TRANSPORTATION COST FUNCTIONS: A MULTIPRODUCT APPROACH

by

SERGIO RODOLFO JARA DIAZ

Ingeniero Civil, Universidad de Chile
(1974)

Magister en Planificación Urbana y Regional,
Universidad Católica de Chile
(1977)

S.M. Massachusetts Institute of Technology
(1980)

SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1981

Signature redacted

Signature of Author ______________________
Department of Civil Engineering

Signature redacted

Certified by ______________________
Ann F. Friedlaender, Clifford Winston, Thesis Supervisors

Signature redacted

Accepted by ______________________
C. Allin Cornell
Chairman, Departmental Committee on Graduate Students
Department of Civil Engineering

TRANSPORTATION COST FUNCTIONS: A MULTIPRODUCT APPROACH

by

SERGIO RODOLFO JARA DIAZ

Submitted to the Department of Civil Engineering on April 30, 1981, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

ABSTRACT

Although major advances have been made in the estimation of transportation cost functions in terms of functional specification and microeconomic properties, available studies present inconsistencies with observed industry behavior, and have been criticized as a reliable basis for policy design. However, a third aspect has received less attention than those already mentioned: the characterization and treatment of transportation output. Virtually all studies up to date have used ton-miles, or similar measures, as the basis for output description. In this work, the product of a transportation system is defined as a vector of origin-destination-commodity-period specific flows. The concept of transportation function is defined and used to derive cost functions for two particular theoretical spatial settings, from which the ambiguity of the aggregate (ton-miles-like) output definition is shown. Most important, economies of spatial scope are shown to be a potential source of merging incentives, in spite of the existence of constant returns to scale. The multioutput concept is applied to the estimation and analysis of cost functions corresponding to the operations of two short-line railroads. In this example, based upon monthly observations, the recently developed theory of the multiproduct firm is applied. The results are compared with those obtained from the aggregate approach, showing that this latter not only destroys the possibility of analyzing production complementarity of any kind, but also fails to correctly estimate economies of scale. Based upon the radial nature of multioutput economies of scale, a procedure to perform non-distorting spatial aggregation is proposed and applied successfully to the example.

Thesis Supervisor: Ann F. Friedlaender
Title: Professor of Economics and Civil Engineering

Thesis Supervisor: Clifford Winston
Title: Assistant Professor of Civil Engineering

Acknowledgements

This thesis is a long answer to a question posed to me by Ann Friedlaender in fall 1979. I have been fortunate in having her not only as my committee chairman, but also as the person who provided academic orientation and encouragement, in its most ample meaning, throughout my studies at MIT. The faith, support and collaboration of Clifford Winston, co-chairman of my thesis committee, were important factors in the development of the thesis. The permanent availability of both supervisors for discussion and advice, played a key role in setting the pace of my work.

I want to thank Yossi Sheffi and Dan McFadden, committee members, for their patience in reading the manuscripts, and in listening and discussing my ideas. They were also sources of helpful suggestions.

The typed version of the thesis is the result of Marc Thorman's efficiency, which includes the correction of eventual flaws in my English phrasing. My thanks to him for the fast and neat job.

Further thanks to the executives of the two short-line railroads that allowed me to obtain the necessary information to perform the applied work contained in Chapter 4.

I want to explicitly acknowledge my fellow students Brendon Hemily, Shaw-er Wang, Sergio Gonzalez and Josue Tanaka, as well as the family Vivaldi-Macho, for bringing me their friendship in good and bad moments. Our conversations and discussions continually helped me feel alive.

Literally, I would have not been able to complete this work withou. the moral support always provided by my parents, particularly during crucial moments while at MIT. Further recognition to my friends and colleagues of the University of Chile, Tristán Gálvez, Jaime Gibson, Sergio González Tagle, Marcelo Farah, and the late Juan Francisco Chernilo, as well as to Joaquin de Cea and Enrique Fernández from the Catholic University of Chile, and to Terry Friesz from the University of Pennsylvania. The distance has not been an obstacle to keep on building our friendship, so important when one is away from home.

Somebody told me once that three theses and one child in less than three years was good "productivity" for a couple. I would put it in a slightly different way: our two children, Pedro and Francisco, have been the source of joy that has kept our good spirits at a level high enough to keep on doing our job.

Finally, but not less important, my gratitude to the Civil Engineering Department of the University of Chile, and to the Organization of American States, for help in funding my doctoral studies at MIT.

TRANSPORTATION COST FUNCTIONS: A MULTIPRODUCT APPROACH

Table of Contents

List of Figures

(List of Figures, continued)

List of Tables

CHAPTER 1. INTRODUCTION

1.1 Objectives and Description of the Research

Public policy toward transportation industries is a topic that has consistently generated conflict and discussion. Operators, users, planners and analysts have always had to take a position, either implicitly or explicitly, actively or passively, with regard to specific transportation policies. One of the most important aspects of public policy concerns its impact on industrial structure and pricing. Competitive, oligopolistic or monopolistic patterns of production will be the desired outcome in terms of industrial organization, depending upon the cost structure of firms within that industry, and upon the size of the market. Historically, defining such a desired pattern in order to develop consistent transportation policies is a problem that has been approached through the estimation of transportation cost functions, i.e., those functions which represent the minimum cost of producing a given level of output.

Econometric estimation of cost functions for different transportation industries has evolved in many different ways in the last decade. Functional specification departed from the linear form toward less restrictive functions, and the microeconomic framework has been incorporated with increasing intensity, taking advantage of the theoretical properties that a cost function should have. However, there is one aspect that has received comparatively little attention: the treatment of transportation output. With no exceptions in the literature, output has taken the form of units-times-distance (UTD), e.g., ton-miles, as a generic description of the production of a transportation firm. Although the pure UTD approach is still widely used, the recent trend has been to add so-called "quality" and "technological" variables to improve output description.

Although the procedures to estimate cost functions in transportation have improved significantly, published work still presents inconsistency in terms of predicting industry behavior. Economists' a priori beliefs and empirical studies support the presence of constant returns to scale in the trucking industry, which appears to be incompatible with the observed merging trend. Thus, as illustrated in Spady and Friedlaender (1978), the pure UTD approach indicates counterintuitive increasing returns, "explaining" mergers and supporting regulation; on the other hand, the "quality adjusted" UTD approach indicates acceptable constant returns, supports deregulation, but does not explain by itself industry behavior. The same type of problem appears in studies of other industires, e.g., airlines. These apparent inconsistencies make policy conclusions potentially unreliable, not only for trucking and airlines, but for all transportation industries; in other words, the bottom line of the problem appears to be a methodological failure. We believe that the kernel of its solution is precisely a correct treatment of transportation output. It is the objective of this thesis to elaborate, justify and apply a new approach to focus on transportation cost functions, based upon a multiproduct view of transportation output.

The use of the cost function and, more generally, of the theory of the firm has been criticized by transportation analysts. It has been argued that "this theory does not apply, as formulated, to transportation" (Manheim, 1980). Although a competitive, single output, market was kept in mind to formulate such criticism, it is true that multioutput microeconomics has not been applied to analyze economic activity as a general rule, although a fairly solid body of knowledge has been built in the last few years. If we accept that a transportation firm generates multiple products (and <u>not</u> one product with many "characteristics"), then the theory of the multiproduct firm is the appropriate reference to perform the analysis and to draw conclusions from an estimated transportation cost function.

We will pursue our objectives from three complementary points of view:

i) through the analysis of transportation processes from the generation of production (transformation) functions, in the style of Vernon Smith (1961), to the derivation of cost functions, in order to gain insights into the kind of misspecification caused by the UTD approach, and also to better understand the multiproduct nature of transportation cost functions;

ii) through a methodological analysis and critique of the procedures to estimate cost functions contained in published work up to date, in relation with transportation industries; and

iii) through the application of the new framework to an actual case, in order to elaborate on the methodological aspects of it, and to actually face the problems arising from its use in empirical work.

The remainder of this chapter presents briefly the main concepts related to cost functions, scale economies and natural monopoly in both the single-output and multioutput contexts, plus all the new aspects, definitions, and concepts which have emerged associated exclusively with multioutput production. The last section provides the basis for the applied analysis of a multioutput cost function in terms of subadditivity (natural monopoly). In Chapter 2 we present and discuss the many approaches taken to derive and estimate cost functions in different transportation modes. In Chapter 3, we begin by defininning what a product of a transportation firm is, and then proceed to develop the concept of a transportation function. This is applied to simplified versions of transportation systems producing one and two outputs; the corresponding cost functions are derived, and the importance of what we define as "economies of spatial scope" is established within the context of the analysis of industrial structure. We conclude this chapter by sketching a framework to estimate transportation cost functions. This framework is applied to the case of short-line railroad operations in Chapter 4, where all the multioutput apparata are used to perform the analysis, which is then compared with the UTD approach. This application helps to establish methodological and practical points, which are highlighted in this chapter and elaborated in Chapter 5, after a retrospective view is presented. Finally, the main conclusions and directions for future work are given.

1.2 Scale Economies, Cost Functions, and Natural Monopoly in the Single Output Case

Let us define:

Input set $x = \{x_1, x_2, \ldots, x_n\}$; x_i is a factor of production

Output y ; scalar

Technology $(x,y) \varepsilon T$; i.e. y can be produced from x.

Production Function $f(x) = \{Max\ y/(x,y)\varepsilon T\}$; optimal technical use of x.

It is usually assumed that $(0,y)\varepsilon T$ if and only if $y = 0$. In addition, an increase in input use will either increase output or leave it unaffected (i.e. $\frac{\partial f(x)}{\partial x_i} \geq 0$). We will say that increasing returns to scale or economies of scale are present in the production of y if a proportional expansion of inputs leads to a more-than-proportional expansion of outputs. Formally, with $\lambda > 1$ we can always write

$$f(\lambda x) = \lambda^m f(x); \quad \text{then} \quad \begin{array}{l} m > 1 \rightarrow \text{increasing returns to scale} \\ m = 1 \rightarrow \text{constant returns to scale} \\ m < 1 \rightarrow \text{decreasing returns to scale} \end{array} \tag{1.1}$$

Alternatively, we may say that economies of scale exist at (x,y) if

$$(x,y)\varepsilon T \rightarrow (\lambda x, \mu y)\varepsilon T \text{ with } 1 < \lambda < \mu. \tag{1.2}$$

Scale economies, then, are defined in terms of technology. The optimal usage of the available technology T, summarized by f(x), can be very simple or extremely complicated.[1]

1/ For a clear explanation of the production function concept, see Vernon Smith (1961).

Let us add the following set of definitions:

Input prices $w = \{w_1, w_2, \ldots, w_m\}$

Total cost wx'

Cost function $c(y,w) = \text{Min } \{wx'/(x,y)\varepsilon T\}$

$= \text{Min } \{wx'/y=f(x)\}$ 2/

Average cost function $AC(y,w) = c(y,w)/y$.

Now we can use the cost function to analyze scale economies by noting that, under constant factor prices,

$$f(\lambda x) = \lambda^m f(x) \rightarrow c(\lambda^m y, w) = w(\lambda x')$$

$$\therefore \quad AC(\lambda^m y) = \frac{w\lambda x'}{\lambda^m y} = \lambda^{1-m}\frac{wx'}{y} = \lambda^{1-m} AC(y). \tag{1.3}$$

Since $\lambda > 1$, we have that

$$\begin{aligned} AC(\lambda^m y) &< AC(y) \quad \text{for } m > 1 \\ AC(\lambda^m y) &= AC(y) \quad \text{for } m = 1 \\ AC(\lambda^m y) &> AC(y) \quad \text{for } m < 1. \end{aligned} \tag{1.4}$$

As $\lambda^m y > y$ for any m, positive, (1.4) and (1.1) imply that if the production of y presents increasing, constant or decreasing returns to scale, then the average cost is decreasing, constant or increasing respectively. Thus, a property of the technology T can be studied through the cost function.

2/ This second expression is not tautological but can be easily proved. Let x^o be such that $C(y_o) = wx^{o'}$. Assume $f(x^o) \neq y_o$. As $\frac{\partial f(x)}{\partial x_i} \geq 0$, $\exists\, x^1/x^1_i \leq x^o_i$ (strict inequality for at least some j) and $f(x^1) = y_o$. Then $wx^{1'} < wx^{o'}$ which contradicts the assumption. Q.E.D.

C(y) will be said to be subadditive at y if for any $y^1, y^2, \ldots, y^k$ such that $\sum_{i=1} y^i = y$, we have $C(y) < \sum_{i=1}^{k} C(y^i)$.[3/] In other words, C(y) subadditive means that one firm can produce y cheaper than two or more firms, i.e. a case for natural monopoly.[4/] Now the relation between returns to scale and natural monopoly becomes clear, because under increasing returns to scale we have $\partial AC/\partial y < 0$, that is, the unit cost of producing less then y is greater than AC(y). Therefore, if

$$Y = \sum_{i=1}^{m} y_i, \text{ then}$$

$$\sum_{i=1}^{m} C(y_i) = \sum_{i=1}^{m} y_i \, AC(y_i) > \sum_{i=1}^{m} y_i \, AC(y) = AC(y) \sum_{i=1}^{m} y_i = AC(y)y = C(y). \quad (1.5)$$

Summarizing, scale economies $\rightarrow$ decreasing average costs $\rightarrow$ natural monopoly (subadditivity), but the converses are not true. It is possible to have subadditivity at y without decreasing average costs at y, and these may exist without scale economies (however, if at $AC(y)/dy < 0$, scale economies exist in the neighborhood of y; Baumol 1976). Figure 1.1 suggests a case where increasing average costs are present at y^a, but it is not possible to split y^a such that multiple firms can produce cheaper than $AC(y^a)\, y^a$.

On the other hand, first-best pricing rules indicated that price (P) should equal marginal cost (MC). However, this would cause losses to the firm under increasing returns because

$$\frac{\partial\, AC(y)}{\partial y} < 0 \quad \text{or} \quad \frac{\partial \left(\frac{C(y)}{y}\right)}{\partial y} < 0 \rightarrow \frac{1}{y^2}\left[y \frac{\partial C(y)}{\partial y} - C(y)\right] < 0; \quad (1.6)$$

3/ The superscript denote vector, as usual. In this section y has only one dimension, of course.

4/ Note that subadditivity is a local measure (at y), but requires global information on C(y), $\forall\, y^i < y$.

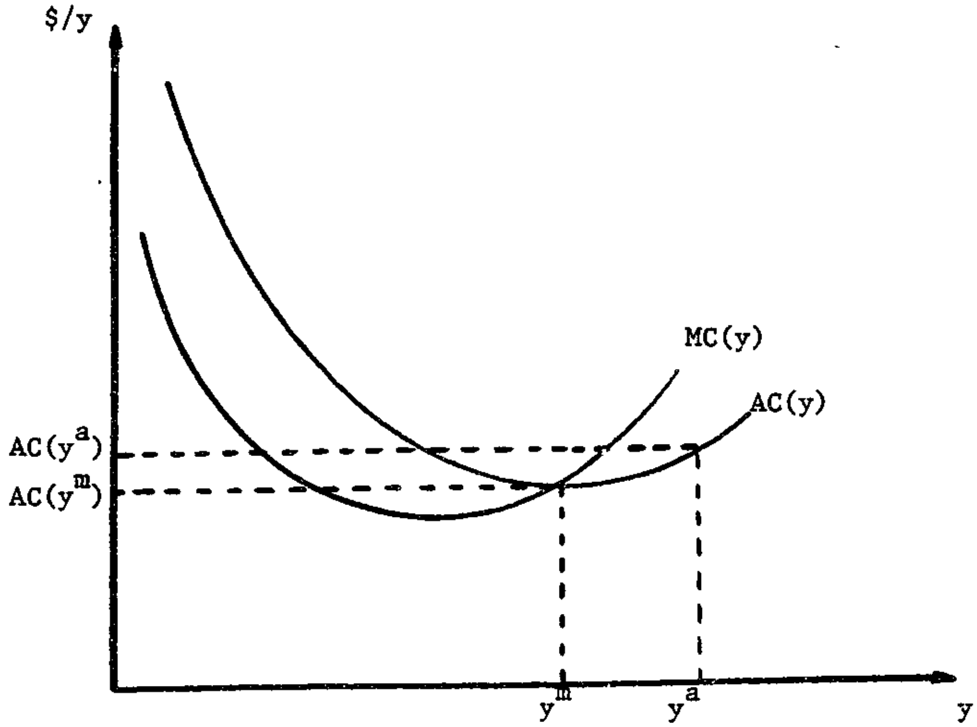


Figure 1.1

Subadditivity with Increasing Average Costs but $\frac{\partial C(y)}{\partial y} = MC(y)$. Therefore,

$$\frac{1}{y}\,[P - AC(y)] < 0, \text{ or } \; Py < AC(y)y. \tag{1.7}$$

This indicates that revenues do not cover total cost under increasing returns and marginal cost pricing. However, a natural monopoly like the one depicted in Figure 1.1 would cover costs under marginal cost pricing, i.e. a natural monopoly may be profitable, but that price would not be sustainable in the sense that some other firm may enter the market producing y^m and charging $AC(y^m)$, thus attracting consumers and covering costs (Panzar and Willig, 1977). This is not possible for any $y < y^m$; average cost pricing for such outcomes is sustainable, i.e., decreasing average costs creates a case for natural monopoly with sustainable (average cost) prices.

The degree of scale economies at y is defined as

$$S = \frac{C(y)}{MC(y)y} = \frac{AC(y)}{MC(y)}, \tag{1.8}$$

which is equal to the ratio of total costs over total revenues that would obtain from marginal cost pricing. Returns to scale are increasing, constant, or decreasing as S is greater, equal, or less than 1.

Summarizing, if total industry output amounts to $y \leq y^m$, there are incentives for firms to merge but first-best pricing is not profitable. The implications in terms of regulatory policies are important (second-best pricing, subsidies, etc.). We will see that the relevance of "economies from output expansion" is even greater when going into the multiple-output formulation, where output can be expanded in scale and/or scope.

1.3 Multiple-Output Natural Monopoly

Let us expand y to be an output vector $y = \{y_1, y_2, \ldots, y_m\}$ and let M be the set of products. Under some regularity conditions on technology,[5] it is possible to define a transformation function F(x,y) such that

$$F(x,y) \geq 0 \text{ if and only if } (x,y) \varepsilon T,$$

where equality holds for efficient input-output combinations.[6] A straight extension of the concept of scale economies given in (1.1) states that the technology T exhibits economies of scale at (x,y) if (Panzar and Willig, 1977)

$$(x,y) \varepsilon T \rightarrow \exists\, n > 1 \mid (\lambda x, \lambda^n y)\ \varepsilon T,\ \lambda > 1 \tag{1. 9}$$

Alternatively, an extension of (1.2) states that T exhibits economies of scale at (x,y) if

$$(x,y) \varepsilon T \rightarrow \exists\, \mu / (\lambda x, \mu y) \varepsilon T,\ 1 < \lambda < \mu \tag{1.10}$$

Global economies of scale are said to be present when the preceding conditions hold for all input and output combinations. In the single output case we have seen a relation between scale economies and average costs, that allowed for an analysis of the former in terms of the cost function. We now face the problem of redefining average costs in the case of production of an output bundle, and this is how the concept of ray average costs (RAC) emerged (Baumol, 1976). RAC are said to be strictly declining at y if

$$\frac{\partial [C(w,vy)/v]}{\partial v} < 0, \tag{1.11}$$

5/ In short they are: 1) $(0,y) \varepsilon T \leftrightarrow y=0$, and 2) increasing some input use allows to either increase or leave unchanged the amount of outputs.

6/ In the single output case, $F(x,y) = f(x) - y = 0$.

with v in the neighborhood of one. RAC are declining everywhere if

$$\frac{C(w,vy)}{v} < C(w,y), \quad v > 1, \ \forall y. \tag{1.12}$$

One of the key aspects in multiple-output analysis is the fact that strict global economies of scale are sufficient but not necessary for ray average costs to be declining, the basic reason being that scale economies require inputs to change proportionately and this will not necessarily minimize the cost of an expansion; however, if unit costs decrease when inputs are increased proportionately, they must certainly decrease along the least-cost expansion path. This can be better understood through the particular case of a production process involving two inputs and two outputs, characterized by a transformation function $F(x_1, x_2, y_1, y_2) = 0$. Let us start with a situation depicted by (x^o, y^o), shown in Figure 1.2 in both input and output spaces. The curve $F(x_1, x_2, y_1^o, y_2^o) = 0$ represents all combinations of x_1 x_2 which are able to produce the output vector (y_1^o, y_2^o); it is, thus, the multiple-output concept of an isoquant. On the other hand, $F(x_1^o, x_2^o, y_1, y_2) = 0$ presents all combinations of y_1 and y_2 which can be produced with the input combination (x_1^o, x_2^o), i.e., the production possibility locus. If inputs are expanded at a scale $v > 1$, the new production possibilities will be represented by $F(vx_1^o, vx_2^o, y_1, y_2) = 0$, where some point will correspond to a proportional expansion from the initial output (y_1^o, y_2^o). This point corresponds to $(\mu y_1^o, \mu y_2^o)$, the intersection between a ray from the origin passing through (y_1^o, y_2^o) and the new production locus. The corresponding "isoquant" in the input space is $F(x_1, x_2, \mu y_1^o, \mu y_2^o) = 0$. If the transformation function is not homothetic, then the minimum cost input combination will differ from (vx_1^o, vx_2^o) and,

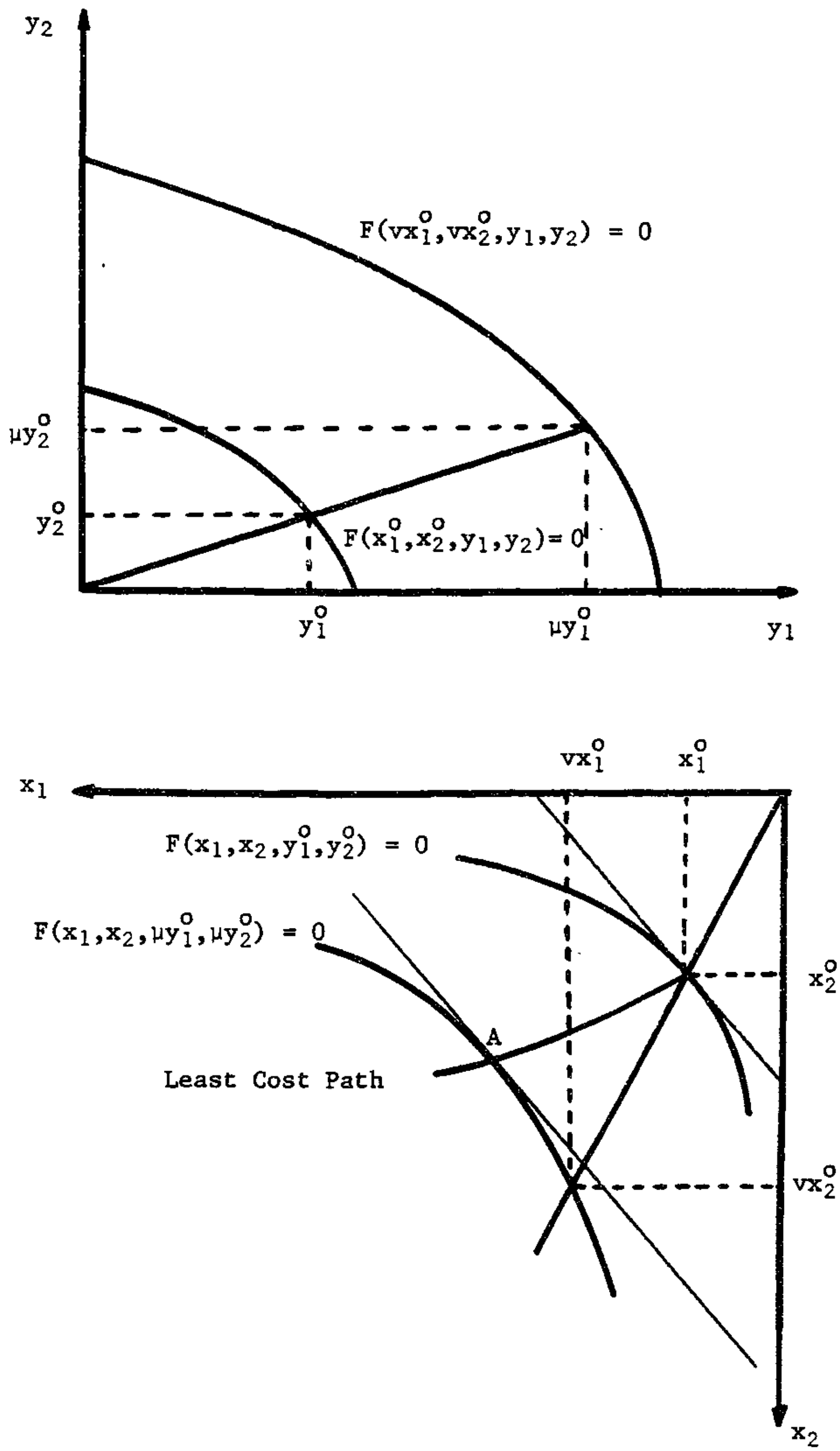


Figure 1.2: Economies of Scale in the Two Input-Two Output Case

therefore, this input combination will not be in general the best (cheapest) way of producing $(\mu y_1^o, \mu y_2^o)$. The optimal point is A in Figure 1.2. Summarizing, we have:

$$C(\mu y_1^o, \mu y_2^o) < w_1 v x_1^o + w_2 v x_2^o = v(w_1 x_1^o + w_2 x_2^o) = vC(y_1^o, y_2^o). \quad (1.13)$$

If, moreover, scale economies are present, i.e., $\mu > v$, then

$$vC(y_1^o, y_2^o) < \mu C(y_1^o, y_2^o) \therefore \frac{C(\mu y_1^o, \mu y_2^o)}{\mu} < C(y_1^o, y_2^o), \quad (1.14)$$

which proves sufficiency (scale economies → declining ray average costs). However, necessity can not be proved, i.e., with a nonhomothetic transformation function, ray average costs may decline even without scale economies, because the least-cost expansion point will imply a different proportion of inputs that may compensate for some degree of diseconomies of scale.

The unprofitability of marginal cost pricing under scale economies (shown in (1.7) for the single-output case), still holds in the multiple-output case. To see this, note that

$$\frac{\partial[C(vy)/v]}{\partial v} = \frac{[\Sigma \frac{\partial C(vy)}{\partial (vy_j)} y_j] v - C(vy)}{v^2} < 0 \quad (1.15)$$

under declining ray average costs; setting $v = 1$ (local measure), we get

$$\Sigma \frac{\partial C(y)}{\partial y_j} y_j - C(y) < 0. \quad (1.16)$$

But $\partial C(y)/\partial y_j$ is the price P_j of output j under marginal cost pricing, therefore total revenues $(\Sigma P_j y_j)$ are less than total costs. Therefore, scale economies → declining ray average costs → unprofitability of marginal cost pricing.

The preceding paragraphs suggests that ray analysis is in fact similar to that of a single output, defining this latter in terms of a basic output bundle, e.g. (y_1^o, y_2^o), in terms of which proportional expansions are studied.[7] A multiproduct generalization of the degree of scale economies S in (1.8) is $S_M(y)$, a local measure given by

$$S_M = \frac{C(y)}{y\nabla C(y)} = \frac{C(y)}{\sum_{i=1}^{m} y_i \frac{\partial C(y)}{\partial y_i}} \qquad (1.17)$$

Under usual regularity conditions on T, S_M is also the maximal proportionate growth rate of outputs along their ray as all inputs are expanded proportionally. Of course, (1.17) is consistent with the unprofitability of marginal cost pricing under scale economies (S>1).

The concept of ray concavity completes the ray-related set of definitions; in short, it means declining marginal costs on the "curve" of costs associated with a ray (Figure 1.3). With this definition, ray concavity and C(0) = 0 (which is a technological assumption) imply declining ray average costs, but the converse is not necessarily true; this is intuitively clear from the single-output case, where declining marginal costs is a sufficient condition for declining average costs, but is not a necessary condition. Neither ray concavity nor declining ray average costs are necessary for ray subadditivity (but declining ray average costs do imply this restricted type of subadditivity).[8] This statement is proved by Figure 1.4, where C(y) is not concave and AC(y) increases over BC; however, C(y) is strictly subadditive because

$$C(y) \leq OA + AD < n \cdot OA \quad (n > 1) \qquad (1.18)$$

for any output y. In short, scale economies are sufficient but not necessary

7/ In fact, ray average cost can be defined as $\frac{C(y)}{v\sum_i y_i^o}$, where y^o is the basic bundle and $\sum_i y_i^o = 1$ by definition.

8/ Formally, ray subadditivity is present at y if $\sum_i C(v_i y) > C(y)$, with $\sum_i v_i = 1$ (Baumol, Panzar and Willig, 1979).

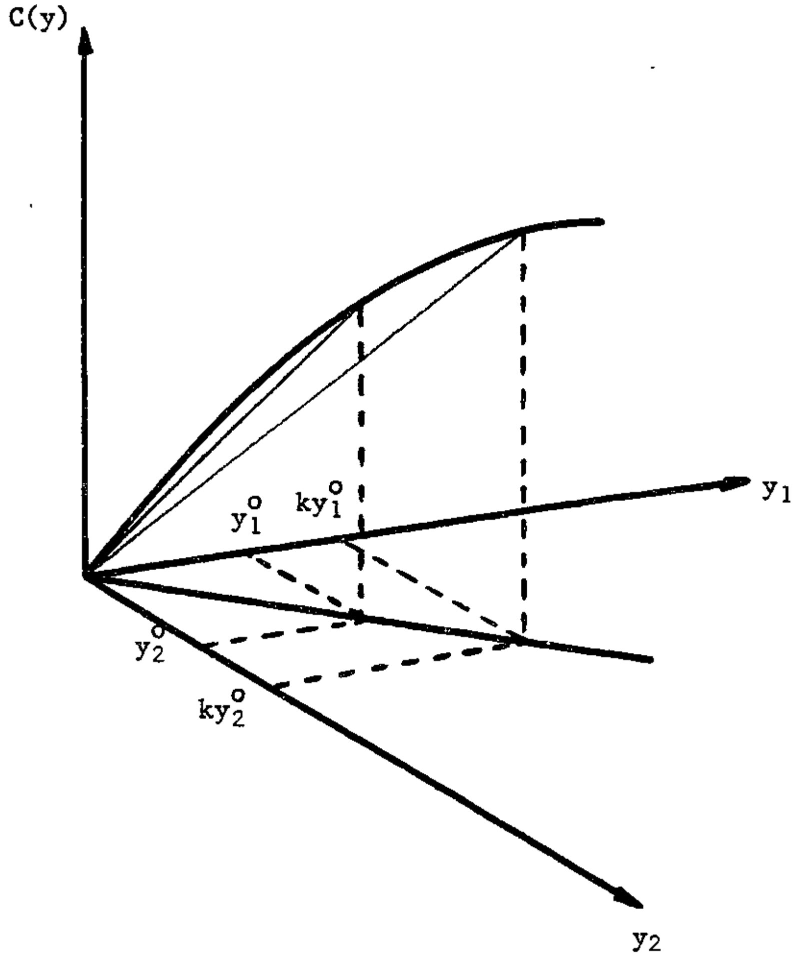


Figure 1.3

Declining Ray Average Cost, Ray Concavity

and Ray Subadditivity

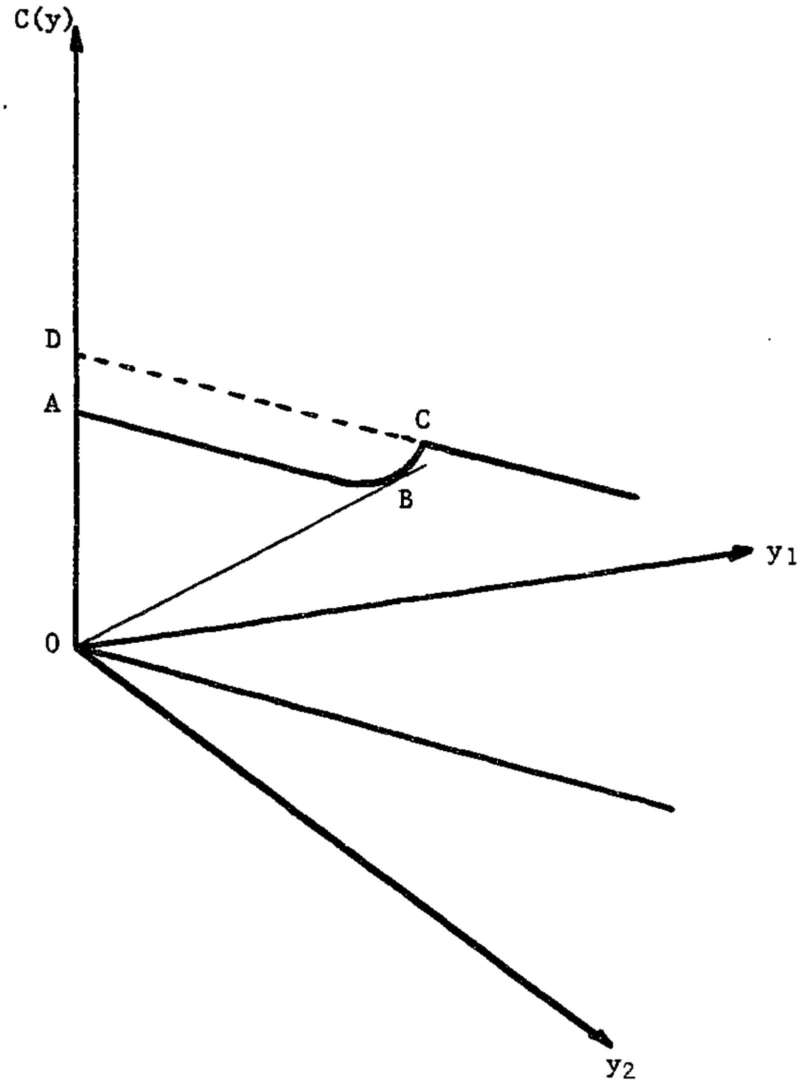


Figure 1.4

Ray Subadditivity Without Ray Concavity

for ray subadditivity.

The preceding paragraphs have shown the necessity of analyzing possible changes in output combinations when expanding production in a multiple-output framework, because we have seen that ray properties of the cost function are not enough to study technology. What is maybe more important is that ray properties alone tell very little about natural monopoly, i.e., whether one firm can produce cheaper than many firms; as scale economies are a ray concept, the need to go beyond is clear.

The analysis of complementarity in production, i.e. the convenience or not of producing two outputs in conjunction, cannot be performed from ray related properties of C(y). One way to depart from ray analysis is to study the behavior of C(y) as the level of production of a particular product y_i varies, keeping the rest of the bundle at some positive level. The incremental cost IC_i is defined as (Panzar and Willig, 1977)

$$IC_i(y) \equiv C(y) - C(y_{M-i}) \equiv C(y) - C(y\ , \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_M),\text{[9]} \quad (1.19)$$

i.e. is the cost of producing y_i, in addition to a given bundle at a given level. The average incremental cost AIC_i and the degree of scale economies specific to y_i at y, $S_i(y)$, are defined as

$$AIC_i(y) = \frac{IC_i(y)}{y_i} \quad (1.20)$$

$$S_i(y) = \frac{IC_i(y)}{y_i \frac{\partial C(y)}{\partial y_i}} = \frac{AIC_i(y)}{\frac{\partial C(y)}{\partial y_i}} \quad (1.21)$$

respectively. Naturally, product specific returns to scale are said to be increasing, constant or decreasing as $S_i(y)$ is greater than, equal to,

9/ In general, we will denote y_L a vector such that $y_i = 0$, $i \epsilon \{M-L\}$.

or less than one, respectively.[10] Figure 1.5 illustrates these concepts; there, $IC_2(y^o) = BC$, $AIC_2(y^o) = \frac{BC}{AB}$, and $S_2(y^o) > 1$. These same concepts can be extended to a subset T of products. In this case, $AIC_T(y)$ is a ray-like concept, but the ray does not go through the origin; rather, components M-T of y are held fixed. The <u>degree of scale economies specific to a subset T of M</u> is given by

$$S_T(y) = \frac{C(y) - C(y_{M-T})}{\sum_{j\varepsilon T} y_j \frac{\partial C(y)}{\partial y_j}} = \frac{IC_T(y)}{\sum_{j\varepsilon T} y_j \frac{\partial C(y)}{\partial y_i}} \tag{1.22}$$

Again, if $S_T(y) > 1$ then marginal cost pricing does not cover incremental costs, as in (1.21).

The main concept related to the convenience of producing output bundles as opposed to isolated outputs, is that of <u>economies of scope</u>. Economies of scope (Panzar and Willig, 1975) are said to exist over the product set M at y if and only if

$$C(y) < \sum_{i=1}^{k} C(y_{T_i}), \quad \underset{i}{\cup} T_i = M, \; T_i \neq M, \; T_i \cap T_j = \phi .\text{[11]} \tag{1.23}$$

This is, economies of scope are present if production of an output bundle by one firm is cheaper than production by many firms of subsets of that bundle at the same level.[12] In Figure 1.6, points A, B, and D belongs to

10/ The presence of increasing product specific returns to scale indicates that at least that product should be produced by one firm (evantually jointly with others).

11/ In other words, $\{T_i\}$ is a non-trivial partition of the product set $M_j y_{T_i}$ is orthogonal to Y_{T_j}, $i \neq j$.

12/ In the two outputs case, economies of scope are present if $C(y_1, y_2) < C(y_1, 0) + C(0, y_2)$. In short, (1.25) means strict orthogonal subadditivity of C(y).

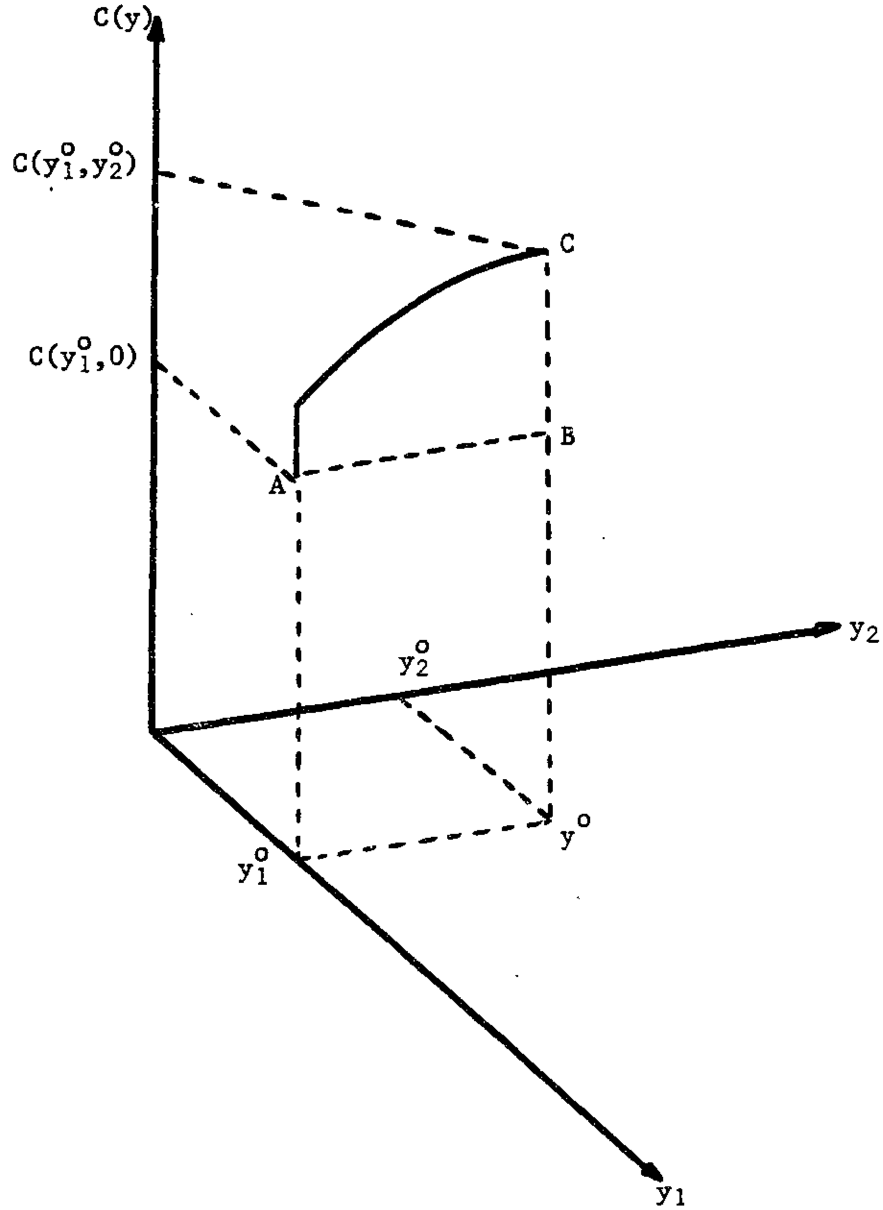


Figure 1.5

Incremental Analysis

the cost surface $C(y_1, y_2)$. It can be easily shown that $E = [y_1^o, y_2^o, C(y_1^o, 0) + C(0, y_2^o)]$ belongs to the plane P (determined by the origin, A and D).[13/] As $B = [y_1^o, y_2^o, C(y_1^o, y_2^o)]$ is below $E \varepsilon P$, we have economies of scope at y^o. The degree of economies of scope (Baumol, Panzar, and Willig, 1979) is defined at y relative to T as

$$SC_T(y) = \frac{C(y_T) + C(y_{N-T}) - C(y)}{C(y)} , \qquad (1.24)$$

such that $SC_T(y) > 0$ implies the presence of economies of scope. Under this definition, it can be proved (Baumol, Panzar, and Willig, 1979) that

$$S_M(y) = \frac{\alpha_T S_T(y) + (1 - \alpha_T) S_{N-T}(y)}{1 - SC_T(y)} , \text{ where} \qquad (1.25)$$

$$\alpha_T = \frac{\sum_{j \varepsilon T} y_j \frac{\partial C(y)}{\partial y_j}}{\sum_{j=1}^{m} y_j \frac{\partial C(y)}{\partial y_j}} . \qquad (1.26)$$

(1.25) clearly indicates that in the absence of economies of scope, overall scale economies would be a weighted average of product specific scale economies. However, economies of scope magnify these latter in the determination of the former. The economic intuition behind the formal relation between scale and scope summarized by (1.25), is that cost advantages of expanding the level of some outputs being produced in isolation, are increased when they are produced jointly and expanded. The traditional idea of complementarity

---

13/ The equation of the plane P (through the origin) is $C + \alpha y_1 + \beta y_2 = 0$. $A \varepsilon P \rightarrow C(y_1^o, 0) + \alpha y_1^o = 0$. $D \varepsilon P \rightarrow C(0, y_2^o) + \beta y_2^o = 0$. At $E = (x, y_1^o, y_2^o)$ we should have $x + \alpha y_1^o + \beta y_2^o = 0$. From the three equalities, $x = C(y_1^o, 0) + C(0, y_2^o)$.

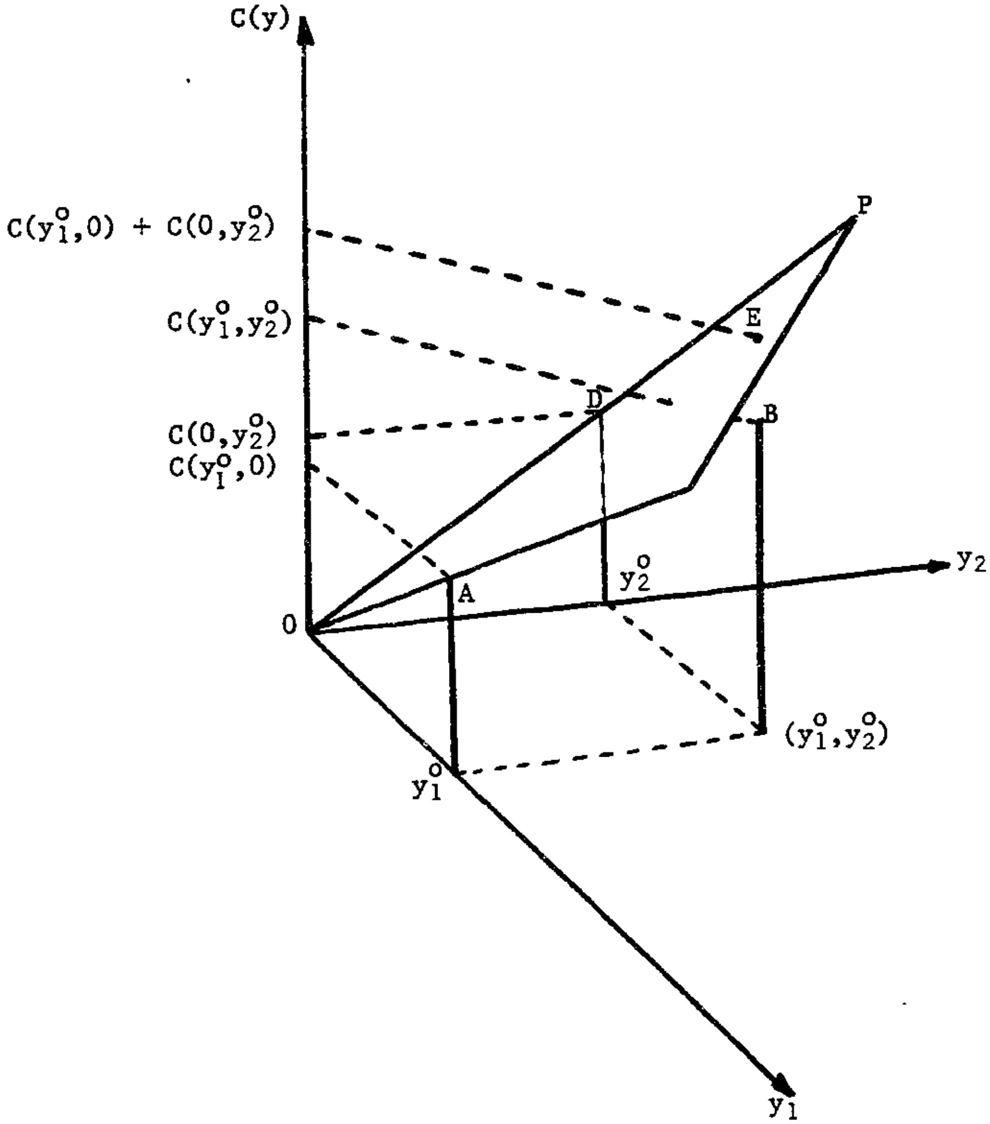


Figure 1.6

Economies of Scope

in production relates to economies of scope, in the sense that cost complementarity over the product set M at y, i.e.

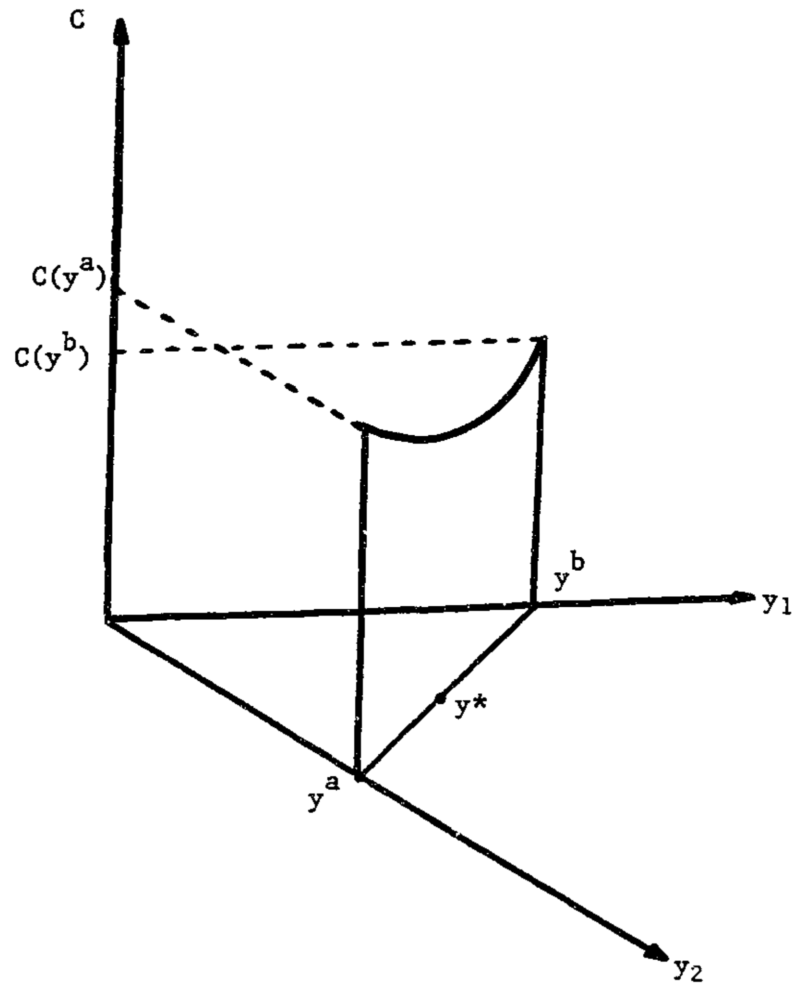
$$\frac{\partial^2 C(y')}{\partial y_i \, \partial y_j} \leq 0 \;, \quad i \neq j, \; y' \leq y \;, \qquad (1.27)$$

is a sufficient condition for economies of scope to be present at y. Finally, there are cases which can be intuitively characterized as presenting economies of scope, as for instance joint production with shared inputs (Panzar and Willig, 1975). The "public input" case is apparent: if production of all $i \epsilon M$ requires a public input P, once P is available for one product, it is available for all and the convenience of joint production is clear. Similarly, the presence of indivisibilities in the plant of the productive enterprise favors the production of other outputs. In both cases, not taking advantage of the possibilities offered by the availability of some input, creates "idle capacity" and economies of joint production (scope).

Another way to deal with cost advantages of output bundles is through the analysis of C(y) on a hyperplane defined by $\Sigma \mu_i y_i = \mu$, $\mu_i > 0$, $\mu > 0$. We will say that a cost function is transray convex at y if

$$C[ky^a + (1-k)\, y^b] \leq k\, C(y^a) + (1-k)\, C(y^b), \; 0 < k < 1 \qquad (1.28)$$

for y^a and y^b contained in a hyperplane through y. Figure 1.7 shows the shape of a transray convex cost function in a two-outputs case; the hyperplane there takes the form of a line of negative slope in the y_1, y_2 plane. The presence of transray convexity favors the production of many outputs by one firm instead of many firms, each one producing a subset of outputs. Therefore, transray convexity works in favor of subadditivity, while concavity works against it. This reinforces the idea that scale economies (a ray concept) are not sufficient for global subadditivity. Somewhat



$y^a = (0, y_2)$

$y^b = (y_1, 0)$

$y^*, y^a, y^b, \in \{y \,|\, \Sigma \mu_i y_i = \mu;\ \mu_i,\ \mu > 0\}$

Figure 1.7

Transray Convexity

related to transray convexity is the concept of quasi-convexity of a cost function, which is itself related to iso-cost surfaces.[14/] C(y) is quasi-convex over y^o if the set $\{y/C(y) \leq C(y^o)\}$ is a convex set. In the two outputs case, $C(y) = C(y^o)$ generates an iso-cost curve in the output space, and quasi-convexity makes these curves concave to the origin as in Figure 1.8; there, cost analysis on a transray hyperplane (line) suggests that C(y) should have the shape of Figure 1.7. However, and somewhat counterintuitively, neither quasi-convexity implies transray convexity, nor the latter implies the former (in formal analytical terms).

The question to be addressed now is, under what conditions is a cost function subadditive? This is, when are we in the presence of a natural monopoly? It should be at this point clear that scale economies as a ray concept are neither necessary (recall Figure 1.4) nor sufficient (because of eventual diseconomies of scope) for natural monopoly. We need to combine ray and cross-ray conditions to ensure subadditivity on C(y). Each of the following sets of conditions

i) C(y) transray convex along a hyperplane, and decreasing RAC up to that hyperplane;

ii) C(y) convex, and decreasing RAC,

can be proved to be sufficient for C(y) subadditive. Intuitively, sufficiency arises in both i) and ii) from the implicit savings associated with proportional expansions of output (decreasing RAC), plus savings from output combinations. Originally, Baumol (1977) stated sufficiency conditions in terms of ray concavity and transray convexity, which are in fact particular to i) (See Figure 1.9.).

14/ Obviously defined as $\{y/C(y) = K\}$

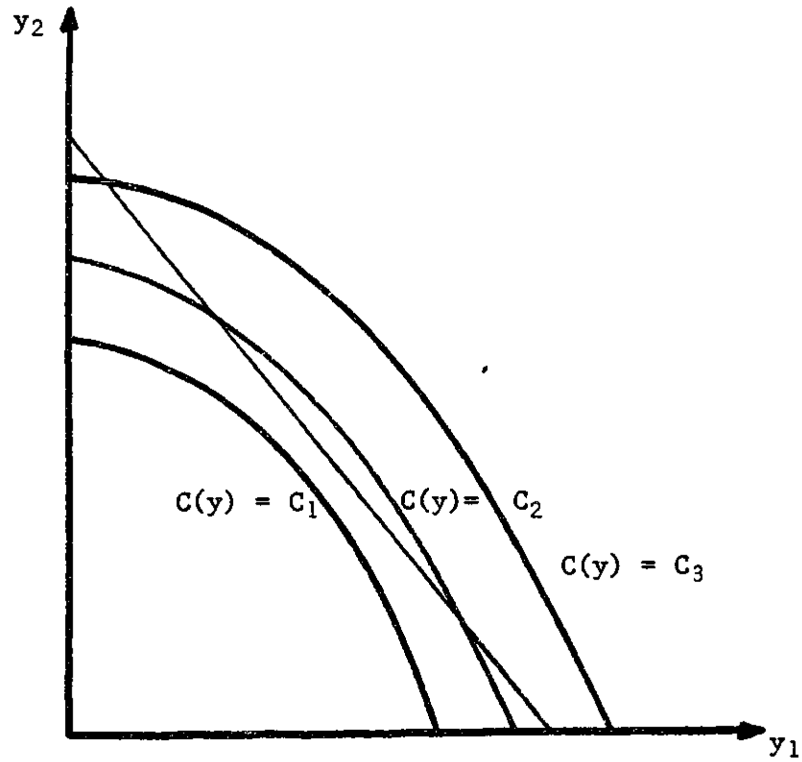


Figure 1.8

Iso-Cost Contours of a Quasi-Convex Cost Function

Although product-specific fixed costs[15] makes C(y) to violate transray convexity, we expect these kind of costs (as well as global fixed costs) to favor subadditivity. In fact, without losing generality a cost function can always be written as

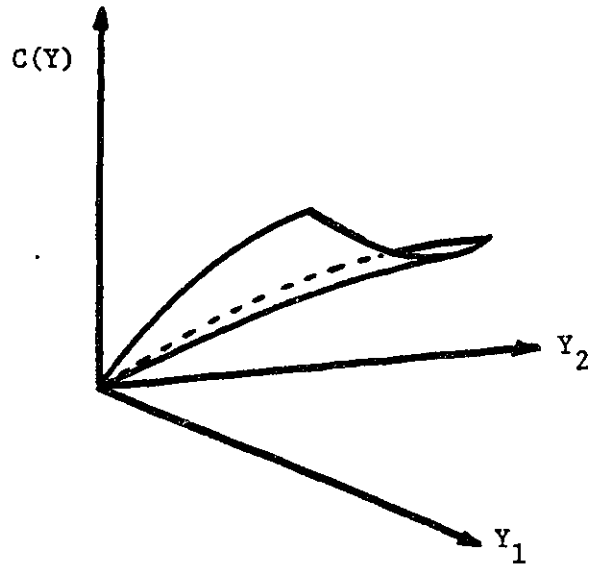
$$C(y) = F(S) + C_1(y) \ , \qquad (1.29)$$

where $S = \{i \varepsilon M / y_i > 0\}$. Here, F(S) includes the case $F(S) = \sum_{i \varepsilon S} F_i$ as a particular one, where F_i is the fixed cost associated to Y_i. The generality of F(S) in (1.31) lies upon the fact that the "fixed" cost depends on the whole set of products actually being produced. It can be shown (and is intuitively clear) that if $F(S \cup T) \leq F(S) + F(T)$, then subadditivity of $C_1(y)$ implies subadditivity of C(y) (Baumol, Panzar, and Willig, 1979). This nice property allows for a restricted analysis in terms of $C_1(y)$ under the required conditions.

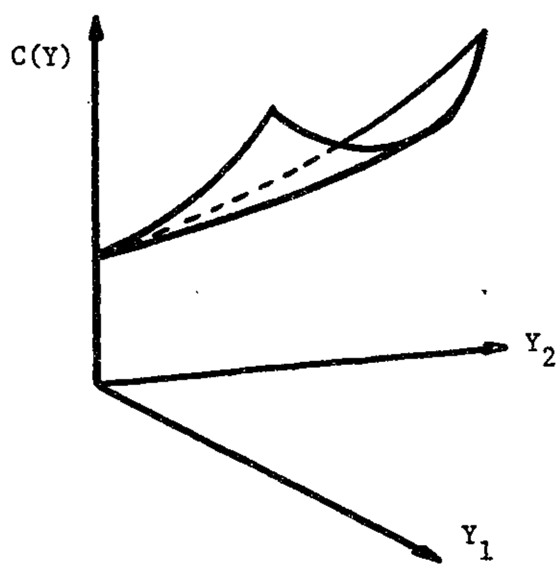
Finally, a general test for multioutput natural monopoly from actually estimated cost functions, has not yet been developed.[16] Economies from proportional expansions of output are detected by $S_M \geq 1$ (multiproduct degree of scale economies greater than 1), which implies ray subadditivity. On the other hand, $\frac{\partial^2 C}{\partial y_i \partial y_j} < 0$ (production complementarity), generates economies from the production of bundles as opposed to isolated goods. The presence of both conditions on C(y) for a product bundle M provides a case for natural monopoly, although they still may constitute too strong an imposition on C(y). However, they constitute an analytically tractable set of conditions, which makes them undoubtedly attractive.

15/ i.e. costs that does not depend on the amount of that product, but on the fact that it has been added to the output bundle.

16/ The work by Baumol and Braunstein (1977) on journal publications only considered two outputs. In this case transray convexity and ray concavity can be easily stated in terms of a single output.



a) Transray Convexity and Ray Concavity



b) Convexity and Decreasing RAC

Figure 1.9

Subadditive Cost Functions

1.4 Toward a Workable Test of Subadditivity

Coupled ray and transray properties of a multioutput cost function have been proved to be sufficient conditions for subadditivity. Both types of properties are related to the curvature of C(Y) along hyperplanes; it is intuitively feasible and practically desirable to state these conditions in an analytically tractable form. Second derivatives of C(Y) should provide all the necessary information for curvature-related analysis. In this section we explore different procedures to analyze sufficiency conditions.

1.4.1 Overall Convexity of the Cost Function

We have already seen that the combination of a convex cost function and diminishing ray average costs until Y, makes C(Y) subadditive at Y. This provides an immediate test for subadditivity. It should be remembered that the multioutput measure of the degree of returns to scale, S_M, is in fact a ray-related quantity; moreover, $S_M > 1$ suffices for diminishing ray average costs. Therefore, the presence of scale economies in a convex cost function suffices for subadditivity.

For this test to be passed, we required a positive definite Hessian of C(Y), and $S_M > 1$. It should be remembered that a positive definite Hessian is equivalent to all the characteristic roots of that matrix of second derivatives, being positive. This in turn implies that all principal minors of the Hessian should be positive, including $C_{ii} = \partial^2 C/\partial Y_i^2$. Therefore, the first thing to do is to check the sign of the diagonal elements of the Hessian; if these are all positive, overall convexity should be analyzed. If this test fails, subadditivity may still be present, but it requires more analysis.

1.4.2 Transray Convexity and Ray Concavity

Transray convexity, as stated in Baumol (1977), is said to be present in C(Y) at Y^* if

$$C[\alpha Y^a + (1-\alpha)Y^b] \leq \alpha C(Y^a) + (1-\alpha)C(Y^b), \ \forall\alpha, \quad 0 < \alpha < 1$$

where Y^a and Y^b are output vectors lying in some hyperplane $\sum_i w_i Y_i = w$ through Y^*, with $w_i > 0 \ \forall i$. This is equivalent to saying that C(Y) is convex along a hyperplane; this condition can be studied from the bordered Hessian corresponding to the problem

$$\begin{aligned} & \text{Min } C(Y) \\ & \text{subject to} \\ & \sum_i w_i Y_i - w = 0 \quad , w_i > 0 \\ & Y_i > 0 \quad , \end{aligned} \tag{1.30}$$

which is given by:

$$H_{BT} = \begin{bmatrix} \frac{\partial^2 C}{\partial Y_1^2} & \frac{\partial^2 C}{\partial Y_1 \ Y_2} & \cdot & \cdot & \frac{\partial^2 C}{\partial Y_1 \partial Y_n} & w_1 \\ \frac{\partial^2 C}{\partial Y_1 \partial Y_2} & \frac{\partial^2 C}{\partial Y_2^2} & \cdot & \cdot & \frac{\partial^2 C}{\partial Y_2 \partial Y_n} & w_2 \\ \cdot & \cdot & & & \cdot & \\ \cdot & \cdot & & & \cdot & \\ \frac{\partial^2 C}{\partial Y_1 \ Y_n} & \frac{\partial^2 C}{\partial Y_2 \partial Y_n} & \cdot & \cdot & \frac{\partial^2 C}{\partial Y_n^2} & w_n \\ w_1 & w_2 & \cdot & \cdot & w_n & 0 \end{bmatrix} . \tag{1.31}$$

Then C(Y) is convex along the hyperplane if the Hessian along this latter is positive definite, which will occur if and only if $d^i < 0$ for $i = 2,\ldots,n$, where

$$d_i = \det \begin{vmatrix} \frac{\partial^2 C}{\partial Y_1^2} & \cdot & \cdot & \cdot & \frac{\partial^2 C}{\partial Y_1 \partial Y_i} & w_1 \\ \cdot & & & & \cdot & \cdot \\ \cdot & & & & \cdot & \cdot \\ \cdot & & & & \cdot & \cdot \\ \frac{\partial^2 C}{\partial Y_1 \partial Y_i} & & \cdot & \cdot & \frac{\partial^2 C}{\partial Y_i^2} & w_i \\ w_1 & \cdot & \cdot & \cdot & w_i & 0 \end{vmatrix} \qquad i = 2,\ldots,n \qquad (1.32)$$

Ray concavity can be analyzed in a somewhat easier way, by recalling that a ray through $Y = \{ Y_1,\ldots,Y_n \}$ determines a direction in R^n. In general, the variation of C(Y) along any direction $U = (U_1, U_2, \ldots, U_n)$ is given by

$$\frac{\partial C}{\partial U} = \frac{\partial C}{\partial Y_1} U_1 + \frac{\partial C}{\partial Y_2} U_2 + \cdot \cdot \cdot + \frac{\partial\ C}{\partial Y_k} U_k \quad , \qquad (1.33)$$

To analyze the curvature of C(Y) along U, we have to study the variation of $\partial C/\partial U$ along the same direction. Then

$$\frac{\partial^2 C}{\partial U^2} = \frac{\partial^2 C}{\partial U \partial Y_1} U_1 + \frac{\partial^2 C}{\partial U \partial Y_2} U_2 + \cdot \cdot \cdot + \frac{\partial^2 C}{\partial U \partial Y_k} U_k \quad , \qquad (1.34)$$

which corresponds to

$$\frac{\partial^2 C}{\partial U^2} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 C}{\partial Y_i \partial Y_j} U_i U_j = UHU' \quad , \tag{1.35}$$

where H is the Hessian of C(Y).[17/]

Therefore C(Y) will be ray concave along a ray through Y if $YHY' < 0$.

Let us apply these concepts to $Y = (Y_1, Y_2)$, the simplest multi-output bundle. The conditions for transray convexity reduce to

$$d_2 = \begin{vmatrix} C_{11} & C_{12} & w_1 \\ C_{12} & C_{22} & w_2 \\ w_1 & w_2 & 0 \end{vmatrix} = 2w_1w_2C_{12} - w_1^2 C_{22} - w_2^2 C_{11} < 0 \; , \tag{1.36}$$

where $C_{ij} = \partial^2 C/\partial Y_i \partial Y_j$. On the other hand, ray concavity requires

$$[y_1 \; Y_2] \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = Y_1^2 C_{11} + Y_2^2 C_{22} + 2Y_1Y_2C_{12} < 0 \; . \tag{1.37}$$

Recalling that $w_i > 0$ and $Y_i > 0$, we can reduce (1.36) and (1.37) to

$$2aC_{12} - a^2C_{22} - C_{11} < 0 \quad , \; a > 0 \tag{1.38}$$

$$2kC_{12} + k^2C_{22} + C_{11} < 0 \quad , \; k > 0 \; . \tag{1.39}$$

Therefore, if (1.38) and (1.39) hold for some finite positive values of a and k, C(Y) is subadditive. In particular, note that if C(Y)

17/ Of course, for $U = (0,0,\ldots,1,\ldots,0)$, with $U_i = 1$, $\partial^2 C/\partial U^2$ reduces to $\partial^2 C/\partial Y_i^2$.

is concave on Y_i, i = 1,2, and $C_{ij} < 0$ (i.e., weak production complementarity is present), then (1.38) holds irrespective of a. The same conditions applied to (1.39), however, suggests that the cross effect should be greater than the (sum of) own effects of outputs in cost.

Baumol and Braunstein (1977) proposed an initial specification in their applied study on journal publication, namely

$$C = b_0 + b_1Y_1 + b_2Y_2 + b_{12}Y_1Y_2 \ . \tag{1.40}$$

(1.38) and (1.39) lead to the same condition on the parameters of (1.40), namely

$$2ab_{12} < 0 \ . \tag{1.41}$$

Therefore, a test on subadditivity reduces to a test on the sign of b_{12}. A slight expansion of (1.40) toward a quadratic form

$$C = b_\theta + b_1Y_1 + b_2Y_2 + b_{11}Y_1^2 + b_{22}Y_2^2 + b_{12}Y_1Y_2 \tag{1.42}$$

leads to conditions

$$2a\ b_{12} - 2a^2b_{22} - 2b_{11} < 0 \tag{1.43}$$

$$2k\ b_{12} + 2k^2b_{22} + 2b_{11} < 0 \ . \tag{1.44}$$

It is clear that a set of values fulfilling

$$b_{12} < 0 \ , \quad b_{11} > 0 \ , \quad b_{22} > 0 \ , \quad \text{and} \tag{1.45}$$

$$-b_{12} > b_{11} + b_{22} \tag{1.46}$$

satisfy (1.38) and (1.39) for $a = k = 1$. These are also satisfied by the alternative set

$$b_{12} < 0 \quad , \quad b_{11} < 0 \quad , \quad b_{22} < 0 \quad , \quad \text{and} \tag{1.47}$$

$$|b_{12}| > |b_{11} + b_{22}| \tag{1.48}$$

On the other hand, the Hessian of C(Y), namely

$$H = \begin{bmatrix} 2b_{11} & b_{12} \\ b_{12} & 2b_{22} \end{bmatrix} \quad , \tag{1.49}$$

has a determinant which under (1.45) and (1.46) is negative. As b_{11} and b_{22} are positive, H is neither positive nor negative definite, i.e., C(Y) is neither concave nor convex. This can also be seen from the calculation of the eigenvalues of H, which under the same conditions have opposite signs.

1.4.3. Transray Convexity by Output Pairs

In section 1.3 we suggested that the presence of weak cost complementarities among pairs of products in an output bundle, plus the presence of increasing returns to scale (i.e., $S_M > 1$), should be sufficient for subadditivity. The rationale behind this proposition is similar to that behind any test of this sort, that is, economies from proportional expansion and economies from product combinations favors subadditivity. $S_M > 1$ provides a simple analytical test for the presence of ray subadditivity in C(Y), which is actually less demanding than ray concavity. This allows us to concentrate on output combinations, and particularly on transray convexity; let us further investigate the role of C_{ij} $(= \partial^2 C/\partial Y_i \partial Y_j)$ on this property of C(Y).

We have seen that transray convexity in the bi-output case corresponds to

$$E = 2a\, C_{ij} - C_{jj} - a^2 C_{ii} < 0 \quad , a > 0 \quad . \tag{1.50}$$

The question to be asked is whether there exists at least some $a > 0$ for which $E < 0$. Let us view E as E(a) and analyze its behavior under different conditions on C_{ij}. To do this, note that $E(a) = 0$ leads to

$$a = \frac{1}{C_{ii}} \left[C_{ij} \pm \sqrt{C_{ij}^2 - C_{ii}C_{jj}}\right] \quad , \tag{1.51}$$

and also note that the curvature of E(a) is given by the sign of $-C_{ii}$. Analyzing shape and roots of E(a), it can be shown that there are only two cases for which no transray plane exists such that C(Y) is convex along it (see Appendix 1.1). Both cases involve negative "own effects" $(C_{ii} < 0,\ C_{jj} < 0)$; if this happens, C(Y) is transray convex along some plane only if $C_{ij} < 0$ and $C_{ij}^2 > C_{ii}C_{jj}$, i.e. only if weak production complementarity is present <u>and</u> it is greater in absolute value than the geometric mean of the own second derivatives of C(Y). If either cost complementarity is absent or $C_{ij}^2 < C_{ii}C_{jj}$, then negative own second derivatives of C(Y) will make transray convexity impossible. The economic intuition behind this is that concavity of C(Y) in Y_i (i.e., $C_{ii} > 0$) indicates a cost advantage when expanding production of Y_i alone, while the absence of production complementarity (i.e.,

$C_{ij} > 0$) indicates a disadvantage when producing Y_i and Y_j together. Therefore, movements toward specialization (at least locally) are advantageous. A first conclusion, then, is that the presence of weak cost complementarity will usually make C(Y) tranray convex, but its absence does not necessarily imply the absence of tranray convexity. Secondly, as the order in which output components are arranged is arbitrary in (1.34), then condition (1.47) should hold for every pair of outputs under transray convexity. This implies that if $C_{ii} < 0$, $C_{jj} < 0$ and $C_{ij} < 0$ for some i,j, then there is no hyperplane $\sum_{i=1}^{k} w_i Y_i = w$ such that C(Y) is convex along it. However, even if this is the case, C(Y) may still be subadditive.

1.4.4 A Procedure to Analyze Quadratic Forms

A quadratic cost function around the mean $\{\overline{Y}_i\}$ has the form

$$C(Y) = A_0 + \sum_{i=1}^{k} A_i(Y_i - \overline{Y}_i) + \sum_{i=1}^{k} A_{ii}(Y_i - \overline{Y}_i)^2 +$$

$$1/2 \sum_{i} \sum_{j \neq 1} A_{ij}(Y_i - \overline{Y}_i)(Y_j - \overline{Y}_j) \quad . \tag{1.52}$$

A sufficient condition for ray subadditivity at $\overline{Y}$ is

$$S_M = \frac{A_0}{\sum_{i=1}^{k} A_i \overline{Y}_i} > 1 \quad . \tag{1.53}$$

The Hessian of C(Y) is given by

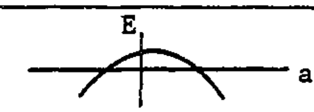
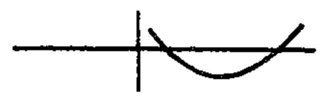
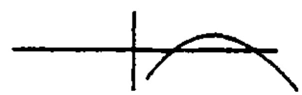
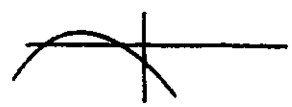
$$H = \begin{bmatrix} 2A_{11} & A_{12} & \cdots & A_{1k} \\ A_{12} & 2A_{22} & \cdots & A_{2k} \\ \vdots & & & \\ A_{1k} & A_{2k} & \cdots & 2A_{kk} \end{bmatrix} . \tag{1.54}$$

If $A_{ii} > 0$ $\forall i$, H may be positive definite. If it is, and (1.53) holds, then C(Y) is subadditive. If it is not, then possible transray convexity should be analyzed. This can be done with the help of conditions (1.32), which in this case correspond to

$$d_i = \det \begin{vmatrix} 2A_{11} & \cdots & A_{1i} & w_1 \\ \vdots & & \vdots & \vdots \\ A_{1i} & \cdots & 2A_{ii} & w_i \\ w_1 & \cdots & w_i & 0 \end{vmatrix} < 0 \quad , \quad i = 2,\ldots,k \quad . \tag{1.55}$$

If a set of positive values $\{w_1,\ldots,w_k\}$ fulfilling (1.55) can be found, then C(Y) is transray convex along the corresponding hyperplane, which together with condition (1.53) would indicate that C(Y) is subadditive. If such a hyperplane can not be found and C(Y) is not transray convex, strong economies of scale may still generate natural monopoly.

Appendix 1.1 Presence of Transray Convexity in the Bi-Output Case

C_{ii}	C_{jj}	C_{ij}	$C_{ij}^2 - C_{ii}C_{jj}$	E(a)	
+	-	any	any		
-	+	any	any		
-	-	+	+		*
-	-	-	+		
-	-	any	-		*
+	+	+	+		
+	+	-	+		
+	+	any	-		

*Under these conditions $\nexists\, a > 0 \,/\, E < 0$.

CHAPTER 2. SCALE ECONOMIES IN TRANSPORTATION: A METHODOLOGICAL REVIEW

A limited although heterogeneous literature on transportation scale economies has emerged in the last decade, from different perspectives. In this chapter we are going to review in some detail the various approaches that have been used to analyze this important aspect of transportation production. We will explicitly emphasize the methodological dimension of that research, paying particular attention to the procedure followed, the specification used and output definition, the treatment of factor prices, the assumptions, the type of data, and all kind of methodological or conceptual comments made by the authors. Naturally, the review will not be mode-specific and therefore policy implications of each study will be mentioned but not emphasized.

Two generations of approaches to the analysis of scale economies through the estimation of cost functions are presented and discussed in the first and second sections, while a discussion and synthesis of the main methodological points are offered in the third section.

2.1 Econometric Approaches: First Generation

The econometric approaches to cost functions are characterized by the procedure of direct estimation of a relation between costs (as dependent variable) and output and factor prices (as independent variables), from available empirical data. Within this gross category we have distinguished two "generations" of studies in a somewhat loose way, although not completely arbitrary. The division was based — mainly by historical sequence — on the degree of complexity of functional forms and on the treatment of independent variables, and is also somewhat related to the degree of internal microeconomic consistency.

Lee and Steedman (1970) performed an analysis of scale economies in bus transport using British data from urban areas. They selected a main dependent variable defined as "annual total working expenses of motor buses less alterations to buildings and other items" per bus-mile, which is close to an average variable cost figure. In addition they pursued the estimation of equations for some cost components in separate regressions, i.e., power costs per bus-mile, traffic operation costs per bus-mile, repair and maintenance costs per bus-mile, and management and general expenses per bus-mile. Annual bus mileage was selected on practical grounds as a measure of output for transport services; however, other measures of size were used to estimate cost components equations, such as average fuel consumption and average fleet size. In addition, the problem of estimating long-run cost functions from cross sectional observations that may not be in long-run equilibrium, was dealt with through the inclusion of a size change variable , defined as the proportional variation in vehicle mileage with respect to the preceding period. Geographical differences and the relative importance of labor were the reasons to include labor price as an independent variable. Similarly, fuel price was used in the estimation of power costs per bus-mile. Differences in the composition and quality of service, which makes the output of different firms heterogeneous, were accounted for through four variables in an effort to identify unexplained variations in cost; only two of them appeared in the final selected equations, namely the percentage of bus mileage on two-man operations and the time distribution of demand for bus services (i.e., incidence of the peak ratio). Population density, vehicle utilization, and average speed of buses in operation were used as physical and traffic environment variables, accounting for variations between geographical areas in terrain and traffic conditions. The conclusions obtained from the many equations "tried" by the authors for total average costs and cost components were far from definitive; in fact, they were somewhat elusive in postulating constant returns to scale, warning about the possible effects of changing the definition of the dependent variable. It is interesting to note, however, the effort in imporving the poor output definition by adding variables related to quality of service, geographical environment, and traffic conditions. This treatment of isolating quality or geographic characteristics, however, makes it difficult to infer anything in terms of scale economies.

Case and Lave (1970) estimated average cost functions for inland waterway transport using quarterly observations on five U.S. firms over a five-year period. The study was motivated by the large number of mergers and firm growth observed previously; the hypothesis to be tested was that returns to scale were present. The measure of output used was equivalent barge-miles (EBM), which is in itself an output index that accounts for waterway and barge equipment characteristics, thus providing a more homogeneous measure of output across firms and seasons than ton-miles. Four variables were tried in separate regressions as measures of firm size: total towboat horsepower, number of towboats, total cargo tons of barge capacity, and number of barges; it turned out that all four yielded practically the same results. A time trend dummy variable was included to account for technological improvements during the period; it also accounted for absolute variations in factor price levels, which were assumed constant relative to each other and therefore not included explicitly. The seasonal variations in navigation conditions were captured by a seasonal dummy, while "variations between firms" were captured by a firm-specific dummy, implying that firm cost functions are identical save for a single factor.[18/] Separate average cost functions were estimated for different cost components and total costs. For each case, a lo g-linear form was specified in terms of average cost as a function of output level, firm size, time, season and (firm-specific) efficiency level. As was expected, all three equations showed a negative, large, and extremely significant coefficient

---

[18/] One may think that the firm size variable already captures this aspect, but the authors' idea was to account for differences in efficiency levels among firms.

of output, which combined with the negative size coefficient in the average total cost regression indicated "considerable returns to scale in the long run." The total cost equation was considered the most relevant because different procedures for dividing expenses between direct and indirect costs across firms were suspect. It should be noted that the use of an output index such as the EBM (recommended by the ICC) allowed for less ambiguous inferences on scale economies than those from the ton-miles concept.[19/]

Koshal (1972) developed what perhaps constitutes the simplest approach to the analysis of scale economies in a transportation industry. He estimated cost functions for the public and private sectors of the Indian trucking industry, as well as for bus transport in the United States. In all three cases total costs were assumed to be a linear function of output. Output was defined as truck-kilometers (public trucking), ton-kilometers (private trucking), and total bus mileage (bus transport). Conclusions on scale economies, thus, were based upon the sign and significanace level of the constant term, which was found to be positive and significant (indicating scale economies) for public trucking and part of the private firms in India; this latter sector was divided into two geographical zones due to differences in road quality and maximum distance covered. Cross sectional data were used except for private trucking firms, and no factor prices were included. Unlike most

[19/] A similar underlying idea will later appear in the work of Spady and Friedlaender (1978) for trucking, defining the concept of effective output.

other studies, Koshal did not discuss the problems of output definition nor did he try to overcome potential difficulties due to the aggregation of heterogeneous components in a single vehicle- or weight-distance unit.

Griliches (1972) performed a serious and deep critique of the procedures used by the ICC in studying the existence of scale economies in the U.S. railroad industry. In essence, the ICC had estimated linear relations between total costs per mile and tons per mile, concluding that the percent variable (ratio of marginal cost to average cost) amounted to 0.8, i.e., economies of scale were present. Griliches discussed the problem of definition and aggregation of output, arguing that it is very difficult to summarize in one type of measure both level and characteristics of output. He pointed out that although cross sectional data should be preferred to time-series, problems arise due to observations "out" of long-run equilibrium, concluding that short-run influences would bias downward the percent variable estimated;[20/] this was the reason to rexamine the problem using five-year averages. Miles of track (M) was criticized as a deflator on two grounds: i) it is a poor measure of size, and ii) if it is suspected that large observations of M are associated with large errors (i.e. heteroscedasticity), then the best estimates are obtained by dividing the whole equation (including the constant term) by an appropriate power of M. Finally, Griliches found inappropriate the assumption of a single linear equation representing

---

[20/] The analytical reasoning on this point followed closely Friedman's critique of the so-called Keynesian consumption function, where he distinguished between permanent and transitory components of income in cross sectional data.

all firms, and he split the sample into small and large roads. Based on all these observations, reestimation of the cost equation(s) following different procedures[21] led Griliches to the rejection of the hypothesis that scale economies exist, provided smaller roads are either not considered or represented by a different function (in which case the percent variable is measured also for larger roads). He concluded that even if one accepts the ICC's definitions of cost and output, results are extremely sensitive to the choice of particular observations and to the statistical procedure used. Perhaps the richest part of Griliches' study lies in the depth of his discussion, which opened the door to future improvements. He stated that even his own conclusions "are based on very questionable definitions of cost and output, and on a very gross aggregation of types of traffic, dimensions of output, regions of the country, and sizes of railroads." Here we reproduce what we consider Griliches' main point of criticism: "There may well be decreasing average costs for some types of traffic, at some times, in some areas. But all the studies examined ask the question, what will happen to average costs if total traffic is expanded on the average in the same proportions and having exactly the same distribution over the various commodities, types, routes, and seasons as the previously handled traffic? There may be very little return to scale from a proportionate increase in all kinds of traffic. Whatever decreasing costs there may be are likely to arise only if one can contemplate disproportionate changes in traffic, changes

[21] Weighted and unweighted linear forms, and log-linear, were estimated for the whole and split samples.

in some kind of traffic but not in others. But that cannot be discovered from such studies as we have examined above. It requires a different and much more ad hoc research program."[22/] As we have already seen, some effort had been developed by that time in terms of including some of the aspects pointed out by Griliches, although in studies related to other transportation industries. However, the bulk of his criticism had general validity and still has at present, as we will verify in this review.

2.2 Econometric Approaches: Second Generation

The work by Keeler (1974) on railroad costs threw a lot of insight into the problem of transportation cost functions from various perspectives, including the short-run—long-run discussion, specification and output treatment, the econometrics involved, and the overall discussion. The basic notion was to depart from basic neoclassical production theory in a consistent way toward a railroad cost function. Two types of output were specified, gross ton-miles of freight service (Y_1), and gross ton-miles of passenger service (Y_2). Two separate production functions relating each type of output to inputs unique to it (assuming truck services can be allocated) were stated in an unrestricted Cobb-Douglas form, i.e.,

---

[22/] The "examined studies" are two by G. H. Borts, in 1952 and 1960, and another by L. R. Klein in 1953, and are referred to as "the only modern econometric studies of returns to scale in the railroad industry."

$$Y_i = A_i T_i^{\alpha_i} R_i^{\beta_i} F_i^{\gamma_i} L_i^{\delta_i} \qquad i = 1,2 \tag{2.1}$$

where T is track-miles, R is rolling stock investment, F is fuel, and L is labor (per unit time). Then Keeler derived a short-run cost function (SRTC) solving

$$\min_{R_i, R_k, L_i} w_T(T_1 + T_2) + w_R(R_1 + R_2) + w_F(F_1 + F_2) + w_L(L_1 + L_2)$$

$$\text{subject to} \qquad (3.14) \quad T_1 + T_2 = T \ , \tag{2.2}$$

that is, adjusting all inputs except T. Assuming all input prices (w_i) constant and also $\alpha_i = \alpha$, $\beta_i + \gamma_i + \delta_i = \varepsilon$, he got the specification

$$SRTC = w_T T + (k_1 Y_1^{a/b} + k_2 Y_2^{a/b})^b T^{(1-b)} \tag{2.3}$$

where $a = 1/\varepsilon$, $b = a/\varepsilon + 1$, $k_1 = C_1^{\frac{\varepsilon}{\alpha+\varepsilon}}$, $k_2 = C_2^{\frac{\varepsilon}{\alpha+\varepsilon}}$. SRTC was estimated from pooled data and cross sectional data correcting for heteroscedasticity, as equation (2.3). Although it was tempting to estimate only the variable part, i.e., SRTC - $w_T T$, Keeler argued that it would have required an arbitrary assumption as to what part of track-related costs were actually fixed (i.e., w_T), and therefore w_T was also estimated. After (2.3) was estimated, a long-run cost function (LRTC) was found by minimizing SRTC with respect to the fixed factor, T, i.e., deriving the envelope of (2.3), which led to a form

$$LRTC = A\ Y_1^{\gamma} + B\ Y_2^{\gamma} \quad . \tag{2.4}$$

Actual values obtained indicated long-run constant returns to scale. In addition, the optimal $T = T^*$ from $\partial SRTC/\partial T = 0$ allowed estimation of excess (or insufficient) capacity, by comparison with actual values observed. Enormous overall excess capacity was found, which suggested abandonment of some lines as a policy to be implemented. It followed that short-run marginal cost pricing fell short of average costs; the percent variable was less than 0.7 for most railroads. Keeler stated finally that accuracy might be improved by using less aggregated data (in terms of different traffic densities), by including interregional cost differences, and by expanding output to more commodity classes and different regions. It is worth noting that the explicit assumptions used by Keeler, i.e., the Cobb-Douglas production functions and the quality of exponents α_i and ε_i, lead to a LRTC which shows no interaction between outputs Y_1 and Y_2 on costs, i.e., $\partial^2 LRTC/\partial Y_1 \partial Y_2 = 0$. In other words, no production complementarity is allowed to exist among products. On the other hand, output specification does not allow one to state which lines should actually be abandoned given the detection of overall excess capacity. In fact, these problems were foreshadowed by Keeler in his final recommendations although not mentioned explictly as issues. It is, however, clear that this study provided much more insight than previous railroad studies and developed a more consistent framework for analyzing transportation cost functions.

The work by Särndal and Statton (1975) which analyzes factors influencing operating costs in the airline industry, contains some interesting points in the perspective of our work, although their study was not intended to produce a cost function. The analytical procedure followed by the authors is somewhat long and we have chosen to describe the general approach, discussion and conclusions. The data base used referred to the U.S. domestic airline industry in 1967/68 (cross section). It was assumed to be a fixed network system, and that the firms adapted to the system by employing a certain technology represented by number and types of aircraft of different characteristics. Two classes of variables were defined: network variables describing the system of routes served by a carrier, and technology variables describing its fleet of aircraft; the study analyzed the potential influence of both kinds of variables on unit costs, and the causal relation between the classes of variables in order to eliminate multicollinearity. The unit cost variable chosen was total operating cost over available ton-miles. The basic network variables were (by carrier) number of cities served, average frequency of weekly departures per city served, and average stage length flown; in addition, the standard deviation and coefficient of variation of both departure frequency and stage length distribution, were calculated and used. The technology variables were (by carrier) the number of aircraft types, the number of aircraft per type, degree of utilization (number of miles flown) per aircraft, average number of revenue tons weighted by miles flown by aircraft type, the average number of engines weighted by time in revenue service by aircraft type, and average age of equipment using the same weight as before. Using techniques of path analysis and partial correlations among variables, a number of causal relations were established indicating that airline size could be effectively expressed as a function of average stage length, average frequency of weekly departures by city, and number of cities served. The analysis of unit cost in terms of network variables led the authors to conclude that average stage length was of major importance, the longer its value the lower unit cost; in addition, it was found that a high coefficient of variation of stage lengths was advantageous from a cost perspective, i.e., a network with a considerable mixture of long and short stages had, ceteris paribus, lower unit costs. From the analysis in terms of technology variables (a somewhat misleading name for fleet characteristics), unit cost advantages were produced by higher available ton capacity and higher miles flown per aircraft; this was to be expected from the network analysis, because one of the authors' conclusions from the path analysis had been that average stage length (a network variable) directly influenced these two fleet variables. It should be noted that all these findings indicate that two types of economies seem to be present in the airline industry, that is, ceteris paribus, the higher the capacity (longer distances, more tons) and/or the more heterogeneous the spatial pattern of services (mixture of long and short stages), the lower the unit costs. The causal pattern found and the type of conclusions reached are very important from a methodological point of view in order to establish both the nature of transportation cost functions and the form of the variables that should be specified. It is interesting to note that in a previous work, Gordon and deNeufville (1973) developed analytical relations among vehicle capacity, fleet size, and network shape to satisfy a given origin-destination flow pattern in airlines. That model, as a whole, could be viewed as an implicit production function that elegantly shows the type of substitution that exists among fleet capacity and network shape, in the satisfaction of a given pattern of traffic.

In his revision of a study on railroad production function by Klein in 1947, Hasenkamp (1976) kept Klein's conceptual approach in terms of output definitions and of the use of both input and output functions to perform aggregation, but improved the econometrics in the estimation process and explored various functional forms. Output was defined as a vector Y where Y_1 was freight service (net ton-miles of freight carried), and Y_2 was passenger service (net passenger-miles). Three inputs were considered, namely labor (man-hours), fuel (ton of coal equivalents) and capital service (car-miles or train-hours), denoted by X_1, X_2, and X_3 respectively. Outputs were classified as exogenous and inputs as endogenous; therefore a cost-minimizing behavior of the firm was proposed. Cross sectional data from two periods were used. The original production function formulated by Klein, namely

$$Y_1 Y_2^{\delta} = A\, X_1^{\alpha_1}\, X_2^{\alpha_2}\, X_3^{\alpha_3} \qquad (2.5)$$

was criticized because the output function was not convex as required on theoretical grounds. Hasenkamp kept the separable form, i.e., $f(Y) = g(X)$ where X is the input vector, but proposed a constant elasticity of transformation (CET) output function

$$f(Y) = (\Sigma\, \delta_i\, Y_i^c)^{1/c} \quad , \tag{2.6}$$

and either a Cobb-Douglas

$$g(X) = A(\Pi\, X_j^{\alpha_j})^r \quad \Sigma\, \alpha_j = 1 \tag{2.7}$$

or a constant elasticity of substitution (CES)

$$g(X) = A(\Sigma\, \alpha_j\, X_j^{\beta})^{r/\beta} \qquad \Sigma\, \alpha_j = 1 \tag{2.8}$$

input function. In both (2.7) and (2.8), r indicates degree of returns to scale. Hasenkamp then derived the cost functions C(w,Y) and the system of input demand functions X(w,Y) corresponding to the proposed separable forms, where w is the input price vector. A logarithmic stochastic formulation for the latter system (and the corresponding C(w,Y)) was used for estimation purposes. Results were reported for both the estimation based on the input demand system and that based on the cost function. The conclusions from the procedure, applied to the different functional forms proposed, were virtually the same as those obtained by Klein, namely increasing returns to scale were found. Moreover, the convexity assumption in f(Y), which corresponds to $c > 1$ in (2.6), was violated by the empirical results, implying that if railroads could choose which outputs to produce then either Y_1 or Y_2 would be the outcome. Klein's implicit a priori assumption was then supported by the new procedure! Hasenkamp's contribution was more oriented toward "internal microeconomic consistency" and econometric correctness, than toward understanding the nature of the transportation problem involved. This work showed, however, that is is possible to obtain conclusions on production complementarity in transportation through appropriate analytical treatment, and without imposing a priori restrictions.

Koenker (1977) estimated a cost function to analyze the U.S. trucking industry using a time-series of annual data of a cross section corresponding to interstate common carriers. For these firms, output is exogenous since they have to carry all requests at predetermined prices, making a cost-minimizing behavior appropriate. The cost function was assumed to be of a separable form, i.e.,

$$C(w,Y) = C(w)\alpha(Y) \tag{2.9}$$

where $\alpha(Y)$ is a scaling function.[23/] Then the output vector Y was defined as $\{g,\ell,h\}$, where $g = s\,\ell\,h$, s was number of shipments made by a firm per year, ℓ was mean load (tons) per trip, and h was mean haul length (miles) per trip. Then the scaling function was proposed as

$$\alpha(g,\ell,h) = g^{\theta}\ell^{\beta_1}h^{\beta_2} \quad , \tag{2.10}$$

and θ was expressed as $\theta_0 + \theta_1 \ln g$. Assuming input prices constant across firms and accounting for one period lag adjustment, the model to be estimated was

$$\ln C_{ft} = A_t + \gamma \ln\left(\frac{q_{ft}}{q_{ft-1}}\right) + \alpha_0 \ln q_{ft-1} + \alpha_1[\ln q_{ft-1}]^2 \tag{2.11}$$

$$+ \beta_1 \ln h_{ft} + \beta_2 \ln \ell_{ft} + \varepsilon_{ft} \quad ,$$

where f stands for firm and t for year. Using the estimated (2.11),

---

[23/] This cost function is dual to a separable production function $f(Y) = g(X)$, where $\alpha(Y) = [f(Y)]^{1/r}$, and $C(w)$ is determined uniquely by $g(X)$. The assumption behind these forms is that factor proportions are independent of firm scale.

a joint estimate of optimal scale was obtained by minimizing average costs C/g with respect to g, for fixed values of h and ℓ (fixed quality, in Koenker's words). A value of 6.69×10^6 ton-miles per year resulted, which compared to observed firm sizes led Koenker to conclude that the trucking industry was dominated by firms larger than optimal. In other words, most firms were in the decreasing returns zone of the average cost curve. It should be pointed out that the debate on the issue of scale economies in trucking has many facets, a priori images and particular interests are somehow present in some studies, and results are available "fitting all tastes." Koenker also reported results of a "static version" of (2.11), represented by a log-linear equation of costs in g, h and ℓ, leading to a greater optimal size of 7.78×10^6 ton-miles/year. An additional conclusion of this study is that C/g "falls dramatically as length of haul and weight of load are increased, since these factors are correlated with firm size; neglect of their influence can lead to faulty inferences about the existence of scale economies." This does not seem to be a conclusive argument at all, in the sense that the existence of such a correlation may indicate that, on technical grounds, only bigger firms can operate with this characteristic, and therefore can better utilize their fleet than small firms. Finally, note that the definition of output is not really a vector; in fact it is nothing but old ton-miles (g).

The work by Harris (1977) on the rail freight industry focused on economies of traffic density, i.e., what happens to average cost as output increases holding the route system (miles of rail line) constant.

Harris claimed that the problem of excess capacity is not related to trackage (Keeler's idea) but to the route system; therefore, he stated, "it is the cost of the basic indivisibility — the length of road required to connect two points — that we should measure." The basic model proposed was

$$C = \beta_0 RTM + \beta_1 \ RFT + \beta_2 MR \quad , \tag{2.12}$$

where RTM is revenue ton-miles, RFT is revenue freight tons, and MR is miles of road. Cross sectional data for two years were used, including only those firms with negligible passenger operations. Equation (2.12) was divided by RTM to correct for heteroscedasticity, obtaining

$$AC = \beta_0 + \beta_1 \frac{1}{ALH} + \beta_2 \frac{1}{D} \quad , \tag{2.13}$$

where ALH is average length of haul (RTM/RFT), and D is density (RTM/MR). In addition, a dummy variable was introduced to account for higher costs of railroads operating in urban areas. This took the form of a "slope affecting" dummy on RFT and MR. Significant economies of traffic density were found (i.e., a significant and high positive value of β_2). Moreover, Harris concluded that this was due to high fixed operating costs per MR, rather than to capital costs.[24/] He went on to compare his results with

[24/] This conclusion was obtained running separate regressions for capital (KC) and operating (OC) costs. KC included capital rental cost in addition to the ICC's "net rents" figure. Capital rental cost was calculated as undepreciated capital accounts for way, structures and equipment, times a unit "cost" $\rho = r(1 - e^{-rL})$, where r is interest rate and L the life of the correspondent capital good.

those obtained by Keeler; after applying a conversion factor to make units compatible, e.g., miles of track = 1.5 MR, results were said to be "nearly identical over the relevant density range."

Pozdena dnd Merewitz (1978) developed a cost function for rail rapid transit which methodologically followed nearly literally Keeler's (1974) procedure. A Cobb-Douglas production function was assumed, with output defined as annual vehicle-miles, and inputs defined as labor L (hours), electricity E (kilowatt-hours), rolling stock R (vehicles), and miles of track T. The short-run dual cost function took the form

$$C(w,Y,T) = P_t T + c\, Q^{b_1} P_\ell^{b_2} P_e^{b_3} T^{b_4} P_r^{b_5} , \qquad (2.14)$$

where P_i is price of input i and b_4 should be negative. However, only operating costs were considered in the estimation of (2.14), replacing P_t by a, representing fixed operating costs. Pooled data consisting of 105 observations of a time series of cross sections was used. P_r was found to be constant across firms and was, therefore, dropped as independent variable. One of the procedures used in actual estimation was based on the linear form obtained by taking log of short-run operating costs (SROC) after subtraction of aT; a was found by iteration, seeking a minumum sum of squared residuals. When estimating (2.14) in its nonlinear form, the sample was divided in small, medium and large properties, replacing a by

$$a = a_0 + a_1 S + Q_2 M \quad , \qquad (2.15)$$

where S and M were dummy variables for small and medium properties, respectively.[25/] Correction for heteroscedasticity was made by dividing through $T^{0.75}$ (decided after standard procedures). The estimated SROC from the linear procedure was used to analyze a particular rapid transit system[26/] after addition of total annual capital costs (KC) in the form

$$KC = \alpha T \tag{2.16}$$

where α was calculated externally.[27/] This way, the short-run total cost function (2.14) was "recovered." The long-run total cost (LRTC) function was obtained optimizing with respect to T, which, after replacing the estimated values, gave

$$LRTC = 7.42\ P_{\ell}^{0.98} P_{e}^{0.48} Y^{0.76} \quad . \tag{2.17}$$

(2.17) indicates long-run scale economies in the provision of rapid transit service. This study offers no significant methodological contribution and seems to accept Keeler's conceptual framework as appropriate. Unfortunately, no discussion on output definition was offered, implicitly accepting vehicle-miles as a sufficiently homogeneous measure. The fact that rail rapid transit operates only in urban areas appears, however, as an argument in favor of this assumption.

[25/] Of course, $a_0 < 0 < a_2 < a_1$ was expected.

[26/] San Francisco's BART.

[27/] This is, $P_t = a + \alpha$.

The work of Spady and Friedlaender (1978) is regarded as the state-of-the-art in estimation of cost functions and analysis of scale economies in the transportation industries.[28/] Spady and Friedlaender used the so-called hedonic approach to study costs in the regulated U.S. trucking industry. The ton-mile output concept was found an "inadequate measure when the commodities hauled are diverse, when average lengths of haul, average shipment sizes, and average loads, and amount and type of area served vary widely from firm to firm" (Spady, 1978). A quality-separable hedonic cost function

$$C = C[\psi(Y,q),w] \tag{2.18}$$

$$\psi = Y\phi(q) \tag{2.19}$$

was proposed to overcome the problems arising from the aforementioned heterogeneity of output, where ψ is "effective output," Y is ton-miles and q is a vector of quality characteristics. The actual components of q used by Spady and Friedlaender were: average shipment size, average length of haul, percentage of tons shipped in less-than-truckload lots, insurance,[29/] and average load. Input prices included labor, fuel, capital, and purchased transportation. The authors regarded the components of q as exogenous to the firm, i.e., beyond the firm's control; they were cautious in this respect, stating that if this was

---

28/ This work should be considered jointly with Spady and Friedlaender (1976), which provides a detailed theoretical basis for the actual cost function specified, and Spady (1978), which offers more details and includes a rail example.

29/ Insurance was explicitly included to capture the difference among types of commodities transported, a high value reflecting valuable and/or fragile goods.

not the case, (2.18) would not be a correct specification.[30/] In addition, the hedonic formulation implied that the cost-minimizing output combination is independent of the composition of effective output. A translog formulation for (2.18) was used in conjunction with the corresponding factor share equations derived from Shephard's Lemma.[31/] A cross section of 168 firms was used to estimate the cost function, assuming they were in long-run equilibrium. Factor prices for capital and labor were calculated from total expenditures attributable to those items divided by some measure of input (e.g., total labor), while regional prices for purchased transportation and fuel were calculated using econometric procedures in per unit terms (i.e. purchased transportation/rented vehicle-mile, and fuel/vehicle-mile). Results were obtained i) from direct estimation of (2.18), ii) from the system of (2.18) and the factor share equations, and iii) using $\phi(q) = 1$, i.e., a nonhedonic form. Among the main conclusions from this study, the following are particularly interesting.

i) for any given number of ton-miles, valuable or fragile less-than-truckload shipments in small loads and short hauls appear costlier to produce than low-value truckload shipments in large loads and long hauls;

[30/] The reason for this to be true is that if firms actually control some of the q_i's, they can operate optimizing with respect to those aspects, which would then "disappear" from the cost function.

[31/] This Lemma states that $\partial C(w,Y)/\partial w_i = X_i$ (demand for factor i).

ii) the nonhedonic specification led to marginal rejection of the assumption of homothetic production and to strong rejection of the assumption of constant returns to scale (implying increasing returns), while the hedonic formulation strongly rejected homotheticity and only marginally rejected constant returns, in fact suggesting decreasing returns. This second conclusion, less intuitive than the first one, implies different policy recommendations in the trucking industry under the different cost specifications. In the final discussion, Spady and Friedlaender offered an explanation for the large number of mergers in trucking in terms of economies of density and utilization and regulatory practices. In particular, they stated that "if smaller firms could operate with the same loads, lengths of haul, and share of less-than-truckload as larger firms, they would have the same costs as the larger firms, and hence there would be little incentive to merge." In addition, as firms are assigned routes, merging allows operating rights on a wider network. A number of questions can be posed under the new evidence presented in this study. First, can small firms technically operate taking advantage of the aforementioned conditions? Second, if merging is convenient among firms serving different routes, are we in the presence of production complementarity? Third, if production is not homothetic, shouldn't economies of scope in fact be studied? These questions seem to indicate the need to incorporate what we can momentarily call the "spatial setting" when searching for a cost function in transportation. In this respect, Spady (1978) adopted a suggestion by McFadden (1978) in terms of including "technological conditions" as an argument in the cost function, writing

$$C = C(w,Y,t) \quad , \tag{2.20}$$

and offering as an example the route structure of a carrier.[32/]

Harmatuck (1979) criticized former railroad studies on grounds of their uselessness "in analyzing merger policy because they are devoid of geographic content and are characterized by a single dimension of output." Harmatuck estimated a railway cost function using a translog formulation and a three-dimensional output composed of gross ton-miles, tons, and traffic composition (proportion of cars moving certain types of cargo to total cars loaded). In addition, five price indexes associated with activities[33/] rather than with inputs were used. Miles of track was included as a fixed factor, and a regional dummy variable was defined. Estimation was based on a cross section of 40 firms using three-year averages for output, prices and track mileage variables; factor share equations[34/] were included, forming a system actually estimated. The rejection of homotheticity in production and the finding of economies of density particularly at small tonnage levels, were among the main conclusions of this study. Perhaps Harmatuck's main methodological comment was that aggregate data make policy implications uncertain.

[32/] We will later discuss this point in terms of the treatment of the route structure as a factor of production. See Chapter 3 .

[33/] "Prices" were associated with maintenance and capital costs of way and structure, maintenance and capital cost of equipment, yard expenses, train expenses, and other expenses. "Prices" were calculated as total expenses per unit "activity measure."

[34/] In fact, they should be called "activity" share equations in this case, because activity and not factor prices were used.

Unfortunately, and this is common to all studies reviewed, the "ideal disaggregation" has never been made explicit.

Although oriented toward the estimation of productivity growth in U.S. railroads, the work by Caves, Christensen and Swanson (1980) relies on the estimation of a cost function that incorporates time as an argument. From a methodological perspective in the transportation cost function specification, two aspects of this study should be mentioned. First, a generalized translog multiproduct formulation was used, which allows for zero levels in the components of Y. Second, the output vector was defined as composed of ton-miles of freight (Y_1), average length of freight haul (Y_2), passenger-miles (Y_3), and average length of passenger trip (Y_4). This treatment of Y is somewhat similar in concept to that of Harmatuck (1979), in the sense that Y_i are not distinct outputs but dimensions of the same output.

Finally, Braeutigam, Daugherty and Turnquist (1980) have developed what they call a "hybrid" approach to the estimation of a railroad cost function. This consists of the inclusion of engineering information in the form of an "overall average velocity" ($\bar{v}$), as part of output description. They used monthly data corresponding to one firm, regarding track, switches, buildings, and land, as fixed factors (K) within the period of observation; cars, fuel, locomotives, crew and noncrew labor were regarded as adjustable inputs (X_i). Then the cost function was specified in a translog form for $C(Y, \bar{v}, P_1, P_2, P_3, P_4, P_5, K)$, where Y is loaded car-miles and P_i is price of X_i. $\bar{v}$ was obtained from engineering information on train speeds, average length of haul, and delay in yards. The authors stated that they "planned to have an output variable associated with each commodity shipped in each direction over every segment of the system," dropping the idea in view of the huge number of parameters to be estimated. Unfortunately, neither justification for such an approach nor discussion of the implications of not carrying it out were offered. Monthly prices were calculated following standard procedures. An index for the amount of fixed factors K was constructed by dividing the miles of high quality track by total track mileage in the system. The exogenous nature assigned by Braeutigam et al. to speed, was justified on the description of the firm under study as a "bridge line" connecting major railroads, $\bar{v}$ essentially determined by the action of these latter. Among the main findings of the study are i) the rejection of the joint hypothesis of separability (transformation function) and homotheticity in production, and ii) short-run average costs exceed six times the respective marginal costs. The inclusion of speed in the specification was judged to significantly improve the model. This latter conclusion, however, was based on econometric testing rather than on a discussion linking production (engineering) functions and cost functions; moreover, the underlying justification was somewhat intuitive as opposed to an answer to applied production theory.

2.3 Synthesis and Discussion

This review of methodological aspects in building transportation cost functions did not discriminate across modes, which has proved quite useful in providing insights into the nature of the problem involved. It is apparent that different aspects are emphasized in different studies: measures of firm size, treatment of fixed factors, importance of network shape, adaptability of fleet, quality related output dimensions, etc.

We have seen that direct econometric estimation of transportation cost functions has been based upon many different specifications of both functional forms and outputs. Table 2.1 summarizes the different approaches in terms of dependent variable used, output definition, functional form, underlying production structure, and other variables.[35/] Although it does not follow directly from the table, our review shows a clear trend toward improving two interrelated aspects in cost function estimation. The first one relates to functional specification and econometric procedure in general; linear and log-linear forms evolved to dual forms corresponding to some underlying production structure, and then to the so-called flexible forms, which do not require a priori assumptions on production structure, and from which that structure can actually be rescued. Secondly, the microeconomic treatment has improved enormously in terms of internal consistency. As examples, we can mention the derivation of long-run cost functions from estimated short-run functions (Keeler) by optimizing with respect to fixed factors; and, most importantly, the use of the derivative property of the cost function to generate additional equations based upon the (derived) factor demands. This last property

---

[35/] Griliches' study was not included because it can be considered more a good criticism than a proposal of any form of cost function.

Table 2.1: Summary of Econometric Approaches

Author(s)	Cost Function Structure	Functional Form	Underlying Production Structure	Mode
Koshal	$C(Y_0)$	linear	—	Truck, Urban bus
Lee-Steedman	$AC(Y_4,q,g,t)$	linear	—	Urban bus
Case-Lave	$AC(\psi_1,\bar{x}_1,T)$	log-linear	—	Inland Waterways
Keeler	$C(Y_0,Y_1,\bar{X}_2)$	dual to →	Cobb-Douglas	Railroad
Särndal-Statton	$AC(Y_2,g,t_0,t_1)$	linear	t_0 dependent on t_1	Airlines
Hasenkamp	$C(Y_0,Y_1,w)$	dual to →	$F(Y) = h(X)$ (CET) (CD or CES)	Railroad
Koenker	$C(Y_0,q)$	log-linear	—	Truck
Harris	$C(Y_0,Y_4,\bar{X}_3)$	linear	—	Railroad
Spady-Friedlaender	$C(\psi_2,w)$	translog	Any	Truck
Spady	$C(\psi_3,Y_0,w,t)$	translog	Any	Railroad
Harmatuck	$C(Y_0,Y_2,Y_3,\bar{X}_2,w)$	translog	Any	Railroad
Braeutigam et al.	$C(Y_0,t_2,w)$	translog	Any	Railroad

Y_0 = ton-miles

Y_1 = passenger-miles

Y_2 = tons

Y_3 = traffic mix

Y_4 = vehicle-miles

T = time

w = factor prices

ψ_1 = equivalent barge-miles

$\psi_2 = Y_0\phi(q)$

$\psi_3 = Y_1\phi(q)$

$\bar{X}_1$ = total barge capacity

$\bar{X}_2$ = track-miles

$\bar{X}_3$ = route-miles

q = quality variables

g = geographical variables

t = technical variables

t_0 = fleet characteristics

t_1 = network characteristics

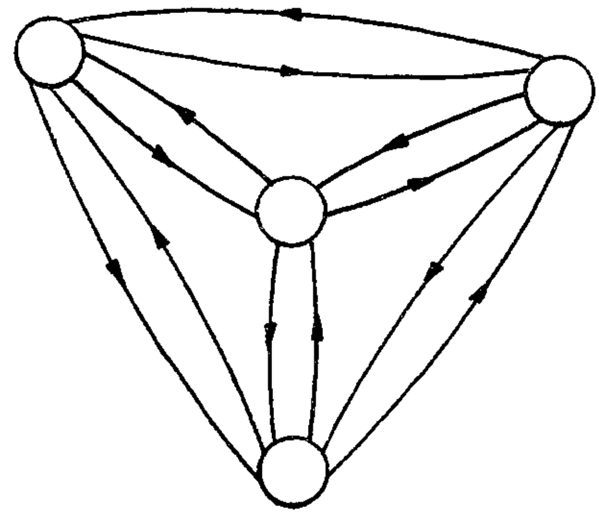
t_2 = mean speed

generates as many new equations as factor prices are involved in the cost function, thus generating a system which improves the efficiency of parameter estimates.

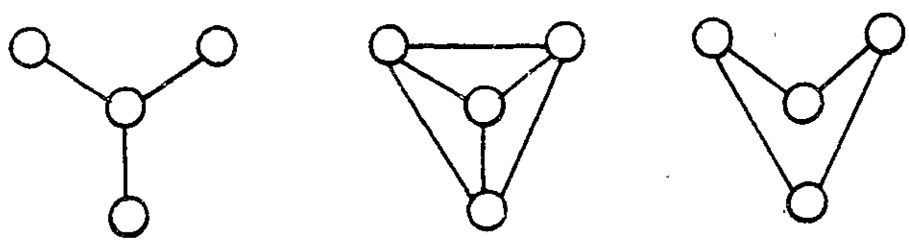
However, when it comes to analyzing the specification and treatment of output in transportation cost functions, there is no clear trend. Output has been characterized in various ways. First we have the single output-single measure definition, in units-times-distance per unit time (UTD), e.g., ton-miles (per month, year, etc.), as in Koshal (1972) and Braeutigam et al. (1980). A second group of studies uses many dimensions of the same "generic" UTD output, like in Harmatuck (1979), Lee and Steedman (1970), and Caves et al. (1980); this can be characterized as a single output-many descriptor approach. Thirdly, the single composite output has been used in Case and Lave's equivalent barge-miles (1970), and in Spady and Friedlaender's hedonic definition (1978), using the UTD-type measure as generic output. Finally, a characterization of transportation output in terms of more than one product is present in Keeler (1974), Hasenkamp (1976) and Spady (1978); in the three (railroad) cases, the distinction has been made between passenger-miles and ton-miles, the former being treated in a hedonic way by Spady. Thus, although the inappropriateness of UTD-type measures of output was recognized by the majority of studies, all of them use UTD as the basic or generic notion of transportation product for cost function estimation. The inclusion of "quality," geographical or technical aspects is actually an effort to account for output heterogeneity. A systematic methodological inconsistency arises, however, in those studiés that include these latter kind of variables as part of output description, in the sense that scale economies are finally analyzed in terms of average costs obtained by division of costs by ton-miles. In this sense, Spady and Friedlaender's study constitutes an exception, for they obtained conclusions on scale economies using the hedonic output index which already included quality adjustments, thus accounting for output heterogeneity in an internally consistent way.

As suggested at the beginning of this section, studies of different modes have emphasized different aspects of transportation. Railroad studies have stressed commodity differentiation in terms of freight and passenger services. The aspect of network shape has been emphasized only in airline studies, as in Särndal and Statton. The key aspect here is that network configuration in terms of actual routes is an answer to the origin-destination flow pattern; this idea, which is present in Gordon and deNeufville (1973), is not specific to the air mode. Following Figure 2.1, the option among different route structures to produce a given O-D pattern can be found in trucking and even in railroads in the long run, and the convenience of each alternative will generally depend on the actual magnitude of flows among all O-D pairs. Thus, route patterns are generally operational answers to a vector of O-D flows, within the boundaries of an actual physical network, which is in turn a long-run answer to those flows.[36/]

[36/] The fact that "network shape" has been emphasized only in some airline studies is probably due to the non-constraining nature of the problem in terms of a physical network.



a. Origin-Destination Flows



b. Possible Route Structures

Figure 2.1: O-D Flows and Route Structure

Trucking studies, particularly Koenker, and Spady and Friedlaender, have stressed aspects like lengths of haul and load size as part of output description. As suggested in our review of both studies, this poses the problem of whether firms decide or not to operate in a certain manner, i.e., are operating characteristics endogenous or exogenous to the firm?

An important aspect which is worth stressing is that even when an output index is used consistently, e.g., a hedonic formulation or Case and Lave's equivalent barge-miles, such an aggregation does not allow for analysis in terms of production complementarity. For example, nothing can be concluded from such approaches on the (cost) convenience of serving different O-D pairs with one or more firms; this is why mergers between firms serving different routes can not be explained by such cost specifications.[37/] In other words, potential economies of spatial scope are not allowed to be examined from the reviewed formulations. Spady's suggestion in terms of specifying C(y,w,t), where t is veiwed as "technological conditions determined by operating rights" (such as "route structure") works toward overcoming this shortcoming. However, the fact that t is specified separately from y keeps the basic problem unresolved. Griliches' comment on "whatever decreasing costs there

---

[37/] Naturally, this also holds for appr ches using more heterogeneous aggregation, e.g., straight ton-miles.

may be are likely to arise only if one can contemplate disproportionate changes in traffic" remains valid. On the other hand, there has been an effort to distinguish among the types of commodities being transported, e.g., Keeler (1974), Hasenkamp (1976), and Spady (1978). While Keeler's formulation implicitly assumes no production complementarity between moving passengers and freight, Hasenkamp's and Spady's do allow for an analysis in this respect. At this point we can conclude that a better formulation of transportation cost functions should make it possible to analyze economies of spatial scope (which has been emphasized in airline studies), as well as economies of commodity scope (which has been emphasized in railroad studies).

It is thus apparent that a number of problems arise in the definition of the arguments of a correctly specified transportation cost function, even before going into the problem of functional form. In Chapter One we emphasized the role of technology in a neo-calssical derivation of a cost function, either in a single-output or a multi-output framework. There should be no inconsistnecy between the engineering involved in the transformation function and the economic analysis from the derived (dual) cost function. Therefore, if engineers are looking for optimal ways to accommodate fleets and routes to produce a given pattern of movements of different commodities between different O-D pairs, the corresponding cost function should reflect the minimum cost of producing this pattern, not ton-miles or quality adjusted ton-miles. The very use of duality properties becomes dubious in this context, e.g. the use of

Shephard's Lemma on cost functions using "problematic" output definitions. The review performed in this chapter allows us to conclude that the econometric techniques to estimate (transportation) cost functions have improved enormously in a relatively short period, but the transportation concepts underlying these formulations are far from being consistent with the operation of transportation systems and, finally, with the engineering involved. Moreover, the conclusions on industry structure obtained from estimated cost functions have been contradicted by actual behavior of transportation firms, particularly in the trucking industry. This inconsistency throws doubts on the policy implications of these studies—which is the final objective of performing them. Although most of this work could be viewed as an effort toward the best use of available information within feasible technical boundaries, it is our opinion that a contribution in terms of providing consistency between the technical and economic analysis is required, taking advantage of the possibilities opened up by improvements in both microeconomic analysis and econometric procedures. It is our strong belief that the kernel of this convergence is correct output definition and aggregation, the understanding of the generation of $T(X,Y) = 0$, and the understanding of the process from $T(X,Y) = 0$ to $C(w,Y)$, or $C(w,Y,\overline{X})$, which is the subject of the next chapter. Quoting Griliches, "a different and much more ad hoc research program" is still needed to improve the reliability of policy conclusions from the analysis of cost functions for the transportation industries.

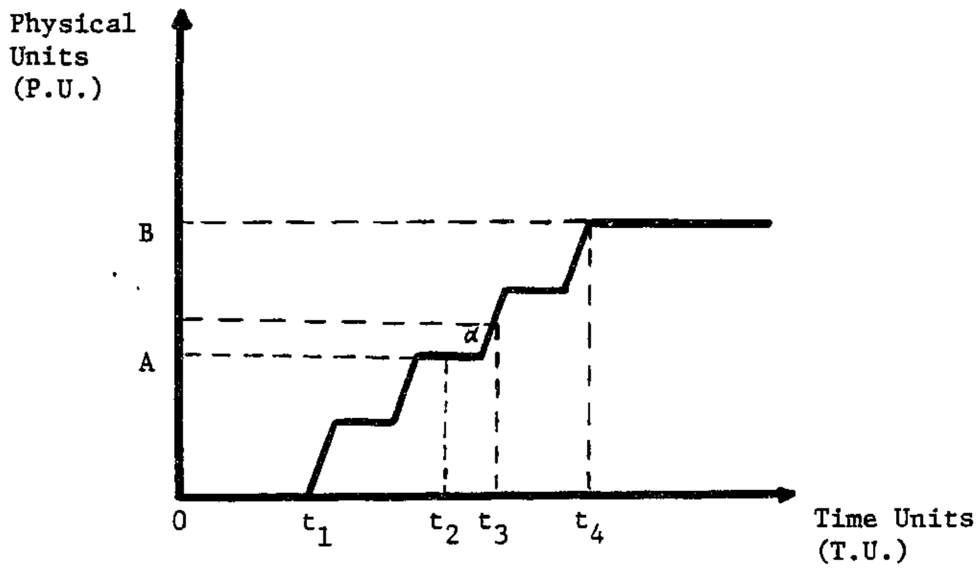
CHAPTER 3. TRANSPORTATION PRODUCT, TRANSPORTATION FUNCTIONS, AND COST FUNCTIONS

The theoretical concepts related to cost functions have been applied in a variety of ways to the estimation of such functions in the different transportation modes. We have found basic inconsistencies between the output treatment in those studies and that implicitly adopted by the underlying engineering analysis. Most importantly, estimated cost functions do not seem to provide a reliable basis to analyze industry structure. This calls for an ad hoc search toward a fundamental redefinition of a transportation cost function. The objective of this chapter is to gain insight into the process of "transportation production," by making use of the framework provided by the microeconomics of the firm, in order to refocus the econometric analysis of transportation cost functions. With this in mind, the generation of a cost function from operational or physical relations will be made explicit, and a critique of the ton-miles concept will be performed on solid grounds. The first section is devoted to the definition of transportation output as a vector of origin-destination-period-commodity specific flows, and to an initial discussion of aggregation. After defining the concept of transportation function as a restricted form of the corresponding economic transformation function, we apply it in sections 2 and 3 to develop the microeconomics of a transportation firm under two spatial settings, which helps show the shortcomings of ton-miles as an output concept, and provides insight into the role of technology and fixed factors under a cost-minimizing behavior. Most important, spatial scope economies are shown to be a potential source of merging in spite of constant multioutput returns to scale. The fourth section addresses the problem of actually estimating transportation cost functions, discussing types of aggregation, role and meaning of scope and scale analysis, treatment of variables, fixed factors and operational characteristics, the nature of required observations, and functional specification.

3.1. Transportation Product

We can understand the concept of transportation process as the result or immediate effect of the action of transporting, i.e., the displacement of some physical entity from a certain origin in space-time to a certain destination in space-time. We can associate this concept with that of "product" in an economic sense, with some reservations. To describe a product we refer to its qualitative characteristics, assigning a name for simplicity (e.g., oranges, shoes, etc.). To measure a product we need a physical unit of reference, and a quantity in terms of these physical units (e.g., five tons of oranges, or a thousand pairs of shoes). When we talk about a production process we need flow units, as opposed to stock units (e.g., a thousand pairs of shoes per week). However, to measure a transportation process we would need a qualitative description of what is being transported, a physical unit of reference, quantity (flow) in terms of these units, and origin and destination in space-time. The need to explicitly establish origin and destination in space-time is the characteristic that distinguishes more clearly a transportation product from the traditional concept. Two additional aspects should be discussed with regard to these concepts. First, eventual changes in quantity and quality of what is being transported may make necessary a description of both dimensions at origin and destination.[38/] Second, because of the time component of origin and destination, two identical transportation processes cannot exist. The first aspect would be conceptually incorporated by treating quantity and quality at origin as an input to the process, while quantity and quality at destination result as a proper output of the process. Alternatively, we can explicitly simplify the analysis by assuming invariability of these two dimensions, thus accepting that relevant technologies in the transportation system keep constant both the nature and amount of what is being moved. The second aspect impedes the addition of transportation processes; however, we may consider as equivalents those processes that coincide in their spatial origin and destination and in the qualitative characteristics of what is being transported, adding over similar physical units. Then, we can define two concepts associated with a particular point in space: flow intensity, which is the derivative with respect to time of a function accounting for the amount of units starting at, passing through, or arriving at, a point; and mean flow intensity, which is the increment in units being transported in a period, divided by the magnitude of this period (see Figure 3.1).

[38/] This may be particularly important in the case of perishable goods.



flow intensity at t_2 = 0 [P.U./T.U.]

flow intensity at t_3 = $\tan\alpha$ [P.U./T.U.]

mean flow intensity $t_1 \rightarrow t_2$ = $A/(t_2-t_1)$[P.U./T.U.]

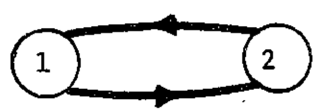
mean flow intensity $t_1 \rightarrow t_4$ = $B/(t_4-t_1)$[P.U./T.U.]

mean flow intensity $0 \rightarrow t_4$ = B/t_4 [P.U./T.U.]

Figure 3.1: Instantaneous and Mean Flow Intensities

a) 

$$Y = \left\{Y_{12}^{11},\ Y_{12}^{21},\ Y_{12}^{12},\ Y_{12}^{22}\right\}$$

b) 

$$Y = \left\{Y_{12}^{11},\ Y_{12}^{21},\ Y_{12}^{12},\ Y_{12}^{22},\ Y_{21}^{11},\ Y_{21}^{21},\ Y_{21}^{12},\ Y_{21}^{22}\right\}$$

Figure 3.2: Transportation Product in Two Periods, Two Commodities

a) one O-D pair

b) two O-D pairs

Taking all the preceding aspects into consideration, we can define the transportation product associated with a particular transportation system as a vector

$$Y = \left\{ Y_{ij}^{kt} \right\} , \tag{3.1}$$

where Y_{ij}^{kt} is the mean flow intensity of product k between origin i and destination j in period t, e.g., a thousand boxes of frozen strawberries per week from Los Angeles to Boston during winter, or 200 people per minute from Chatelet to Gar Montparnasse between 4 P.M. and 5 P.M. on Monday. Depending on the transportation system of reference and on the level of aggregation over the different dimensions involved, the dimensions (number of elements) of Y in (3.1) will vary. This leads us directly into the problem of aggregation, where we can differentiate between three basic types: aggregation over commodities, time, and space. In addition, combined aggregative schemes are also possible.

Aggregation over commodities can be performed (and has been implicitly done) by transforming commodity units into common units of weight or volume, e.g., tons or liters, and then adding across, i.e.,

$$Y_{ij}^{at} = \sum_{k=1}^{k_1} \alpha_k \, Y_{ij}^{kt} \tag{3.2}$$

where a stands for commodity class involving products 1, 2,.,,,k_1, and α_k converts units of commodity k into common units. It should be noted that no movements are "lost" in the aggregation. Following this procedure, the number of commodity classes can be reduced to the limit of one, generally in weight units. This was the rule in the studies reviewed in Chapter 2, with the exceptions of Keeler (1974) and Hasenkamp (1975).[39]

Total aggregation over time, i.e.

$$Y_{ij}^{kT} = \sum_{t=t_1}^{t_r} Y_{ij}^{kt} \times t \ , \tag{3.3}$$

where $\Sigma t_i = T$, $t_i \cap t_j = \phi$ and T is the period of observation, has been the usual procedure in all studies reviewed in the preceding chapter.[40] Again, no processes are lost in the aggregation. (3.3) is equivalent to calculating the amount of units of k from i to j in period T, divided by T.

A procedure to aggregate over space is perhaps the most controversial aspect of aggregation. A first idea would be to "consolidate" or "nuclearize" adjacent nodes, thus diminishing the number of O-D pairs as shown in Figure 3.3. This can be analytically written as

$$Y_{AB}^{kt} = \sum_{i=1}^{3} \sum_{j=4}^{5} Y_{ij}^{kt} \tag{3.4}$$

$$Y_{BA}^{kt} = \sum_{i=4}^{5} \sum_{j=1}^{3} Y_{ij}^{kt} \ . \tag{3.5}$$

However, this procedure as it stands omits some movements, i.e., those flows Y_{ij}^{kt} where i = 1,2,3, j = 1,2,3, and Y_{ij}^{kt} where i = 4,5, j = 4,5 are not accounted for either in (3.4) or (3.5). This creates

---

[39] Harmatuck's traffic mix variable should be included as an effort to deal with commodity aggregation.

[40] Case and Lave's seasonal dummy tries to account for different periods.

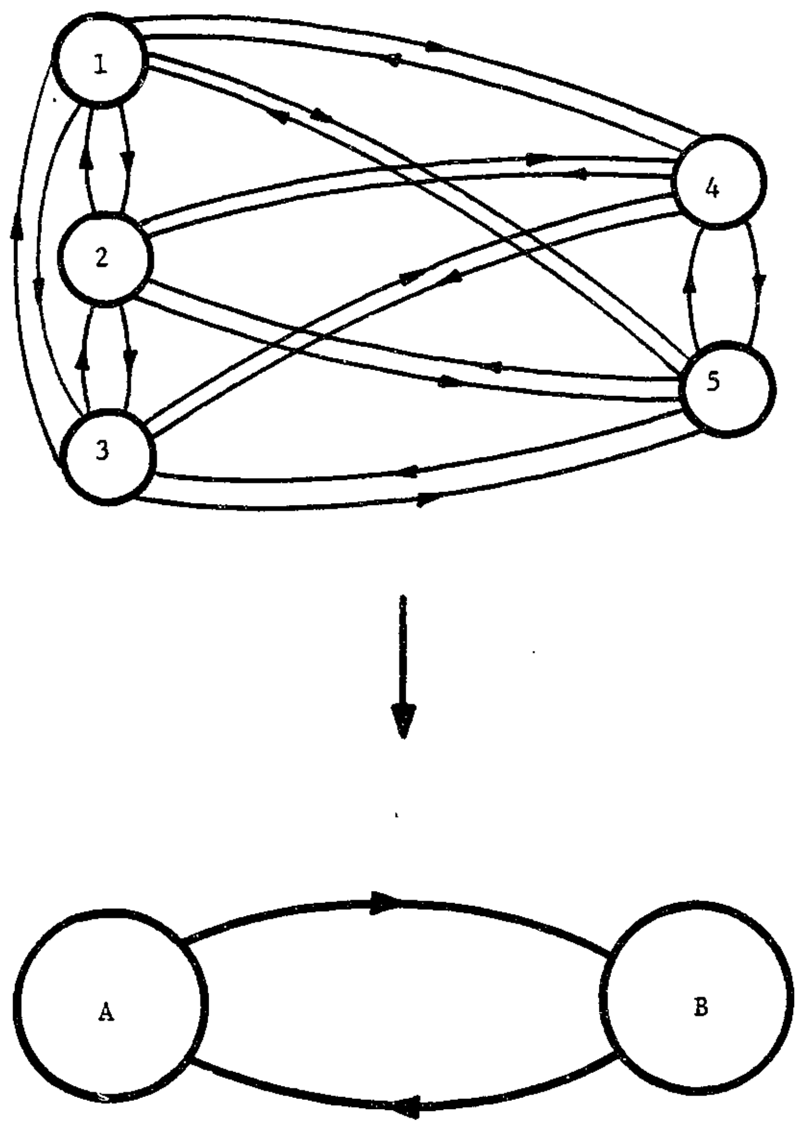


Figure 3.3: Spatial Aggregation

the need to generate additional variables for these "intra-nodal" flows. One possible approach would be to generate Y_A^{kt} and Y_B^{kt} as

$$Y_A^{kt} = \sum_{i=1}^{3} \sum_{j=1}^{3} w_{ij} \, Y_{ij}^{kt} \qquad i \neq j \tag{3.6}$$

$$Y_B^{kt} = \sum_{i=4}^{5} \sum_{j=4}^{5} w_{ij} \, Y_{ij}^{kt} \qquad i \neq j \quad , \tag{3.7}$$

where w_{ij} are weights attached to the movements from i to j. Two kinds of weights have been used to perform this aggregation, under different perspectives. First, $w_{ij} = 1$ which reproduces the procedure of zonal division and generation of O-D matrices of common use in transportation, particularly in demand analysis. Second, $w_{ij} = d_{ij}$ = distance traveled between nodes, which reproduces the usual procedure in the estimation of cost functions, applied to the entire space, and resulting in the well-known ton- or passenger-miles.

Summarizing, the basic definition of transportation product associated with a particular transportation system, is a vector $Y = \{Y_{ij}^{kt}\}$. Y_{ij}^{kt} represents the mean flow intensity of commodity k during period t from origin i to destination j. The dimension of Y can be reduced through aggregation over commodities, time and/or space, a procedure which involves the loss of some information associated with the transportation process generated by the system in reference. Total aggregation over the three dimensions has been usually done in an implicit way as

$$\overline{Y} = \sum_k \sum_t \sum_i \sum_j \alpha_k \, d_{ij} \, Y_{ij}^{kt} \times t \quad , \tag{3.8}$$

thus generating the single output $\bar{Y}$ in common units times distance, per period of observation, e.g., ton-miles per year, quarter or month. In other words, the ton-miles concept can be interpreted as the result of total aggregation over the three generic dimensions of transportation output.

From the perspective of estimating transportation cost functions, the single output generated as in (3.8) has been accepted either directly or as the basis for output definition in all studies to date. It has been explicitly recognized, however, that it does not represent "an unambiguous measure of output." On the other hand, the unambiguous measure of output has not been proposed and, therefore, the problem of how appropriate and how ambiguous $\bar{Y}$ or modifications of $\bar{Y}$ are, has not been systematically analyzed. We face, then, various problems to study. First, how does aggregation (particularly spatial) affect appropriate estimation of transportation cost functions? Second, if these effects are relevant, how to deal with them in the estimation process? In the next sections we will analyze two types of transportation systems in order to gain insight into these and other aspects of transportation cost functions.

3.2 From Transportation Functions to Cost Functions

In any productive system the amount of products (output) is related to the amount of factors (inputs) through a production or transformation function, which summarizes technology and implies a technological optimum within the boundaries of this technology. In each particular field of production, however, there are some technical relations that come out as a result of the analysis of the corresponding engineer, and there are some other relations which are in fact given to the process, and which the "technical expert" cannot influence or decide. Then, although the transformation function relates product(s) to inputs such as labor, capital, land, raw materials, etc., the core of the engineering work is focused on optimizing the direct physical process. "Economists tend to center their attention on capital-labor substitution rates, while engineers have tended to simply rely upon 'well-known' formulas to calculate labor needs and labor costs after the physical processes are fully specified. . In formulating engineering models for processes where labor is readily substitutable for other inputs, there is indeed a gap to be closed between the engineering and economic formulations. . . But, in the more highly technical processes, which are becoming more and more prevalent, where labor does not enter as a substitutable input, the engineering formulation is directly applicable." (Marsden, Pingry and Whinston, 1974; emphasis added).

In the case of transportation, the basic relation (or set of relations) which are the main concern of the transportation engineer, is that which directly associates transportation processes with characteristics of vehicles, terminals and rights-of-way. In other words, these basic relations associate the transportation product Y of a system (as defined in (3.1)), with distances, fleet size, speed, capacities, etc. We will name this set of relations, which also imply technical optima, the transportation function of the system, after Gálvez (1978). By adding other functions, which as such are not under the control of the transportation engineer, to the transportation function, an economic transformation function can be generated, on which basis a cost function can be derived by minimizing the sum of input prices times inputs, subject to the whole set of physical (technical) relations. In what follows we will apply these concepts to a particular but useful setting.

Let us define a system which can be characterized as discrete, in the sense that what is being transported is concentrated in some points along the trajectory as quanta that coincide with vehicles;[41] as cyclical of fixed frequency, and where vehicles are identical and interchangeable. We will assume one origin (1), one destination (2), and a unique product (or aggregate product) of a continuous nature. We will denote the mean flow intensity of this product by y_{12}, measured in physical units (PU) per unit time (UT). Define

B : fleet size (number of vehicles)

K : capacity per vehicle, in PU

k : load per vehicle, in PU

μ^+ : loading capacity at origin, in PU/UT

μ^- : unloading capacity at destination, in PU/UT

$t_{ij}(k)$: travel time from i to j as a function of k, in UT

d_{ij} : distance travelled from i to j, in distance units (DU)

$v(k)$: speed of each vehicle as a function of k, in DU/UT

η : proportion of vehicles in service

f : frequency of trips, in UT^{-1}.

[41] This is not the case for pipelines, for instance.

Thus, the cycle time of one vehicle is given by

$$t_c = t_{12}(k) + t_{21}(0) + \frac{k}{\mu^+} + \frac{k}{\mu^-} \tag{3.9}$$

The frequency needed to satisfy y_{12} is given by

$$f^1 = \frac{y_{12}}{k} \quad , \tag{3.10}$$

while the system can produce a frequency

$$f = \frac{\eta B}{t_c} \quad . \tag{3.11}$$

From (3.9) through (3.11) we obtain

$$y_{12} = \frac{\eta Bk}{t_{12}(k) + t_{21}(0) + k(\frac{1}{\mu^+} + \frac{1}{\mu^-})} \quad , \text{ or} \tag{3.12}$$

$$y_{12} = \frac{\eta Bk}{\frac{d_{12}}{v(k)} + \frac{d_{21}}{v(0)} + k(\frac{1}{\mu^+} + \frac{1}{\mu^-})} \quad . \tag{3.13}$$

Naturally, $k \leq K$. In addition, $\partial y_{12}/\partial k > 0$.[42/] Therefore, the maximum y_{12} the system can produce is given by

$$y_{12} = \frac{\eta BK}{\frac{d_{12}}{v(k)} + \frac{d_{21}}{v(0)} + K(\frac{1}{\mu^+} + \frac{1}{\mu^-})} \quad , \tag{3.14}$$

42/ Provided $\frac{\partial v(k)}{\partial k}$ is small.

which can be postulated as the transportation function for the defined system. Its simplicity is derived from that of the system.

To go from the transportation function to the cost function, we need to introduce the relations between other inputs and the parameters of the transportation system, as an intermediate step. Before doing so, we will make the following simplifications without affecting the conceptual analysis:[43/]

$$\begin{aligned} \eta &= 1 \\ \mu^{+} = \mu^{-} &= \mu \\ v(K) = v(0) &= v \\ d_{12} = d_{21} &= d \quad . \end{aligned} \tag{3.15}$$

(3.14) then reduces to

$$y_{12} = \frac{BK}{2(\frac{d}{v} + \frac{K}{\mu})} \quad . \tag{3.16}$$

Let us define:

- g : gas consumption per vehicle per DU, in volume units (VU)/DU
- L : labor consumption, in men × UT
- ε : labor needed to operate one vehicle, in men/vehicle
- θ : labor needed to operate one loading or unloading site, in men/site.

43/ From a cost function perspective, it would have been desirable to interpret η as a variable which represents the quality or efficiency of vehicle maintenance (which has a cost).

Then we define the relations

$$g = F(v,K) \tag{3.17}$$

$$L = \varepsilon B + \theta \frac{2y_{12}}{\mu} \ . \tag{3.18}$$

(3.17) represents gas consumption as a function of speed and vehicle size. This relation is exogenous to the transportation engineer, provided vehicles of known characteristics are available in the market (i.e., the transportation engineer does not design the vehicle). Similarly, ε and θ are "fixed coefficients" like parameters which correspond to the available technology.[44/] $2y_{12}/\mu$ represents the total number of loading and unloading sites necessary to operate a y_{12} flow.

Let us set T as the period of observation, and calculate expenditures on each factor accordingly. Let us define the following prices:

- P_g : price of gas, in monetary units (MU)/VU
- $P(K)$: price of one vehicle[45/] as a function of capacity, in MU/vehicle
- P_d : price of one "unit" of road length,[46/] in MU/DU
- w : wage rate (period T), in MU/man
- $P(\mu)$: price of one loading-unloading site[47/] as a function of capacity, in MU/site

---

[44/] We expect F to be convex in v (and to have a minimum), and increasing in K.

[45/] Note that we have assumed ε and θ independent of the capacities K and μ of vehicles and loading-unloading sites; this is not very restrictive.

[46/] Depreciated to account for a T period, and including maintenance.

[47/] Rental price in a T period.

Then total cost of operating the system in period T is given by

$$C = 2P_d d + P(K)B + P(\mu)\,\frac{2y_{12}}{\mu} + wL + P_g gB\left[\frac{d}{\frac{d}{v} + \frac{K}{\mu}}\right] \quad , \tag{3.19}$$

where $d/(\frac{d}{v} + \frac{K}{\mu})$ is the actual distance travelled by one vehicle in period T (or, in other words, is the mean overall speed in units of T).[48/] The long-run cost function, i.e., possibly adjusting all factors, corresponds to the solution of

$$\text{Min } C = 2P_d d + P(K)B + P(\mu)\,\frac{2y_{12}}{\mu} + wL + P_g gB\left[\frac{d}{\frac{d}{v} + \frac{K}{\mu}}\right]$$
$$\text{subject to (3.16) through (3.18).} \tag{3.20}$$

However, we can assume even in an abstract case, that distance travelled cannot be adjusted during T and is in fact exogenous and given $(\overline{d})$. In addition, available sizes of vehicles and loading-unloading sites rank from 0 to a certain upper bound. Taking this into account, and after replacing (3.16) through (3.18) in (3.19), the problem in (3.20) can be stated as (rearranging terms)

$$\underset{(K,v,\mu)}{\text{Min}}\ C = 2P_d\overline{d} + P(K)\,\frac{2y_{12}}{K}\left(\frac{\overline{d}}{v} + \frac{K}{\mu}\right) + w2y_{12}\left[\frac{\varepsilon\overline{d}}{kv} + \frac{\varepsilon+\theta}{\mu}\right]$$
$$+ P_g\,\frac{2y_{12}}{K}\,\overline{d}\,F(v,K) + P(\mu)\,\frac{2y_{12}}{\mu} \tag{3.21}$$

$$0 < K < K\ \text{max}$$
$$0 < \mu < \mu\ \text{max}\ .$$

[48/] The orthodoxial statement of $C = \Sigma w_i x_i$ is in terms of constant prices, which does not seem to be the case in (3.19) because of P(K) and P(μ). However, we will see that in fact prices obtain when these functions are explicity introduced. This formulation assumes that vehicles and sites of all sizes are available.

In (3.21) we are implicitly stating that the firm is able to adjust K, v and μ (and implicity fleet size) in order to produce y_{12}. This may well be true in a fairly dynamic renting system, or in a situation of stable flow.[49] Under these conditions, the cost function corresponds to the solution of the (equivalent to (3.21)) problem

$$C = 2P_d\bar{d} + 2y_{12} \underset{(K,v,\mu)}{\text{Min}} \left[\frac{P(K)}{K}\left(\frac{\bar{d}}{v} + \frac{K}{\mu}\right) + w\left(\frac{\varepsilon\bar{d}}{Kv} + \frac{\varepsilon+\theta}{\mu}\right) + \frac{P_g\bar{d}}{K} F(v,K) + \frac{P(\mu)}{\mu}\right] \quad . \tag{3.22}$$

$$0 < K < K\text{max}$$

$$0 < \mu < \mu\text{max}$$

At this point, it is important to make one observation which relates to the traditional output definition, e.g., ton-miles. The function in brackets, call it M, does not have the form $\bar{d} \times N$ because of the presence of loading-unloading effects (represented by μ). If loading-unloading was instantaneous at a finite price, then (3.22) would reduce to

$$C = 2P_d\bar{d} + 2y_{12}\bar{d} \underset{(K,v,\mu)}{\text{Min}} \left[\frac{P(\bar{K})}{Kv} + \frac{w\varepsilon}{Kv} + \frac{P_g}{K} F(v,K)\right] \quad , \tag{3.23}$$

$$0 < K < K\text{max}$$

$$0 < \mu < \mu\text{max}$$

[49] This observation has importance in actual estimation of cost functions. We will come back to it in section 3.4.

and minimizing the cost of producing a mean flow intensity of y_{12} units per T would be equivalent to the cost of generating $y_{12}\ \bar{d}$ units ×DU per T. Given the importance of this aspect in the later discussion on estimating transportation cost functions, we will explore somewhat further the implications.

Let us use the folloiwng forms for P(K), P(μ) and F(v,K):

$$P(K) = P_b + P_k\ K \tag{3.24}$$

$$P(\mu) = P_o + P_\mu\ \mu \tag{3.25}$$

$$F(v,K) = A + G\ K + E(v-v_o)^2\ . \tag{3.26}$$

Thus, P_b is the "basic" price of a vehicle while P_k is the price of an additional capacity unit. Similarly, P_o is the fixed component of the price of a loading-unloading site, while P_μ is the price of additional capacity. (3.26) states that gas consumption per mile increases linearly with vehicle capacity, but there is an optimum speed v_o at which gas consumption is minimum irrespective of capacity.[50/] This is graphically shown in Figure 3.4. We will assume, only for expository purposes, that v_o is well beyond the limit exogenously imposed by another standard (e.g., safety). Let us denote this "imposed" speed by $\bar{v}$. Under these conditions, minimizing the expression (M) in brackets in (3.22) is equivalent to minimizing

50/ Note that this does not mean that v should be set at v_o. It may well be worthwhile to increase v in order to increase frequency and diminish K, for instance.

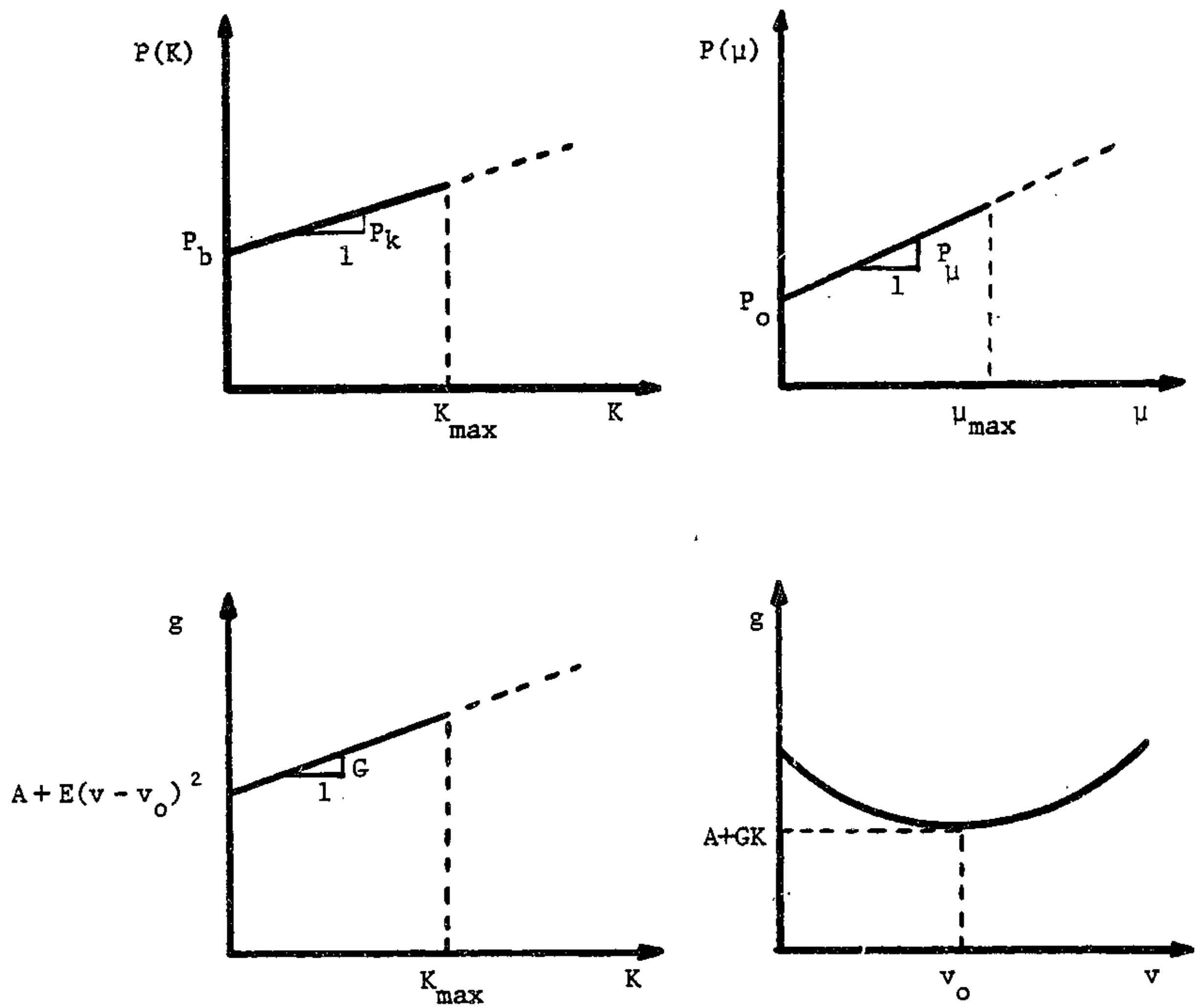


Figure 3.4: Vehicle Price, Loading-Unloading Site Price, and Gas Consumption

$$M = \frac{P_b\bar{d}}{K\bar{v}} + \frac{P_k\bar{d}}{\bar{v}} + \frac{P_b + P_k K}{\mu} + w(\frac{\varepsilon\bar{d}}{K\bar{v}} + \frac{\varepsilon+\theta}{\mu}) + \frac{P_g\bar{d}A}{K} + P_g\bar{d}G + \frac{P_g\bar{d}E}{K}(\bar{v} - v_o)^2$$

$$+ \frac{P_o}{\mu} + P_\mu \quad . \tag{3.27}$$

After dropping constant terms and rearranging, min M → min Q, where

$$Q(K,\mu) = \frac{1}{K}\left\{\frac{\bar{d}}{\bar{v}}(P_b + w\varepsilon) + P_g\bar{d}[A + E(\bar{v} - v_o)^2]\right\} + \frac{1}{\mu}[P_b + w(\varepsilon+\theta) + P_o]$$

$$+ P_k \frac{K}{\mu} \quad . \tag{3.28}$$

$$0 < K < K_{max}$$

$$0 < \mu < \mu_{max}$$

Under our assumptions, the coefficients of $1/K$ and $1/\mu$ are constant. Let us call them Ψ and Ω respectively. Then (3.28) becomes

$$\text{Min } Q(K,\mu) = \frac{1}{K}\Psi + \frac{1}{\mu}\Omega + P_k\frac{K}{\mu} \quad . \tag{3.29}$$

$$0 < K < K_{max}$$

$$0 < \mu < \mu_{max}$$

A contour of (3.29) is obtained setting $Q = Q_i$ and expressing μ as a function of K (or vice versa). We obtain

$$\mu = \frac{-K(\Omega + P_k K)}{\Psi - \Omega Q_i} \quad , \tag{3.30}$$

which corresponds to one branch of a hypoerbole, that branch for which $K > \Psi/Q_i$ in order to preserve μ positive. It can be shown that $\partial^2\mu/\partial K^2 > 0$, that μ has a minimum corresponding to a certain $K^*(Q_i)$, that $\partial K^*/\partial Q_i < 0$, and that contour lines representing different levels of Q_i look like the ones shown in Figure 3.5. Therefore, the minimum of Q within the feasible region determined by $0 < \mu < \mu_{max}$ and $0 < K < K_{max}$, is found at $\mu = \mu_{max}$, but not necessarily at $K = K_{max}$. A solution like (1) indicates that the optimum level of K is $K^*(Q_4)$. In general, solutions like this obtain if the locus L of minimum points intersects the horizontal portion of the feasible region. Otherwise, a point like (2) would represent optimality. To find the optimum K (K_{opt}), we have to solve the reduced problem $\text{Min}Q(\mu_{max},K)$ over $0 < K < K_{max}$ which is a "line search" problem. Setting $\partial Q/\partial K = 0$ we obtain K_o

$$K_o = \sqrt{\frac{\mu_{max}\bar{d}}{P_k}\left\{\frac{P_b}{\bar{v}} + \frac{w\varepsilon}{\bar{v}} + P_g[A + E(\bar{v} - v_o)^2]\right\}}\ , \tag{3.31}$$

which can be written as $K_o = \sqrt{\bar{d}} \times h(P_b, P_k, w, P_g, \bar{v})$; the remaining parameters in the h function are technical constants (i.e., μ_m, ε, v_o, A, and E).[51] Therefore, K_{opt} is given by

$$K_{opt} = \begin{cases} K_o & \text{if } K_o \le K_{max} \\ K_{max} & \text{otherwise} \end{cases}\ . \tag{3.32}$$

---

[51] These constants are characterized by the available technology and, in this sense, differ in concept from a technical value such as $\bar{v}$, which has been assumed fixed as an "operational imposition."

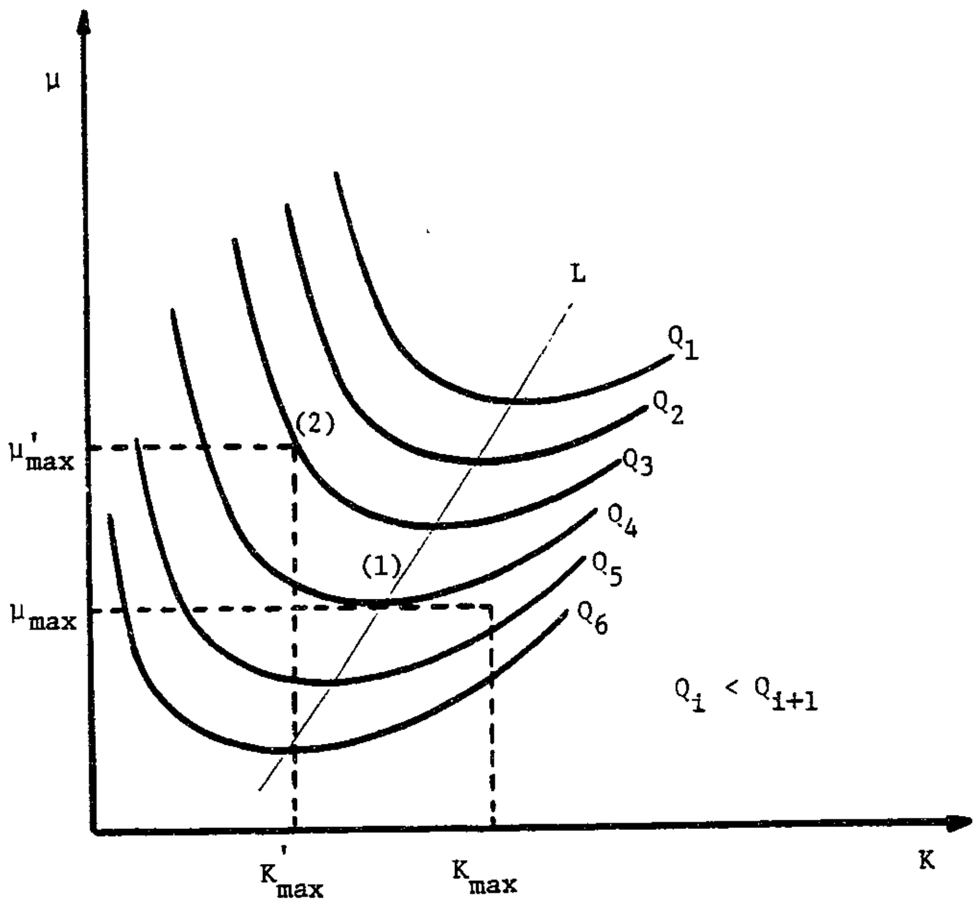


Figure 3.5: Optimum Vehicle and Site Capacities

From M in (3.27), we have that minM = M^* can be obtained by substituting for K and μ K_{opt} and μ_{max} respectively. It can be shown that

$$M^* = \begin{cases} I(P_g,P_b,P_k,\overline{v})\overline{d} + J(P_b,P_k,P_o,P_\mu,w) & \text{if } K_{opt} = K_{max} \\ N(P_b,P_k,P_g,w,\overline{v})\sqrt{\overline{d}} + S(P_k,P_g,\overline{v})\overline{d} + J(P_b,P_k,P_o,P_\mu,w) & \text{otherwise.} \end{cases} \quad (3.33)$$

From (3.22), the cost function corresponding to the transportation system under analysis takes the form

$$C(P_d,P_b,P_k,P_o,P_\mu,P_g,w,y_{12},\overline{v},\overline{d}) = \begin{cases} 2P_d\overline{d} + 2I\,\overline{d}y_{12} + 2J\,y_{12} & \text{if } K_{opt} = K_{max} \\ 2P_d\overline{d} + 2S\overline{d}y_{12} + 2N\sqrt{\overline{d}}\,y_{12} + 2Jy_{12} & K_{opt} < K_{max} \end{cases} \quad (3.34)$$

The form of the cost function in (3.34) was derived assuming a single product, type of vehicle, and O-D pair. Moreover, a "steady state" type of operation was assumed, under a cost-minimizing behavior. Therefore, no aggregation was needed,[52/] nor is it necessary to create a hedonic output index to account for different characteristics of different movements. Under these conditions, all econometric studies reviewed in Chapter 2 would have reduced to a linear speci-

[52/] Not even over time, because the mean flow intensity is constant.

fication in ton-miles,[53/] i.e., $C = \alpha + \beta(y_{12} \times \bar{d})$. But (3.34) shows that even in this case it constitutes a specification error. This is important to stress because the usual critique of ton-miles is that the minimum cost of moving a tons/hour b miles is generally different with respect to that of moving b tons/hour a miles, although both generate the same amount of ton-miles per hour. In the system depicted in this section, this ambiguity does not exist, in spite of which the use of ton-miles is still inappropriate due to the existence of terminal operations, as mentioned before. We will come back to this point in the fourth section, after discussing a two-dimensional output system. It is convenient to call attention to the fact that C in (3.34) resulted in a function of input prices, level of output, level of fixed factor $\bar{d}$, and on the value of a fixed technical parameter $\bar{v}$; note that $\bar{d}$ also plays this role in addition to that of a fixed factor.

When analyzing returns to scale in (3.34), we realize the importance of the fixed cost $2P_d\bar{d}$. If the cost of the right-of-way is paid directly by the firm, (3.34) holds literally, returns to scale are present, and natural monopoly in the geographical context described arises. If the right-of-way is not paid by the firm, we have marginal cost = average cost and constant returns to scale. The first case can be associated with railroads while the second is close to trucking, airlines and shipping (somewhat in accordance

---

[53/] Under prices, $\bar{v}$, $\bar{d}$, invariant. These conclusions do not depend on the inclusion of factor prices in C.

with economic wisdom); however, a basic assumption for the constant returns case is the possibility of adjusting B, K and μ to flow requirements. Of course, these remarks can not be freely extrapolated to different spatial settings; doing so may be extremely misleading as will be seen in the next section.

3.3 The Production Possibility Frontier and Spatial Complementarity

From the discussion in the preceding section it should be clear at this point that the transportation function is the basis for the optimal usage of technology inherent in a transformation function. It is so because the remaining relations involve information associated with the design of elements of vehicles, terminals and rights-of-way, design that enters as an input not subject to change. It is not surprising then that the technically feasible output corresponding to a certain system can be analyzed without necessarily generating a transformation function. In other words, the production possibility frontier could be constructed through the analysis of the operation of the system.

The production possibility frontier represents the maximum level of a certain output component, given the level of the other output components and inputs. Let us analyze the two output components version of the system depicted in the preceding section, i.e., include the possibility of flow with origin at node 2 and destination at node 1, namely y_{21}. Now the output of the system is $Y = (y_{12},y_{21})$. Let us construct the transportation function and the production possibility frontier corresponding to this setting.

In what follows we will keep the previous notation, adding sub-indexes of the i-j type, indicating an O-D specific variable. Let us preserve unique frequency as an operating rule. Therefore,

$$f = \mathrm{Max}\{f_{12}^{'}, f_{21}^{'}\} \quad . \tag{3.35}$$

From this we obtain vehicle load as

$$k_{12} = \frac{y_{12}}{f} \qquad\qquad k_{21} = \frac{y_{21}}{f} \quad . \tag{3.36}$$

Cycle time of one vehicle is given by

$$t_c = t_{12}(k_{12}) + \frac{2k_{12}}{\mu} + \frac{2k_{21}}{\mu} + t_{21}(k_{21}) \quad , \tag{3.37}$$

under $\mu^+ = \mu^- = \mu$. The fleet size needed is

$$B = \eta\, f\, t_c \quad . \tag{3.38}$$

We have to study two cases in terms of relative frequency. Let us first analyze $f_{12}^{'} > f_{21}^{'}$. Recalling that frequency is in terms of trips, we have

$$f = f_{12}^{'} = \frac{y_{12}}{K} \quad , \tag{3.39}$$

and from (3.36),

$$k_{12} = K \qquad \text{and} \qquad k_{21} = \frac{y_{21}}{y_{12}} K \quad . \tag{3.40}$$

From equations (3.37) through (3.40) we get

$$\eta BK = y_{12}[t_{12}(K) + \frac{2K}{\mu} + \frac{2y_{21}}{\mu y_{12}} K + t_{21}(\frac{y_{21}}{y_{12}} K)] \quad . \tag{3.41}$$

As we are interested in relating the $\{Y_{ij}^{kt}\}$ output concept to the ton-miles idea, it is convenient to express (3.41) in terms of distances and speeds. In addition we will assume, as before, that actual speed is independent of vehicle load. Rearranging, (3.41) reduces to

$$\eta BK = y_{12}[\frac{d_{12}}{v} + \frac{2K}{\mu} + \frac{d_{21}}{v}] + \frac{2K}{\mu} y_{21} \quad . \tag{3.42}$$ [54]

It should be noted that if v(k) were a straight linear function of 1/k, then $v(k_{21})$ would be given by $\alpha \frac{y_{12}}{y_{21}K}$, and (3.42) would generate two terms in flow_{ij}—distance_{ij} (ton-miles) units, i.e., $y_{12}d_{12}$ and $y_{21}d_{21}$. However, two terms in pure flow terms would remain due to the loading-unloading effect, as in the one-dimensional output case.

Equation (3.42) is valid for $f'_{12} \geq f'_{21}$. Recalling conditions (3.10), this is equivalent to limit its validity to $y_{12} \geq y_{21}$. Similarly, we have a symmetric expression for the case $y_{21} \geq y_{12}$. Rearranging (3.42) we get the result

$$y_{21} = \frac{\eta\mu B}{2} - [\frac{\mu(d_{12} + d_{21})}{2Kv} + 1]y_{12} \qquad \text{for } y_{12} \geq y_{21} \tag{3.43}$$

$$y_{12} = \frac{\eta\mu B}{2} - [\frac{\mu(d_{12} + d_{21})}{2Kv} + 1]y_{21} \qquad \text{for } y_{21} \geq y_{12} \quad . \tag{3.44}$$

Noting that the slope of $y_{21} = f(y_{12})$ is negative and less than -1, the graphical representation of the system (3.43) — (3.44) looks like Figure 3.6. These two equations represent the transportation function

---

[54] In fact, the analysis can be done directly in terms of load-dependent travel times.

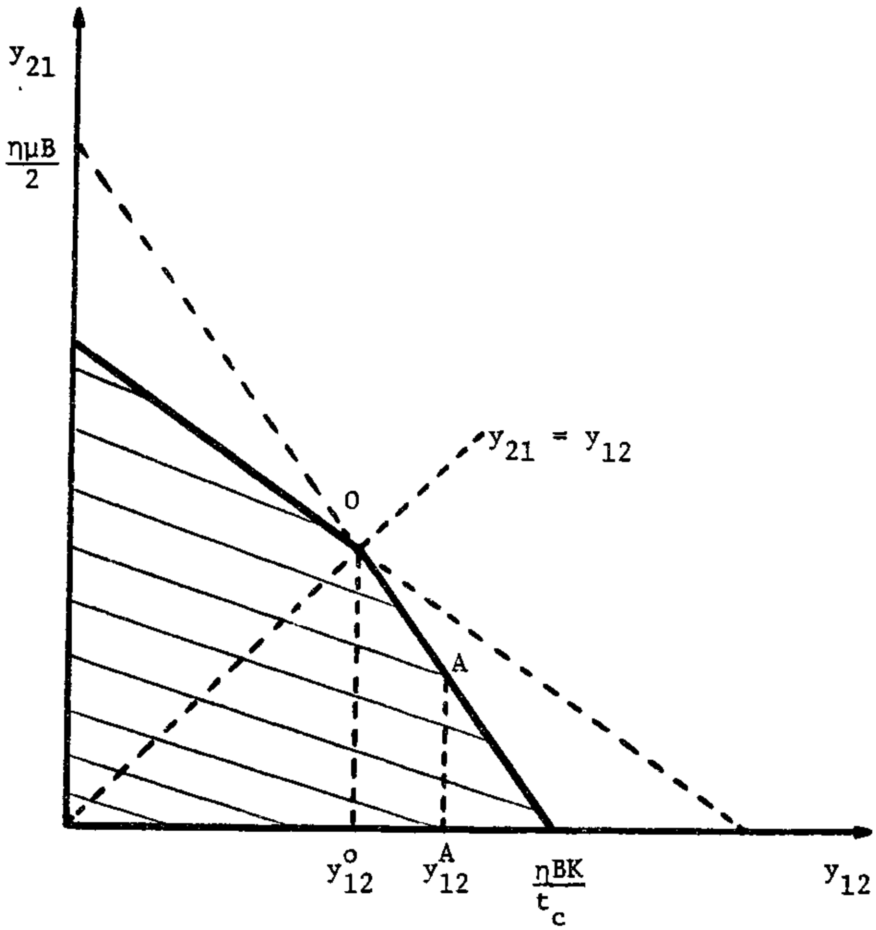


Figure 3.6: Production Possibility Frontier with a Two-Dimensional Transportation Output

of the system, and the shaded area in the figure represents all the vectors (y_{12},y_{21}) that can be produced with a given fleet B, and capcities μ and K, but only the boundary represents optimal usage. The boundary, then, is the production possibility frontier, whose symmetry is derived from the assumption of load independence of speed.

It is convenient to analyze Figure 3.6 to a certain extent. At first sight one may ask why a flow like y_{12}^{A} is associated with a maximum $y_{21} < y_{12}^{A}$; the system described suggests that a similar flow could be "returned" from 2 to 1. The answer is that for a given fleet size, vehicle capacity and site capacity, only full capacity operation in both directions allows for flow equality. Any other point (like A) would require an imbalance between loading (or unloading) times associated with each flow. For instance, y_{12}^{A} would require a longer stay in node 2 than y_{12}^{o} and, therefore, a shorter stay in node 1 in order to keep $f = y_{12}^{A}/K$. If loading-unloading times were zero, only points like O would be generated. Given that the number of sites (loading-unloading) is proportional to the mean flow intensity being produced if site capacity is kept constant, we can do some qualitative although restricted analysis of iso-cost curves in the (y_{12},y_{21}) space, in terms of the expenditure in vehicles and sites. Holding d, v, K, and μ constant, gas expenditure will be proportional to B, and labor associated with vehicles and sites can be incorporated into costs through inclusive prices, i.e.,

$$P(K) + w\varepsilon + P_g g \frac{d_{12} + d_{21}}{\frac{d_{12} + d_{21}}{v} + \frac{2K}{\mu}}$$

as a vehicle price P_B, and $P(\mu) + w\theta$ as a site price P_S [see (3.20) and (3.21)].

Under this setting,

$$C(y_{12},y_{21}) = C_o + P_B B(y_{12},y_{21}) + P_S S(y_{12},y_{21}) \quad , \tag{3.45}$$

where S is the total number of sites given by $2/\mu(y_{12}+y_{21})$, and C_o should be associated with right-of-way costs. In Figure 3.7, AED and HGI are production possibility frontiers corresponding to B_2 and B_1 fleet sizes respecitvely, with $B_2 > B_1$. The number of sites is constant on JEL; call it S_2. Similarly, $S = S_1 < S_2$ on AGD. Therefore, we can establish the following relations (C_i denoting cost at point i):

$$C_A = C_D \quad (B = B_2,\ S = S_1)$$

$$C_G < C_A \quad (B_G = B_1 < B_2,\ S = S_1)$$

$$C_E > C_A \quad (B = B_2,\ S_E = S_2 > S_1 = S_A) \ .$$

Therefore, there is some point F between E and G such that $B_1 < B_F < B_2$ and $S_1 < S_F < S_2$, and $C_F = C_A = C_D$. In addition, the iso-cost locus is symmetric with respect to the $y_{12} = y_{21}$ line, due to the symmetry of both the production possibility frontier and the "iso-sites" loca. The iso-cost locus then looks like the DFA curve in Figure 3.7, concave to the origin. Thus, $C(y_{12},y_{21})$ is quasi-convex (as defined in Chapter 1). Piece-wise linearity arises because of linearity of both B and S on y_{12} and y_{21}.

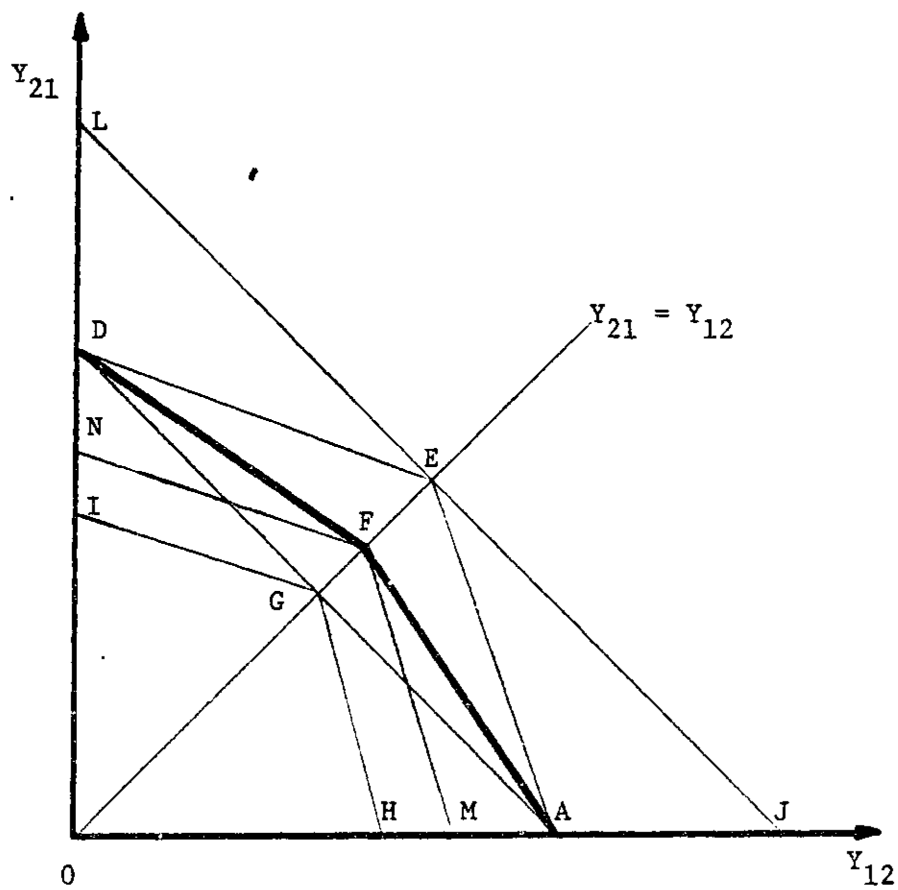


Figure 3.7: Iso-Cost Locus with a Two-Dimensional Transportation Output

The next question to be addressed is whether the cost analysis in terms of an aggregate output defined in PU times DU units, yields an unambiguous answer. Such an output (e.g., ton-miles), would be generated as a particular case of (3.8), namely

$$\overline{Y} = y_{12}d_{12} + y_{21}d_{21}\left[\frac{P.U.}{U.T.}\ D.U.\right] . \tag{3.46}$$

$\overline{Y}$ appears in the (y_{12},y_{21}) space as a straight line with slope $-d_{12}/d_{21}$. Following Figure 3.8, the line representing $\overline{Y}_0$ (associated with $d_{12} > d_{21}$) intersects different iso-cost curves and therefore cannot be associated with a single cost figure. Not even in the $\overline{Y}_1$ case ($d_{12} = d_{21}$) is this correspondence possible, although the variation of cost level along $\overline{Y}_1$ is less than along $\overline{Y}_0$. Again, the formulation of $C = C(\overline{Y})$ would constitute a specification error because it is inconsistent with the underlying technical transportation analysis.

In order to address the subadditivity problem, we can formulate $C(y_{12},y_{21})$ analytically by replacing B as a function of output from (3.42) and S by its value $2/\mu(y_{k2}+y_{21})$. Then, rearranging terms (3.45) becomes

$$C(y_{12},y_{21}) = \begin{cases} C_o + y_{12}\left[\frac{P_B}{\eta}\left(\frac{d_{12}+d_{21}}{vK} + \frac{2}{\mu}\right) + \frac{2P_S}{\mu}\right] + y_{21}\,\frac{2}{\mu}\left(\frac{P_B}{\eta} + P_S\right) & y_{12} \geq y_{21} \quad (3.47) \\ C_o + y_{21}\left[\frac{P_B}{\eta}\left(\frac{d_{12}+d_{21}}{vK} + \frac{2}{\mu}\right) + \frac{2P_S}{\mu}\right] + y_{12}\,\frac{2}{\mu}\left(\frac{P_B}{\eta} + P_S\right) & y_{21} \geq y_{12} . \quad (3.48) \end{cases}$$

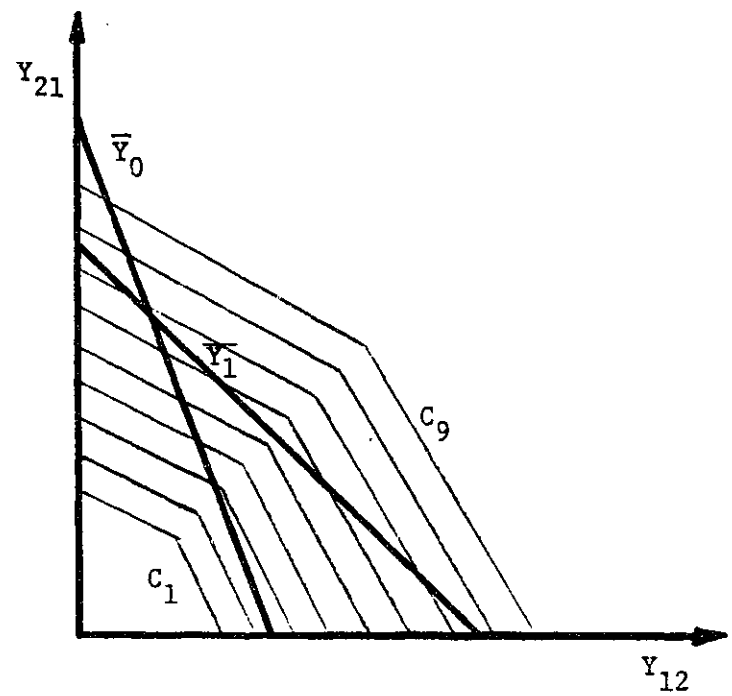


Figure 3.8: Cost Ambiguity of Aggregate Output

It can easily be seen that the degree of scale economies S is given by

$$S = C(y_{12},y_{21})/(y_{12}\frac{\partial C}{\partial y_{12}} + y_{21}\frac{\partial C}{\partial y_{21}}) = C(y_{12},y_{21})/[C(y_{12},y_{21}) - C_o] ,$$

$$\forall (y_{12},y_{21}) . \qquad (3.49)$$

In particular, S = 1 for $C_o = 0$ (the "trucking" case). In other words, in the absence of costs associated with the right-of-way, proportional expansion of the output vector requires proportional expansion of inputs. Naturally, this can also be seen from the fact that ray average costs are constant for $C_o = 0$.[55] Thus, constant returns to scale are present, which would indicate that a competitive structure would be efficient. However, more careful analysis of subadditivity complements this aspect. We have to compare $C(y_{12},y_{21})$ with $\sum_{i=1}^{n} C(y^i_{12},y^i_{21})$, where $\sum_{i=1}^{n} y^i_{12} = y_{12}$ and $\sum_{i=1}^{n} y^i_{21} = y_{21}$. Let us take n = 2 and compare $C(y_{12},y_{21})$ with $C(y_{12},0) + C(0,y_{21})$, which is the analysis for economies of scope in the two-product case. $C(y_{12},0)$ is given by (3.47) with $y_{21} = 0$. $C(0,y_{21})$ is given by (3.48) with $y_{12} = 0$. Then the sum gives

$$C(y_{12},0) + C(0,y_{21}) = 2C_o + (y_{12}+y_{21})\left[\frac{P_B}{\eta}(\frac{d_{12}+d_{21}}{vK} + \frac{2}{\mu}) + \frac{2P_S}{\mu}\right].$$

(3.50)

---

[55] Proof: from (3.47) and (3.48) and $C_o = 0$, $\frac{1}{k} C(ky_{12},ky_{21}) = C(y_{12},y_{21})$, $\forall k$, $\forall (y_{12},y_{21})$.

This should be compared to (3.47) and (3.48), which leads to

$$C(y_{12},0) + C(0,y_{21}) - C(y_{12},y_{21}) = \begin{cases} C_o + y_{21}P_B\left(\frac{d_{12}+d_{21}}{\eta vK}\right) > 0 \;, & y_{12} \geq y_{21} \\ C_o + y_{12}P_B\left(\frac{d_{12}+d_{21}}{\eta vK}\right) > 0 \;, & y_{21} \geq y_{12} \;. \end{cases} \tag{3.51}$$

This indicates that there are economies of scope in the production of (y_{12},y_{21}). Therefore, even under the case of no direct costs for right-of-way ($C_o = 0$), production of (y_{12},y_{21}) is cheaper with one firm than with two (or more) firms producing orthogonal partitions of that output bundle.[56/] Keeping in mind that the kind of complementarity between the production of y_{12} and y_{21} corresponds to spatial complementarity, from this we conclude that in spite of constant ray average costs for $C_o = 0$, merging is convenient due to economies of spatial scope. It is important to recognize that these kinds of economies are present due to the existence of idle capacity in the case analyzed in (3.2) (i.e., backhaul capacity), which should be remembered as one cause of economies of scope in general as seen in Chapter 1. Figure 3.9 represents the two-output cost function corresponding to (3.47) and (3.48) for both $C_o = 0$ and

56/ Note that the extra expenditure is associated with the purchase of additional vehicles with respect to those needed by one firm.

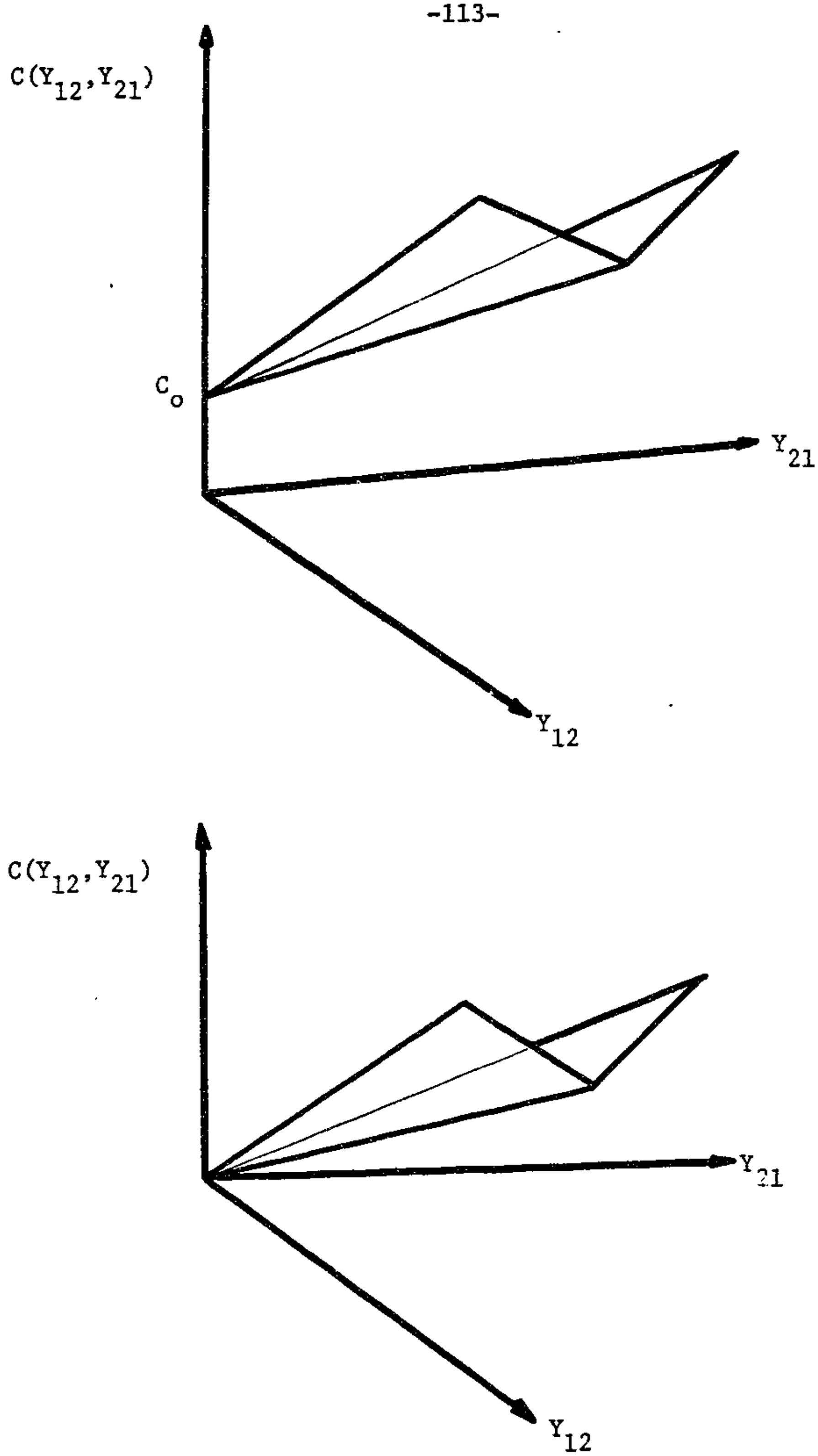


Figure 3.9: A Two-Output Transportation Cost Function

a) $C_o \neq 0$; diminishing RAC, transray convexity

b) $C_o = 0$; constant RAC, transray convexity

$C_o \neq 0$ cases; from this, transray convexity can be clearly seen.

Let us go back now to equation (3.42) which forms the basis for the production possibility frontier, and constitutes the generic form of the transportation function of our system. We have stated that under a certain form of $v(k_{ij})$, namely $v(k_{ij}) = \alpha/k_{ij}$, terms like $y_{ij}d_{ij}$ would be generated. In fact, recalling that $k_{12} = K$ and $k_{21} = Ky_{21}/y_{12}$ $(y_{12} > y_{21})$, (3.42) would become

$$\eta BK = \frac{K}{\alpha}(y_{12}d_{12} + y_{21}d_{21}) + \frac{2K}{\mu}(y_{12} + y_{21}) \quad , \qquad (3.52)$$

which would also hold for $y_{21} > y_{12}$ given its symmetry. It is not difficult to conclude that even in this case the $\bar{Y}$ concept is ambiguous, but we should realize that, as in the cost function in (3.34), the non-distance-weighted flow terms arise due to loading-unloading activities,[57] i.e., μ. This suggests that the ton-miles concept is more inappropriate the more important terminal operations are, and the smaller the relation between speed and vehicle load.

3.4. The Estimation of Transportation Cost Functions

We have seen that the definition of transportation output consistent with the technical analysis behind the performance of transportation systems, is a central aspect when defining a cost function for that system, if any meaningful set of policy conclusions is to be established from this function. However, the specification of Y as a vector $\{Y_{ij}^{kt}\}$ generally makes any attempt to estimate such a function econometrically from actually observed data infeasible. This makes some kind of aggregation necessary. On the other hand,

---

[57] See (3.33) to check that the coefficient of the "pure flow" term in (3.34) is the only place where $P(\mu)$ enters the cost function.

there are three other types of variables that should theoretically appear in C, namely input prices, fixed factors, and those technical parameters or variables which are exogenous to the firm (i.e., those that the firm cannot optimize or modify). In this section, we will discuss output aggregation, the role of fixed factors and technical parameters, and the relation between engineering transportation models and estimated cost functions. To analyze these three aspects, we will profit from the initial discussion on aggregation given in section 3.1, and from the insights provided by the development of the two cases in sections 3.2 and 3.3. In addition, we will identify some generic cases which can be analyzed with reasonable aggregation; we will select those specifications which seem appropriate for policy analysis; and we will summarize the advantages of the new approach.

3.4.1. Aggregation in the Analysis of Complex Systems

As stated in section 3.1, aggregation of output over any dimension (commodity, time or space) involves losing information associated with the transportation processes generated by the system in reference. At this point, we are interested in viewing this loss of information from the perspective of the cost function of the system.

Let us begin by discussing the most unclear type of aggregation: spatial, which has been our implicit preoccupation in the analytical development in sections 3.2 and 3.3 As evident, spatial aggregation destroys the information on the geographical context of the origin-desination system in which a transportation system operates. We have already seen that nuclearizing adjacent origins and destinations may be considered a first step in spatial aggregation. However, this procedure should be complemented by some other procedure to account for intranodal movements, i.e., movements within the new, aggregated, nodes. If all distances were equal among pairs of basic nodes within an aggregate, equations like (3.52) suggest that simple summation of intranodal flows would create an appropriate flow variable to "represent" that aggregate in the cost function, i.e.

$$Y_{N_i}^{kt} = \sum_{\ell,m \in N_i} Y_{\ell m}^{kt} \quad , \tag{3.53}$$

where $N_i = \{\ell,m,\ldots\}$ represents a "collection" of adjacent nodes. However, distances will generally not be equal. In this case, the discussion in the preceding sections, particularly those parts related to the ton-miles criticism, suggest that a more appropriate aggregation of intranodal flows from a cost function perspective would have the form

$$Y_{N_i}^{kt} = \sum_{\ell,m} d_{\ell m} Y_{\ell m}^{kt} + \alpha \sum_{\ell,m} Y_{\ell m}^{kt} \quad , \quad \ell,m \in N_i \tag{3.54}$$

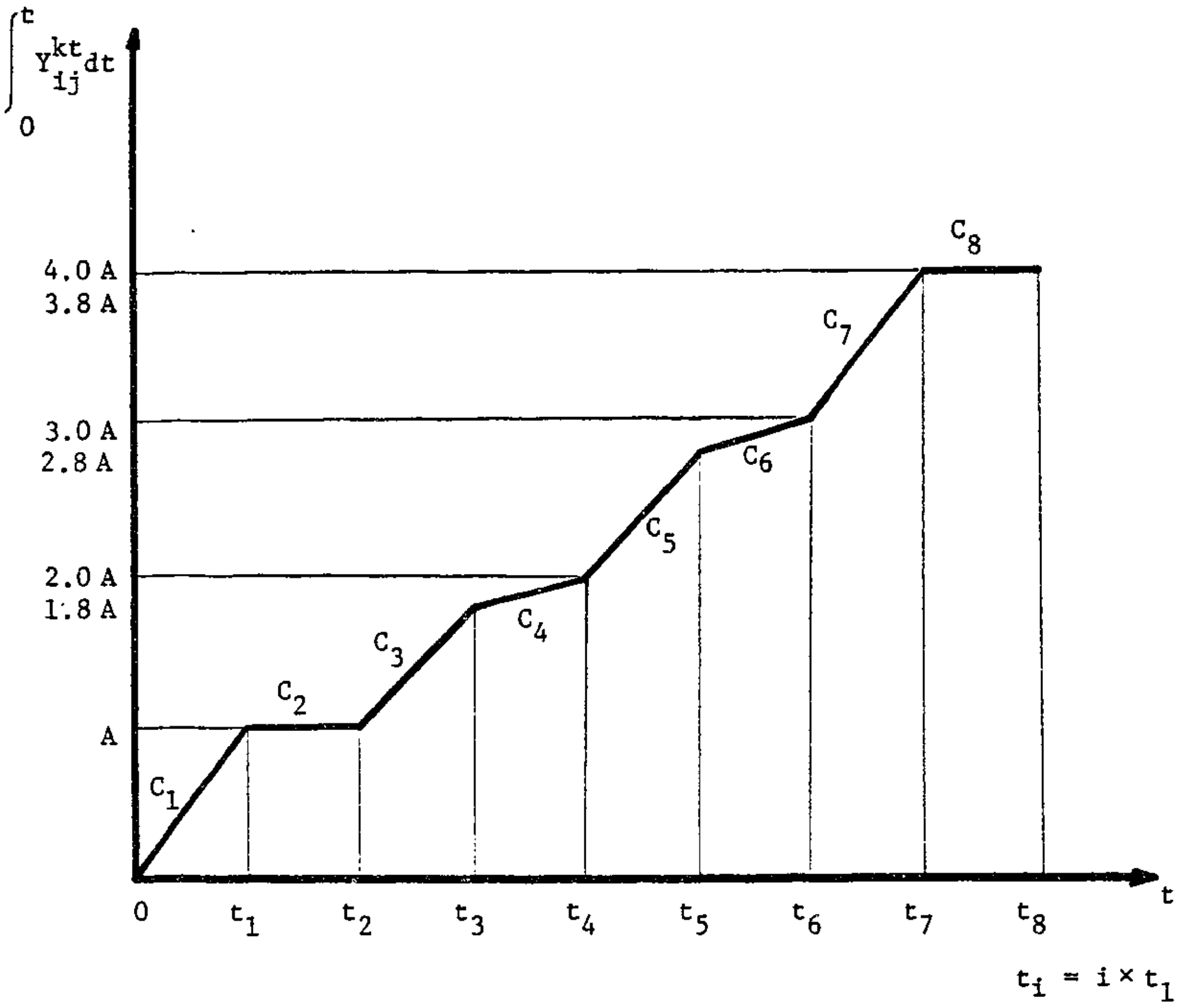
The distance-weighted sum of flows accounts for actual movements in space, while the unweighted sum of flows accounts for loading-unloading activities, as suggested by (3.34) and (3.52).[58/] It is worth stressing that flows are associated with O-D pairs and not with links; d_{ij} is distance traveled to move Y_{ij}^{kt} (and may eventually be d_{ij}^{kt}!). The problem of distance itself and the link-network

---

[58/] This rationale behind proposition of (3.52) can be grasped through the association of a system like the one in section 3.3, to international shipping, railroads, trucking, inter-city buses, etc.: the loading-unloading part of costs may be highly relevant and is directly associated with flow and not with distance-weighted flow.

problem will be again analyzed when discussing fixed factors and operational parameters. Following the procedure summarized in (3.54), however, will not generally allow one to derive any conclusions in terms of marginal costs, scale and/or scope within a macro-node, and should be seen more as a device for spatial cost allocation in a sort of nested analysis after which intranodal costs could eventually be studied.

Aggregation of output over time, as in (3.3), may cause distortions when estimating cost functions if periods of distinctive mean flow intensities are being averaged. At a microscopic level, we lose information on the kinematics of the processes; this can be visualized by associating Figure 3.1 to arrivals and unloading of trucks at a point. At a macroscopic level, we lose information on the flow pattern in relevant periods. This may cause some ambiguity in cost analysis because of two aspects; first, two observations involving very different time flow patterns but with the same mean flow intensities will be counted as producing the same output, but their associated costs may differ substantially. As an example, $Y_{ij}^{k(0 \rightarrow t_2)} = Y_{ij}^{k(t_2 \rightarrow t_4)}$ in Figure 3.10, but $C_1 + C_2 \neq C_3 + C_4$ in general. A second problem relates to the operating conditions prevailing in two different periods; if these conditions are different, the production of the <u>same</u> pattern of flow intensities will generally have different associated costs. As an example, annual observations of costs and flows in a waterway system with total time aggregation (i.e., $Y = \{Y_{ij}^k\}$) would weight equally winter and summer movements; for the same mean flow intensities in two years, we would expect the



$$Y_{ij}^{k(0 \to t_2)} = Y_{ij}^{k(t_2 \to t_4)} = Y_{ij}^{k(t_4 \to t_6)} = Y_{ij}^{k(t_6 \to t_8)} = 0.5 \frac{A}{t_1}$$

$$Y_{ij}^{k(0 \to t_1)} = Y_{ij}^{k(t_6 \to t_7)} = \frac{A}{t_1} \qquad Y_{ij}^{k(t_1 \to t_2)} = Y_{ij}^{k(t_7 \to t_8)} = 0$$

$$Y_{ij}^{k(t_2 \to t_3)} = Y_{ij}^{k(t_4 \to t_5)} = 0.8 \frac{A}{t_1} \qquad Y_{ij}^{k(t_3 \to t_4)} = Y_{ij}^{k(t_5 \to t_6)} = 0.2 \frac{A}{t_1}$$

Figure 3.10: The Problem of Time Aggregation

year with higher winter movements to be associated with the higher cost. Associating t_i's in Figure 3.10 with quarters, we expect $\sum_{i=1}^{4} C_i \neq \sum_{i=5}^{8} C_i$. Therefore, in both the different time flow patterns and different operating conditions cases, ceteris paribus time aggregation would cause the expected value of the error term to be different from zero in the stochastic specification of the cost function, and estimators will be biased. To avoid this, we would like to specify output accordingly, e.g., $Y = (Y_{12}^{11}, Y_{12}^{12}, Y_{12}^{13}, Y_{12}^{14}, \ldots, Y_{ij}^{k1}, \ldots, Y_{ij}^{k4}, \ldots)$ in Figure 3.10, and then specify annual costs $C = C(Y)$. Whenever this procedure cannot be followed because of data availability problems or estimation capacities, description of the operating conditions may help describe output, but policy conclusions do not follow easily from the results.

Finally, commodity aggregation may affect cost estimation since the (minimum) cost of moving the same aggregate weight or volume will generally depend on the composition of that output. Moving frozen strawberries, coal or gasoline requires different technologies; and moving five tons of strawberries and ten of coal per month from i to j in a given year will certainly have a different cost with respect to 15 tons of gasoline per month in the same spatial setting. In general, physical and chemical characteristics of commodities provide the necessary information to judge both the compatability among them (i.e. carrying them together), and the need for different equipment. It is clear that bags of potatoes and apples can be carried together and that appropriate equipment is similar, while gasoline and strawberries cannot be mixed and equipment is different. Thus, in general some aggregation can be done without causing too much cost ambiguity.

In summary, the loss of information due to aggregation over any dimension may cause serious problems of coefficient interpretation when estimating a cost function. Because of the very nature of the problem, each particular case should be carefully analyzed in order to seek the appropriate aggregation over space, time, or commodities, or at least to have an idea of the type of distortion introduced when undesirable aggregation has to be done. It is convenient to have a perception of what is being lost in the aggregation process in any case; this allows for both a better analysis of the particular case under study, and for more relevant policy conclusions.

It is interesting to note that there has been an attempt in some of the studies in Chapter 2 to use surrogate procedures to "rescue" the information implicitly lost when defining a grossly aggregated transportation output. Thus, seasonal and "traffic condition" dummies are in fact trying to capture the effect of the implicit time aggregation on costs.[59/] Similarly, variables like traffic mix or insurance value try to grasp commodity aggregation. Finally, the only effort to somehow counterbalance spatial aggregation has been the use of mean haul length as part of output description. However, as we have stated before, this kind of procedure darkens the interpretation and analysis of the cost function. In particular, the meanings of marginal costs, scale economies, and production complementarity

---

[59/] In the same category should be included a variable like frequency (as part of the output description); the reason for this will be explained later although it could be foreshadowed at this point.

remain completely obscure.

3.4.2 The Role of Fixed Factors and Technical Parameters

We know that cost minimization subject to a transformation function and some fixed factors $\overline{X}$, gives rise to a restricted or short-run cost function $C(w,Y,\overline{X})$. The fixed nature of $\overline{X}$ arises because of the impossibility of adapting the amount of those factors following changes in the (exogenous) level of Y. $\overline{X}$ generates a fixed cost $w_{\overline{X}}\overline{X}'$ and also influences the variable part of C, i.e., $C = w_{\overline{X}}\overline{X}' + C_1(w,Y,\overline{X})$. Fixed factors in transportation will depend on the particular case under study. For instance, miles of track in railroad analysis will always play this role; number of loading-unloading sites in shipping, and even fleet size and capacity, may well be another example.
We have seen in Chapter 1 that a cost function can always be seen (without loss of generality) as the sum of a function F(S) and $C_1(Y)$, where F(S) is the total value of "fixed" costs and depends upon the precise set of goods of which <u>strictly</u> positive quantities are produced.[60/] With our definition of transportation output we can associate this concept with the fixed cost nature of the right-of-way as follows: the fixed cost of producing a positive flow Y_{ij} between i and j is the miles of road (track) necessary to connect i and j, unless other pairs i-k, k-j are being served already

---

[60/] Remember that $F(S) = \sum_{i \in S} F(i)$ is a <u>particular</u> case of F(S), F(i) being <u>the</u> fixed cost associated with product i. F(i)'s magnitude does <u>not</u> depend on other products' amounts.

(i.e., $Y_{ik} \neq 0$, $Y_{kj} \neq 0$) such that this cost may be avoided.[61] In other words, the fixed cost associated with the provision of a flow $Y_{ij} \neq 0$ will be at most the cost of physically connecting these two points i and j. Therefore in the transportation case in general the property $F(S \cup T) \leq F(S) + F(T)$ holds, which allows us to analyze only the $C_1(Y)$ part of the cost function when studying subadditivity.[62] This is a very nice property which facilitates policy analysis through cost functions in transportation because, in some cases, only $C_1(Y)$ need be estimated.

A related but different aspect in transportation cost functions has to do with what we can call technical parameters, like speed or route structure. In (3.22) for example, cost minimization requires optimizing with respect to v; as long as the firm has the choice of setting v as desired, it will not appear as an argument in C. Whenever the technical parameter is exogenously imposed and no choice is possible, it should in fact enter the specification of C. Particularly important to understand is the case of route structure, which is strongly connected with the inappropriateness of the ton-miles concept foreshadowed at the end of Chapter 2. The main point here is that in fact the distance covered by a firm in seeking to serve a pattern of flow during a given period, is in general a decision

---

[61] Of course, this is going to depend on the geographical context (topography and location).

[62] It should be remembered that if $C_1(Y)$ is subadditive and F(S) fulfills the aforementioned condition, then C(Y) is subadditive. This is to be expected because fixed costs favor subadditivity, although sometimes destroy transray convexity of C(Y).

of the firm. Take for example the O-D system depicted in Figure 3.11, where commodities should flow from all origins to all destinations. Even aggregating over time and commodities, this particular O-D system generates 30 variables or components of the transportation output vector Y. Many route configurations can be used to produce a given flow system, as shown in the same figure. Given the value of the Y components, each configuration can be associated with appropriate (optimal) fleet size, vehicle capacity, loading-unloading capacities, etc. The combination of all these factors generates the cost of each alternative and the minimum cost structure corresponding to a given Y would generally be a decision of the firm. In other words, cost minimization for a given flow vector Y generates a certain route structure as part of a firm's decisions, unless this structure is fixed in the short run because of investment lumpiness or institutional constraints. In the absence of these latter, we can even draw a difference between right-of-way as a fixed factor and route structure, this being an answer to a given Y within the boundaries of the former.[63/] Therefore, distance plays a dual role in the behavior of the cost-minimizing transportation firm: as a fixed factor associated with the payment for right-of-way (e.g., our $P_d\overline{d}$ in (3.2)), and as a technical variable associated with the firm's route structure

[63/] The reason we view the work by Gordon and deNeufville (1973) as important from this perspective is precisely because they pointed out the relation between fleet size and "network shape" when producing a given flow pattern (see our Chapter 2). In a later work, they explicitly faced the problem of output definition, although turning their attention to the quality description of aggregate output (Gordon and de Neufville (1977)).

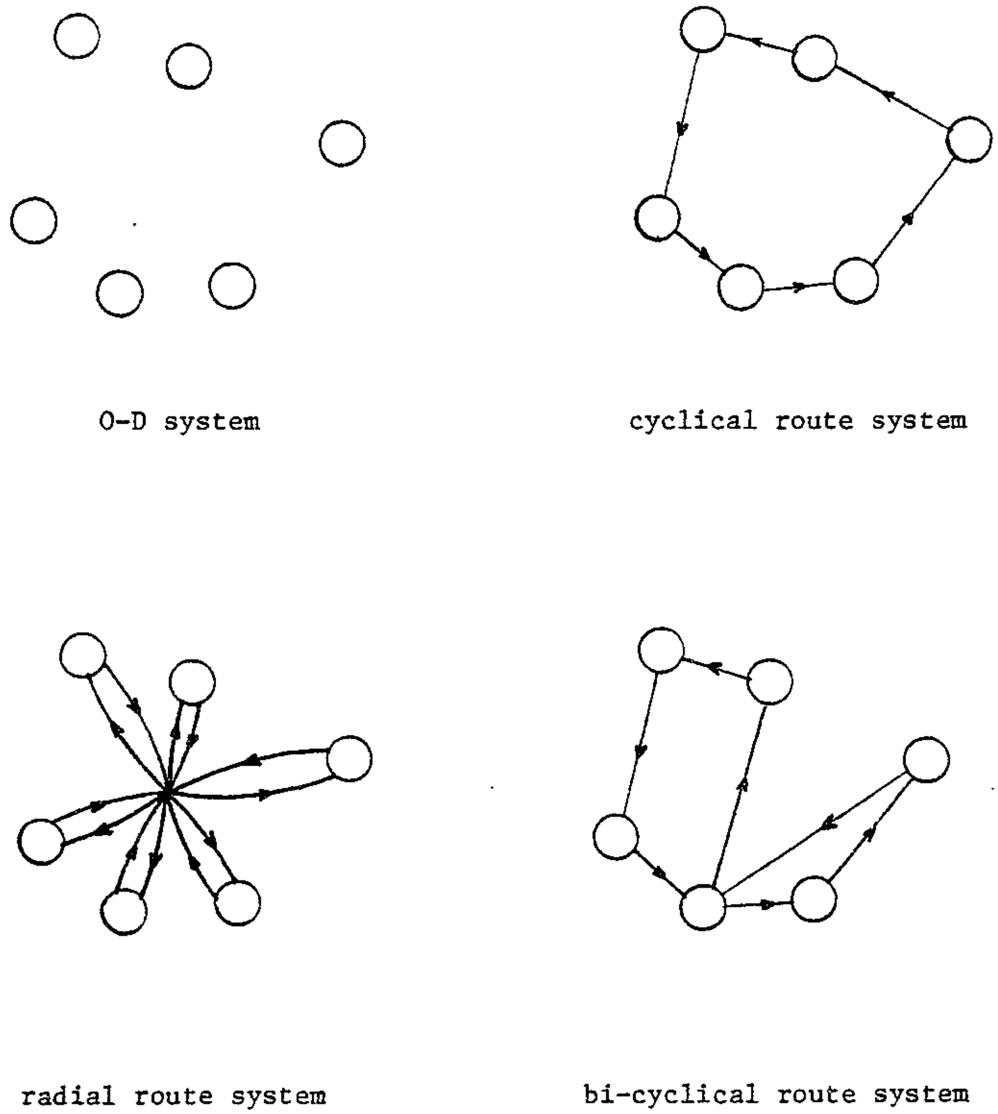


Figure 3.11: Different Route Systems Associated with a Given O-D System

choice. It may well be the case that optimal route structures within the boundaries of existing physical networks, vary in shape depending on the level of the (exogenous) components of Y.

Although each particular case should be analyzed separately, we may expect a priori that elements like speed, fleet size, vehicle capacity and route structure, would enter in an endogenous way in the airline's decisions (and even in trucking because of the high density of the physical networks).[64/] The amount of track would be a fixed factor in railroads, and generally would make the route structure associated with a given O-D system somewhat rigid. Thus, the manner in which each transportation system operates provides enough information to judge the exogeneity or endogeneity of technical or operational parameters.[65/] Finally, it is worth noting that the relevant reference period to judge the adaptability of factor amounts to variations in output, is just the period of observation. In general, operating decisions (i.e. how factors are used) are easier to modify than the amount of factors used within that period.

64/ Here we refer to a firm's behavior as a private entity taking decisions, not to social costs.

65/ In this sense, the mean speed variable ($\bar{v}$) used by Braeutigam et al. (1980) is not playing the role of a technical parameter (it is not train speed) but the one of surrogate desciptor of mean flow intensity. The "bridge line" nature of the railroad firm is not an argument for speed exogeneity, but actually for the exogeneity of our Y.

3.4.3 Comparative Improvements of the New Approach

Output definition and measurement constitute the core of our discussion of a new approach to estimate transportation cost functions. Our Y_{ij}^{kt} vector restates the problem from the beginning, following a bottom-up procedure. Under this perspective, the ton-miles or similar concepts are identified as the result of aggregating over time, commodities and space. Up to this point we have offerred a systematic critique of the ton-mile output concept, a critique that flows from the very roots of transportation on engineering and economic grounds. The problem of output measure has been intensively discussed for more than a decade, but we have observed that the conceptual aspect has been strongly distorted by accepting as the departure point the ton-miles concept, which has acted as the "sin of youth" of transportation economics. We have found no single paper that has not used the ton-miles idea as the basis for output definition, or as a "generic" measure. The limitations of this concept have been foreshadowed, however, by many researchers in the transportation field. Serious attempts to restate the problem can be found in the literature; the work by Steger (1966) on policy-sensitive output measures, and the work by Gordon and de Neufville (1977) on the contradiction between scale economies analysis and the actual merging in the airline industry, can be seen as good examples of this assertion.

Although interrelated, comparison among different output definitions should be understood in our study in the context of cost function analysis in transportation. From this point of view, we can grossly classify output treatment in four categories:

i) physical units times distance (ton-miles, passenger-miles, etc.): Y_1; ii) Y_1 plus other variables accounting for different periods, commodities, or location; iii) quality modified Y_1, vector or hedonic; iv) vector of mean flow intensities of different commodities, among different O-D pairs, in different periods.

The main characteristic of the second approach is that, in addition to ton-miles, variables like seasonal dummies, traffic mix, regional dummy, traffic conditions, etc., are introduced in the cost function specification. This procedure can in fact be recognized as one way to deal with the three aspects inherent in each component of the vector in iv). The third approach describes the output associated with each observation in terms of Y_1 and some descriptors of what is being moved, distances, and how things are moved.[66/] Again, this approach accounts for commodity and spatial heterogeneity among observations; however, it introduces a highly polemical aspect by describing output using operational dimensions. Whether these dimensions are exogenous or not is an extremely delicate point.[67/] In general, as seen in section 3.4.2, it is convenient to clearly distinguish between proper output and the way output is produced.

66/ For instance, in Spady and Friedlaender (1978), insurance and shipment size account for what is being moved, mean length of haul accounts for distances, and less-than-truckload lots and average load account for how things are moved.

67/ Take for example the proposition of including frequency as part of the output description in Gordon and de Neufville (1977).

When the cost function is specified in terms of an ambiguous definition of output, then the analytical interpretation of that function becomes delicate, and frequently ambiguous too. The meaning of marginal cost as the cost of producing an additional output unit certainly gets lost when using either of the three first approaches, because this cost depends on where and when the additional ton-mile is generated. Similarly, the "first-best price" interpretation of marginal cost makes no sense in this context. On the other hand, the expression $\partial C(Y,w)/\partial Y_{ij}^{kt}$ is unambiguous in its meaning, i.e., it represents the cost of moving an additional unit of k between i and j during period t. It follows that the first-best price meaning is perfectly adequate, identifying commodity, origin-destination, and period. A second aspect related to the analytical capabilities of the cost function deals directly with the important aspect of economies of scope. This is an extremely important point, that has been neglected in the past, provoking a serious difficulty in the analysis of transportation industries. Is it possible to explain merging in tranpsortation through cost analysis in terms of ton-miles or modified ton-miles? The negative answer can be clearly obtained by comparing the implications from estimated transportation "cost functions" using the ton-miles output, and actual behavior of firms within the corresponding industry. Constant returns to scale have been postulated in those industries where the carrier does not own its right-of-way, as airlines and trucking. However, in both cases "paradoxical" merging and/or enlargement has been observed as firms' behavior. The airline case has been described and analyzed in Gordon and de Neufville (1977);

resolution of the paradox is explained in terms of the quality components, e.g. flight frequency, of the seat-miles figure, and introducing the concept of "output value" (V_T) which depends on quality. The explanation then states that a process described by "constant returns," i.e.,

$$C = KY_1 \tag{3.55}$$

is perfectly compatible with a modified concept of scale economies, i.e.,

$$\frac{\partial (C/V_T)}{\partial Y_1} < 0 \; ; \tag{3.56}$$

where $V_T = Y_1 V_u(\text{quality})$. (3.57)

The idea is that quality (frequency), and therefore value V_u, increases when quantity (Y_1) increases. Therefore, the cost per output-value diminishes when more seat-miles are produced. This interesting approach to explain firm enlargement even in one O-D pair, should be understood as the combination of two effects under the Y_{ij}^{kt} multi-output definition (which, in this case of one O-D and passengers, reduces to Y^t): the existence of economies of time scope, and a demand effect. The first aspect relates to the fact that it is convenient for one airline to produce transportation in period t_i, given it is already producing in period t_j ($j \neq i$). The second aspect relates to the convenience of producing more due to the effect on demand via the generalized price effect (e.g., money price + value of time); this effect, however, makes the product (demand) endogenous and proper estimation of a cost function should include the demand side as part of the system to be analyzed. Under these circumstances (3.56) cannot be called scale economies of any kind. The trucking case on merging and scale economies was studied by Spady and Friedlaender (1978) under a similar approach. In this industry, economists believe that constant returns are present, but the straight ton-miles approach normally indicates increasing returns to scale (e.g. a significant constant term appearing in a linear form). Spady and Friedlaender redefined output in a way very similar to (3.57).[68] This so-called effective output ψ was then used to analyze returns to scale in the usual way. The hypothesis of constant returns could not be rejected. Actual merging in the industry was explained in terms of regulatory incentives, economies of density and utilization. The fact is that merging among firms serving different O-D pairs because of actual cost savings reflects economies of spatial scope, i.e., reflects the convenience of serving other O-D pairs given that some O-D pairs are being served. The reason for these kinds of economies being present can be found in the flexibility of schedules, backhauls, etc., as in the case developed in section 3.3 (better usage of the fleet). These kinds of economies will certainly be present in some spatial O-D patterns for any mode, and can be detected only by properly specifying output. Spatially aggregated output does not allow for the analysis of this essential aspect of transportation production. It should be remembered that the analytical development contained in equations (3.47) through (3.50) shows that even in the trucking case, and against conventional economic wisdom, incentives for mergers may appear

---

[68] See our Chapter 2 for a description of and comments on this paper.

due to spatial complementarity in spite of constant ray average costs In fact, the modified or hedonic ton-miles (as ton-miles themselves) allow in the best case only for analysis in terms of proportional changes in flow components, as foreshadowed by Griliches (1972); the "composite commodity" concept from multi-output theory is, in this context, applied to the vector of flows in a weighted fashion.[69/] We view the possibility of the presence of economies of spatial scope as the most important aspect systematically neglected in the literature on cost functions, natural monopoly, and regulation in transportation; its relevance comes from the very nature of transportation as a spatial phenomenon, and properly accounting for it may drastically change policy conclusions.

In summary, the $C(Y_{ij}^{kt})$ formulation of a transportation cost function is consistent with the operational aspects of transportation systems, incorporating time, commodity and spatial dimensions. It explicitly identifies the kind of information that is lost due to aggregation, and recognizes the inclusion of geographical, operational and/or commodity and time related variables, as surrogates to account for aggregation over different dimensions. This formulation rescues the original meaning of marginal cost and its first-best price interpretation. As an extremely relevant point, it allows for the analysis of economies of time scope, commodity scope, and most importantly,

---

69/ In addition, flow components do not vary proportionally across observations (either cross section or time series). Therefore, conclusions in terms of returns to scale from ton-miles or related output treatments do not necessarily represent ray behavior.

spatial scope, which is essential to the study of natural monopoly, merging and regulation; production complementarity is placed as a key aspect in policy analysis, in addition to scale economies related to proportional output expansions. In spite of the operational difficulties clearly associated with the estimation of such transportation cost functions, the amount of insight it provides is enormous.
In addition, the theoretical considerations that flow from the very concept of $C(Y_{ij}^{kt})$ heavily influence policy analysis and conclusions, like the property of product-specific fixed costs pointed out in section 3.4.2 which permit consideration of the variable part only of the multiproduct cost function when studying subadditivity.
Finally, this approach is consistent with demand analysis in terms of output dimensions, and opens the door to the study of general equilibrium in transportation.

3.4.4 From Engineering Models to Cost Functions

Specific techniques and sometimes very complex analytical tools are being used by engineers to solve the problem of how to provide capacity to be able to produce a certain flow pattern in systems of increasing complexity; mathematical programming, flow theory, graph theory, queueing theory, etc., are among the tools that are being applied in search of solutions to a variety of problems related to transportation functions (fleet assignment to routes, network design, layout of terminals, scheduling, etc.). A question arises in terms of using engineering models (when available) to actually generate "observations" by inputing different sets of flows and inputs prices, obtaining the associated costs. Usually these models are only operational and assume some factors as given, e.g., physical network, thus generating short-run observations. Although this procedure should not be dismissed as a possibility for policy analysis (particularly in proposed systems), there are some reasons to prefer actually observed data to estimate cost functions.

Let us assume there is a true, ideal transformation function generated by a hidden optimal transportation function plus other technical relations. Let us call it $T'(X,Y) = 0$. Engineers search for the optimal way to combine the elements of a transportation system in order to generate a set of flows Y. Their analytical capability would ideally lead to $T'(X,Y) = 0$, but in general the limitations posed by available (although advanced) analytical tools will produce a $T''(X,Y) = 0$, for instance a model, normally operationally oriented. At the next level, transportation managers and operators will try to implement what technical analysis indicates as the best to do. As things never work as planned, and many resource requirements are not explicitly accounted for in operational models, the actual combination of inputs and outputs that results from the whole firm's activity will generate a $T'''(X,Y) = 0$. We may say, thus, that analytical and managerial aspects are part of the technical problem, and do enter $T'''(X,Y) = 0$ in addition to the basic technology available. We can conclude then that a transportation cost function is the actual result of minimizing expenditures within the context of the optimal available ways of combining resources generated by engineering capabilities, with the aim of producing a given flow pattern. Naturally, our proposed economic approach is consistent with this view.

3.4.5 Desirable Conditions and Specification

The actual estimation of a transportation cost function from observed data under the proposed output treatment, involves a series of apsects that should be taken into account. The first one is output itself. A fully disaggregated specification of Y by commodities, periods, and O-D pairs may generate (and generally will) a huge number of parameters to estimate under any reasonable functional specification. Data limitations (form and quantity) are going to play an important role in any attempt to obtain relevant conclusions or inferences from econometrically estimated functions. The conflict arises, then, between feasible estimation of some form of cost function through appropriate aggregation of output, and the degree of relevance of the results in terms of policy conclusions. At this point we should stress the fact that the trend in econometrically estimated transportation cost functions has been to accept the ton-miles-per-unit-time concept as a basic descriptor of output, adding other variables to improve such description; it has never been the case that explicit aggregation to make estimation feasible was performed. In other words, a top-down instead of a bottom-up procedure has been followed. The limitations of inferences from estimated cost functions comes neatly to the surface when the (implicit or explicit) aggregation involved is explicitly recognized.

A second important aspect is the exogeneity of output. This has caused some degree of confusion to a certain extent in the analysis of size. The exogeneity assumption which is implicit in the cost-minimizing behavior, implies that the firm has a priori knowledge of the flow pattern it has to produce, a pattern that is invariant with respect to the actual way it is produced. If in fact demand changes due to operational aspects of the transportation system, output is not exogenous and the appropriate analytical treatment to estimate a cost function should include the demand aspect (e.g., through a system of equations or instrumental variables). A surrogate to this procedure is to specify the flow pattern Y in terms of firm capacity, but this causes ambiguity in many cases.

Input prices are assumed exogenous in the cost-minimizing context. This means that the firm's purchasing power in each factor market is relatively small, i.e. the firm has no monopoly power when buying inputs. If input prices vary across observations, they should be included in the cost specification, i.e., $C(w,Y)$, as theory indicates. We have seen that usually input prices have been calculated in an ad-hoc way, e.g., total expenditure in some item divided by a measure of the associated activity, or price indexes, etc. When factor price variability is present and output is truly exogenous, one of the basic properties of the cost function, Shephard's Lemma, can be applied in order to improve coefficient estimation. Shephard's Lemma basically states that the partial derivative of $C(w,Y)$ with respect to the i^{th} factor price w_i yields the conditional factor demand X_i.[70/]

With respect to exogenously determined technical or operational parameters, they should theoretically enter the cost formulation. This is a somewhat specialized aspect which requires careful analysis

---

70/ In other words, the cost-minimizing input vector $X^* = \{X_i^*\}$ equals $\{\partial C(w,Y)/\partial w_i\}$.

in each particular case; it is generally risky to classify an operational parameter as exogenous unless explicit rules have been established with respect to the particular mode or firm under study. It may well be argued that the estimation process itself will throw some light in this respect, as done by Keeler (1974) in the case of the track price (which was in fact part of his results).

Probably the less "objective" part of the process of cost function estimation is the specification of an actual functional form. In this sense, it should be remembered that we are specifying a proper multioutput function which not only is consistent with the basic technological analysis, but which also allows for the study of production complementarity in addition to scale effects. Therefore we do not want to specify functional forms which destroy these aspects. For instance, the linear in outputs form implies no interaction among output components, i.e., $\partial C(w,Y)/\partial Y_i \partial Y_j = 0$. Our main interest is in analyzing economies of scale and scope, and natural monopoly. It is clear that we do not want to impose through the functional form any a priori restrictions from the perspective of concavity or convexity of the cost function with respect to output components; the specification should be such that the answer arises from the values of the estimated parameters. In other words, we want a nonrestricted Hessian in the sense that its components should flow from the estimation procedure. Naturally, both the quadratic and translog (log quadratic) formulations fulfill this condition. The quadratic in output components form can be looked on as a second order approximation to the true cost function, and generates a constant Hessian which facilitates analysis. Finally, not all specifications with unrestricted Hessian are useful. For instance, a Cobb-Douglas form for variable costs as

$$C = Y_1^{B_1} Y_2^{B_2} \cdots Y_m^{B_m} \tag{3.58}$$

can be shown to lead to

$$\frac{\partial C}{\partial Y_i} = \frac{B_i}{Y_i} C \quad , \quad \text{and} \tag{3.59}$$

$$\frac{\partial C}{\partial Y_i \partial Y_j} = \frac{B_i B_j}{Y_i Y_j} C \quad , \forall i \neq j \quad . \tag{3.60}$$

In general, marginal cost (3.57) should be non-negative at any value of Y; a negative value of (3.60), i.e., weak cost complementarity among Y_i and Y_j, would violate this condition because it would require some B_i to be negative. Thus, the Hessian may result in negative components, but that would be inconsistent with a priori beliefs about the production process. Both the quadratic and translog forms present no problem in this respect. The product-specific fixed cost property of transportation systems which require physical networks (e.g., railways)[71/] makes the quadratic form even more attractive because discontinuities at $Y_{ij} = 0$ would be very difficult to treat,[72/] and a second order approximation to the variable part $C_1(Y)$ seems extremely reasonable.

[71/] See section 3.4.2.

[72/] In fact, this advantage is true for every continuous form of C(Y).

CHAPTER 4. APPLIED MULTIPRODUCT TRANSPORTATION ANALYSIS: AN EXAMPLE ON RAILROAD OPERATIONS

In this chapter we apply the framework built in Chapter 3 to the analysis of railroad operations through the estimation of cost functions. Two short-line railroads operating over simple origin-destination networks (nevertheless generating 4 and 6 output components) are separately studied, using time-series data composed of monthly observations along 5 consecutive years. Aggregation over time and commodities is justified on empirical grounds, while full spatial disaggregation is preserved. Although the analysis concentrates on operational costs, comparison with the results obtained from a similar specification using a single-fully aggregated-output, leads to serious discrepancies in terms of estimated degrees of returns to scale, in addition to the apparent loss of insights in terms of spatial complementarity and origin-destination-specific analysis.

Section 1 describes the conditions under which both railroads operate, supporting cost minimization as the most appropriate description of firms' behavior. A restricted cost system is proposed and justified for actual estimation. Sections 2 and 3 present each specific case, from a description of the physical system to the presentation and analysis of results, which are further discussed in section 4. Emphasis is placed on the methodology followed, and on the comparison of results with the aggregate approach.

4.1 The Transportation System, Data, and Cost Function Specification

The econometric estimation of a transportation cost function following the approach described and justified in the previous chapter, requires the collection of data in a fairly disaggregate way. Empirical investigation of the misspecification caused by the aggregate treatment of output, in terms of scale (ray-related) economies and production complementarity, requires a "clean" set of data if relevant points are to be established. On practical grounds, this implies that we would like to analyze a transportation system that moves a limited set of commodities over few relevant periods on a relatively simple network, with flow components truly exogenous to the firms.

Following these lines, short-line railroad operations seemed to be an adequate example to be developed on empirical grounds.[73/] The operations of these kinds of railroads follow a relatively simple general pattern; they usually connect with one or more major lines at a certain point ("point of connection"), and they deliver freight carried by those lines to their final destination, and/or they carry freight from its origins to the point of connection. In other words, the origin-destination system of a short-line railroad includes the connecting point, and points of final delivery and initial (generic) origins.

[73/] Two other cases were identified as potentially appropriate for empirical analysis under this perspective: passenger intercity bus services in a developing country, and international shipping. Unfortunately we did not succeed in getting data from these sources.

The two railroads studied here have some common characteristics in terms of their operations and general behavior. A first important aspect is that they move freight over an origin-desination system that has few O-D pairs, and flows are completely exogenous to the firm on a daily basis, i.e., they have to carry all freight needed to be transported from and to the connecting point, at given pre-specified rates. From a cost function perspective, this makes output exogenous. Secondly, because of firm size, their labor is non-unionized; this does not translate in labor adjustment to output requirements. On the contrary, this makes the firms keep a fixed number of workers which is below the requirements of peak periods. The rationale behind this behavior is that they maintain a labor force which can perform all kinds of jobs; thus, in periods of low traffic, labor is assigned to do track and equipment maintenance, while in periods of peak activity people just have to work a little harder. Summarizing, firms believe it is in their interest to keep a constant non-specialized labor force which is able to perform all kinds of jobs. Naturally, this is institutionally feasible due to the absence of union requirements. As a corollary, monthly maintenance activity is unrelated to monthly traffic; they just do maintenance when they can. A third important characteristic relates to equipment used; both firms do not own but rent cars from the major line (or lines) to which they are connected. However, they do own locomotives. Finally, the physical network that connects the O-D system has been kept constant throughout the whole period being studied.

Data were obtained on a monthly basis, directly from the firms' records for a period of five and a half years (1975-1980). Monthly costs were gathered grouped in several items: car rental (per-diem), maintenance material, fuel expenses, other material, maintenance labor, other labor, and payments due to usage of joint facilities and TOFC services (when applicable). From the analysis of these data, based on the characteristics described in the preceding paragraph, it was clearly established that labor expenses present no relevant variation in real terms across observations, and that expenditures on maintenance and other materials is unrelated to traffic.[74/] Thus, we postulated as a maintained hypothesis that both labor and materials only contributed as a fixed portion to the total operating costs. It was judged that their inclusion in any cost function specification would only contribute to create "noise" in the analysis. Flow data were gathered as (monthly) O-D specific movements, implicitly aggregating over commodities and time. Time aggregation, i.e., monthly averages of daily movements, was exogenously imposed by the form of the data that were available. Commodity aggregation was decided upon the relatively homogeneous traffic mix by O-D pair; in other words, spatial disaggregation

---

[74/] In fact, annual monthly averages on these items were practically constant, while monthly observations presented huge variations. This suggested that a maintenance cycle of one year could be postulated, thus assigning to each month a fixed amount throughout the analyzed period. In short, annual maintenance is related to annual traffic, and variation of this latter is not enough to cause variation in the former.

also accounts for type of commodity carried.

Under the conditions already described, a restricted operating cost function was formulated, including as an argument (in addition to flows) the only factor price that presented variation across observations, i.e., fuel price. It is worthwhile stressing the fact that physical amounts of labor and materials could not be included as playing their role of fixed factors in the operating cost function, because labor presented no variation across observations, and something similar could be stated in terms of materials when proper assignment to months of maintenance and other expenses was performed (see footnote 74). The restricted operating cost function, then, has the form

$$C_R = C_0 - \sum_j w_j x_j = C_R(Y_{01}, Y_{10}, \ldots, Y_{0n}, Y_{no}, w_F) \ , \qquad (4.1)$$

where j stands for factors judged to contribute only to the fixed part of operating costs C_0, Y_{ij} represents flow from origin i to destination j in tons per month, w_F is fuel price and C_R is the sum of expenditures on fuel,[75] car rental, usage of joint facilities and TOFC operations (when applicable). The application of Shephard's Lemma to (4.1) generates a second equation, i.e., firms' fuel demand, F, which generically corresponds to

$$\frac{\partial C_R}{\partial w_F} = F(Y_{01}, Y_{10}, \ldots, Y_{0n}, Y_{n0}, w_F) \ . \qquad (4.2)$$

Given the type of data available, the restricted cost system formed by (4.1) and (4.2) was restated by putting

[75] Monthly fuel expenditure was calculated from observed amounts and dates of purchase, using a simple inventory model to perform allocation.

$$\frac{\partial C_R}{\partial w_F} = \frac{1}{P_0} \frac{\partial C_R}{\partial (P_F/P)}$$

where P_F is the fuel price index, P is the general price index,[76/] and P_0 is the actual price of fuel in the base year (1967). Then, the deflated observed fuel expenditure is given by

$$C_F = P_0 \frac{P_F}{P} F = P_0 \frac{P_F}{P} [\frac{1}{P_0} \frac{\partial C_R}{\partial (P_F/P)}] = \frac{P_F}{P} \frac{\partial C_R}{\partial (P_F/P)} \quad . \qquad (4.5)$$

C_R was specified in quadratic form around the mean values of flows and price index, using Y_{ij}'s and P_F/P as arguments. This (fairly flexible) form was chosen because of its attractive interpretation as a second order approximation (Taylor's expansion) around the mean flows and price to any functional form underlying C_R, thus providing information on both marginal costs and curvature. A second reason for this choice was the straight interpretation of estimated coefficients in terms of marginal cost, production complementarity (product interaction terms) and price effects at the point of approximation.[77/] Formally, we formulate

---

[76/] In fact, we used the producers' price index. Of course, $w_F = P_0 P_F/P$.

[77/] An alternative to this is the translog form, which is adequate to visualize elasticities of all kinds. However, second order properties of the translog approximation are not that nice. We actually used the translog form in both cases, and in both the results were worse than the quadratic.

$$C_R = A_0 + \sum_{i=1}^{k} A_i(Y_i - \overline{Y}_i) + \sum_{i=1}^{k} A_{ii}(Y_i - \overline{Y}_i)^2 +$$

$$+ 1/2 \sum_{i=1}^{k} \sum_{j \neq i}^{k} A_{ij}(Y_i - \overline{Y}_i)(Y_j - \overline{Y}_j) + A_P(I_F - \overline{I}_F) +$$

$$+ A_{PP}(I_F - \overline{I}_F)^2 + \sum_{i=1}^{k} A_{iP}(I_F - \overline{I}_F)(Y_i - \overline{Y}_i) + \varepsilon \ , \tag{4.6}$$

where k is the number of O-D pairs, Y_i is the corresponding O-D flow, I_F is the fuel price index in real terms (P_F/P), and ε is the error term. Naturally, $A_{ij} = A_{ji}$. The associated fuel expenditure equation (see (4.5)) is

$$C_F = I_F[A_P + 2A_{PP} + \sum_{i=1}^{k} A_{iP}(Y_i - \overline{Y}_i)] + \mu \ . \tag{4.7}$$

The restricted system (4.6) — (4.7) forms the basis for the estimation of C_R,[78] which was carried out using Zellner's seemingly unrelated equations procedure, implemented in TROLL-GREMLIN. Besides symmetry of A_{ij}, no additional restrictions are imposed on the parameters. However, it should be noted that, at the mean values of the arguments (point of approximation),

---

[78] Estimates from the system are more efficient than those from the single equation (4.6). The intuitive explanation is that the derived equation (4.7) "adds" observations, through the use of a component of C_R, namely C_F.

$$C_R = A_0 \tag{4.8}$$

$$\frac{\partial C_R}{\partial Y_i} = A_i \qquad i = 1,\ldots,k \tag{4.9}$$

$$\frac{\partial^2 C_R}{\partial Y_i^2} = 2A_{ii} \qquad i = 1,\ldots,k \tag{4.10}$$

$$\frac{\partial C_R}{\partial Y_i \partial Y_j} = A_{ij} \qquad i = 1,\ldots,k; \quad j = 1,\ldots,k; \; j \neq i \tag{4.11}$$

$$\frac{\partial C_R}{\partial Y_i \partial I_F} = A_{iP} \qquad i = 1,\ldots,k \tag{4.12}$$

$$\frac{\partial C_R}{\partial I_F} = A_P \tag{4.13}$$

$$\frac{\partial C_R}{\partial I_F^2} = 2A_{PP} \ . \tag{4.14}$$

Therefore, we have a priori expectations in terms of the signs of the coefficients. For instance, concavity in prices (a property that any cost function should have) implies $A_{PP} \leq 0$. We also expect non-negative marginal costs ($A_i \geq 0$), non-negative fuel price effect ($A_P \geq 0$), and non-negative effect of price variation on product-specific marginal costs ($A_{iP} \geq 0$).

Finally, the fact that C_R is a restricted operating cost function indicates that the analysis of scale and scope economies, and ultimately of natural monopoly, should be understood in association with these activities. That is to say, as fixed factors are constant in amount all through the analyzed period (i.e., labor, track, locomotives), C_R involves fuel and car rental expenses (line-haul associated), plus terminal operations.

Therefore, C_R can be associated with the transportation function of the system, as defined in Chapter 3.

4.2 Case I: Description, Results and Analysis

Short-line railroad I operates in a 4 origin-destination pairs system, as indicated in Figure 4.1.a. The physical network corresponding to that system, looks like Figure 4.1.b, where the arrow indicates increasing grade. Node a represents the point of connection with the major line, and nodes b and c are stations. d_{ij} indicates the distance between nodes i and j in miles. Let us define

Y_1 : monthly flow from a to b

Y_2 : monthly flow from b to a

Y_3 : monthly flow from a to c

Y_4 : monthly flow from c to a

Y_1 and Y_3 are movements associated with the same product A, with periods of high and low activity (but <u>not</u> seasonal). Y_2 and Y_4 are both movements associated with the same two kinds of products B and C, in a nearly constant proportion. These allow for the treatment of O-D specific flows Y_i as also commodity-specific without causing too much ambiguity.

C_R includes fuel expenses, per-diem, and operations at the nodes (usage of joint facilities at a, plus TOFC operations). The mean values and standard deviations of Y_i, C_R, C_F and I_F are shown in Table 4.1. Only 53 observations were available with complete information.

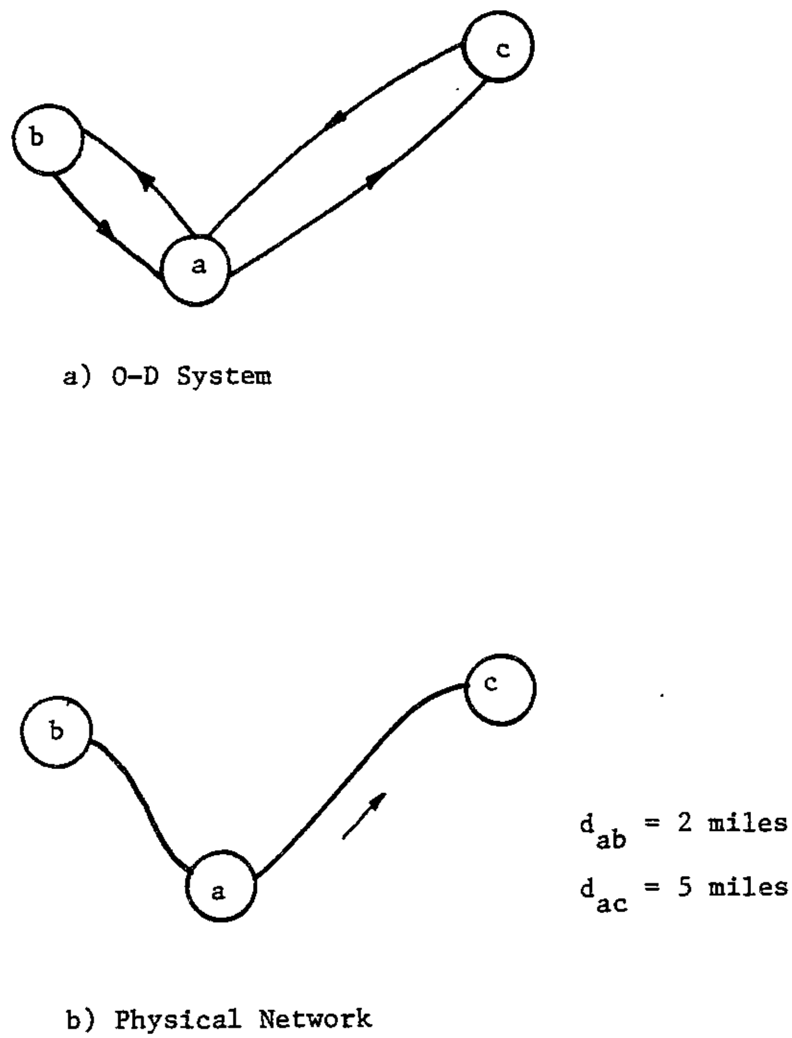


Figure 4.1: Origin-Destination System and Physical Network. Case I.

Table 4.1: Mean Values and Standard Deviations of Flows and Costs.* Case I.

Variable	Mean	Standard Deviation
Y_1	5727.4	6047.2
Y_2	3321.0	1915.1
Y_3	5006.7	3671.3
Y_4	2924.6	1400.8
C_R	3446.0	986.1
C_F	1010.1	486.7
I_F	1.7661	0.3588

*Values in thousand tons and real dollars, respectively.

Equation (4.6) generated 21 parameters to be estimated, 6 of which appear also in equation (4.7). Thus, the system generated around 85 degrees of freedom if we account for the fact that not all "observations" (i.e., eq. (4.7)) involve the whole set of parameters.

The estimated values of the coefficients appear in Table 4.2. The R squared for the cost equation is 0.60 while for fuel demand is 0.30. The Durbin-Watson statistic corresponding to the ordinary least squares estimation of the cost equation (preliminary regression in Zellner's procedure) is 1.996, indicating no serial correlation of the errors. The estimated coefficients have the expected signs. The multiproduct degree of scale economies at the point of approximation is given by

$$\hat{S}_M = \frac{A_0}{\sum_{i=1}^{4} A_i \bar{Y}_i} = \frac{3493.09}{1229.97} = 2.84 \;, \qquad (4.15)$$

which indicates (locally) increasing returns in the operation of the system. It should be noticed that the marginal costs associated with different O-D pairs are different, even on a per-mile basis. The highest marginal cost corresponds to the flow associated with the longest distance and unfavorable grade, while the lowest corresponds to the shortest distance, which is intuitively correct. It should be emphasized that the marginal cost $C_i = A_i$ represents the additional cost of moving 1000 tons in O-D pair i, including terminal operations; therefore, a per ton-mile figure at the O-D pair level would also be misleading as foreshadowed in Chapter 3.

Table 4.2: Coefficient Estimates. Case I.

	Parameter	Value		Standard Error	
*	A_0	3493.09		171.977	
*	A_1	0.09874		0.029428	
	A_2	0.001897		0.088334	
*	A_3	0.101451		0.028202	
	A_4	0.051363		0.070143	
	A_{11}	-2.51788	10^{-6}	4.99142	10^{-6}
	A_{22}	-31.17153	10^{-6}	47.41192	10^{-6}
o	A_{33}	7.30029	10^{-6}	6.77333	10^{-6}
	A_{44}	-6.30618	10^{-6}	30.45272	10^{-6}
	A_{12}	15.61140	10^{-6}	27.86983	10^{-6}
	A_{13}	-4.58711	10^{-6}	6.17440	10^{-6}
	A_{14}	10.16084	10^{-6}	16.21719	10^{-6}
	A_{23}	11.33343	10^{-6}	20.52418	10^{-6}
	A_{24}	-4.41613	10^{-6}	60.99980	10^{-6}
	A_{34}	-4.28328	10^{-6}	21.81751	10^{-6}
*	A_P	627.392		33.685	
*	A_{PP}	-315.535		69.946	
	A_{1P}	0.001553		0.009304	
	A_{2P}	0.00659		0.030224	
†	A_{3P}	0.023675		0.01235	
	A_{4P}	0.010354		0.029593	

* significant at 1%

† significant at 6%

o significant at 28%

A very low price elasticity of demand for fuel (η_F) is expected, due to the small degree of substitution between fuel and other inputs (expected but not obtainable from our model). It can be easily shown that an estimate for η_F at the point of approximation is given by

$$\eta_F = 2A_{PP} \frac{\overline{I}_F^2 P_0}{\overline{C}_F} = -0.187 \ , \qquad (4.16)$$

where P_0 = 0.0961 dollars per gallon (1967).

The effect of fuel price on the marginal cost of the different flows can be studied from the estimated values of A_{iP}. Although the only significant one is A_{3P}, the highest values are in accordance with the network configuration, that is, variations in fuel price affect more heavily the marginal cost of flows associated with long distances, particularly that with unfavorable grade.

The estimated Hessian of C(Y) at the point of approximation is given by

$$\hat{H} = \begin{bmatrix} -5.03 & 15.61 & -4.59 & 10.16 \\ 15.61 & -62.34 & 11.33 & -4.42 \\ -4.59 & 11.33 & 14.6 & -4.28 \\ 10.16 & -4.42 & -4.28 & -12.6 \end{bmatrix} 10^{-6} \quad . \qquad (4.17)$$

The sign and relative magnitude of the diagonal terms of $\hat{H}$ deserve some comments. If we view $C_R(Y)$ as a function of Y_i keeping all other flows constant, we may expect a one-output-like behavior, i.e., costs increasing with Y_i at a decreasing rate up to a certain point (concavity in Y_i) and increasing thereafter (convexity), as in the cost curves in the elementary textbooks. In our case, C(Y) presents concavity in both flows associated with short haul movements at the point of approximation, but $C_{22} < C_{11}$ as expected from $\overline{Y}_2 < \overline{Y}_1$. Similarly in the long haul, $C_{44} < C_{33}$ with $\overline{Y}_4 < \overline{Y}_3$.

Nearly all the estimated elements of the Hessian are highly insignificant, with the exception of C_{33}. Naturally, this makes any inference on complementarity and transray convexity highly uncertain.[79] In spite of this, we can accept $\hat{H}$ as the best approximation to the true Hessian for our restricted operating cost function C(Y), in order to carry out the analysis on the presence of transray convexity as described in Chapter 1. First, we note the presence of weak cost complementarity between Y_1 and Y_3, between Y_2 and Y_4, and between Y_3 and Y_4. Secondly, we note that $\hat{H}$ is not positive definite, by inspection of the signs of the diagonal terms; therefore C(Y) is not convex. Thirdly, the analysis by output pairs indicate that the conditions for transray convexity fail in all cases where C_{ii} and C_{ij} are negative in (4.17), i.e., (Y_1,Y_2), (Y_1,Y_4) and (Y_2,Y_4). Therefore, there is no transray hyperplane

$$\sum_{i=1}^{4} w_i Y_i = w, \quad w_i > 0, \; w > 0,$$

such that C_R is convex along it. This does not preclude subadditivity in C_R, because increasing (multioutput) returns to scale and transray convexity are only sufficient conditions. C_R may still be subadditive

[79] Actually we could not reject the hypothesis of all the elements of H being 0 at any sensible level with an F test.

if scale economies are sufficiently strong, as stated in Baumol (1977), which is in fact the case as shown by $\hat{S}_M = 2.84$. Secondly, in doing this analysis we are implicitly accepting that the properties of the estimated C_R, valid at the point of approximation, hold for the relevant range of outputs. In this respect, we have a priori expectations in terms of firm's behavior at low levels of output; given the exogenous nature of output on a daily basis, the firm should be permanently prepared to produce any required flow, particularly in terms of available fuel. In other words, we expect some fixed costs to appear in C_R. Actually the estimated C_R gives $\hat{C}_R(0) = 2384$, which looks higher than expected but reinforces the idea of decreasing ray average costs suggested by $\hat{S}_M > 1$ if we accept the usual one-output-like behavior of $C_R(Y)$ along a ray (see Figure 4.2).[80/] Our result is consistent with Harris (1977), who found that economies of traffic density are due to high fixed operating costs per mile of road, rather than to capital costs.

It is interesting to calculate some product-specific degree of returns to scale (S_i) at the point of approximation. S_i is given by the ratio of incremental costs (IC_i) to (product-specific) revenues from marginal cost pricing. Equivalently, S_i is the ratio of average incremental cost to marginal cost. We obtain

---

[80/] Intuitively, concavity along a ray through $\{\bar{Y}_i\}$ becoming stronger when approaching the origin, is consistent with $\hat{C}_R(0)$ overestimating $C_R(0)$.

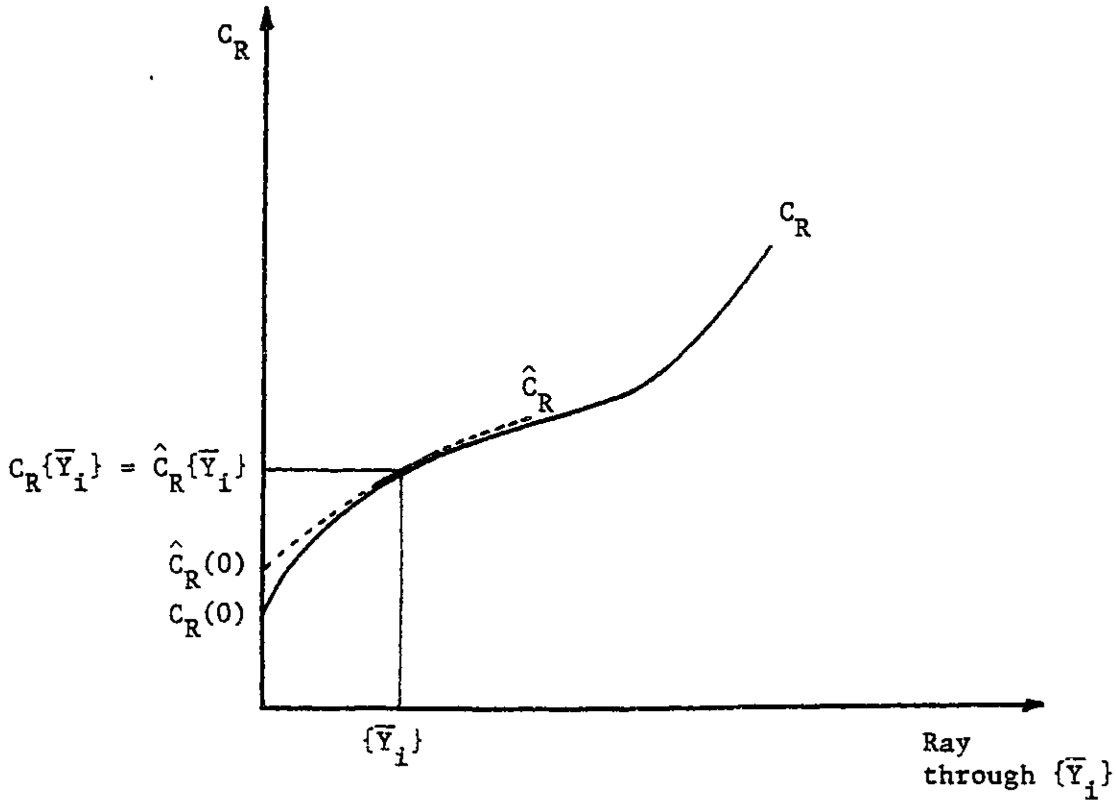


Figure 4.2: Ray Behavior of $\hat{C}_R$ and C_R

$$\hat{S}_3 = \frac{IC_3}{A_3\overline{Y}_3} = \frac{C_R(\overline{Y}_1,\overline{Y}_2,\overline{Y}_3,\overline{Y}_4) - C_R(\overline{Y}_1,\overline{Y}_2,0,\overline{Y}_4)}{A_3\overline{Y}_3} = 0.78 \qquad (4.18)$$

$$\hat{S}_4 = \frac{IC_4}{A_4\overline{Y}_4} = \frac{C_R(Y_1,Y_2,Y_3,Y_4) - C_R(Y_1,Y_2,Y_3,0)}{A_4\overline{Y}_4} = 3.44 \qquad (4.19)$$

(4.18) shows that charging the marginal cost to the flow that goes from the point of connection to station c, would at least cover the additional operating expenses due to the production of that flow <u>in addition</u> to $\overline{Y}_1$, $\overline{Y}_2$ and $\overline{Y}_4$. This result is in accordance with expectations, in the sense that $\overline{Y}_3$ is about 2 million tons greater than $\overline{Y}_4$. Accordingly, the average cost of adding $\overline{Y}_4$ to the firm's activity (already producing $\overline{Y}_1$, $\overline{Y}_2$ and $\overline{Y}_3$), is higher than its marginal cost. It should also be noted that marginal cost of moving $\overline{Y}_3$ from a to c is higher when no movements are made in the opposite direction, than when producing the backhaul. This is

$$\left.\frac{\partial C_R}{\partial Y_3}\right|_{\substack{Y_4 = 0 \\ Y_i \neq 4 = \overline{Y}_i}} = 0.114 > 0.101 = \left.\frac{\partial C_R}{\partial Y_3}\right|_{Y_i = \overline{Y}_i} .$$

Similarly,

$$\left.\frac{\partial C_R}{\partial Y_4}\right|_{\substack{Y_3 = 0 \\ Y_i \neq 3 = \overline{Y}_i}} = 0.073 > 0.051 = \left.\frac{\partial C_R}{\partial Y_4}\right|_{Y_i = \overline{Y}_i} .$$

This is also consistent with expectations, given that the advantages from backhaul operations are in terms of fuel consumption. Within the range of movements involved in Case I, it appears that fuel consumption and similar terminal operations play a positive role in terms of favoring production complementarity, while mixed terminal operations (i.e. loading <u>and</u> unloading) play a negative one. This would explain the signs of the interaction terms C_{ij} $i \neq j$ on the Hessian. As already seen, Y_3 and Y_4 present weak cost complementarity at $\{\overline{Y}_i\}$, which would indicate that the advantage in terms of fuel savings when producing one flow given that the other is being produced, would outweigh the disadvantage arising from loading-unloading at point c. However, the contrary seems to occur between Y_1 and Y_2, where terminal activities are relatively more important given the shorter distance between a and b. Accordingly, $C_{12} > 0$. When fuel plays no role in interaction, complementarity between flows would be determined by the complementarity between the associated terminal operations. Accordingly, $C_{13} < 0$ (only loading at a) and $C_{24} < 0$ (only unloading at b).

A restricted analysis of C_R in terms of Y_3 and Y_4 (the long-haul flows), shows that $C_R(\overline{Y}_1,\overline{Y}_2,Y_3,Y_4,\overline{I}_P)$ presents transray convexity along some plane. In particular, this function is convex along the transray plane (line) $Y_3 + Y_4 = \overline{Y}_3 + \overline{Y}_4$ through $(\overline{Y}_3,\overline{Y}_4)$. This can be checked by analyzing the corresponding estimated Hessian[81/]

[81/] An F test performed on $\beta = [A_{33}\ A_{44}\ A_{34}]$ using the corresponding elements V_β of the variance-covariance matrix, gave $F = \beta V_\beta^{-1} \beta' = 1.37$ for the null hypothesis $\beta = [0\ 0\ 0]$. We could not reject H_0 at a 10-percent level.

$$\hat{H}_{Y3,Y4} = \begin{bmatrix} 14.6 & -4.28 \\ -4.28 & -12.6 \end{bmatrix} 10^{-6} \quad . \tag{4.22}$$

Accepting (4.22) as the best estimate of $H_{Y3,Y4}$ at $\{\overline{Y}_i\}$, the bordered Hessian along $Y_3 + Y_4 = \overline{Y}_3 + \overline{Y}_4$ is positive definite, i.e., C_R is transray convex. However, $[\overline{Y}_3\ \overline{Y}_4]\ \hat{H}\ [\overline{Y}_3\ \overline{Y}_4]'$ is positive, which indicates convexity on a ray direction through $\{\overline{Y}_i\}$. The bi-output version of C_R looks like Figure 4.3. Given ray convexity, we expect (local) decreasing returns to scale specific to Y_3 and Y_4 at $(\overline{Y}_3,\overline{Y}_4)$. In fact,

$$\hat{S}_{(3,4)} = \frac{C_R(\overline{Y}_1,\overline{Y}_2,\overline{Y}_3,\overline{Y}_4) - C_R(\overline{Y}_1,\overline{Y}_2,0,0)}{A_3\overline{Y}_3 + A_4\overline{Y}_4} = 0.90 \quad . \tag{4.23}$$

$\hat{S}_{(3,4)}$ looks well below $\hat{S}_M$. Actually $\hat{S}_{(3,4)} < 1$ indicates that two firms producing $(\overline{Y}_1,\overline{Y}_2,k\overline{Y}_3,k\overline{Y}_4)$ and $[\overline{Y}_1,\overline{Y}_2,(1-k)\overline{Y}_3,(1-k)\overline{Y}_4]$, respectively, where $0 < k < 1$, would be less costly <u>operatively</u> than one firm.[82] But $\hat{S}_M > 1$, which shows the contrary. These apparently contradictory conclusions are in fact explained by the existence of economies of scope. First, note that $\hat{S}_{(1,2)}$ is also less than $\hat{S}_M$,

$$\hat{S}_{(1,2)} = \frac{C_R(\overline{Y}_1,\overline{Y}_2,\overline{Y}_3,\overline{Y}_4) - C(0,0,\overline{Y}_3,\overline{Y}_4)}{A_1\overline{Y}_1 + A_2\overline{Y}_2} = 1.23 \quad . \tag{4.24}$$

---

[82] It should be remembered that the cost of track is not included.

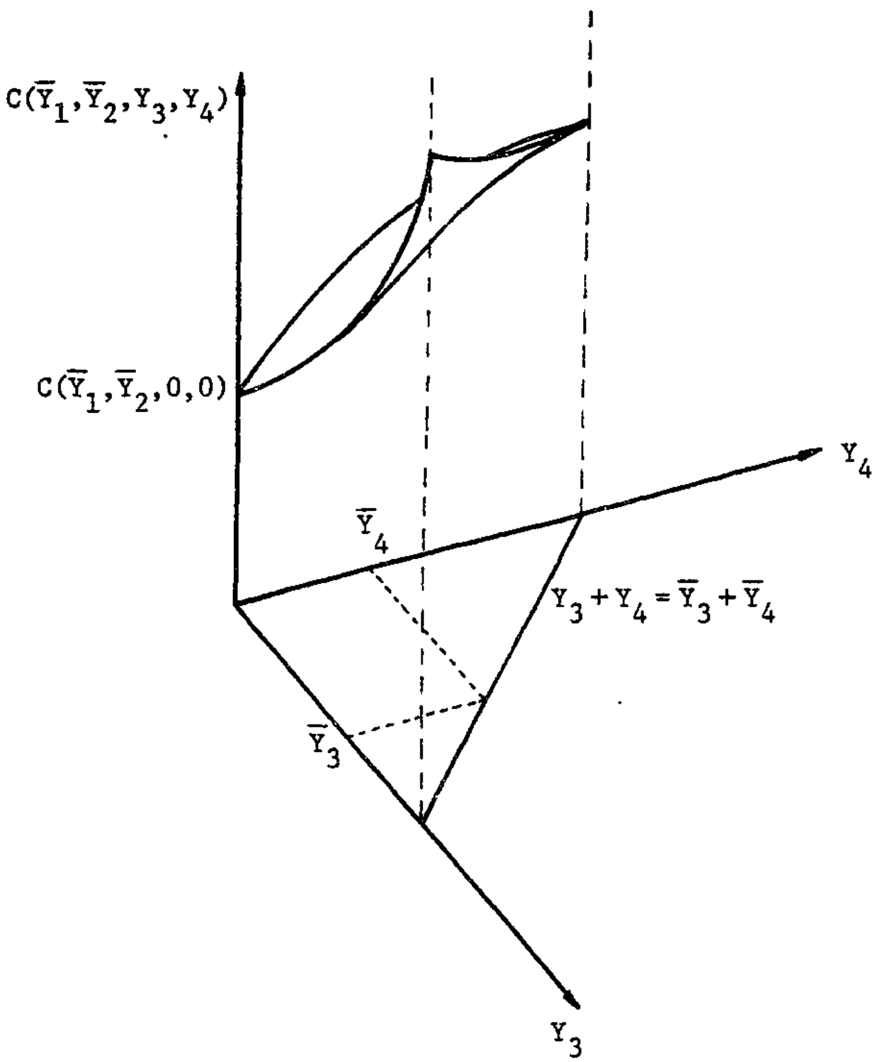


Figure 4.3: Restricted Operating Costs as a Function of Long-Haul Flows in Case I.

The degree of economies of scope relative to (Y_3,Y_4) at $\overline{Y}_i$ is given by[83/]

$$\hat{SC}_{(3,4)} = \frac{C_R(0,0,\overline{Y}_3,\overline{Y}_4) + C_R(\overline{Y}_1,\overline{Y}_2,0,0) - C(\overline{Y}_1,\overline{Y}_2,\overline{Y}_3,\overline{Y}_4)}{C(\overline{Y}_1,\overline{Y}_2,\overline{Y}_3,\overline{Y}_4)} =$$

$$= 0.63 = \hat{SC}_{(1,2)} \quad . \tag{4.25}$$

The coefficient $\alpha_{(3,4)}$ represents the proportion of revenues corresponding to Y_3 and Y_4, with respect to total revenues, under marginal cost pricing. Its value is, then,

$$\alpha_{(3,4)} = \frac{A_3\overline{Y}_3 + A_4\overline{Y}_4}{\sum_{i=1}^{4} A_i\overline{Y}_i} = 0.535. \tag{4.26}$$

Intuitively, multiproduct scale economies for the output bundle (Y_1,Y_2,Y_3,Y_4) would be a weighted average of the multiproduct scale economies associated with (Y_1,Y_2) and (Y_3,Y_4), all of them measured at $\{\overline{Y}_i\}$. But if economies of scope are present among the "sub-bundles," overall scale economies are magnified. In fact, applying (1.27),

$$\hat{S}_M = \frac{\alpha_{(3,4)}S_{(3,4)} + (1 - \alpha_{(3,4)})S_{(1,2)}}{1 - SC_{(3,4)}} = 2.84 \quad , \tag{4.27}$$

which explains the higher value of $\hat{S}_M$.

The next step in our example on railroad operations is to analyze the results obtained from complete aggregation of the transportation

[83/] Recall that economies of scope relative to T are present if $SC_T > 0$.

product. First we generate

$$Y_M = \sum_{i=1}^{4} Y_i d_i \quad . \tag{4.28}$$

Next we specify a quadratic form for $C_R(Y_M, I_F)$, around the mean of observations, i.e.,

$$C_R = B_0 + B_1(Y_M - \overline{Y}_M) + B_{11}(Y_M - \overline{Y}_M)^2 + B_P(I_F - \overline{I}_F) + B_{PP}(I_F - \overline{I}_F)^2 + + B_{1P}(Y_M - \overline{Y}_M)(I_F - \overline{I}_F) + \varepsilon \quad . \tag{4.29}$$

The corresponding (derived) fuel expenditure equation is

$$C_F = I_F[B_P + 2B_{PP} + B_{1P}(Y_M - \overline{Y}_M)] + \mu \quad . \tag{4.30}$$

The estimated coefficients are shown in Table 4.3. All of them have the expected signs, and actually show very similar results in terms of the value of C_R at the point of approximation and the price effects. However, when it comes to analyzing the output related coefficients, conclusions are different. It should be remembered that, as stated in Chapter 1, the degree of scale economies in multioutput production is a ray-related measure. In other words, it considers the output bundle as a composite output where the components enter in fixed proportion, i.e., only scale varies. Of course, ray analysis of a multioutput cost function does not require observed output bundles to vary porportionally across observations! The aggregate output Y_M

Table 4.3: Coefficient Estimates from Aggregate Output, Case I

Parameter	Value	Standard Error
B_0	3376.51	114.443
B_1	0.025403	0.00344
B_{11}	0.132 10^{-6}	8.39 10^{-8}
B_P	632.497	33.3531
B_{PP}	-317.228	44.2628
B_{1P}	0.003211	0.00128

is also a composite commodity whose components vary in scale and proportion across observations. Therefore, estimates of scale economies from a multioutput transportation cost function will generally differ from the same estimate obtained from a priori aggregated output. This is indeed the case in the analyzed system, where flow components are very far from varying proportionally from month to month with the exception of Y_1 and Y_2. Accordingly, the estimated degree of scale economies at $\{\overline{Y}_i\}$ from the model using Y_M, is

$$\hat{\hat{S}}_M = \frac{B_0}{B_1 \overline{Y}_M} = 2.30 \quad , \tag{4.31}$$

deviating from $\hat{S}_M$ by about 20 percent. The standard error of $\hat{\hat{S}}_M$ turns out to be 0.3047.[84/] This implies that the (correct) multioutput measure of the degree of scale economies (2.84) falls outside the 70-percent confidence interval of $\hat{\hat{S}}_M$ (i.e., ± one standard error). We have to specify as wide a confidence region as 95 percent to barely include 2.84. This reinforces the theoretical observation that the aggregate treatment of output not only prevents from analyzing complementarity and output specific properties, but also appears as an unreliable approach

---

84/ A Taylor expansion of $\hat{\hat{S}}_M(B_0, B_1)$ around the estimated values of B_0 and B_1 allows one to express the variance of $\hat{\hat{S}}_M$ as a function of the elements of the variance-covariance matrix of (B_0, B_1). It can be shown that the standard error is given by S.E. = $\hat{\hat{S}}_M \sqrt{V(B_0)/B_0^2 + V(B_1)/B_1^2 - 2\,Cov(B_0 B_1)/(B_0 B_1)}$

to analyze scale economies even under conditions of highly homogeneous time and commodity dimensions.[85/] This leaves spatial aggregation as an issue in the estimation of transportation cost functions and in the corresponding analysis of scale, (spatial) scope and natural monopoly in complex settings. In this sense, the theoretical analysis developed in Chapter 3 gave some insight. One possible procedure is to isolate part of the O-D system and to create a "summary" variable to represent that part, asking this sub-system to be somewhat homogeneous in a loose sense. When time-series data over a fixed network are being used, summation of flows over a spatial sub-system involving similar distances appears as a reasonable procedure to analyze the remaining system. Most important, the preceding discussion suggests that aggregation over flows which vary more or less proportionally across observations would actually "simulate" the behavior of that sub-bundle along a ray in the corresponding restricted output space. In our case, only Y_1 and Y_2 move approximately along a ray across observations (0.8 correlation). Let us define

$$Y_A = Y_1 + Y_2 \quad , \tag{4.32}$$

and create a cost system based on a quadratic around the mean specification of C_R, with Y_A, Y_3, Y_4 and I_F as independent variables.

In Chapter 3 we were prevented from getting any conclusion involving the

---

[85/] Differences between $\hat{S}_M$ and $\hat{\hat{S}}_M$ are likely to be higher in more complex settings. It should be noted that the standard errors of the parameters from the aggregate model are smaller than in the disaggregate version. Of course, one can not conclude from this that the estimates in Table 4.3 are "more reliable" or "significant"!

summary (aggregated) variable; comparison with the fully disaggregated estimation should be done in terms of the remaining parameters, as is done in Table 4.4. The partially aggregated system gives reasonably accurate results, particularly in terms of marginal costs and price effects, from which an analysis in terms of Y_3 and Y_4 can actually be carried out. It is sensible to ask whether the estimates of marginal costs from the partially aggregated system (PA) are statistically different from those obtained from the fully disaggregated model (FD). An F test performed on $\hat{\hat{A}}_3$ and $\hat{\hat{A}}_4$ from PA indicated that we could not reject the hypothesis that $(\hat{\hat{A}}_3,\hat{\hat{A}}_4) = (\hat{A}_3,\hat{A}_4)$ at the five-percent level,[86] where the $\hat{A}_i$'s taken as constants come from FD.[87] Moreover, the point estimate of the multiproduct degree of scale economies turns out to be 2.59 from the PA model, with a standard error of 0.4844. Thus, $\hat{S}_M = 2.84$ falls well within the 70% confidence interval.

Finally, it is convenient to stress the fact that C_R is a restricted operating cost function. Under the conditions prevailing in Case I (already described) and given the range of variation of the outputs, a number of fixed cost components should be added to C_R to raise the cost to total costs. In addition to labor, maintenance and overhead, other perhaps more traditional fixed cost items should be added, like track and equipment (locomotives). As all these components are truly fixed across observations and they amount to a relatively high magnitude, the multioutput degree of returns to scale is actually much higher

---

[86] The calculated F is 0.007, while F_0 from tables is 3.15.

[87] In addition, a χ^2 test on the same set of parameters (specification test) indicated that we could not reject the hypothesis that $(\hat{\hat{A}}_3,\hat{\hat{A}}_4)$ is statistically equal to the (random variable) $(\hat{A}_3,\hat{A}_4)$. The calculated statistic is 0.298, while $\chi_0 = 5.99$ (5%).

Table 4.4: Comparison of Coefficient Estimates
(Partial Aggregation)

Parameter	Fully Specified Output Model	Partially Aggregated	% Variation
A_0	3493.09	3400.16	2.7
A_3	0.101451	0.101694	-0.2
A_4	0.051363	0.045753	10.9
A_{33}	$7.30029\ 10^{-6}$	$6.64\ 10^{-6}$	9.0
A_{44}	$-6.30618\ 10^{-6}$	$-7.79\ 10^{-6}$	-23.5
A_{34}	$-4.28328\ 10^{-6}$	$-5.07\ 10^{-6}$	-18.3
A_P	627.392	626.942	0.07
A_{PP}	-315.535	-311.63	1.2
A_{3P}	0.023675	0.023927	-1.1
A_{4P}	0.010354	0.009028	12.8

than our estimated $\hat{S}_M$. However, our estimates of marginal costs, single output concavity or convexity, and inter-output complementarity, do try to capture the corresponding true values, which are not affected by a higher fixed cost. A complete analysis in terms of scope, however, should account for the fact that there are product-bundle-specific fixed costs (e.g., 5 miles of track if either $Y_3 > 0$ or $Y_4 > 0$). The analysis we have developed in this section in fact avoids this aspect based on the properties of cost functions with product-specific fixed costs, described in Chapter 1 and justified in the transportation case in Chapter 3.

4.3. Case II: Description and Results

The origin-destination system and physical network associated with the second short-line railroad is described by Figure 4.4 As usual, node a represents the point of connection (with two major lines), while nodes b, c and d are stations. Let us define

Y_1 : monthly flow from a to b

Y_2 : monthly flow from b to a

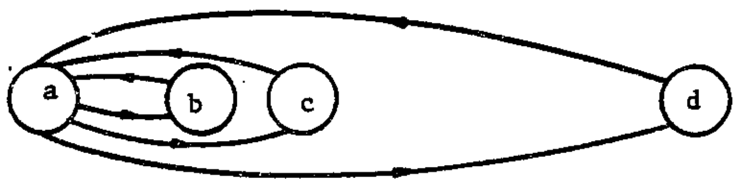
Y_3 : monthly flow from a to c

Y_4 : monthly flow from c to a

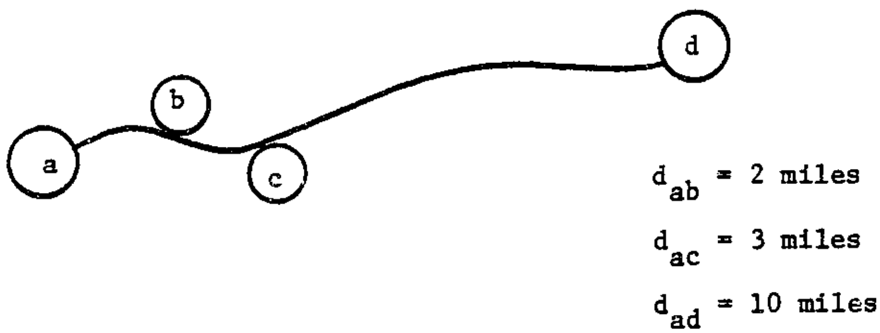
Y_5 : monthly flow from a to d

Y_6 : monthly flow from d to a

Y_1 and Y_2 involve movements of the same type of commodity (although density is much lower for the second), which is also the case for Y_5 and Y_6. Both Y_3 and Y_4 are associated with different combinations of bulk commodities, Again, then, movements are highly homogeneous within each O-D pair.



a) O-D system



b) Physical Network

Figure 4.4: Origin-Destination System and Physical Network, Case II.

C_R now represents expenditure on fuel, car rental and terminal operations (this railroad does not operate TOFC). Table 4.5 shows the mean values and standard deviations for the observed values of C_R, C_F, flows and I_F. The relatively small standard deviations indicate that observations are "nuclearized" around the mean, which will make highly inappropriate any extrapolation using coefficient estimates. The amount of observations with complete information is 68. Equation (4.6) generates now 36 parameters to be estimated, 8 of which also appear in equation (4.7). Roughly, the degrees of freedom generated are 100, although the second equation (fuel expenditure) involves only part of the parameters.

Table 4.6 shows the estimated values of the coefficients, obtained from the system of equations in C_R and C_F. The R squared for the cost equation is 0.50 and for the fuel demand is 0.55. The Durbin-Watson test applied to the ordinary least squares version of the cost equation is inconclusive with respect to serial correlation. 15 coefficients were expected to have some sign a priori, 13 of which agree with expectations. The pair of unexpected signs are related to Y_1 (A_1 and A_{1P}) and only one (A_{1P}) happens to be significant.[88/] At the point of approximation, the multiproduct degree of scale economies can be estimated as

---

[88/] $A_{1P} < 0$ implies than an increase in the price of fuel makes the marginal cost of outbound movements in the short haul diminish. This can actually happen.

Table 4.5: Mean Values and Standard Deviations of Flows and Costs.*

Case II.

Variable	Mean	Standard Deviation
Y_1	1151.78	303.922
Y_2	1335.4	353.973
Y_3	1356.49	493.685
Y_4	9694.09	2356.41
Y_5	4165.91	1933.18
Y_6	313.015	303.516
C_R	1173.58	357.676
C_F	282.87	91.007
I_F	1.93502	0.467127

*Values in thousand tons and real dollars, respectively.

Table 4.6: Coefficient Estimates. Case II.

	Parameter	Value	Standard Error
*	A_0	1241.5	69.386
	A_1	-0.02679	0.1488
o	A_2	0.268205	0.1250
	A_3	0.00327	0.0780
†	A_4	0.027138	0.01793
*	A_5	0.077037	0.0226
	A_6	0.206362	0.1930
o	A_{11}	$-927 \quad 10^{-6}$	$394 \quad 10^{-6}$
	A_{22}	$-272 \quad 10^{-6}$	$327 \quad 10^{-6}$
	A_{33}	$46.565 \quad 10^{-6}$	$145 \quad 10^{-6}$
†	A_{44}	$7.652 \quad 10^{-6}$	$5.28 \quad 10^{-6}$
	A_{55}	$-3.587 \quad 10^{-6}$	$7.53 \quad 10^{-6}$
	A_{66}	$68.149 \quad 10^{-6}$	$420 \quad 10^{-6}$
	A_{12}	$6.141 \quad 10^{-6}$	$447 \quad 10^{-6}$
	A_{13}	$-265 \quad 10^{-6}$	$225 \quad 10^{-6}$
	A_{14}	$30.648 \quad 10^{-6}$	$52.086 \quad 10^{-6}$
	A_{15}	$112 \quad 10^{-6}$	$106 \quad 10^{-6}$
	A_{16}	$305 \quad 10^{-6}$	$657 \quad 10^{-6}$
	A_{23}	$-137 \quad 10^{-6}$	$281 \quad 10^{-6}$
	A_{24}	$40.309 \quad 10^{-6}$	$64.861 \quad 10^{-6}$
†	A_{25}	$-102 \quad 10^{-6}$	$81.503 \quad 10^{-6}$
	A_{26}	$36.641 \quad 10^{-6}$	$370 \quad 10^{-6}$
†	A_{34}	$-48.74 \quad 10^{-6}$	$37.377 \quad 10^{-6}$
	A_{35}	$-23.623 \quad 10^{-6}$	$42.937 \quad 10^{-6}$

Table 4.6, continued

	A_{36}	395	10^{-6}	335	10^{-6}
	A_{45}	-6.582	10^{-6}	11.80	10^{-6}
	A_{46}	20.464	10^{-6}	68.83	10^{-6}
†	A_{56}	-122	10^{-6}	96.36	10^{-6}
*	A_P	148.62		4.1427	
†	A_{PP}	-6.59708		5.37475	
	A_{1P}	-0.051127		0.015101	
o	A_{2P}	0.027818		0.012898	
*	A_{3P}	0.027482		0.009491	
	A_{4P}	0.001595		0.001816	
*	A_{5P}	0.009071		0.002168	
	A_{6P}	0.014236		0.018693	

* significant at 1%

o significant at 5%

† significant at 20%

$$\hat{S}_M = \frac{A_0}{\sum_{i=1}^{6} A_i \bar{Y}_i} = 1.27 \quad , \tag{4.33}$$

i.e., local increasing returns are present in the operation of the system. Again, marginal costs vary across O-D pairs, and the price elasticity of demand for fuel is very small (practically inelastic demand) as expected. The estimate for this latter at the point of approximation is

$$\eta_F = -0.017 \quad . \tag{4.34}$$

The low density of the commodity in Y_2 helps explain the high marginal cost and high effect of fuel price on marginal cost, in spite of the short distance. In general, the interpretation of the estimated coefficients requires an analysis in terms of both level of outputs, and the spatial interrelations. For instance, the same argument given in Case I helps in understanding the positive value of C_{12} $(=A_{12})$ and the presence of weak cost complementarity between Y_3 and Y_4, and between Y_5 and Y_6. The estimated Hessian is

$$\hat{H} = \begin{bmatrix} -1854 & 6.141 & -265 & 30.648 & 112 & 305 \\ 6.141 & -544 & -137 & 40.309 & -102 & 36.641 \\ -265 & -137 & 93.13 & -48.74 & -23.623 & 395 \\ 30.648 & 40.309 & -48.75 & 15.304 & -6.582 & 20.464 \\ 112 & -102 & -23.723 & -6.582 & -7.174 & -122 \\ 305 & 36.641 & 395 & 20.464 & -122 & 136.298 \end{bmatrix} 10^{-6} \tag{4.35}$$

$\hat{H}$ is not positive definite, which is to say that $C_R(Y)$ is not convex. In addition, reference to Appendix 1.1 helps show that the bi-output analysis of (Y_1,Y_2) and (Y_1,Y_5) indicates that $C_R(Y)$ is not convex along any transray hyperplane. $\{\overline{Y}_i\}$ H $\{\overline{Y}_i\}'$ is negative, which indicates that $C_R(Y)$ is (locally) ray convex at the point of approximation. However, $\hat{C}_R(0) < 0$, which shows a strange ray behavior. As stated before, the concentration around the mean of observations makes any inference very unreliable.

Finally, the results from the aggregated version of the cost system appear in Table 4.7. Although signs and values of the coefficients appear within expectations, the estimated value of the degree of returns to scale is

$$\hat{\hat{S}}_M = 1.56 \quad , \qquad (4.36)$$

about 25 percent <u>higher</u> than the multioutput counterpart. The standard error of $\hat{\hat{S}}_M$ is 0.3467, which indicates that the (correct) multioutput degree of returns to scale (1.27) barely lies within the 70-percent confidence interval of $\hat{\hat{S}}_M$.

4.4. Some Comments

The results obtained from the application of the multiproduct framework to estimate transportation cost functions deserve some qualifications in terms of their analysis and interpretation.[89/] First, it should be kept in mind that an approximation of C(Y) around a point is accurate as a description in that neighborhood. Accuracy

[89/] Here we will refer mainly to Case I, which presents more intuitively correct results. We will postpone the presentation of methodological and policy conclusions until Chapter 5.

Table 4.7: Coefficient Estimates from Aggregate Output. Case II

Parameter	Value	Standard Error
B_0	1186.93	45.35
B_1	0.009171	0.002097
B_{11}	$-3.2 \quad 10^{-8}$	$6.18 \quad 10^{-8}$
B_P	147.791	4.63
B_{PP}	-11.1698	4.65
B_{1P}	$8.95 \quad 10^{-4}$	$2.03 \quad 10^{-4}$

diminishes when we move away from that point. This becomes a problem particularly when analyzing overall and product-specific fixed costs. In Case I, flow observations are very far from the origin, in spite of some points with zero O-D specific flow. Although it does not seem to be a serious problem in our case, eventual product-specific fixed cost may cause some difficulty. This can be exemplified with a two-output picture as the one in Figure 4.5. There, solid lines represent the true cost function C(Y).[90/] If a continuous (flexible) cost function is specified and observations involve enough pairs like $(Y_1,0)$ and $(0,Y_2)$ at different levels of the non-zero output component, the estimated cost function would look like the dotted-line surface $\hat{C}(Y)$, erroneously indicating transray concavity when in fact C(Y) is tranray convex for $Y_i > 0$. In general, such a shape will be the rule more than the exception in the transportation case, e.g., cost of the right-of-way. Even in our operating example I, $Y_3 = Y_4 = 0$ would imply no movement of equipment on link a-c, while $Y_i \neq 0 (i = 3 \text{ or } 4)$ requires a minimum expenditure equal to fuel consumption necessary to go a-c-a empty. As stated before, this will not be relevant in that example, but will undoubtedly be important in bigger systems even on a purely operational basis.

A second necessary comment is in regard to the inconclusiveness of our tests on subadditivity. We should stress that our C_R is an operating cost function which does not include any (clearly or suspected)

[90/] This form has been named "Transylvanian" in the multioutput literature, due to the bat-like form of a transray-convex function with fixed product-specific costs.

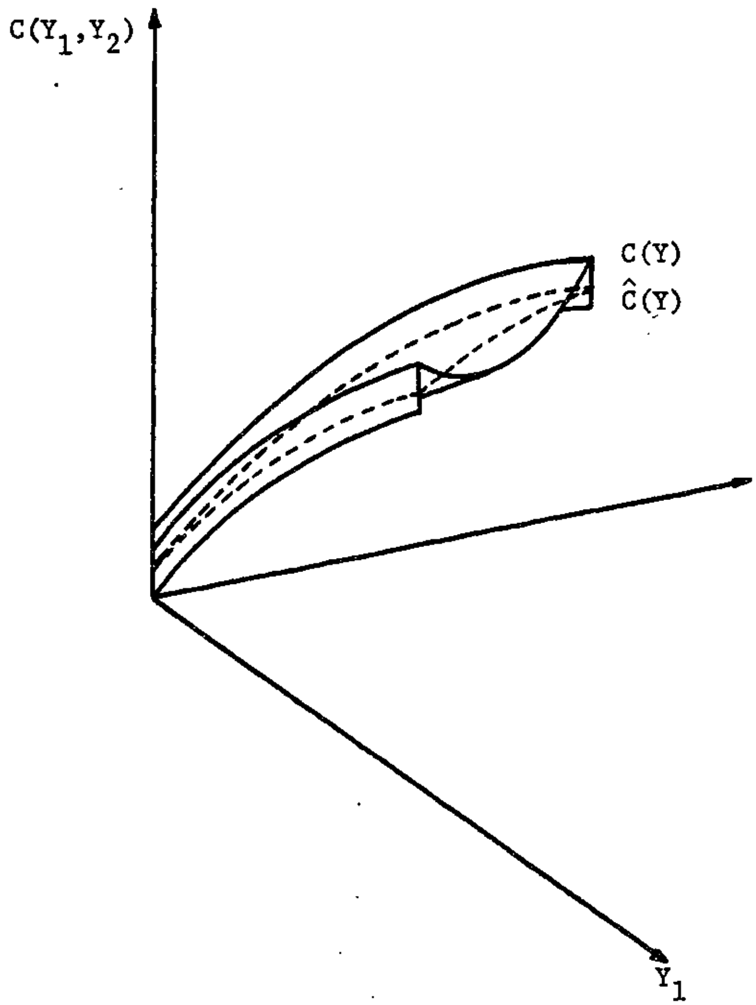


Figure 4.5: Estimation Problems in the Presence of Product-Specific Fixed Costs.

fixed cost information. The level at which fixed costs enter the picture is relatively high.[91] Therefore, the complete cost function is undoubtedly subadditive due to the very high fixed costs.[92] In this sense, our $\hat{S}_M > 1$ should be interpreted as the detection of operating economies of scale over a fixed route structure, or as a restricted version of economies of density which does not include some fixed expenditures. As an important methodological point, deletion of those cost components which are judged to contribute in a fixed amount to C(Y) (either because of data analysis, knowledge of the system, or boundaries of output levels), is a procedure that can be recommended in order to diminish the noise in the estimation of both product-specific characteristics and interproduct cost complementarity.

A third qualification of our results is related to output aggregation. The "forced by circumstances" time aggregation introduces some ambiguity in the interpretation of results due to reasons that follow very closely the analysis of Figure 3.10 in Chapter 3. In short, two identical monthly observations on (Y_1, Y_2, Y_3, Y_4) may be generated by drastically different daily patterns, eventually generating different (operating) costs in turn. The fact that fuel expenses were usually performed one day during the month, only allows for a reasonable allocation of that expenditure on a monthly basis, never

---

[91] For instance, the sum of labor, maintenance material and overhead in Case II adds up to ten times the mean of C_R. This goes even higher when including track and locomotives.

[92] In the long run, product-specific fixed costs are linked to track. At this level of output, decreasing average incremental costs will be present for every output. This plus increasing returns are sufficient for subadditivity.

on a daily basis. Thus, even if daily flows were available, time aggregation would be compulsory. The ambiguity introduced by commodity aggregation is reduced by the associated O-D-commodity type in our case. However, this association holds for the bulk of the movements and is not a hundred percent accurate. In spite of these caveats, we have seen that preserving the full spatial characteristics of output increases enormously the amount of insights on the operation of the system. This leads us to our fourth observation, in relation to a procedure to perform spatial aggregation. Although it is difficult to establish at this point <u>the</u> procedure to aggregate, we can at least point out the following:

i) we would like to add over flow components that vary somewhat porportionally, in order to at least preserve the ray behavior of the aggregated sub-bundle;

ii) we would like to add over flow components involving similar distances; otherwise it would be necessary to distinguish the haul-related cost from the terminal-related cost, following the idea $Y_0 = \Sigma d_i Y_i + \alpha \Sigma Y_i$ proposed in Chapter 3, thus introducing more ambiguity and eventual multicollinearity.

These observations were taken into account in our "experiment" in Case I, which generated results highly consistent with full disaggregation, particularly in terms of first order magnitudes (i.e., marginal cost and price effect) of the non-aggregated sub-bundle. This suggests a procedure to deal with large scale networks through partially aggregated analysis.

Finally, our examples confirm what we stated in Chapter 3 in the sense that the multiproduct approach not only allows for complementarity (scope) analysis, but actually poses the problem of scale economies in the appropriate context, i.e., proper estimation of scale economies as a ray concept involving proportional variations of output, can only be performed from an explicit multioutput form of the cost function. Estimates of scale economies from any a priori aggregation on output will not allow for such a ray analysis unless output components do actually vary proportionally across observations. These considerations make our commodity aggregation acceptable, but make spatial aggregation inappropriate even for the simple network considered. Therefore, the spatial characteristics of the O-D system being served play a central role in transportation cost functions.

CHAPTER 5. RECAPITULATION AND CONCLUSIONS

In this chapter we review the major points established throughout this work from the perspective of an ex-post analysis. The main weaknesses of the available approaches to the analysis of cost functions and scale economies in transportation are restated, and a multioutput formulation of a transportation cost function is advocated as an alternative, both on theoretical and empirical grounds, based upon our results in chapters 2 through 4. The main conclusions from our work are summarized in section 2, while additional comments and directions for future research are formulated in the last section.

5.1 Recapitulation

Many different approaches have been taken to analyze scale economies in transportation by means of somehow estimating a cost function, i.e., a function describing the minimum cost to produce a given transportation output. After a few isolated efforts oriented to actually derive a cost function from information on technical characteristics and input prices (e.g., De Salvo, 1969), a number of studies faced the problem from an econometric perspective, i.e., trying to unveil the underlying relation between cost and output in transportation from the analysis of data corresponding to given transportation systems. Major advances have been made in two dimensions of econometric studies: i) functional specification, ranking from the simplistic linear form to the flexible quadratic and translog forms; and ii) microeconomic basis, where the properties of well defined cost functions have been increasingly incorporated. This has been quite clear in Keeler's railroad study, where a long-run cost function is derived from an estimated short-run one, by optimizing with respect to a fixed factor. Lately, the addition of factor demand equations derived from the (proposed) cost function using Shephard's lemma has generated what is called a cost system.

However, a third aspect has received less attention than those already mentioned: the treatment of transportation output. Systematically, the "units-times-distance" (e.g., ton-miles) concept has been used as the basis for output definition in all studies, including those which incorporate basic technical information instead of econometric techniques ("engineering" studies). In this respect, output has been treated in various forms: straight "units-times-distance" (UTD), UTD plus "quality" and/or geographical variables, UTD by type of commodity, UTD plus "technical" variables, and the hedonic treatment of UTD. In all cases except the latter, the addition of other variables to improve output definition has been done inconsistently, in the sense that conclusions in terms of scale economies have been established by looking only at UTD.

In spite of the use of advanced econometric techniques and elegant microeconomic treatment, available studies present inconsistencies when it comes to the prediction of industry behavior. In particular, both in the airlines and the trucking cases, estimated cost functions (and economic wisdom) support the idea of the presence of constant returns to scale at the relevant level of output. Accordingly, economists have advocated deregulation in these industries. However, "paradoxical" merging has been observed in both industries in the same periods covered by the studies. We believe that what appears to be a contradiction arises due solely to an ambiguous treatment of output.

We have advocated for the use of a vector $Y = \{Y_{ij}^{kt}\}$ as the output of a transportation system, where Y_{ij}^{kt} is the mean flow intensity of commodity k between origin i and destination j during period t, in commodity units per unit time. Accordingly, the cost function should be specified as $C = C(Y)$. Some of the previous studies claim to use a multioutput approach. Actually only those that use UTD by type of commodity can be classified as such; the hedonic output is in fact a single output approach using a scalar value determined by various "characteristics" of that output, while other approaches simply treat as separate variables different dimensions of the same "output" (e.g., Harmatuck's ton-miles, tons, and traffic mix). Under the vector definition, the UTD measure results from aggregation over periods, commodities and space.

For a transportation firm facing an exogenously given transportation output, the choice of a technical optimum (i.e., the generation of a transformation function) is essentially centered around equipment and on the <u>operation</u> of the system. We define the <u>transportation function</u> of a system as the technically optimal relation between output and characteristics of vehicles, terminals, and rights-of-way. Thus, by adding other relations between inputs (usually of the fixed-proportions type), an economic transformation function can be generated. Following these lines, we have developed the transportation function and the corresponding cost function associated with the production of a one-component output (one O-D pair, one commodity, one period), and also for a two-components output (two O-D pairs, one commodity, one period). This analysis, in the same spirit as Vernon Smith's earlier work, has allowed us not only to state the misspecification caused by the UTD formulation of the cost function, but also to show that the analysis of economies of scope, particularly spatial scope, is essential in the economic study of transportation systems, if any relevant policy conclusions in terms of industry structure are to be established. Using the same example, we were able to reconcile the "economists'" view of the trucking industry (constant returns), with the "truckers'" view (merging advantages), by pointing out that both can co-exist due to the presence of economies of spatial scope.

The application of the multiproduct approach to the analysis of the operations of a Class III railroad proved quite useful in the development of a methodology, supporting some previous points, and suggesting new ones. The richness of the approach can be detected only if the corresponding cost function is properly treated. Our usage of the quadratic-around-the-mean specification turned out to be very appropriate in this sense. Such a formulation gives an approximation to the value of the function, the gradient, and the Hessian, for any underlying cost function, at the mean values of the independent variables. From this information, ray and transray analysis can be performed in order to get conclusions on scale and scope. It is apparent that

this type of information is undoubtedly richer than that usually obtained; in addition, it explicitly recognizes the limitations in terms of extrapolation of results beyond reasonable limits. In this context, the multiproduct version of a transportation cost function in our example proved not only very insightful, but also consistent with the technological aspects of transport operations. In other words, $C\{Y_{ij}^{kt}\}$ seems to be the appropriate microeconomic counterpart of the underlying transformation function corresponding to a given origin-desination system. First, we can identify flow-specific marginal costs, which in our example happened to vary sensibly across O-D pairs. Secondly, interaction terms provide information on how flow-specific marginal costs vary when <u>other</u> flows vary. In addition to the implications in terms of subadditivity, this is valuable information for the development of efficient pricing policies. In this respect, our example also helped us show that an estimated multioutput transportation cost function should be treated cautiously when drawing conclusions on product-specific scale economies, and when comparing marginal cost pricing with (product-specific) average incremental costs; the basic reason for this is the possible existence of product-bundle-specific fixed costs, or better, of discontinuities of $C(Y)$ at $Y_i = 0$.
We expected the estimation of the multiproduct degree of returns to scale to differ with respect to the estimation given by the UTD approach, because this latter does not correspond to a ray measure. This was confirmed by our example. Finally, partial aggregation across flows varying somewhat proportionally and involving the same distance gave reasonable results in terms of estimation of marginal costs and

price effects associated with the remaining flows.

5.2 Conclusions

There is a comment systematically made in nearly every paper on scale economies in transportation, which in unified language would read something like "although the units-times-distance (UTD) measure of transportation output is somewhat ambiguous, it is commonly accepted as the basic description of that output, and we will use it." What we have tried to do is to redefine the problem from the start in a very explicit way, in order to overcome ambiguities, and to establish a new perspective from which to focus transportation cost functions in a way that is consistent with the underlying technology, and from which policy conclusions follow unambiguously. In summary, the main conclusions from our critical, theoretical and empirical work can be stated as follows:

i) The multiproduct approach to a transportation cost function, namely $C(Y_{ij}^{kt})$, is more consistent with the underlying technology than previous output measures; this has been established on both theoretical and empirical grounds. We propose it as the microeconomic counterpart of the transformation function corresponding to a transportation system.

ii) Such an approach indicates that transportation should be viewed as a joint production process, where the interrelation among products plays an important role (e.g. production complementarity). In particular, we can distinguish three types of cost complementarity: in terms of commodities, in terms of time, and in terms of space, reflecting the convenience (or inconvenience) of moving different types of goods, of producing services during different periods, or carrying things between different O-D pairs, respectively. By extension, the concepts of economies of commodity scope, time scope, and spatial scope are simultaneously introduced.

iii) As scale economies involve proportional variations of output components, the presence of constant returns is perfectly compatible with merging, due to the presence of some type of complementarity among some products.

iv) The straight UTD approach should be viewed as the result of aggregation over space, time and commodities, across the components of the output vector. Thus, the behavior of the UTD variable does not correspond to movements along a ray in the output space, because in general the components of the output vector will not vary proportionally. In this sense, aggregated analysis can not even give a correct answer in terms of economies of scale.

v) When the transportation output is viewed as a function of a "quality" vector, e.g., the hedonic approach, it is still a single output approach that tries to account for the different dimensions that have been "swallowed" by UTD. It does not reproduce ray behavior in the $\{Y_{ij}^{kt}\}$ space; therefore, measures of scale economies from this approach are still potentially misleading.

vi) The analysis of economies of spatial scope is a key aspect to properly understand scale economies and natural monopoly in transportation. In general, it will be crucial in any process involving networks (e.g., telecommunications).

vii) The multiproduct transportation cost function $C\{Y_{ij}^{kt}\}$ will present product-specific or bundle-specific fixed costs associated with the right-of-way, in those cases where firms pay this item (e.g., railroads). In general, if we express

$$C\ \{Y_{ij}^{kt}\} = F(S) + C_1\{Y_{ij}^{kt}\} \quad , \tag{5.1}$$

where $S = \{\text{set of outputs } Y_{ij}^{kt}/\sum_k \sum_t Y_{ij}^{kt} > 0\}$, then it holds that

$$F(S \cup T) \leq F(S) + F(T).^{93/} \tag{5.2}$$

Property (5.2) allows for an analysis in terms of $C_1\{Y_{ij}^{kt}\}$. If $C_1(Y)$ is subadditive, C(Y) is subadditive (but the former is not a necessary condition). In general, F(S) will also have an operating component in addition to right-of-way expenditures.

viii) In view of the preceding property, the estimation of operating cost functions appears as the relevant part when searching for the existence of interproduct complementarity. Deletion of any kind of fixed costs, either overall or product-bundle-specific, will improve accuracy in the estimation of $C_1(Y)$, in the sense that the results obtained from the specification of C(Y) as a continuous differentiable function will be distorted by actual discontinuities. Any kind of fixed costs can be more appropriately introduced after $C_1(Y)$ has been estimated.

93/ The presence of overall fixed costs only reinforces this property.

ix) Estimation of C(Y) in complex origin-desination systems will require some spatial aggregation. Feasible and appropriate econometric estimation can be done through the isolation of sub-systems of interest, and aggregation of the remaining flows into "summary" flows trying to at least capture the ray behavior of the aggregated bundles. We suggest aggregating over flows involving similar distances, and that vary more or less proportionally across observations, i.e., flows that move along a ray defined on a sub-space of the output space. This procedure will at least allow for a better estimation of the (multioutput) degree of scale economies.

x) The application of the multiproduct framework to the analysis of short-line railroad operations shows that economies of density are far from being exhausted, and that scale economies are present even on a purely operational basis. Cost complementarity between different O-D flows seems to be favored by fuel savings from cyclical operations and by similar terminal activities, while mixed terminal activities seem to act against it. However, economies of spatial scope tend to be present (in the short and long run) due to the existence of product-specific fixed costs. O-D-pair-specific marginal costs differ among each other, depending both on topography and characteristics of the network, as well as on the level at which other flows are being produced. Long-run product-specific returns to scale are clearly present, but not necessarily in the short run (i.e., the average operating cost of adding an O-D-specific flow in addition to the remaining bundle may or may not exceed its marginal cost). However, the existence of economies of spatial scope magnifies sub-bundle-specific scale economies, leading to overall operating scale economies.

5.3 Final Comments and Directions for Research

It is impossible to avoid the temptation of adding some statements which do not necessarily flow from what has been presented so far, but rather correspond to opinions on some aspects of the estimation of transportation cost functions, which have been built as the result of discussions with many persons in the course of this work.

It is our impression that there is some sort of confusion around the concept of scale economies and natural monopoly in transportation, in the sense that demand aspects are usually involved in discussions on this topic; the confusion arises from the inclusion of sustainability as a surrogate for natural monopoly. For instance, we have seen that one firm is the cheapest way to serve the two O-D pair system (backhaul) depicted in Chapter 3; however, somebody may claim (as somebody did) that users will prefer two firms because service will be faster (i.e. frequency will be higher). Implicit in this view of the system is the image of a person arriving to one of the terminals carrying a box of potatoes to deliver, and being asked "what would you prefer, higher or lower frequency?". Naturally, this picture does not correspond to the problem solved by minimizing C(Y), because their output is exogenous; in other words, users want things to be carried from a to b in, say, a day. The firm adapts frequency according to the (exogenous) demand. If demand is (as it is in many cases) dependent on travel time of the trip, of course one firm could be "cut" by a second offering faster service. This is exactly the problem of sustainability, where travel time plays a price role. The estimation of cost functions in cases where output is not exogenous would require demand information, and this has not been the procedure in usual practice, even when this phenomenon of demand-cost interrelationship is likely to be present. Instead, we have detected a trend toward a "synthetic" analysis, including users' perceptions as part of cost functions.

A second comment relates to data availability as an obstacle to correctly analyze subadditivity in some transportation industries. In this respect, data are not necessary either to establish the failure of the aggregate analysis, or to establish the correct way to do it in order to investigate the "deviation from the truth" corresponding to that analysis. The $\{Y_{ij}^{kt}\}$ definition has never been used for actual estimation, but neither has it been used to explain the limitations and actual inaccuracy of other approaches when analyzing industry structure. It is our intuition, based on the correct usage of the $\{Y_{ij}^{kt}\}$ concept, that the old controversy around the "trucking case" is going to be settled by the findings of constant (multioutput) returns to scale and the detection of spatial complementarity, or economies of spatial scope, in some way or another, over limited spatial settings. This will reconcile economic wisdom with industry behavior. Also related to data availability, the problem is not so much that this type of information (flows and costs) is not available in the required form, but instead is a problem of published data and the reluctance of firms to release unpublished data. The fact is that transportation firms do not analyze operations in terms of ton-miles, but in terms of O-D flows, so the required information is there. We believe that the very fact of stating that data should be used in the manner we have advocated in this work should generate action in the direction of actually generating the required information, from both firms and regulatory agencies.

It is always possible to propose as future research to investigate the (unbounded) complement of the existing knowledge. We want to explicitly suggest future research that we believe immediately follows from our work. This research may take many (non-overlapping) directions:

a) Application of the framework developed in this work to estimation of transportation cost functions of complex systems, not only in terms of product definition and multiproduct analysis, but also in terms of bundle-specific fixed costs and spatial aggregation.

b) Development of new forms of aggregation, with a clear interpretation of the resulting output treatment in terms of analysis of scale and/or scope in its various forms.

c) Development of procedures to perform a consistent multioutput analysis from cross section data corresponding to firms operating in different spatial settings.

d) Derivation of analytical transportation functions and their corresponding cost functions for other basic O-D systems, in order to gain insight into the different forms of spatial complementarity.

Finally, let us recall that the concept of C(Y) as the minimum cost of producing a vector of O-D-, commodity-, and period-specific flows, is not unimodal. The minimum cost for particular bundles may well result from a combination of modes. Where to set a limit to the analysis is also a matter of discussion and research.

BIBLIOGRAPHY

Baumol, William J., "Scale Economies, Average Cost, and the Profitability of Marginal Cost Pricing," in Ronald E. Grieson (ed.), Public and Urban Economics (Essays in Honor of W. S. Vickrey), Lexington, Massachusetts: Lexington Books, D. C. Heath & Co., 1976.

Baumol, William J., "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry," American Economic Review, Vol. 67, No. 5, 1977.

Baumol, William J., Elizabeth Bailey and Robert D. Willig, "Weak Invisible Hand Theorems on the Sustainability of Multiproduct Natural Monopoly," American Economic Review, Vol. 67, No. 3, 1977.

Baumol, William J. and Yale M. Braunstein, "Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication," Journal of Political Economy, Vol. 85, No. 5, 1977.

Baumol, William J., John C. Panzar and Robert D. Willig, Unpublished Manuscript, Chapters 2, 3, 4 and 7 (Draft), July 1979.

Braeutigam, Ronald R., Andrew Daughety and Mark Turnquist, "The Estimation of Hybrid Cost Functions for a Railroad Firm," Document 425-07, Northwestern University, April 1980.

Case, Leland, and Lester B. Lave, "Cost Functions for Inland Waterways Transport in the United States," Journal of Transport Economics and Policy, May 1970.

Caves, Douglas W., Laurits R. Christensen and Joseph A. Swanson, "Productivity in U.S. Railroads, 1951-1974," Bell Journal of Economics, Spring 1980.

DeSalvo, Joseph S., "A Process Function for Rail Linehaul Operations," Journal of Transport Economics and Policy, January 1969.

Friedlaender, Ann F., The Dilemma of Frieght Transport Regulation, Washington, D.C.: The Brookings Institution, 1969.

Fuss, Melvyn, Daniel McFadden and Yair Mundlock, "A Survey of Functional Forms in the Economic Analysis of Production," in M. Fuss and D. McFadden (eds.), Production Economics: A Dual Approach to Theory and Application, Amsterdam: North-Holland Press, 1978.

Gálvez, Tristán, "Análisis de Operaciones en Sistemas de Transporte," Publication ST-INV/04/78, Departamento de Obras Civiles, Universidad de Chile, 1978.

Gordon, Steven, and Richard de Neufville, "Design of Air Transportation Networks," Transportation Research, Vol. 7, pp. 207-222, 1973.

Gordon, Steven, and Richard de Neufville,"Rationalization of the European Air Network", Transportation Research, Vol. 11, pp.235-244, 1977.

Griliches, Zvi, "Cost Allocation in Railroad Regulation," Bell Journal of Economics, Spring 1972.

Harmatuck, Donald J., "A Policy-Sensitive Railway Cost Function," The Logistics and Transportation Review, Vol. 15, No. 2, 1979.

Harris, Robert G., "Economies of Traffic Density in the Rail Freight Industry," The Bell Journal of Economics, Autumn 1977.

Hasenkamp, Georg, "A Study of Multiple-Output Production Functions, Klein's Railroad Study Revisited," Journal of Econometrics, Vol. 4, pp. 253-262, 1976.

Hasenkamp, Georg, Specification and Estimation of Multiple-Output Production Functions, Springer-Verlag, 1978.

Jara Díaz, Sergio, "El Problema de Transporte: Análisis y Síntesis," Publication ST-INV/02/76, Departamento de Obras Civiles, Universidad de Chile, 1976.

Keeler, Theodore E., "Railroad Costs, Returns to Scale and Excess Capacity," Review of Economics and Statistics, Vol. 56, pp. 201-208, 1974.

Kneafsey, James T., The Economics of the Transportation Firm, Lexington, Massachusetts: Lexington Books, D. C. Heath & Co., 1974.

Koenker, Roger, "Optimal Scale and the Size Distribution of American Trucking Firms," Journal of Transport Economics and Policy, January 1977.

Koshal, Rojindar K., "Economies of Scale. I, The Cost of Trucking: Econometric Analysis. II, Bus Transport: Some United States Experience," Journal of Transport Economics and Policy, May 1972.

Lee, N. and J. Steedman, "Economies of Scale in Bus Transport. I, Some British Municipal Results," Journal of Transport Economics and Policy, January 1970.

Malinvaud, Edmond, Lecons de Théorie Microéconomique, Dunod, 1969.

Manheim, Marvin L., Fundamentals of Transportation Systems Analysis, Volume 1: Basic Concepts, Cambridge, Massachusetts: MIT Press, 1979.

Manheim, Marvin L., "Understanding 'Supply' in Transportation Systems," Transportation Research, Vol. 14A, pp. 119-135, 1980.

Marsden, James, David Pingry and Andrew Whinston, "Engineering Foundations of Production Functions," _Journal of Economic Theory_, Vol. 9, 1974.

McFadden, Daniel, "Cost, Revenue and Profit Functions," in M. Fuss and D. McFadden (eds.), _Production Economics: A Dual Approach to Theory and Application_, Amsterdam: North-Holladn Press, 1978.

Morlok, Edward K., _Introduction to Transportation Engineering and Planning_, McGraw-Hill, 1978.

Mohring, Herbert, _Transportation Economics_, Ballinger Books, 1976.

Panzar, John C., and Robert D. Willig, "Economies of Scale in Multi-Output Production," _Quarterly Journal of Economics_, Vol. 91, 1977.

Panzar, John C., and Robert D. Willig, "Economies of Scale and Economies of Scope in Multi-Output Production," Bell Laboratories Economic Discussion Paper #33, August 1975.

Panzar, John C., and Robert D. Willig, "Free Entry and the Sustainability of Natural Monopoly," _Bell Journal of Economics_, Spring 1977.

Pozdena, Randall J., and Leonard Merewitz, "Estimating Cost Functions for Rail Rapid Transit Properties," _Transportation Research_, Vol. 12, pp. 73-78, 1978.

Särndal, Carl-Erik, and W. Brent Statton, "Factors Influencing Operating Cost in the Airline Industry," _Journal of Transport Economics and Policy_, January 1975.

Sharkey, William W., and Lester G. Telser, "Supportable Cost Functions for the Multiproduct Firm" _Journal of Economic Theory_, Vol. 18, 1978.

Smith, Vernon, _Investment and Production: A Study in the Theory of the Capital-Using Enterprise_, Cambridge, Massachusetts: Harvard University Press, 1961.

Spady, Richard H., _Econometric Estimation of Cost Functions for the Regulated Transportation Industries_, Garland Press, 1979.

Spady, Richard H., and Ann F. Friedlaender, "Econometric Estimation of Cost Functions in the Transportation Industries," Massachusetts Institute of Technology, Center for Transportation Studies Report 76-13, 1976.

Spady, Richard H., and Ann F. Friedlaender, "Hedonic Cost Functions for the Regulated Trucking Industry," _The Bell Journal of Economics_, Spring 1978.

Steger, Wilbur A., "Transportation Output Measures: Needs for Decision-Making," Transportation Research Forum, 1966.

Thompson, J. M., Modern Transport Economics, London: Penguin Books, 1974.

Varian, Hall, Microeconomic Analysis, Norton Press, 1978.