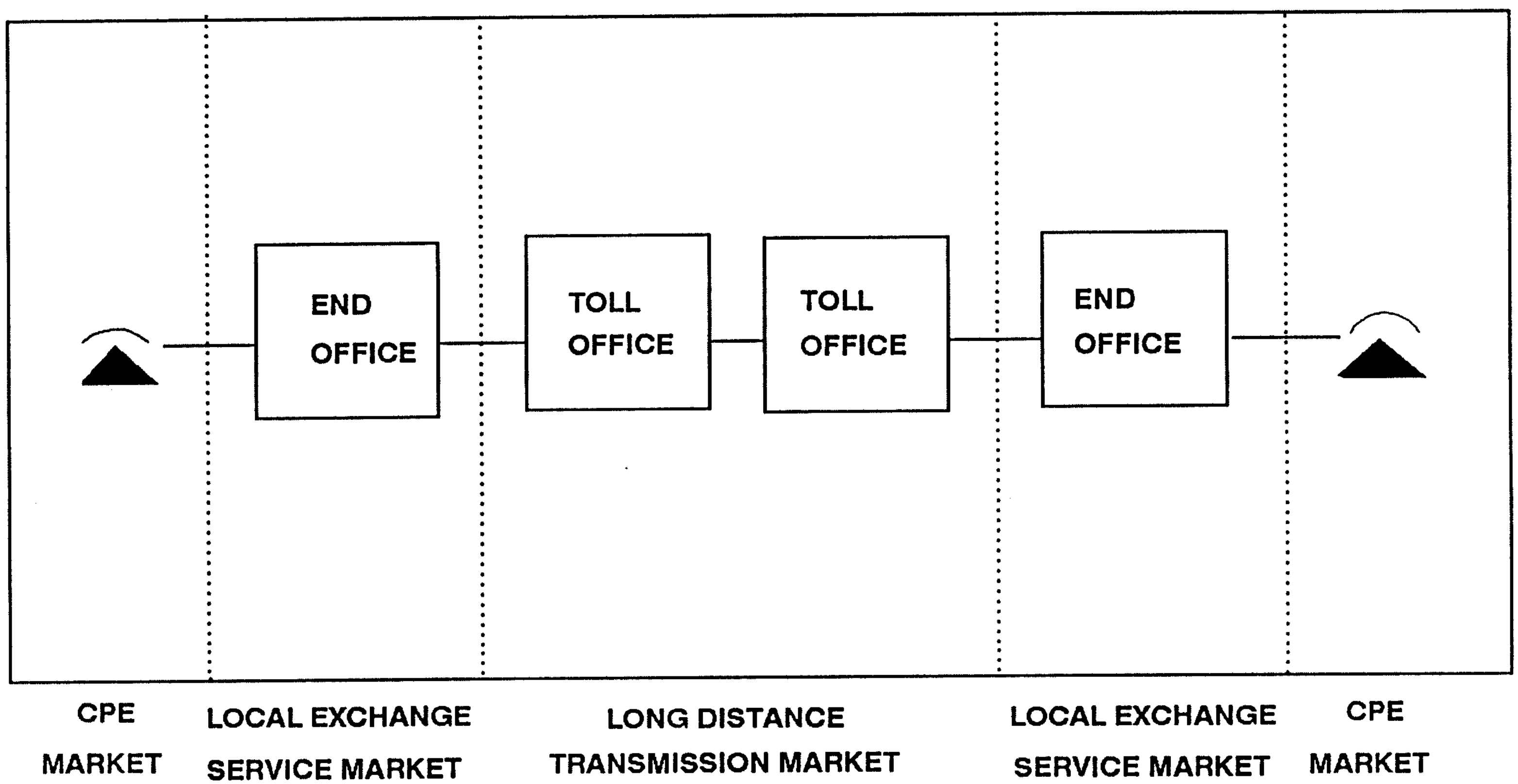


FIGURE 3-3

DISCRETE MARKET MODEL OF TELEPHONE INDUSTRY



Footnotes

[1] Legette (1985) makes this point and the studies by Fuss and Waverman (1978, 1981) infer this.

[2] A discussion on the telecommunications industry is incomplete without addressing the pricing issue, since traditionally toll services have been priced above costs to enable subsidies to flow to local service - a result of national policy objectives. In the absence of marginal cost pricing, Professor Alfred E. Kahn and W. B. Shew have advocated the Ramsey Pricing approach. (See Current Issues in Telecommunications Regulation: Pricing, Yale Journal On Regulation, Vol. 4; 191, 1987). The approach suggests that the common costs could be recovered through the services with the most inelastic demand. The authors argue that such an approach would provide for the most economically efficient pricing solution, in that the output reductions resulting from price increases would be the lowest in the most inelastic market segments. As a result, where in the current environment prices for services are either above or under costs, a case could be made to recover common costs through some flat rate pricing structure. However, the Ramsey rule is mostly concerned with the pricing of services, whereas that of economies of scale and scope are concerned with the production aspects.

[3] Economies of scale are said to exist when an *x%* increase in all inputs result in a greater than *x%* increase in output.

[4] Where the firm is a multi-product producer, cross elasiticities of demand with respect to price can be important. For example, higher cross-elasticities increase the potential that the natural monopoly is likely to be more vulnerable to competitors whose intentions are to *cream skim* the best markets.

[5] The degree of scale economies is frequently measured by the scale elasticity which is the inverse of the cost elasticity. This is discussed in more detail in Chapter 4. The older industrial organization literature defined scale elasticity as the percentage change in output resulting from a common percentage change in all inputs. In general, these measures are not equal because, in order to produce a given percentage change in output, it might be inefficient to change all inputs by the same percentage.

[6] Pioneering work on production theory was done by Diewert (1974).

[7] A generalized translog cost function has been proposed by Fuss and Waverman (1980), Brock and Evans (1983) and Kiss (1983).

[8] Symmetry means that the second order derivatives of cost function with respect to the independent variables are invariant with respect to the order of differentiation.

[9] The restriction to first degree homogeneity ensures that when all input prices are changed by the same percentage and outputs and technology remain unchanged, the resulting percentage change in total cost will be equal to the equiproportional input price change.

[10] These studies included those of Dobell et. al. (1972), Davis et. al. (1972), Vinoid (1972), Sudit (1973), Mantell (1975), Vinoid (1975,1976), Corbo (1976), Corbo et. al. (1978), Corbo et. al. (1979) and Smith and Corbo (1979).

[11] A conclusion arrived at by Kiss and Lefebvre (1987).

[12] This is actually the same study. In 1981, a detail account of the study was presented at a conference at l'Ecole des Hautes Etudes Commerciales in Montreal. In 1983, an abridged version was published which discussed the estimates of economies of scale and scope.

[13] Evans and Heckman published several reports of their study. The most comprehensive was in a volume of studies by CERA Economic Consultants, published under the title Breaking Up Bell by North-Holland in 1983. A subsequent article was published in American Economic Review (September 1984).

[14] Global subadditivity is defined in Chapter 3. However, one major difficulty with testing for global subadditivity is that many of the possible output combinations may lie some distance away from output observations of the cost function being estimated. Some of these output combinations may included zero and since the output variable is expressed in logarithms, the translog function cannot produce a cost estimate when a least one output is zero. (log0 is not defined).

[15] Recall from chapter 4 that if the hypothesis of separability is rejected a consistent aggregate of output cannot be constructed, therefore, a multi-product cost function is the appropiate functional specification.

[16] Between 1977 and 1979, Denny et. al. at the Institute of Policy Analysis, University of Toronto, produced five studies for Bell Canada. The models estimated were all three-output cost models. The report to the Department of Communications of the Government of Canada, 1979, by Denny et. al. may regarded as the end product of this investigation.

[17] An observation made by Kiss and Lefebvre (1987).

[18] Kiss and Lefebvre (1983) disagree with the assumption that toll service output is endogenous. They argue that toll output is much closer to being exogenous since the majority of toll rates are set by the regulator. This assumption

implies monopolistic profit maximizing, whereas, the assumption that toll service output being exogenous implies cost minimizing behaviour of the firm. Kiss and Lefebvre re-estimated a two-output cost model and a three-output cost model, similar to the Fuss and Waverman (1981) model, for both Bell Canada and AT&T adopting the assumption that toll service output was exogenous. They found that the resulting parametric estimates and the estimated economic properties were insensitive to the two forms of specification.

[19] Kiss and Lefebvre (1987) question the reality of the finding of diseconomies of scope between local and toll service. They suugest that independent production of local and toll services would have led to a duplication of a large portion of the telephone network. However, one would have to address the issue of toll service distribution by customer before coming to the conclusion that Kiss and Lefebvre arrived at.

References

Baumol, W. J., J. C. Panzar and R. D. Willig (1982), Contestable Markets and the Theory of Industry Structure, New York: Harcourt Brace Jovanvich, Inc.

Bernstein, J. I. (1980), " A Corporate Economic Model of the British Columbia Telephone Company", Working Paper No.80-7, Center for the Study of Regulated Industries, McGill University, Montreal.

Bernstein, J. I. (1986), " Dynamic Factor Demands and Adjustment Costs: An Analysis of Bell Canada's Technology", Carleton Industrial Organization Research Unit, Carleton University, Ottawa.

Bowers, D. A. and W. F. Lovejoy (1965), " Disequilibrium and Increasing Costs: A Study of Local Telephone Service", Land Economics, February, pp. 31-40.

Breslaw, J. and J. B. Smith (1979,1980), " Efficiency, Equity and Regulation: An Econometric Model of Bell Canada", Report to the Department of Communications, Interim Report, December 1979; Final Report, March 1980.

Carr, J. L. (1972), " A Suggestion for the Treatment of Serial Correlation: A Case in Point", Canadian Journal of Economics, May, pp. 301-306.Center For Communications Management (1983), The Anatomy of Local Services, New Jersey.

References

Christensen, L. R., D. Cummings and P. E. Schoech (1983), "Econometric Estimation of Scale Economies in Telecommunications", in L. Courville,

deFontenay, A. and R. Dobell (eds.) Economic Analysis of Telecommunications: Theoy and Applications, Amsterdam: North-Holland Publishing Company.

Denny, M., M. Fuss and L. Waverman (1981a), " The Measurement and Interpretation of Total Factor Productivity in Regulated Industries, with an Application to Canadian Telecommunications", in T. G. Cowling and R. Stevenson (eds.), Productivity Measurement in Regulated Industries, New York: Academic Press, pp. 170-218.

Denny, M., C. Everson, M. Fuss and L. Waverman (1981b), " Estimating the Effects of Diffusion of Technological Innovations in Telecommunication: The Production Structure of Bell Canada", Canadian Journal of Economics, February, pp. 25-43.

Denny, M., M. Fuss and L. Waverman (1979), "The Measurement and Interpretation of Total Factor Productivity in Regulated Industries, with an Application to Canadian Telecommunications", paper presented at the National Science Foundation Conference on Productivity Measurement in Regulated Industries, Madison, Wisconsin, April 30 - May 1, 1979.

References

Dobell, A. R., L. D. Taylor, L. Waverman, T. H. Liu, and M. D. G. Copeland (1972), "Communications in Canada", The Bell Journal of Economics and Management Science, Spring, pp. 179-219.

Evans, D.S. and S. J. Grossman (1983), " Integration ", in David S. Evans (ed.) Breaking Up Bell: Essays on Industrial Organization and Regulation, Amsterdam: North-Holland Publishing Company.

Evans, D. S. and J. J. Heckman (1983), " Natural Monopoly ", in David S. Evans (ed.) Breaking Up Bell: Essays on Industrial Organization and Regulation, Amsterdam: North-Holland Publishing Company.

______, " Multiproduct Cost Function Estimates and Natural Monopoly Tests for the Bell System", in David S. Evans (ed.) Breaking Up Bell: Essays on Industrial Organization and Regulation, Amsterdam: North-Holland Publishing Company.

Fuss, M. and L. Waverman (1978), " Multi-product, Multi-input Cost Functions for a Regulated Utility: The Case of Telecommunications in Canada", Institute for Policy Analysis, Working Paper No. 7810, revision of Fuss and Waverman (1977). Also under title " Regulation and the Multiproduct Firm: The Case of Telecommunications in Canada", in G. Fromm (ed.) Studies in Public Regulation (Cambridge: M. I. T. Press, 1981)

References

Fuss, M. and L. Waverman (1981), " The Regulation of Telecommunications in Canada", Technical Report No. 7, Economic Council of Canada, March, 1981.

Fuss, M. A. (1983), " A Survey of Recent Results in the Analysis of Production Conditions in Telecommunications", in L. Courville, A. deFontenay and R. Dobell (eds.) Economic Analysis of Telecommunications: Theory and Applications, Amterdam: North-Holland Publishing Company.

Gentzoglanis, Anastassios and Robert D. Cairns (1989). "Cost Comparisons and Natural Monopoly Issues For Two Canadian Telecommunications Carriers: Bell Canada and Alberta Government Telephones. Presented at the Bellcore-Bell Canada Conference on Telecommunications Costing in a Dynamic Environment, San Diego, April 1989.

Kiss, F. and B. J. LeFebvre (1985), " Economies of Scale in Common Carriers", presentation at the Bellcore Economic Cost Modeling Forum, Atlantic City (September).

Kiss, F. and B. J. LeFebvre (1987), "Econometric Models of Telecommunications Firms: A Survey", Revue Economique,No. 2, March, 1987, pp. 307-374.

Kiss, F., S. Karabadjian and B. J. LeFebvre (1983), " Economies of Scale and Scope in Bell Canada", in L. Courville, A. deFontenay and R. Dobell (eds.) Economic

References

Analysis of Telecommunications: Theory and Applications, Amterdam: North-Holland Publishing Company.

Meyer, J. R. et. al. (1980), The Economics of Competition in the Telecommunications Industry, Cambridge, MA: Oelgenschalager, Gunn and Hain Publishers, Inc.

McFadden, D. (1978), " Cost, Revenue and Profit Functions", Chapter 1 in M. Fuss and D. McFadden (eds.), Production Economics: A Dual Approach to Theory and Applications,Amsterdam: North-Holland Publishing Company.

Ngo Hoi (1990). "Testing For Cost Subadditivity: The Case For Bell Canada", Presentation at the 1990 Meetings of the Canadian Economics Association (University of Victoria, B.C.) and at the Ninth Annual Conference of the Advanced Workshop in Regulation and Public Utilities Economics (New Platz, New York).

Panzar, J. C. and R. D. Willig (1977), " Free Entry and the Sustainability of Natural Monopoly". The Bell Journal of Economics, vol. 8, no. 1, pp. 1-22.

Pryor, T. M. and C. G. K. Weaver (1987), " The Future of Competition in the Telecommunications Industry ", Public Utilities Forthnightly, March 5, pp. 28-32.

References

Sharkey, William W. (1987), The theory of natural monopoly, New York: Cambridge University Press.

Smith, J. B. and V. Corbo (1979), "Economics of Scale and Economies of Scope in Bell Canada", Final Report to the Department of Communications, March, Ottawa.

Varian, H. R. (1978), Microeconomic Analysis, New York: W. W. Norton & Co.

Williamson, O. E. (1981), "The Modern Corporation: Origins, Evolution, Attributes", Journal of Economic Literature,Vol. XIX, December.

Williamson, O.E. (1975), Markets and Hierarchies: Analysis and Antitrust Implications, New York: The Free Press.