DOT-RSPA-DPB-50-79-31

UNIVERSITY RESEARCH

Development of Hybrid Cost Functions from Engineering and Statistical Techniques: The Case of Rail

DECEMBER 1979

FINAL REPORT

Under Contract: DOT-OS-70061

Document is available to the U.S. public through The National Technical Information Service, Springfield, Virginia 22161

U.S. DEPARTMENT OF TRANSPORTATION
Research & Special Programs Administration
Office of University Research
Washington, D.C. 20590

Technical R

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/RSPA/DPB-50/79/31		

4. Title and Subtitle	5. Report Date
DEVELOPMENT OF HYBRID COST FUNCTIONS FROM ENGINEERING AND STATISTICAL TECHNIQUES: THE CASE OF RAIL - FIRST YEAR FINAL REPORT	JANUARY 1980
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Andrew F. Daughety and Mark A. Turnquist	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Northwestern University The Transportation Center Evanston, IL 60201	11. Contract or Grant No. DOT-OS-70061
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered First Year Final Report
U.S. Department of Transportation Research and Special Programs Administration Office of University Research Washington, D.C. 20590	14. Sponsoring Agency Code DPB-50

15. Supplementary Notes

Technical Monitor: Joel Palley, RPD-32

16. Abstract

Meaningful economic analysis of public policy and resource allocation in the transportation industries requires empirical understanding of costs. Two traditional methods of acquiring such knowledge are: 1) engineering techniques based on detailed models of operations, and 2) statistical analysis based on expenditure and output records of firms. The objective of this research is to develop a method for combinining these approaches, in order to provide accurate, meaningful models of a cost function for a railroad. The resulting hybrid cost function incorporates detailed information on both operations and non-operations activities so as to provide a more complete picture of the relation of costs to outputs produced and inputs purchased and used by the rail firm. Thus hybrid models reflect costs associated with yard and linehaul activity as well as marketing, planning and other non-operations elements of the firm. This links together management, train crews, yard activity, maintenance, financial considerations, etc. as elements of cost generation in the firm.

A general methodology for integrating engineering and economic approaches to cost analysis is developed and applied to data collected from a rail firm. A multi-output, short-run variable cost function is estimated and examined. Concepts of economies of size, configuration and density are clarified and used to examine theoretical effects of regulation on firm maintenance policy.

17. Key Words	18. Distribution Statement
transportation cost modelling railroad cost analysis scale economies railroad operations analysis	Document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
UNCLASSIFIED	UNCLASSIFIED		

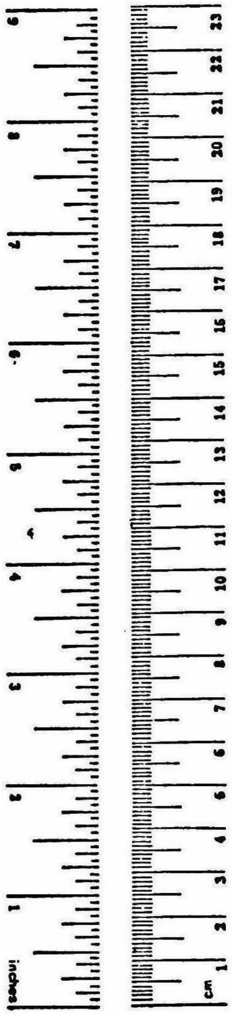
Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

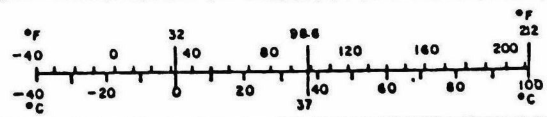
Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.



Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F



EXECUTIVE SUMMARY

INTRODUCTION

An understanding of the nature of costs of production is important in every regulated industry, both for individual firms and their regulators. At the most basic level a firm will require cost data for corporate planning. For example, a firm may wish to know what size plant to build, whether to upgrade the quality of plant or whether, at an existing tariff, the revenues for a service cover the incremental cost of providing the service.

Regulators and other policy makers also have many reasons to seek improved information about costs. When examined correctly, cost data can be used to determine whether there are in fact economies of scale in production, and whether regulation is a necessary tool of social control in a given industry. Regulators often ask whether a service is being subsidized by other services of a multiproduct firm, is subsidizing other services, and whether the provision of service by one mode will eliminate another mode over a given route.

PROBLEM STUDIED

Previous railroad cost studies typically have examined a cross section of Class I railroads, using ICC data, and most have assumed a single product, usually total ton-miles. Several aspects of these studies have served to limit the inferences that can be drawn. They rely on data from the ICC accounts rather than on raw data from the firm. With few exceptions, they have specified a relatively simple functional form for costs, and assert that the form is appropriate without a test of that assertion. Few adjust for quality of service, and more importantly, many do not account for the multiproduct nature of virtually every rail firm. Finally, they do not attempt to adjust for the fact that some railroads operate with a more complicated network than others.

Our own research on railroad transport costs represents a strikingly different approach to the problem for a number of reasons.

1) Our analysis begins at the level of an individual firm, and uses cost and production data obtained directly from the firm rather than from the ICC. This has a number of important advantages, including the avoidance of arbitrary cost allocations of the sort often found in the ICC accounts. We employ a time series analysis for a single firm rather than a cross-sectional analysis for a particular year.
2) The multiproduct nature of the firm is incorporated into the analysis. Output will be characterized both by the volume of freight hauled and by the average speed of a shipment through the system. Models with disaggregated volume (by commodity type) have been estimated, as well as with aggregate data. We explicitly recognize that speed of service is an important determinant of rail costs, and include this in our estimates.

3) We use information about the underlying technological production process, developed through engineering process functions, to better specify the nature of technology and to improve the efficiency of our estimates.

In several respects the last point is particularly novel. Historically, most econometric estimates of cost functions have ignored valuable information generated from an analysis of engineering process functions to provide observations of service-related variables. We have labeled this a "hybrid" approach for that reason, and we believe that important new insights can be gained from applications of this technique to other modes, as well as in rail transport.

RESULTS ACHIEVED

A short-run variable cost function incorporating commodity flow information, service characteristics, factor prices and a measure of the plant quality was developed and estimated from data for a railroad. Engineering models of linehaul and yard activity were used to provide information on the average speed of a shipment through the system. The models of the linehaul allowed for grade differences on linehaul sections, track quality, trailing load of the train, amount of available locomotive power and delays due to congestion on single-track sections. The yard activity model predicted expected waiting time in the yard. This information, along with information on amounts of commodities moved, quality of plant and prices of factors such as cars, locomotives, crews, non-crew labor and fuel were used to provide an estimated cost model.

Three questions were examined, based on the estimated model:

1) Does engineering model information contribute significantly to model performance and quality?

2) How do the various terms in the cost functions influence predicted short-run costs?

3) Does the data support the use of some of the simpler production models often resorted to in earlier analyses?

First, a test of the value of the engineering information was constructed and performed. The result was that the introduction of engineering information significantly improved the model. This acts as a test of the value of a hybrid approach to cost analysis and confirms its superiority to traditional economic or engineering models.

Second, the impact on cost of changes in factor prices, commodity flow level, speed and plant quality were examined. The elasticities of cost with respect to factor prices reveal the somewhat surprising result that the firm cannot easily substitute away from non-crew labor into capital as the non-crew labor price rises. This reflects the need for further computerization of the firm. Increases in speed would reduce short-run variable costs. These would largely come about from improvements in linehaul condition, which was also seen via the elasticity of cost with respect to plant quality which was negative. Marginal commodity flow costs were positive, as expected.

Third, the model structure admitted testing of various sub-cases such as Cobb-Douglas production technology, which were rejected. Thus, previous studies that have started out assuming such models, are likely to be misspecified and produce biased results. This finding is consistent with very recent literature in the area of cost and production theory and estimation.

The project also provided two theoretical results. First, a procedure for properly integrating economic and engineering models was developed. The procedure provides the general form of the cost function to be estimated and classifies the model variables as to whether or not they are to be provided by engineering process models. This is a general procedure which can be used in many other analyses.

The second theoretical result concerns a clarification of economies of scale. Economies (and diseconomies) of scale are discussed in terms of economies (and diseconomies) of size, configuration and density. These concepts are defined and used to relate firm-level maintenance decisions to regulatory constraints on service abandonment.

UTILIZATION OF RESULTS

Potential users of the research results include both government agencies and private railroad firms. In fact, officials of railroad "X", which cooperated with us on the empirical work in this project, have already expressed a desire to use our results with respect to marginal cost computations as a basis for submitting a proposed rate. Certainly this is evidence of the immediate applicability of the work to problems faced by rail firms. It is not, however, the only way in which the results could be used by railroads. Significant insights have also been gained with respect to cost elasticities for various input factors. These elasticities have important implications for corporate planning.

From a government perspective, this work provides an important step toward operationalizing the concept of "incremental" costs as a basis for policy-making and regulatory proceedings. This concept plays a central role in the regulatory reform legislation currently in Congress. By making the concept operational (specifying data requirements and analysis procedures) regulatory proceedings and policy-making can better incorporate economic principles.

CONCLUSION

A procedure for integrating economic and engineering approaches to cost analysis was developed and employed to use data from a medium size railroad to estimate a model relating short-run variable costs to commodity flows, speed, factor prices and plant quality. Engineering models of yard and linehaul activity were used in a general cost model incorporating financial and engineering data. The data requirements for the study were not unreasonable; almost everything that was needed was already maintained by the firm. The results of the research are useful to both railroad firms and policy makers. The cooperating railroad in this study is presently evaluating the estimation results for possible inclusion in a rate proceeding. Moreover, this study is a step toward providing an operational definition of incremental costs.

Acknowledgments

Major contributions to the inception and development phases of this project were made by Ronald R. Braeutigam (at this writing, with Northwestern University, on leave to California Institute of Technology).

Special thanks to the executives and staff of Railroad X who provided access to data and help in understanding what we had stumbled upon. A number of them put in no small amount of time digging out and assembling information, without which this project would not have been possible. Thanks also to William Delaney who, while with the Federal Railroad Administration, was our contract monitor, and was without par.

We also thank Robert Baesemann, formerly of Northwestern University, who was instrumental in starting the project and Greg Duncan, who spent a short, but productive period beating on data requirements issues with us. Joseph Swanson of Northwestern has been invaluable as a sounding board for ideas and as a source of information on calculating prices of capital goods.

The Transportation Center contributed significantly with financial support to augment the original contract from USDOT. Without this help we would still be transcribing data.

Thanks to Ken Hurdle of the AAR for providing some very needed data, and especial thanks to Melissa Bartlett, John Gray, Victor Liu and Hossain Poorzahedy, who slaved over much of the data set helping piece it together.

Finally, thanks to all those who contributed to the typing of the various papers and parts that have gone into this project, especially Theresa Bonk who plowed through pages upon pages of manuscript, providing her usual excellent typing and editing job. This work was sponsored by the U. S. DOT, Research and Special Programs Administration, Office of University Research under contract DOT-OS-70061.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

An understanding of the nature of costs of production is important in every regulated industry, both for individual firms and their regulators. At the most basic level a firm will require cost data for corporate planning. For example, a firm may wish to know what size plant to build, whether to upgrade the quality of plant or whether, at an existing tariff, the revenues for a service cover the incremental cost of providing the service. Cost data may be used to argue for a change in tariffs. A firm may want to know how a change in the level of output of one service affects the costs of providing another service, and it may rely in part on cost data to determine whether it would be profitable to discontinue a service, introduce a new service, or attempt to merge with another firm.

Regulators and other policy makers also have many reasons to seek improved information about costs. When examined correctly, cost data can be used to determine whether there are in fact economies of scale in production, and whether regulation is a necessary tool of social control in a given industry. Regulators often ask whether a service is being subsidized by other services of a multiproduct firm, is subsidizing other services, and whether the provision of service by one mode will eliminate another mode over a given route. If regulators are interested in setting tariffs that allocate economic resources efficiently, they will require information about costs. Generally speaking, then, regulators need cost information to determine how their policies will affect market structure and economic performance. These comments apply without exception to the railroad industry.

1.1. Other Railroad Cost Estimates

A number of studies have examined costs in the railroad industry. The early work in this area attempted to characterize the output of railroads as a single product, usually ton-miles. These studies typically have examined a cross section of Class I railroads, using ICC data, to test whether there are economies of scale in rail transport. The results have generally been mixed. For example, Klein [50] used 1936 data to find economies of scale that were statistically significant, though modest. On the other hand, estimates by Borts [8] and Griliches [38] have suggested that, while there may be economies of scale for smaller railroads, scale economies are not prevalent for larger Class I railroads.

Several aspects of these studies have served to limit the inferences that can be drawn. They rely on data from the ICC accounts rather than on raw data from the firms. They typically specify a relatively simple functional form for costs, and assert that the form is appropriate without a test of that assertion. They do not adjust for quality of service, and more importantly, they do not account for the multiproduct nature of virtually every rail firm. And, they do not attempt to adjust for the fact that some railroads operate with a more complicated network than others.

Keeler [49] and Hasenkamp [42] used approaches grounded in production theory to examine multi-product aspects of railroad activities, distinguishing between freight and passenger activities. Using more sophisticated analysis Brown, Caves and Christensen [9] and Friedlaender and Spady [32] develop models that allow multiple outputs and do not enforce spearability of inputs and outputs. Caves, Christensen and Swanson [13] have also used such techniques to examine productivity growth in U.S. railroads. In all these cases cross-section data drawn from ICC reports or based on Klein's work [50] has been used. Thus railroads with rates-of-return varying between -10% and +40%, facing different geography, having different mixes of equipment, customers and managerial perspectives were mixed together in the estimation process. No real use of service variables such as speed could be used since such data is firm-specific and not usually published. While the above studies have represented important advances in understanding of costs, more work is needed, especially at the level of the individual firm.

1.2. A Time Series Estimate of a Hybrid Cost Function

Our own research on railroad transport costs represents a strikingly different approach to the problem for a number of reasons.

1) Our analysis begins at the level of an individual firm, and uses cost and production data obtained directly from the firm rather than from the ICC. This has a number of important advantages, including the avoidance of arbitrary cost allocations of the sort often found in the ICC accounts. (For a discussion of the kinds of problems arising from the use of ICC data, see, for example, Friedlaender [30], Appendix A.) We employ a time series analysis for a single firm rather than a cross sectional analysis for a particular year.
2) The multiproduct nature of the firm is incorporated into the analysis. Output will be characterized both by the volume of freight hauled and by the average speed of a shipment through the system. Models with disaggregated volume (by commodity type) have been estimated, as well as with aggregate data. We explicitly recognize that speed of service is an important determinant of rail costs, and include this in our estimates.

3) We use information about the underlying technological production process, developed through engineering process functions, to better specify the nature of technology and to improve the efficiency of our estimates.

In several respects the last point is particularly novel. Historically, most econometric estimates of cost functions have ignored valuable information generated from an analysis of engineering process functions to provide observations of service related variables. We have labeled this a "hybrid" approach for that reason, and we believe that important new insights can be gained from applications of this technique to other modes, as well as in rail transport.

The plan of this report is as follows. Chapter two represents the extension of the theory of production and cost (reviewed in Appendix B) to the problem of integrating (or hybridizing) economic and engineering approaches. Here we examine concepts of economies of size, configuration and density, a new view of scale economies that comes about from economic/engineering insights gained in the project. These insights also provide the general method for constructing a hybrid cost function developed in the chapter. Finally, the chapter provides a complete analysis of the engineering models to be used.

The third chapter provides an overview of the data used while the fourth chapter presents the estimation resuits (other estimation results on a disaggregate volume model are presented in Appendix A). Chapter five analyzes the results and draws out implications for rail cost analysis. Plans for second year activity on the project are discussed in Chapter 6.

CHAPTER 2

HYBRID ANALYSIS

2.1 Problems to be Solved

In Appendix B we review the general elements of production and cost theory. The analysis contained therein is for a general firm. When, however, we become more specific in the type of firm that we wish to analyze, we can then refine and expand those notions. In our case we will consider a rail firm, and though most (if not all) of what we develop will be immediately transferable to other modes, we will couch most of our discussion in terms of a railroad.

The purpose of this project has been to examine the feasibility and value of linking economic and engineering approaches to cost estimation and apply it in a case study of a railroad; chapters three and four consider the case study. Before proceeding to the case study two fundamental issues must be addressed.

1. Can one be more specific about the relationships between long and short-run cost functions ? What are the natural factors to consider as constant in the short-run and how can one discuss economies of scale?
2. How should engineering information be merged with economic information: is there a way to structure the integration so that a general procedure is developed?

These questions are interrelated: the first question raises the issue of what is the short-run cost function and how does it relate to the long-run cost function; we shall see that insights from the engineering perspective provide the clue. The second question concerns the integration or hybridization of engineering information into the economic model of cost: we shall see that insights from the economic perspective provide the procedure.

2.2. Technological Economies: Size, Configuration and Density Economies of Scale

Economies of scale have been of interest to economists and policy makers for centuries. Adam Smith spent three chapters of his Wealth of Nations discussing specialization and the division of labor in production. A number of categorizations and definitions of economies of scale have been developed. Generally, the view was that economies of scale existed if average costs fell as output was expanded. This raises, however, the disturbing problem of what is an average cost in a multi-product firm if we can not refer to any one of the outputs as *the* output (or if no aggregation function on output exists).

The result is that economies-of-scale definitions for multi-product firms now deal with the technological description and not the cost function. Note that this ignores *pecuniary* economies of scale [72] that are associated with the ability of a large firm to affect the prices at which it purchases inputs. This type of economy is ignored based on the usual presumption of given factor prices (see section B.2), i.e. the firm faces perfectly competitive factor markets.

Panzer and Willig [64] provide the following definition of economies of scale.

> A technology τ with associated transformation function $T(z,x)$ has economies of scale at (x,z) if there exist $r > 1$ and $\delta > 1$ such that
>
> $$T(\lambda^r z, \lambda x) \leq 0 \qquad \text{for } 1 \leq \lambda \leq \delta$$

This is a local definition, i.e. for the point (x,z). The exponent r may be a function of x and z. All that is required is that the point $(\lambda x, \lambda^r z)$ be in the technology for a small neighborhood of λ. Figure 1 shows a technology τ and two points (x^1,z^1) and (x^2,z^2). There are no economies of scale at (x^1, z^1), while there are economies at (x^2, z^2). It is this definition that provides the measure of scale economies S in section B.2.2.3.

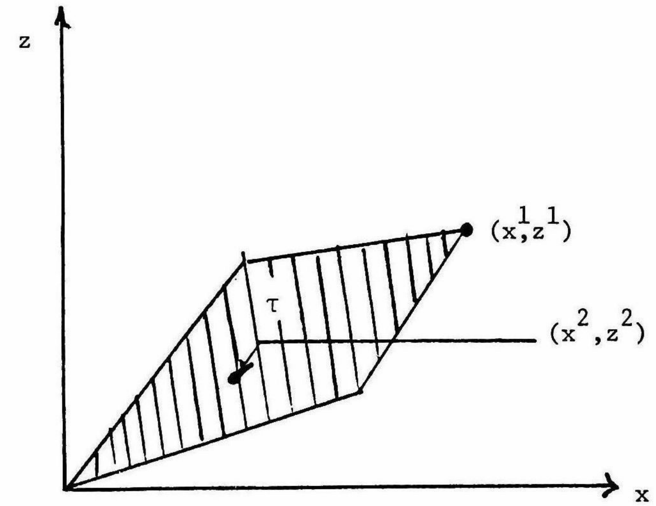


Figure 1 - ECONOMIES OF SCALE

When we become specific, however, about the application of the theory it is important to differentiate between different mechanisms that give rise to technological economies (as opposed to pecuniary economies) of scale. In this section we will examine, and attempt to relate, concepts of scale economies due to size, to configuration and to density.

Scale economies due to size come about from the distant and varied geographical points that a transport system (such as a railroad or a modern motor common carrier firm) connects. Measures such as average length of haul tend to be used to reflect this type of size. Size, in this sense, is important since it opens many markets to the firm and shippers who must send goods long distances usually prefer to work with as few transport firms as possible so as to expedite claims on loss and damage. Large size may, or may not, be accompanied by intensive utilization (or high traffic density) of the system. In between the issues of size and density are the firm's policies on configuration and system maintenance.

For any system of a given size, there may be many ways to configure the actual system itself. The same set of demand points can be serviced by a minimally connected network (e.g. a tree network wherein each demand point is connected to at most two other points, and for n demand points there are $(n-1)$ "arcs" connecting the points) a hub-and-spoke network (e.g. a yard connected directly to each demand point) or a completely interconnected system (each demand point connected to each other demand point). There are, of course, many other possible configurations of a system. Changes in configuration can occasion changes in operating policies (such as blocking policies) and changes in capital utilization (such as the use of cars). Some of the changes will result in economies of scale, some in diseconomies. It is important to note that in general such changes are discrete in nature: there may be severe lumpiness in such changes which may not be smoothable by other input changes. To see this we consider Figures 2 and 3. Here we have assumed one output. Figure 2 shows the classical treatment of the long-run average cost curves (LRAC) as the envelope of the short-run average cost curves. The short-run curves are labelled x^{f_i} where this represents different levels of the fixed inputs. Because the fixed inputs are assumed to vary smoothly then LRAC is also smooth. Equation (B.11) of section B.2.1

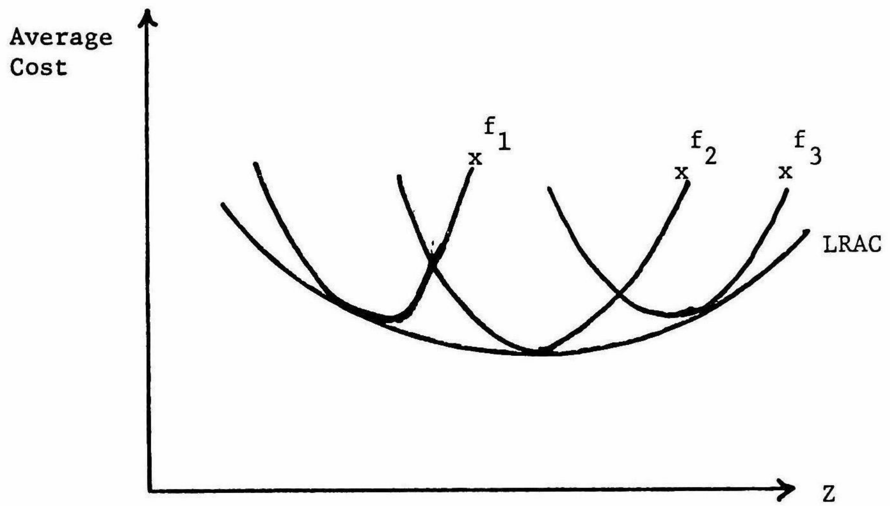


Figure 2 - CLASSICAL LONG-RUN AVERAGE COST

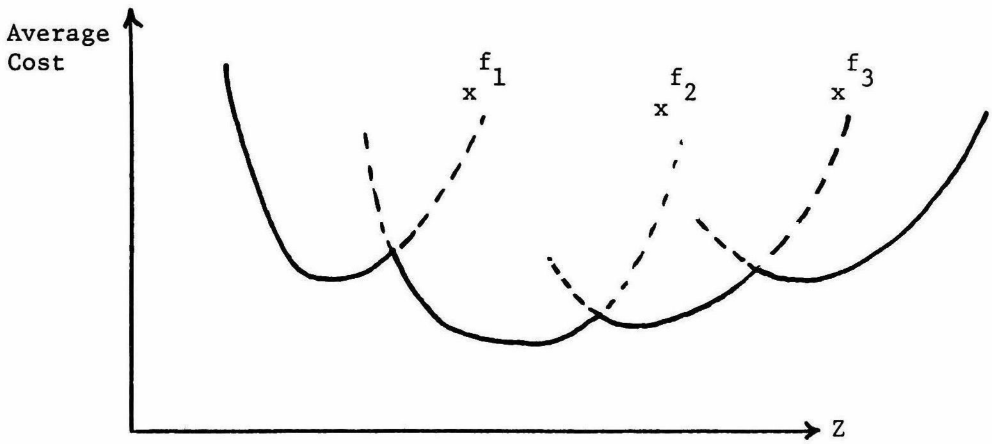


Figure 3 - SCALLOPED LONG-RUN AVERAGE COST

shows the use of the envelope theorem to derive the long-run total cost curve from the short-run curves.

If the fixed factors are lumpy then we still have an LRAC curve, but it may be scalloped, as shown in Figure 3. Here there is no intermediate value of x^f between x^{f_1}, x^{f_2} and x^{f_3} (or below x^{f_1}). The dotted curves represent those portions of the short-run average cost curves that are not part of LRAC. Notice also that if we had incorrectly assumed x^f to be continuously variable then our estimated LRAC curve would almost always underestimate long-run average costs, since only the tangencies between the short- and long-run curves would not be underestimated. Clearly, as the possible values of x^f becomes denser (i.e. as x^f can be varied in smaller jumps) then the underestimation becomes less pronounced. Since, however, configuration changes can imply significant land acquisitions or disposals (and other types of lumpy inputs) we expect that such changes are quite discrete in nature and that Figure 3 properly represents the long-run average cost curve.

Finally, for a given size and configuration there may be economies of density. Stigler [77] observes that there may be times when inputs would be more fully utilized, but are not, producing excess capacity:

> "There may be some unavoidable 'excess capacity' of some inputs. A railroad has a tunnel which is essential for given traffic, but can handle twice as much traffic. The emphasis here is on 'unavoidable.' If the railroad has unused locomotives, in the long run they can be sold or worn out, and hence do <u>not</u> give rise to increasing returns." (p. 153).

In the case of transport systems, especially rail, the density of traffic on the line-haul portion may be low relative to the line-haul capacity. Keeler [49] and Harris [41] have found significant economies of density for U.S. railroads due to excess track capacity (in both these articles, economies of scale are broken down into size and density only). Thus, for example, the elimination of double tracking on some line-haul segment (a change in system configuration) might result in increased traffic density (especially if the original traffic density is light). Density economies can be realized by increases in traffic density for a fixed configuration. Changes in configuration that result in increasing the density of traffic on a particular piece of track (without increasing the flow through the end-points) would appear to be appropriately labelled configuration economies rather than density economies. On the other hand, changes in operating policies that result in high traffic density for the same configuration should be associated with economies of density.

In summary, then economies of size can come from changes in size holding the <u>nature</u> of the output fixed (i.e. we ignore diversification of the firm into other markets than that for transport services). Economies of configuration can come from changes in system configuration (number of yards, location, interconnectedness, double vs. single line-haul tracking) while economies of density arise for varying traffic levels within a configuration.

2.3. Hybrid Cost Models and the Use of Engineering Information

In Appendix B we explain how a cost function $C(z,p)$ is derived from the cost minimization problem:

$$\text{(CMP)} \qquad \min \; p'x \qquad \text{s.t.} \; T(z,x) \leq 0$$

where x is an n-vector of inputs, z an m-vector of outputs, p an n-vector of given prices and $T(z,x)$ a transformation function. It is useful to characterize the output of transportation firms in general, and railroads in particular, in terms of both the physical units of <u>flow</u> over the system (e.g. car-miles of various commodities) and measures of the <u>quality of service</u> provided (e.g. average speed, reliability, loss and damage, etc.) In mathematical terms, let $z = (y,\pi)$ where y is an m_1-vector of flows and π is an m_2-vector of service characteristics with $m_1 + m_2 = m$. Let $T(z,x) \equiv T(y,\pi,x)$ be of an arbitrary general form, presently unspecified.

Most previous cost analyses have fallen into one of two categories. On one hand, economists have developed cost models of firms or industries that have generally ignored models of the physical process associated with the firms activities (an exception is [58]). The strength of the economists's models lay in the recognition that non-operations activities (such as planning, sales, etc.) contributed to output and cost. These things were, to some degree, captured by the economist's model. On the other hand, engineering models of the operations aspect of a firm provide excellent information, but an incomplete picture of the firm. These process models (see [20], [21] for examples in the transportation area) specify certain relationships between inputs (x), flows (y) and characteristics (π). For example, one of the process models to be presented in section 2.4 relates trailing load (a y-variable) to locomotive horsepower (an x-variable) and speed (a π-variable). Thus, another way to view the overall production process of the firm is to "layer" physical relationships as constraints on some very general transformation function. The transformation function is the "glue" that holds the system together, including inputs and outputs that are not definable in process functions. Notationally, we then have the following description of technology.

$$T(y,\pi,x) \leq 0 \qquad (1)$$

$$g^i(y,\pi,x) \leq 0 \qquad i = 1,\ldots,I \qquad (2)$$

where some parts of the y, π and x vectors in the g^i functions may not appear at all; i.e., each g^i function is a process function and as such may not use all the variables. It may also occur that some of the inequalities are really equalities.

Suppose we dropped $T(y,\pi,x)$ and partitioned the x-vector into subsets, x^i, so that $x = (x^1,x^2,\ldots,x^I)$, each $g^i(y,\pi,x)$ was a function of only one element of the output vector and one subset of the input vector,

$$g^i(z_i,x^i) = 0 \qquad\qquad i = 1,\ldots,I \tag{3}$$

we would have the case of non-joint production (see [40]). We will not assume this extreme case. Instead, possibly overlapping subvectors of y, π and x appear in the g^i functions and the entire vectors appear in T. So, for convenience, let $H(y,\pi,x)$ be defined as the set of y, π and x values that satisfy the joint conditions. We will then write the system as

$$H(y,\pi,x) \leq 0 \tag{4}$$

As a simple example, let us consider each vector to have only one element and assume a transformation function:

$$T(Y,\Pi,X) = Y + \Pi + X-1 \tag{5}$$

and one process function:

$$g(Y,\Pi,X) = \Pi - .5 \ . \tag{6}$$

This system is illustrated in Figure 4. The result is the trapazoidal solid in Figure 5 which we will call $H(Y,\Pi,X)$. Notice two points: 1) $H(Y,\Pi,X)$ is more refined than $T(Y,\Pi,X)$ since we have added the information contained in $g(Y,\Pi,X)$; 2) $H(Y,\Pi,X)$ is more comprehensive than $g(Y,\Pi,X)$ since we have not neglected relationships between variables not addressed by $g(Y,\Pi,X)$.

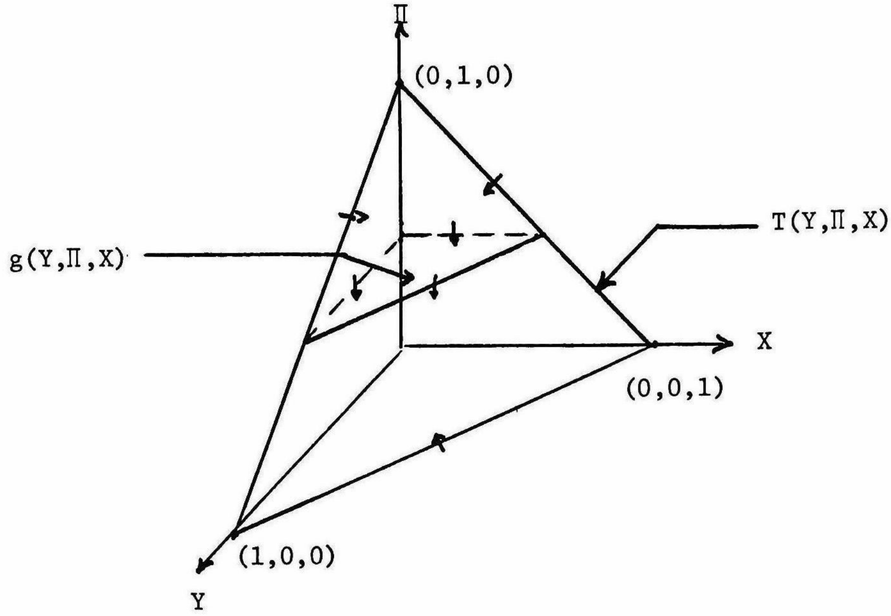


Figure 4 - CONSTRUCTING A HYBRID PRODUCTION SURFACE

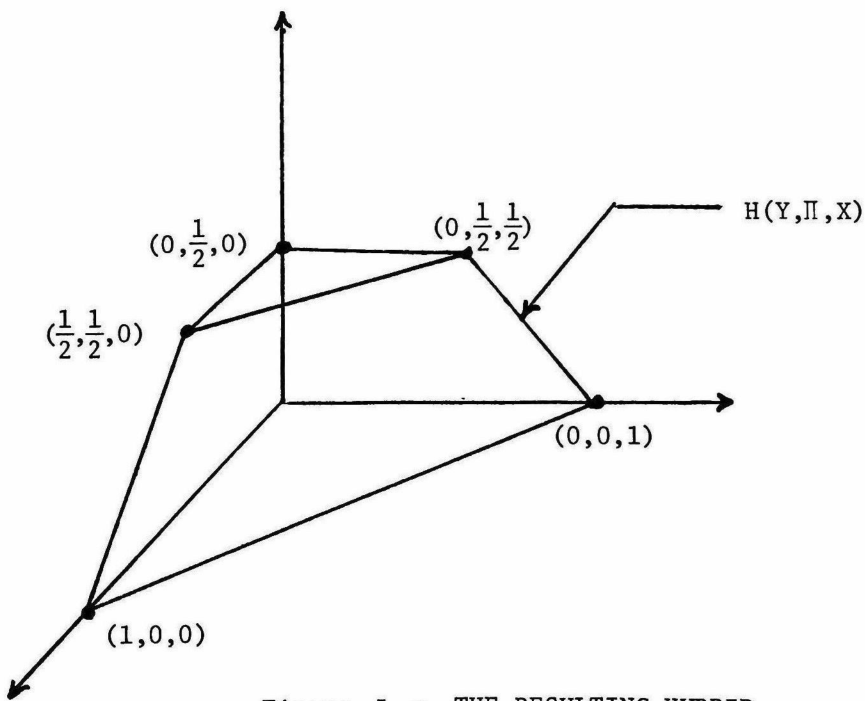


Figure 5 - THE RESULTING HYBRID PRODUCTION SURFACE

Therefore, engineering relationships help us refine a very general structure into a more specific structure, not by making functional assumptions (Cobb-Douglas, CES, etc.) but by the use of physical principles to restrict the possible relationships among the variables. As more and more g^i functions are added, the technology becomes better defined.

What does this mean for the analysis at hand? Consider the cost function dual to $\min \{p'x | H(y,\pi,x) \leq 0\}$, namely $C(y,\pi,p)$. We can estimate this cost function by assuming, for example, a flexible functional form that approximates an arbitrary production function. The variables in the function would be flows of goods, prices of inputs and characteristics of the service provided. It is the last category especially to which we now turn.

The output of a rail firm is stochastic, just like many other firms. This is especially true of the characteristics of service, such as speed. While we would not expect to see significant variation from month-to-month in total aggregate demand (flow), we very well could expect significant variation in such outputs as speed of a shipment through the system. The effect of this on our model of costs is very significant. Let $f(x,\omega)$ be a stochastic production function with random variable ω. Thus the output Z:

$$Z = f(x,\omega) \tag{7}$$

is now a random variable. Let the distribution of ω be $G(\omega)$ with density function $g(\omega)$. In [16] it is shown that the profit-maximizing firm will, implicitly, solve the following cost minimization problem (see section B.2):

(CMP)

$$\min \quad p'x$$
$$\text{s.t.} \quad E(f(x,\omega)) = u$$

where p is given, u is expected output and the expectation is taken with respect to G on its domain. Thus, the cost function is

$$C(E(Z),p)$$

where the expectation here is over the distribution on Z (which can be derived from $f(x,\omega)$ and G). If, on the other hand, we had simply formed the cost function on outputs and prices we would have

$$E(C(Z,p))$$

It can be shown [16] that these two cost functions are consistent (for unknown G) if and only if $f(x,\omega)$ is homogeneous of degree one in x (see Appendix B, section B.1.2.1). Since we do not wish to make this an implicit assumption in the analysis, we do not want to use Z as a variable in our cost function. Rather, we should use a model that predicts the expected value of Z for the observed values of the non-random variables. This is what an engineering process model can give us. Hence, for those stochastic variables in the output vector (y,π) we will use a process model to provide the expected value of the variable. Thus engineering models, in fact, provide more useful information than the raw observations themselves.

In general, then, we see that the role of engineering models is two-fold:

1. They provide useful information on physical relationships between the model variables, thereby properly restricting the model of production.
2. They provide the proper variables for inclusion in cost models, especially when some of the output variables are stochastic.

By using engineering process models to define variables in the cost model we implicitly refine the technology to which the cost model is dual. Therefore the result is a more efficient estimate of a better specified cost model.

This study has concentrated on speed of a shipment through the system as the engineering input; thus the π vector has one variable. Speed is a stochastic variable since time of day, of month, season in the year, etc. all contribute to significant variation in the time it takes a shipment to pass through the yards and over the appropriate segments of the line-haul. Therefore the engineering work will concentrate on process models that relate speed to some of the inputs and other outputs.

What might we expect to see for a fixed configuration? At low density there is little relationship between speed and short-run variable costs (due to union work rules, only radical changes in speed over reasonably long distances would affect costs). At high density, congestion effects act to reduce speed and increase costs. Therefore, one would expect to see a negative relationship between speed and short-run variable costs. Clearly, changes in configuration change the density level at which this occurs. These engineering relationships illustrate the importance of speed. With this in mind, we turn next to formulating the process models for the hybrid cost function.

2.4. Engineering Process Models of Linehaul Train Movement and Yard Operations

2.4.1. Introduction

A basic premise of this research project is that models representing the basic engineering processes involved in railroad operations can contribute significantly to the estimation of rail cost functions. The integration of these process models with econometric estimation of cost functions results in what may be termed a "hybrid" model. This section describes in some detail the set of process models used in the project.

The models cover three important areas of railroad operations:

1) line-haul movement of trains;
2) line-haul delays due to interactions among trains;
3) classification yard operations.

The models presented here draw heavily on the work of previous researchers. The major contribution of the current project is synthesis of these component models into a workable system for use in cost function estimation. The types of models involved, and their interactions, are illustrated schematically in figures 6 and 7. The chapter is organized into three major parts, describing the three models separately.

2.4.2. Line-Haul Train Movement

The line-haul train movement model developed for this project is based on six fundamental characteristics to insure flexibility in application. These characteristics are as follows:

1. The basic form of the model is derived from physical relationships, with corrections for technical conditions separated from the primary function. This allows adjustment for further technical change without modification of the basic form.
2. The model is applicable over a wide range of train types and physical line configurations.
3. It uses a minimum number of variables.
4. The model is designed to use variables defined so as to be commensurate with expected data available (e.g. trailing load, rather than individual car weight statistics.)
5. The variables can be readily related to specific cost data.
6. It provides a base which is adaptable to experimentation with differing types of operational and investment policies.

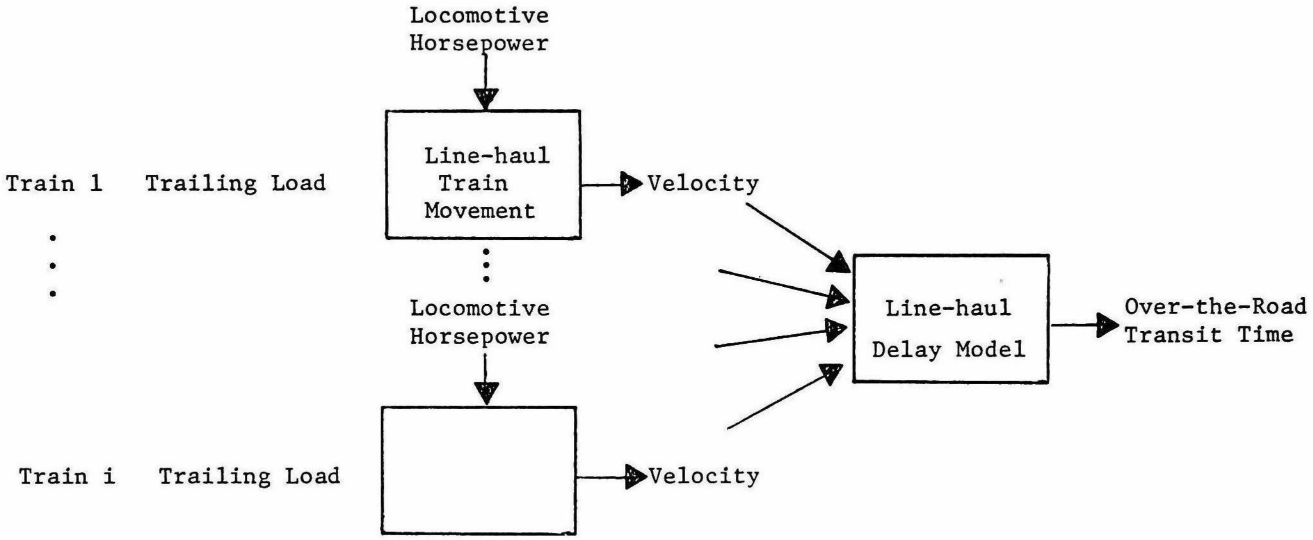


Figure 6. LINE-HAUL MODELS.

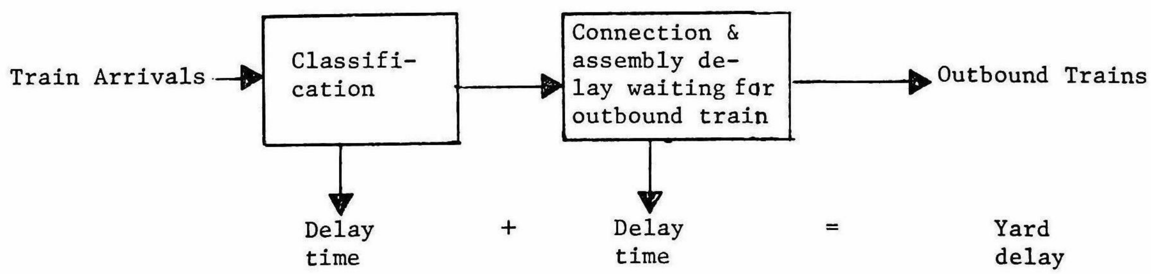


Figure 7. CLASSIFICATION YARD MODELS

The first requirement of the model is to establish a relationship among distance, time, weight of the train, and horsepower provided. Obviously time and distance can be further reduced in part to the velocity over each track segment.

The analysis is based on establishing the physical relationship at a constant velocity. For simplicity, acceleration and deceleration will be ignored. Thus

$$F = m\frac{dv}{dt} = m(0) = 0 \tag{8}$$

where: F = force exerted by the train system

m = mass

dv/dt = acceleration

There are two basic forces at work on the train -- the tractive force exerted by the locomotive and the resistance of the train mass acting in the opposite direction. Thus the basic balance equation is:

$$F = F_t - F_R = 0 \tag{9}$$

where: F_t = tractive force of the locomotive

F_R = resistance force of the train

Tractive force of a locomotive is a function of its horsepower and the velocity at which it is operated. Locomotive manufacturers publish graphs of velocity versus tractive effort for all of their locomotive types. (See, for example, [26].) These curves are essentially rectangular hyperbolas within the velocity range of interest (10 mph - 65 mph). Exceptions to this rule have the form shown in figure 8 and fit the normal form to the right of point a.

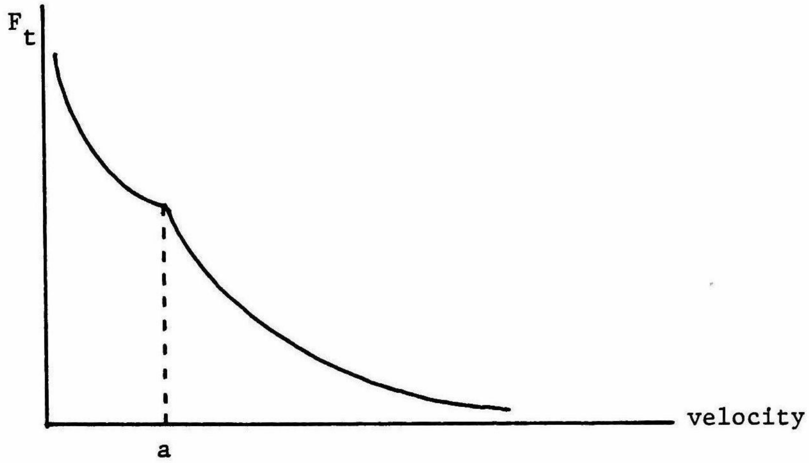


Figure 8 - TRACTIVE EFFORT AS A FUNCTION OF VELOCITY

The value of a is usually no higher than 15 mph. For all locomotives F_t can be defined as follows:

$$F_t = 375 \text{ HP } (e/v) \tag{10}$$

where: HP = locomotive horsepower

e = machine efficiency, (.825 for most North American applications)

v = velocity (miles per hour)

F_t = tractive force (pounds)

This equation describes the relationship between F_t and v indicated by manufacturer data to within about 2% for both types of curves. This indicates that curves of the type shown in figure 8 reflect distortions in the lower velocity ranges rather than basic functional changes. Thus, the form of F_t shown will be used for all locomotives operating in line haul service.

The resistive force, F_R, may be broken into two components: F_R^{ℓ}, the resistance to motion by the locomotive, and F_R^{c} the resistance to motion by the cars. The equation used to represent both of these quantities is the time-honored Davis Formula. A complete discussion of its component parts may be found in references [17], [18], and [43]. For the present purposes, it is sufficient to note that it is based on considerable empirical evidence gathered over the last half century, and it has been shown to provide consistently useful results. It is based on a resultant resistance determined by summing the components of journal, flange, air, track, and grade resistance with appropriate constants. The form used in this work is that presented in Hennes and Ekse [45] which is taken from Davis's earlier work [18]:

$$R = \frac{9.4}{w^{1/2}} + \frac{12.5}{w} + JV + \frac{KAV^2}{wn} + 20s \qquad (11)$$

where: R = resistance to movement, in pounds per ton of train weight

w = average weight per axle in tons

V = speed, in miles per hour

n = number of axles per item of equipment

J,K = constants (for locomotives: J = 0.03, K = 0.0024; for cars: J = 0.045, K = 0.0005)

A = cross sectional area, in ft^2 (for locomotives: A = 120; for cars: A = 87.5)

s = grade encountered, in % .

In the case of locomotives, a value of 32 can be used for w. A check of mainline locomotive weights since the introduction of the General Motors FT diesel in 1939 shows that axle loadings have remained remarkably consistent since that time [11]. While some roads own heavy duty locomotives with axle loads as high as 35 tons and a few older units are as light as 30 tons, this variation is of little consequence (less than 1% in most cases) when viewed against the total tractive effort exerted by the locomotive. The 32 ton weight is also highly representative of the major builders standard models as supplied to railroads not requiring extensive extra features. The number of axles, n, is the total for the locomotive, not the number on each unit. Likewise, in the horsepower-tractive effort equation the horsepower is the total for a locomotive, not that of individual units.

Thus, after substitution of constants and performing the implied multiplications, the total resistance for locomotives is as given in equation 12.

$$F_R = [2.05 + .03V + \frac{.29V^2}{32n} + 20s]\ 32n \qquad (12)$$

$$= 65.60n + .96Vn + .29V^2 + 640sn$$

The car resistance function is a bit more complex. The additional problems are caused by the inability to find a constant value for w. However, it is possible to designate two constant w's (one for loads, one for empties) for a particular railroad's circumstances. This can be shown in the following way.

Examination of the resistance function indicates that three of the five terms are inversely related to the weight of the equipment. Thus an empty car will have a higher resistance per ton than will one which is loaded. This reveals the physical reason behind a phenomena which is well known in railroading: a given locomotive can pull a heavier train if it is composed primarily of loaded rather than empty cars. This is reflected in the policy which a railroad adopts in assigning tonnage ratings to its locomotives. If it rates its locomotives for the 80 to 100 ton cars which are becoming prevalent, or for a higher percentage of loaded rather than empty cars, then it will assign relatively high tonnages. On the other hand, if it bases its ratings on 50 to 70 ton cars, or a low percentage of loads, then lower capabilities will be assumed. In many cases the lighter cars are assumed as the more stringent conditions represent a more conservative view and also allow for other factors, such as adverse weather conditions, recovery of lost time when delays occur on the road and the fact that many cars are not loaded to capacity [26]. In view of this, locomotive tonnage tables published by General Motors for its locomotives [26] are in terms of 50-ton cars with an assumed light weight of 20 tons and 50% empty cars. Thus the general form of <u>car resistance</u> is (assuming n=4 in all cases):

$$F_R^{ec} = \left[\frac{9.4}{w_e^{1/2}} + \frac{12.5}{w_e} + .045V + \frac{.0109V^2}{w_e} + 20s \right] 4w_e$$

$$= 37.6w_e^{1/2} + 50 + .18w_e + .0436V^2 + 80W_e s \qquad (13)$$

and

$$F_R^{fc} = 37.6w_f^{1/2} + 50 + .18Vw_f + .0436V^2 + 80w_f s \qquad (14)$$

for <u>empty</u> and <u>loaded cars</u> respectively, where w_e and w_f are the appropriate axle weights. If the conservative assumptions outlined above are used, these two equations become:

$$F_R^{ec} = 149.22 + 1.26V + .043V^2 + 560s \qquad (15)$$

$$F_R^{fc} = 216.04 + 3.51V + .0436V^2 + 1560s \qquad (16)$$

As noted earlier, the effects of loaded cars should be considered. On most railroads, tonnage ratings are based on the assumption that 50% of the cars in a train will be empty. This even split makes it possible to consider resistance in terms of the sum of the resistance of one loaded and one empty car taken together. This quantity ($F_R^{ec} + F_R^{fc}$) will be denoted a "car unit." Thus the trailing load for a particular locomotive can be expressed as the net tractive effort (total tractive effort minus locomotive resistance) divided by the resistance of a car unit with the resultant multiplied by the weight of a car unit. Obviously, the proportion of loads to empties may be changed by appropriately weighting the two resistance and weight quantities of the car unit. Thus, trailing load is:

$$TL = \left[\frac{F_t - F_R}{F_R^{ec} + F_R^{fc}} \right] 4(w_e + w_f). \tag{17}$$

When expanded this becomes:

$$TL = \left[\frac{309HP/V - (65.6n + .96Vn + .29V^2 + 640sn)}{37.6(w_e^{1/2} + w_f^{1/2}) + .18V(w_e + w_f) + .087V^2 + 80s(w_e + w_f) + 100} \right] 4(w_e + w_f). \tag{18}$$

If standard values of $w_e = 7$ and $w_f = 19.5$ are assumed, as outlined previously, equation (18) becomes:

$$TL = 106 \left[\frac{309HP/V - (65.6n + .96Vn + .29V^2 + 640sn)}{(364.81 + 4.77V + .087V^2 + 2120s)} \right]. \tag{19}$$

Other values of w_e and w_f may be substituted depending on the policy on which tonnage ratings are based. This function is stated in terms of only five quantities; velocity, locomotive horsepower, number of locomotive axles, tonnage trailing the locomotive, and gradient. All of these items represent readily available data for any particular railroad.

2.4.3 Application of the Model to Determine Running Time

To determine the running time for a given train between two yards, equation (19) would be applied in two states. First, given the trailing load and the most restrictive conditions the train will encounter (usually termed the ruling grade), this relationship can be used to determine the horsepower which must be assigned to the train. Then, given the horsepower and trailing load, the resulting velocity for any other track segment can be determined. Once this has been done for all segments, the total running time for the entire route can be computed. Details of this procedure are illustrated by the following example.

Consider a line whose profile is as shown in Figure 9. The total line is 50 miles long, and can be divided into four segments. The first segment is 20 miles long with no grade; the second segment is the ruling grade, with a 1.0% grade over a 5 mile section; the third segment is 10 miles long, with a 0.5% grade; and the last section is 15 miles of level track.

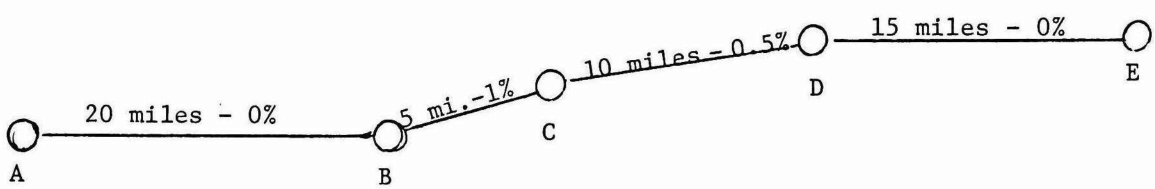


Figure 9. EXAMPLE TRACK PROFILE

Assume that the trailing load is 5000 tons, and that we wish to maintain a velocity of 20 miles per hour on the ruling grade. Let us also assume that the locomotives available are of the four-axle variety.

The first step is to find the required horsepower for the train. Equation (19) can be rewritten as follows to solve for the value of HP:

$$HP = [(0.11 + .065s)V + .00015V^2 + (2.66\times10^{-6})V^3]TL + [(.21 + 2.07s)V + .003V^2]n + .0009V^3 \quad (20)$$

Inserting values of s = 1.0, V = 20 and TL = 5000, we obtain the required horsepower to be

$$HP = 8200$$

In any given situation, we may not be able to assign exactly this amount of horsepower to the train. The actual assignment would reflect the numbers and sizes of locomotives available. For example, if the available locomotives are all 3000 horsepower units, the assignment would be three of these units, with a total of 9000 horsepower. For the purposes of this example application of the model, we will assume that exactly 8200 horsepower are assigned. Thus, we know that the velocity on segment BC will be 20 miles per hour.

The second stage of the analysis involves finding the velocities on the remaining track sections. For this purpose equation (19) can be written as a cubic equation in V, as follows:

$$[(2.55 \times 10^{-6})TL + .0009]V^3 + (.003n + .00015TL)V^2 + [(.011 + .065s)TL + (.21 + 2107s)n]V - HP = 0 \quad (21)$$

The roots of this equation can be found using standard formulae. For the values TL = 5000, n = 4, s =0, HP = 8200, this equation has one real root and two conjugate imaginary roots. The real root is approximately V =74. Thus, on the level track segments, the predicted attainable velocity is 74 miles per hour. In practice however, due to gearing, the locomotives may not actually be able to run this fast. In addition, there may be speed restrictions on the line, so that the expected velocity of the moving train would be given by:

$$V_a = \min[V_t, V_g, V_{TL}] \quad (22)$$

where: V_a = expected velocity

V_t = speed restriction on line imposed by timetable

V_g = maximum achievable locomotive speed

V_{TL} = speed attainable with given trailing load .

For the purposes of this example, let us assume a speed limit of 60 miles per hour along the line, so $V_a = 60$ for the level segments, AB and DE.

For the segment CD, we insert values S = 0.5, TL = 5000, n = 4, HP = 8200 into equation (21). Once again, there is one real root and two imaginary roots. The real root is approximately V = 50.

Table 1 summarizes the results of the computations, and indicates the overall predicted running time.

Table 1. Summary results for example.

Segment	Length (mi.)	V_{TL} (mph)	V_a (mph)	Running Time (hrs.)
AB	20	74	60	0.33
BC	5	20	20	0.25
CD	10	50	50	0.20
DE	15	74	60	0.25
TOTAL	50			1.03

2.4.4. Modeling Delays Enroute

Total origin-destination time for a train is not generally composed of running time alone. In fact, there is always some pre-departure time at the origin yard and some post-arrival time at the destination which must be recognized. However, in addition, there are often delays enroute due to switching and/or interactions (meets or passes) between trains.

Trains are often delayed enroute due to passing or being passed by other trains going in the same direction, or on single track mainline, meeting trains going in the opposite direction. Detailed simulation models are often used by railroads to evaluate train congestion. However, for the purposes of this project, it is desirable to have a simpler, analytic model which can be incorporated more readily into the specification of a production function for cost estimation. The model proposed here draws heavily on work done by Petersen [66]. It is also similar to methods of analysis for low density highway traffic with passing delays (see, for example, Haight [39]).

The situation to be considered is illustrated in Figure 10, showing two yards, A and B, connected by a main line track segment which may be either single or double track.

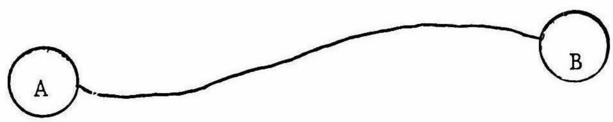


Figure 10. TWO YARDS CONNECTED BY MAINLINE.

Classes of trains in each direction will be defined by different average running times (or speeds). In practice, of course, each train will have a somewhat different running time, as determined by the model in the previous section. Some aggregation will normally take place, since we are interested in identifying classes of trains. For example, we may have local freights at an average of 20 miles per hour, regular through freights averaging 40 miles per hour, and special high priority trains averaging 60 miles per hour.

If we consider ourselves to be located at A, we will assume that there are K different inbound train (speed) classes, and L different outbound classes. We will also adopt the convention that outbound speeds are negative, for algebraic convenience.

Define M_{ij} as the expected number of encounters (meets or overtakes) between a single train of class i and all trains of class j, on its trip between yards. If we assume that each encounter results in an average delay, D_{ij}, to train i, we can write the expression for average transit time, including delays, as follows:

$$W_i = R_i + S_i + \sum_j D_{ij} M_{ij} \tag{23}$$

where W_i = average total transit time for train class i

R_i = average running time for train class i

S_i = average delays enroute from all other occurences.

The values for R_i are determined from the line-haul train movement model described in Section 2.4.2.

In order to utilize the model in equation (23), the expected number of encounters between trains, M_{ij}, must be expressed in terms of quantities available. These include traffic density of trains of different classes, their speeds, and dispatching policies through time. As an example of the derivation, let us consider the expected number of meets between trains going in opposite directions.

If an outbound train of speed i leaves at time t=0, it will encounter inbound trains already on the line at t=0, and those dispatched before the outbound train arrives at the other end, at time $t=W_i$. For inbound trains of speed j, this will include all trains dispatched between $t= -W_j$ and $t=W_i$.

If we assume, as Petersen did, that train departures from either end of the line are independent and uniformly distributed through time, the expected number of meets is

$$M_{ij} = N_j(W_i + W_j) \tag{24}$$

where N_j = rate of dispatching of train class j (trains/unit time).

In a similar manner, we can derive the result that train i will overtake trains of a slower class, j, that depart between $t = -(W_j - W_i)$ and t=0. Furthermore, train i will be overtaken by faster trains, j, which depart between t=0 and $t = W_i - W_j$. If we then let I be the set of inbound train classes, 0_f of the set of outbound train classes of higher speed than train i and 0_s the set of outbound train classes of lower speed than i, we can write equation (23) as

$$W_i = R_i + S_i + \sum_{j\varepsilon I} D_{ij}N_j(W_i + W_j) + \sum_{j\varepsilon 0_s} D_{ij}N_j(W_j - W_i)$$

$$+ \sum_{j\varepsilon 0_f} D_{ij}N_j(W_i - W_j). \tag{25}$$

Since we have an equation of this type for each speed class, both inbound and outbound, this defines a set of K+L simultaneous linear equations that can be solved for the K+L unknowns, W_i. Of course, if the line under study is double track the term delay due to meets vanishes.

The model described above is essentially that developed by Petersen [66]. Several extensions of this basic model are possible. English [27] has made modifications so that it reflects operations on high density lines more accurately. These modifications account for multiple meets and delays induced by signal systems in very high density operations. For lower density operations typical of most lines, an extension is possible to account for the fact that trains are often not dispatched at random times with a constant mean interdispatch time throughout the day. Such a modification is described in the following section.

2.4.4.1 Extension to Other Dispatching Strategies

The model described above assumes that trains are dispatched independently, at random times with a constant rate throughout the day. The implication of this assumption is that times between successive trains will be exponentially distributed. Thus the probability density function is given by f(t):

$$f(t) = \lambda e^{-\lambda t}, \ t > 0 \qquad (26)$$

where t = interdispatch time

λ = average rate of dispatching (=1/mean time between trains).

In many situations however, train dispatches can be scheduled somewhat more regularly, and line-haul delays due to meets can be reduced. In such cases, the times between trains must be characterized by a more general probability distribution. A useful generalization is to characterize these times by the Erlang-k distribution. By varying the value of k, a wide range of distributions can be represented. When k=1, this distribution is equivalent to the exponential model. As k increases, the variance of the distribution decreases, reflecting more regular dispatches. In the limit, as $k \to \infty$, the distribution becomes a spike at a given value, reflecting constant times between dispatches of trains. A few members of this family are illustrated in Figure 11.

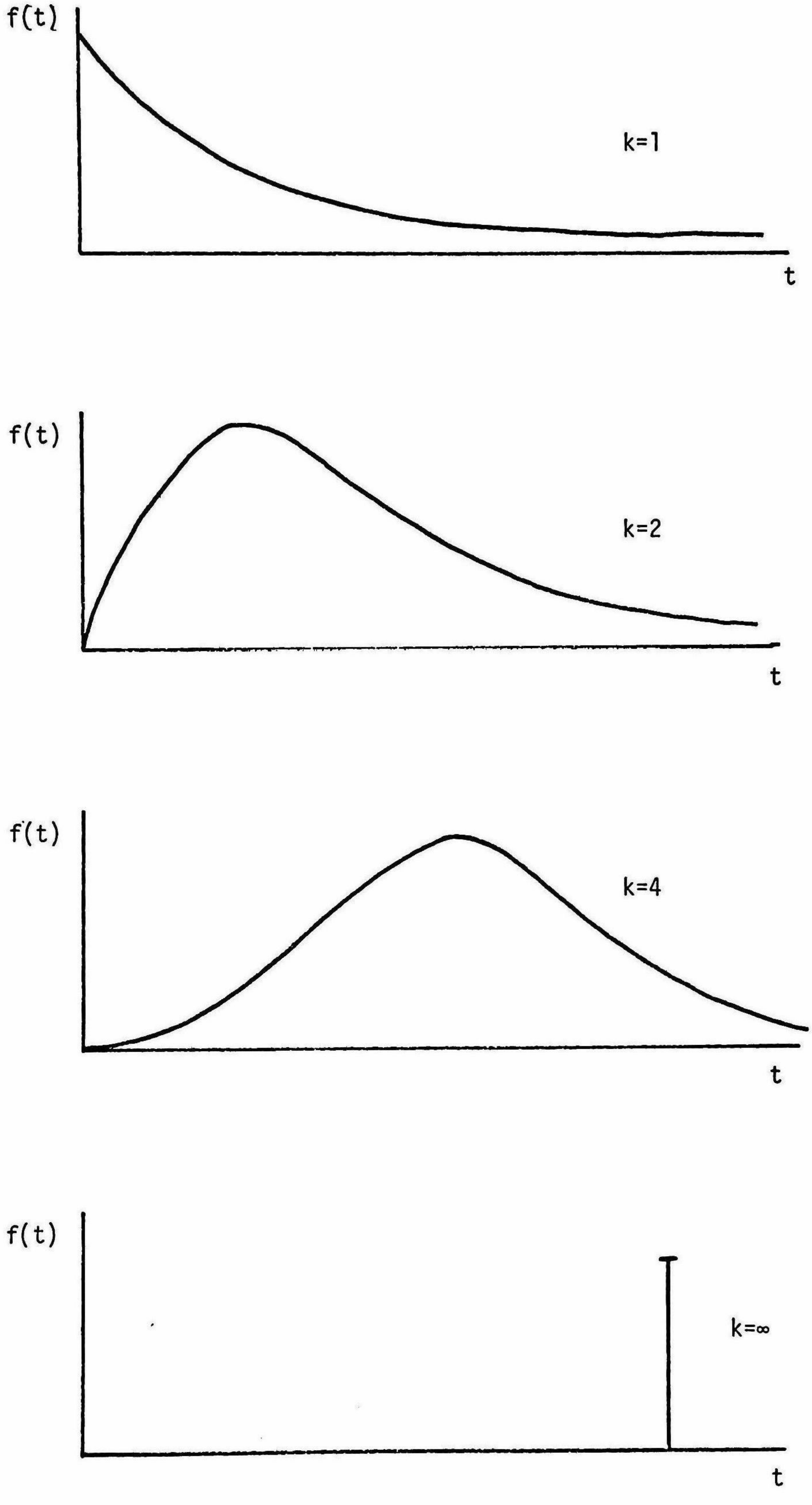


Figure 11. EXAMPLES OF ERLANG DISTRIBUTIONS FOR DIFFERENT VALUES OF k.

The general form of the probability density function for an Erlang-k random variable is:

$$f(t) = \frac{\lambda(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} \quad . \tag{27}$$

If interdispatch times of train class j are distributed Erlang-k, we can derive the expected number of encounters of a train of class i with all trains of class j as follows. If we denote the number of dispatches in a given period as a random variable Y_j, the probability of observing X dispatches in a period of length W is:

$$P(Y_j = X) = \sum_{i=0}^{k-1} \frac{e^{-\lambda W}(\lambda W)^{Xk+i}}{(Xk+i)!} \quad . \tag{28}$$

This result arises from considering an Erlang-k random variable to be a sum of k exponential random variables with common parameter, λ. Thus, the probability of observing exactly X occurrences of the event described by the Erlang distribution is the probability of observing between kX and k(X+1)-1 fundamental exponentially distributed events. This is the probability represented by the sum in equation (28).

Equation (28) defines the probability mass function for the number of encounters of a single train i with all trains of class j. The expected number of these encounters is then

$$E(Y_j) = \sum_{X=0}^{\infty} XP(Y_j = X) \quad . \tag{29}$$

If we define a function p(z,t) as

$$p(z,t) = \sum_{i=z}^{\infty} \frac{e^{-t} t^{i}}{i!} \tag{30}$$

equation (28) can be rewritten as follows:

$$P(Y_j = X) = p(Xk,\lambda W) - p[(X+1)k, \lambda W] . \quad (31)$$

Equation (29) then becomes

$$E(Y_j) = \sum_{X=0}^{\infty} p(k,\lambda W) - \sum_{X=0}^{\infty} Xp[(X+1)k,\lambda W]$$

$$= \sum_{X=1}^{\infty} p(Xk,\lambda W) \quad . \quad (32)$$

Values of the function p(z,t) are tabled (see, for example, Molinas [61].)

The value for $E(Y_j)$ can be substituted into equation (23) to replace the expression for M_{ij} given in equation (24). The basic delay model thus becomes

$$W_i = R_i + S_i + \sum_j D_{ij} \sum_{X=1}^{\infty} p[Xk, (W_i + W_j)] \quad . \quad (33)$$

Equation (33) defines a set of K+L non-linear equations in the K+L unknowns, W_I. These must be solved using iterative solution techniques, but they can be used to provide a more general solution for line-haul delays.

2.4.5. Models of Yard Activity

According to data gathered by Reebie Associates [69] the average rail car spends only 16% of its time actually moving in trains. An additional 56% is spent in classification yards. This underscores the importance of representing classification yard activities if we are to reflect railroad operations with any reasonable degree of accuracy.

While in a railyard, a car undergoes four basic operations:

1) inbound inspection
2) classification
3) assembly into outbound train
4) outbound inspection.

It is quite natural to think of these as a series of queues through which the rail car passes. This perspective is adopted here.

Inbound and outbound inspections consume a relatively small amount of time for each car, and the amounts of time required are not highly variable. For these reasons, they are not analyzed in detail here. However, explicit queuing models have been constructed for the remaining elements: classification and assembly. Average time in the yard is predicted as the sum of delays due to classification and assembly, as shown in equation (34):

$$T_y = T_c + T_a \qquad (34)$$

where T_y = average time in yard

T_c = average delay for classification

T_a = average connection delay before assembly into outbound train.

2.4.5.1 Classification Delay

There are a number of different queuing models which could be suggested for the classification operation. The major previous work along these lines has been done by Petersen [67].

He suggests several possible models, including:

M/G/1: Poisson arrivals of cars on trains, a general service time distribution, and one server;

M/M/s: Poisson arrivals, exponential service times, and s servers;

M/D/s: Poisson arrivals, deterministic (constant) service times, and s servers.

It should be noted that Petersen considers the basic units of arrival to the system to be trains, not individual cars, and thus he derives parameters for service time to classify an entire inbound train. While this simplifies representation of some elements of the system, it leads to some confusion about the relationship of the output process of one queue to the input of another. For this reason, the models developed here are based on individual railcars.

As a result, we should recognize the fact that individual railcars arrive in batches on trains. This fact dictates use of a more general batch arrival model, denoted as:

$M^{(X)}/G/1$: Poisson arrivals in batches of size, X; arbitrary service times; and 1 server.

In this case, X is a random variable corresponding to train length. A solution for such a model, yielding average delay time, has been developed by Gaver [34]. A concise summary of the results is available in Saaty [71]. Average delay (time in queue plus service) is given as follows:

$$T_c = \left[\frac{\rho}{2(1-\rho)} \left(\frac{\delta_2}{\delta_1} + \mu^2\sigma^2 \right) + 1 \right] \Big/ \mu \qquad (35)$$

where δ_1 = average train length (cars)

δ_2 = second moment of train length

$\rho = \tilde{\lambda}\delta_1/\mu$ = traffic intensity of system

$\tilde{\lambda}$ = arrival rate of trains (trains/hr.)

μ = average service rate (cars/hr.)

σ^2 = variance of service time distribution.

The distribution of service times for classifying cars depends greatly on the physical layout and operating plan of a particular yard. Probably the most important distinction is between hump yards and flat yards. In hump yards, the classification service is quite straight-forward. A switch engine pushes the string of cars over the hump at essentially a constant speed (generally 1.5- 2.0 miles per hour). As each car reaches the hump crest, it is decoupled and rolls down into the classification bowl. The only variations in time per car are due to variations in length of individual cars. Such variations are relatively minor, and a deterministic service time distribution is an appropriate model. In this case, $\sigma^2 = 0$ in equation (35), and the model can be denoted $M^{(X)} / D / 1$.

Models for flat yards are somewhat more complex, since the switching operations for classifying trains are not as simple as hump operation. An inbound train comprises a set of cuts (groups of cars with common origin and destination that move together through the yard) that are to be sorted into outbound "blocks" on classification tracks. Each cut will be switched as a unit, and if successive cuts are to be placed in the same block, one switching operation can handle multiple cuts. Thus, rigorous derivation of a service time distribution for individual cars would require incorporation of the distribution of number of cars per cut and the likelihood of successive cuts having common block designations, as well as the distribution of time to complete a particular classification switch.

Since good empirical data were not available on all of these characteristics, our approach has been somewhat less detailed. A sample of flat switching operations were observed in one yard, and the total time required and number of cars switched were recorded. For each of these observations an equivalent "minutes/car" value was then computed. Finally, a gamma distribution was fit to these values. The probability density function of a gamma distribution with parameters α and β is as follows:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad 0 \leq x < \infty . \qquad (36)$$

The mean value of the gamma random variable, x, is α/β and the variance is α/β^2.

Maximum likelihood estimates of α and β were computed as 1.3 and .28, respectively, using the observed data. This corresponds to a mean switching time of 4.6 minutes/car, and a variance of 16.6 minutes^2/car. Previously reported estimates for average switch-engine-minutes/car have varied widely, depending upon the number of cars per cut and the degree of congestion present in the yard facility. For example, Wright [84] reported estimates of 3.2 minutes/car for single car cuts, but average values below 2 minutes/car for multiple car cuts. Martland and Rennicke [59] report average values from 3 to 10 minutes/car for different levels of workload in two yards on the Boston and Maine. Thus, it appears that our estimate is well within the range of plausible values.

Values for variance of the switching time have not been widely reported, so it is difficult to verify our estimates based on previous results. However, the wide range in previously reported average values tends to support the contention that the process is highly variable. Thus, our finding that the service time process is nearly exponential is not surprising (note: an exponential service distribution would have α =1.0). In general, it appears that our estimates of parameters of the service time distribution for flat switching are quite consistent with earlier reported results.

As an example of the values produced by the model, data from one yard studied show an average arrival rate of 5.33 trains/day, with an average length of 45.2 cars and a second moment of length equal to 2876. Combined with the estimated service time parameters, these data result in an estimated utilization rate of .77. Substituting values into equation (35), we obtain a mean delay for classification of 8.2 hours.

Available data for the yard under study did not include detail on time spent waiting for classification, connection delay, etc. As a result, it is difficult to verify this model directly. However, the predicted delay of 8.2 hours is well within the range of observed data presented by Folk [28], Beckmann, et al [3], and Gentzel [35] for various terminal facilities. Values for mean classification delay between 4.6 and 22.4 hours have been reported by these authors for different yards at different times.

2.4.5.2 Assembly into Outbound Trains

Once cars have been classified, they must wait for assembly and dispatch of the appropriate outbound train before they leave the yard. Operationally.

we can think of this process as being one in which cars arrive on the classification tracks, either singly or in small groups (cuts), and wait for the designated outbound train to be "called." At this point, all the cars for this train are assembled, and when made up, the train departs. In terms of a queuing model, we may think of this as a batch-service system in which the "server" is the outbound train. Service for a batch of cars begins when the appropriate outbound train is called for assembly, and the service time is the time between successive outbound trains on which a given cut of cars may be dispatched. The delay time for connection with the outbound train is then the waiting time in queue derived from such a queuing model.

It should be noted that this perspective on modeling the system places principal emphasis on the outbound train schedule as the source of delay for cars following classification. Delays in assembly due to insufficient numbers of switch engines and crews are not considered directly. This effect is only represented indirectly, in terms of late departures of outbound trains, for example. The emphasis on schedule is in keeping with the findings of several previous researchers, and has been recognized by a rail industry task force on reliability studies [29].

The average delay for a simple batch-service queue of this type can be derived easily. Let us assume that individual cars arrive randomly in time (i.e. as a Poisson process) from the classification operation, and that the outbound train takes all cars available at the time it is assembled. This second assumption means that train length constraints on the outbound trains are ignored, for the time being. We will return to this issue following the basic derivation.

Define a probability density function, $g(t)$, $0 \leq t < \infty$, which describes the distribution of time intervals between successive outbound trains for a given block of cars. If cars arrive randomly in time on the classification tracks, and the interval between two trains is a particular value, t_o, the average delay time will be

$$T_a(t_o) = \frac{t_o}{2} \; . \tag{37}$$

The expected number of cars arriving during any interval is proportional to the length of that interval. That is, the expected number of cars arriving in an interval of length t_o is

$$n(t_o) = k \cdot t_o \tag{38}$$

where k is the arrival rate. The total expected delay time for all cars in an interval of length t_o will then be

$$S(t_o) = n(t_o) \cdot T_a(t_o) \; . \tag{39}$$

The unconditional total expected wait time may be obtained by integrating over the density function, $g(t)$:

$$S = \int_o^\infty S(t)g(t)dt \; . \tag{40}$$

In like fashion, the unconditional expected number of cars in an interval is

$$n = \int_o^\infty n(t)g(t)dt \; . \tag{41}$$

Finally, the unconditional expected delay for cars is simply

$$T_a = \frac{S}{n} = \frac{\frac{k}{2}\int_o^\infty t^2 g(t)dt}{k\int_o^\infty t\, g(t)dt} = \frac{E(t^2)}{2E(t)} \quad . \tag{42}$$

If desired, equation (42) may be rewritten as

$$T_a = \frac{E(t)}{2} + \frac{\sigma_t^{\;2}}{2E(t)} \tag{43}$$

where $\sigma_t^{\;2}$ is the variance in the time interval between successive departures. Note that if departures are completely regular ($\sigma_t^{\;2} = 0$), the second term vanishes, and the expected delay is one-half the interval between trains (e.g., 12 hours for trains dispatched once per day.) On the other hand, if dispatches occur very irregularly, the second term indicates that expected delay to cars will increase.

Equation (42) is analagous to a result widely used in studies of urban mass transit systems, expressing the mean waiting time of passengers at a bus stop. Derivations of the result in the mass transit context can be found in Welding [82], Osuna and Newell [63] or Kulash [51].

The derivation of equation (42) assumed that outbound train length is unlimited, or in queuing terms, that that batch size is infinite. In practical terms, this assumption is not really true, since there are limits to the length of train which can be dispatched. Such limits can be the result of mainline track configuration, power availability, etc. More sophisticated batch-service queuing models can be constructed to reflect these constraints, but for batch sizes in excess of 25-30, the numerical results are essentially the same as for infinite batch size. (See Petersen [65].) Since train length constraints would typically be well in excess of these values, use of a simpler, infinite-batch-size model is appropriate.

Using the queuing models for classification delay and connection/ assembly delay described in the previous two sections, we can predict total delay in the yard by simply summing the results from equations (35) and (42), as indicated in equation (34).

2.4.6 Estimation of Average Shipment Velocity

Together, the line-haul train movement and classification yard models provide the means for estimating the average speed of a shipment through the system. This value will be used as a single index of service quality in the cost model to be estimated.

From the running time and delay models described in section 2.4.2 and 2.4.4 average transit time for trains over each mainline track segment can be computed. Since we know the length of each segment, this can be converted to an average velocity. By aggregating over track segments, an overall average velocity of trains is determined. Let us denote this velocity, V_a.

The classification yard model predicts total delay (in hours) to cars passing through a yard. We have denoted this delay by T_y, as shown in equation (34). Since there is effectively no distance involved in this segment of the trip, however, this time value is not directly expressable as a velocity.

To obtain an overall velocity figure, we require one additional piece of information, the average length of haul. We can then obtain average velocity (miles per hour) by dividing average length of haul (miles) by average total time in system (hours moving in trains and waiting in yard.) If we denote average length of haul by L, overall average velocity, $\bar{V}$, is computed as shown in equation (44):

$$\bar{V} = \frac{L}{\frac{L}{V_a} + T_y} = \frac{LV_a}{L + V_a T_y} \quad . \tag{44}$$

Equation (44) reflects two major simplifying assumptions which are justified by the uncomplicated nature of operations on the railroad under study. The first of these is that each shipment passes through one classification yard. This is essentially accurate for the system examined in the case study, but would require modification for more complicated rail operations. Secondly, in equation (44) average time spent in trains is computed as L/V_a, rather than by observing which line segments would be crossed by a given shipment, summing those transit times, and then computing a weighted average based on relative frequency of various origin-destination pairs. Again, the simpler computation used in equation (44) is a reflection of the simplicity of the rail network under study and the relatively limited set of origin-destination pairs. For this case, the simpler computation is quite adequate, but it would have to be modified in a more complex setting.

CHAPTER 3

DESCRIPTION OF THE DATA FOR RAILROAD X

3.1. General Overview of Data Needs

The analysis of the previous chapter provides the following list of data items needed by our model:

1) Flows of various commodity types-

$$(y_1,\ldots,y_{m_1});$$

2) Prices of input factors-

$$(p_1^v,\ldots,p_{n_1}^v);$$

3) Levels of fixed factors -

$$(x_1^f,\ldots,x_{n-n_1}^f);$$

4) Levels of variable factors -

$$(x_1^v,\ldots,x_{n_1}^v);$$

5) Engineering Data ;

6) Cost - C .

Each of these areas will be addressed in turn. To provide a specific cost function for the discussion, chapter four will present a translog model of the cost function

$$C(Y,S,P_1,P_2,P_3,P_4,P_5;QK)$$

where 1) Y is total flow (loaded car-miles); 2) S is speed; 3) P_i is the price of cars, fuel, crews, locomotives, and non-crew labor respectively; 4) QK is a quality of plant representing the fixed factors. Appendix A presents the estimation results for the translog model of $C(Y_1,Y_2,Y_3,Y_4,S,P_1,P_2,P_3,P_4,P_5;QK)$ where the Y_1 are subaggregates with respect to commodity type. Observations for both models are monthly observations from 1969 through 1977 (108 observations).

3.2 Flow

Historically, ton-miles has been used as the measure of output of a transport firm. This measure suffers in many ways:

1. Shippers do not often buy tons of a commodity moved; they usually buy in car-loads, which can vary in weight by the type of commodity.
2. A ton-mile can be misleading: is 100 ton-miles the movement of 100 tons over 1 mile or the movement of 1 ton over 100 miles or something in between? These outputs are clearly not the same.
3. The cost in providing service includes the movement of empty cars to be repositioned so as to be available to make a revenue-generating move. Thus empty car movements are an intermediate product and <u>not</u> a final output of the firm.

Flow data from the firm was of two sorts:

1. Monthly listings by type-of-move (see below) by seven-digit STCC (Standard Transportation Commodity Code) of total loaded cars moved and total tons, from 1969 to 1977.

2. Two years (1972, 1977) of records by type of move (see below) of every move made on the system: origination, destination, commodity. We will refer to these as distance profiles.

Because shippers basically buy loaded cars rather than tons, we decided to use this measure. Originally we planned to have as disaggregate a model as possible, allowing for flows by line-haul segment and direction. This became impossible to compute and counter-productive to the basic goal of producing an integrated model. Since, in a translog model, log(ab) = log a + log b, then if we included miles with a move we would implicitly have:

$$\log(\text{loaded cars}) + \log(\text{miles}) = \log(\text{loaded car-miles}).$$

Therefore we proceeded to use loaded car-miles as the output measure. The construction of the loaded car-miles was based on using the distance profile information to get an average distance traveled by commodity and by type-of-move. There are four types of move: local (L), forwarded (F), received (R), and intermediate (I). They are defined as follows:

L: origin on-line, destination on-line

F: origin on-line, destination off-line

R: origin off-line, destination on-line

I: origin off-line, destination off-line.

The model presented in the next chapter uses total loaded car-miles to represent flow; a model using one possible disaggregation of flow is presented in Appendix A. Many such representations of flow are possible allowing not only for disaggregate commodity types but also for distinctions such as unit train, etc. Issues of this type will be investigated in the second phase of the work (see section 6.2).

3.3. Prices of Variable Factors

Prices were constructed for the following factors:

1) Cars;
2) Rail;
3) Ties;
4) Fuel;
5) Locomotives;
6) Train Crews; and
7) Non-Crew Labor.

These will be discussed in turn below. It should be remembered that prices are the marginal cost of another unit of the factor in question. As such they should be constructed from national or regional market data. Since most such data for capital items (cars, rail, ties, locomotives) is cost of purchase, we used the interest rate for the firm, which was provided by the firm's main bank. The rate, while evidencing some fluctuation, appears to have been reasonably stable during the period of estimation (1969-1977) given the nature of the economy.

Thus, for capital items the following price construction was used to provide monthly prices:

$$P_{it} = (r_t + \delta_i) \text{ Unit Cost}_{i,t}/12$$

where r_t is the nominal interest rate for year t, δ_i is a depreciation rate, unit cost i,t is the unit cost of factor i in year t (when data was found indicating changes in costs during the year, the change was used to split the year and associate unit costs with months). Of course, this leads to some uniformity of some of the capital prices during a year.

3.3.1. Price of Cars

Lease information on cars from car-leasing concerns was unavailable for most of the period studied. The firm's car leases were used to construct yearly profiles of numbers and types of cars. Average costs [83] of new freight-train cars installed in the years 1959-1978 by type of car were used in conjunction with the yearly car profiles to construct a unit-cost for a representative car, i.e. a car representing the mix of existing stock at each point in time. A depreciation rate of six percent [78] was used.

3.3.2. Price of Rail and Ties

Data from the Association of American Railroads on per-ton rail costs and per-tie tie costs was used. Turnover rates on the railroad under study established depreciation rates of .02 and .03 respectively. It was found that the prices were almost perfectly correlated (.987) and thus the price of ties was dropped. Later analysis proved that the price of rail was correlated with the prices of cars and locomotives (both > .9). Dropping the price of rail removed some serious multicollinearity in the model in that almost all other variables had much lower correlations.

3.3.3. Price of Fuel

Firm records provide monthly purchases of fuel. Price was taken as amount paid over quantity purchased.

3.3.4. Locomotives

Firm records were again used to provide month-by-month profiles of types of locomotives and numbers in use. The firm has only recently started renting locomotives. Data from ICC Transport Statistics in the U.S., Table 37 (23 in 1974) was used to provide average costs, by type of locomotive. Again, a composite locomotive unit cost was constructed based on the locomotive mix at each point in time. Missing data (the ICC has stopped issuing the table apparently) was filled by using the few leases the firm did have, which happened to be in the missing data years. The depreciation rate was again .06 [78].

3.3.5. Labor: Crew and Non-Crew Costs

Firm records on monthly payments to various categories of labor were used. Crews were taken as a unit and all other labor (executives on-down) were taken as the unit "non-crew". Prices we computed by taking total hours paid for and dividing by total hours actually worked. This is important since credit is often given for time not actually worked. Per hour supplemental wage payments were added into the wage rates to provide final wage rates (prices).

3.3.6. Deflation of Nominal Prices

The prices given above are nominal prices. They were deflated to 1969 by use of the AAR Charge-Out Indices [85]. Fuel was deflated by the fuel index, cars and locomotives by the materials index and labor by the labor index.

3.4. Fixed Factor Levels

The fixed factor is the system configuration. In the case of railroad X, a system configuration change occurred in 1976. The change was a simple one involving the addition of a stretch of track which had previously been used by another firm to ship goods onto railroad X's system. A number of possible measures of the system configuration are possible. Since railroad X consists mainly of line-haul, we elected to measure the system quality and configuration via a vector of four variables representing the number of miles in each of the Federal Railroad Administration's Track Classification categories Clearly, for a given configuration, the elements of such a vector are correlated. Thus the vector was converted into a scalar measure:

$$\text{Track Quality} = \frac{\text{miles in category four}}{\text{total miles}}$$

where category four represents the best quality. Because the FRA classifications have associated speed limits, and these speed limits affect the speed of a shipment over the system, this index of system quality is a very effective measure of the fixed factor (though clearly not the only one possible).

3.5. Data Used in the Engineering Analysis

The data used in estimation of the short-run cost functions for the railroad under study include information on the physical characteristics of the railroad's lines and yards and operating records of train and car movements.

3.5.1. Physical Characteristics of the Rail Plant

Detailed information on the nature of the rail plant is used in three specific places in the model. First, the track profile (grades) and speed limits have been used in computing running times for trains over various segments of mainline track, as described in section 2.4.2 . Second, the number and locations of passing sidings are used in the calculation of delays on single-track lines, as described in section 2.4.4 . All of this information on the physical nature of the plant was obtained from track charts and annotations supplied by the Vice President - Operations of the railroad studied.

3.5.2. Operating Records of Train Movement

Records of train movements were sampled for each month of the study period (January, 1969 - December, 1977) in order to estimate numbers of trains operated over each major line segment, and characteristics of those trains including number of cars, locomotive horsepower and total trailing load (in tons). The number of trains operated is an important input to the calculation of line-haul delays, as described in section 2.4.4. The horsepower and total trailing load are important determinants of line-haul train velocity and are also used in line-haul delay calculations. The number of trains and their lengths (in cars) are important input for the classification yard delay models.

All of this information was obtained from dispatchers' records of train movements. These are generally large sheets in which all movements of trains on a given day over a particular section of track are noted. These sheets include a good deal of information in addition to the data we required, and constitute the most detailed records retained regarding train movements. Because the records comprise many individual entries made manually by dispatchers through the day, and because there is one sheet for each day's operation, extraction of the relevant information was a laborious, time-consuming process.

A sampling procedure was devised to allow extraction of a reasonable statistical sample of this data for each month of the study period. This sampling procedure involved taking detailed information on 8-12 trains per month. This detailed data included origin and destination of run, total horsepower, numbers of loaded and empty cars, trailing tons and delays encountered. Sample trains were selected so as to cover all directions of movement and various days of the week, in order to avoid obvious biases.

The sample data for each month was then aggregated to obtain characteristics of a "typical" train for that month over each line-haul segment. The horsepower and trailing load for this typical train were then used to compute average line-haul velocity for a shipment during that month.

In addition to this sampling of detailed train movement information, exhaustive samples of numbers and lengths of inbound trains to yard facilities for classification were recorded for 15 days per month. This provided a sufficient sample to estimate the arrival rate of trains and the first two moments of the train length distribution for each month. These three values were then used in the formula (35) to estimate classification delay in yard operations.

The number and departure times of trains sampled from the train movement records were also used to construct estimates of the mean and variance of times between successive outbound trains. This is information needed to compute connection/assembly delay times in classification yards.

In summary, the bulk of the basic operating data for the engineering process models has come from two major sources. The first is track charts providing the physical characteristics of the mainline track segments. The second is dispatchers' records of train movement. These records include detail on train characteristics and movement from which input values for the line-haul and yard models can be derived.

3.6 Cost

Monthly records from 1969 through 1977 on operating costs were provided by the firm. These records are used as the basis for ICC submissions. In general, however, such costs do not include implicit capital costs on cars and locomotives, i.e. the costs did not reflect the economic costs of the two major short-run variable capital factors. An estimate of the missing costs was made by using the car and locomotive prices and levels. These costs were then added to the operating cost to provide short-run variable cost. Data on credits, committed funds for leases, etc. were purposely excluded. In general, one would expect that the main part of property-related taxes would be assessed on the firm's plant (rather than equipment) and since this is fixed in our model, we do not include such taxes in the short-run variable costs. Income taxes are on profits, and thus are also excluded from short-run variable costs. This is because the cost function is homogeneous of degree one in prices (see section B.2.1). Thus, the short-run variable costs cover such items as maintenance, fuel, crews cars, locomotives, staff, supplies, etc.

Costs were deflated by using the AAR charge-out index (aggregate). Autocorrelation analyses indicated that deflating the costs removed most of the autocorrelation present in the cost observations. A weak yearly auto-correlation persisted, but was small enough that we felt it could be ignored in the econometric estimation.

CHAPTER 4

THE MODEL TO BE ESTIMATED AND THE ESTIMATION RESULTS

4.1 The Model

The following model was estimated:

COST:

$$C = \alpha_0 + \alpha_{10}\ PCAR + \alpha_{20}\ PFUEL + \alpha_{30}\ PCREW + \alpha_{40}\ PLOCO + \alpha_{50}\ PMNGT$$

$$+ \beta_{10}\ Y + \gamma_{10}\ S + \delta_{10}\ QK$$

$$+ \frac{1}{2}\alpha_{11}\ (PCAR)^2 + \alpha_{12}\ PCAR \cdot PFUEL + \alpha_{13}\ PCAR \cdot PCREW + \alpha_{14}\ PCAR \cdot PLOCO + \alpha_{15}\ PCAR \cdot PMNGT$$

$$+ \frac{1}{2}\alpha_{22}\ (PFUEL)^2 + \alpha_{23}\ PFUEL \cdot PCREW + \alpha_{24}\ PFUEL \cdot PLOCO + \alpha_{25}\ PFUEL \cdot PMNGT$$

$$+ \frac{1}{2}\alpha_{33}\ (PCREW)^2 + \alpha_{34}\ PCREW \cdot PLOCO + \alpha_{35}\ PCREW \cdot PMNGT$$

$$+ \frac{1}{2}\alpha_{44}\ (PLOCO)^2 + \alpha_{45}\ PLOCO \cdot PMNGT$$

$$+ \frac{1}{2}\alpha_{55}\ (PMNGT)^2$$

$$+ \frac{1}{2}\beta_{11}(Y)^2 + \frac{1}{2}\gamma_{11}(S)^2 + \frac{1}{2}\tau_{11}\ Y \cdot S$$

$$+ \theta_{11}\ PCAR \cdot Y + \theta_{21}\ PFUEL \cdot Y + \theta_{31}\ PCREW \cdot Y + \theta_{41}\ PLOCO \cdot Y + \theta_{51}\ PMNGT \cdot Y$$

$$+ \sigma_{11}\ PCAR \cdot S + \sigma_{21}\ PFUELS \cdot S + \sigma_{31}\ PCREW \cdot S + \sigma_{41}\ PLOCO \cdot S + \sigma_{51}\ PMNGT \cdot S$$

$$+ \delta_{11}(QK)^2 + \mu_{11}QK \cdot Y + \varepsilon_{11}QK \cdot S$$

$$+ \eta_{11}PCAR \cdot QK + \eta_{21}\ PFUEL \cdot QK + \eta_{31}\ PCREW \cdot QK$$

$$+ \eta_{41}\ PLOCO \cdot QK + \eta_{51}\ PMNGT \cdot QK$$

where: C = ℓn (Cost/Average Cost)

PCAR = ℓn (Price of Cars/Average Price of Cars)

PFUEL = ℓn (Price of Fuel/Average Price of Fuel)

PCREW = ℓn (Price of Crews/Average Price of Crews)

PLOCO = ℓn (Price of Locos/Average Price of Locos)

PMNGT = ℓn (Price of Non-crews/Average Price of Non-crews)

Y = ℓn (Loaded Car-miles/Average Loaded Car-miles)

S = ℓn (Speed/Average Speed)

QK = ℓn (FRA Category Four Percentage/Average FRA Category Four Percentage)

There are five prices, two outputs and one fixed factor thus resulting in forty -five coefficients to be computed (see section B.2.4.1). To improve the efficiency of the estimation process we will append the following factor share equations:

FUEL:

$$XFUEL = \alpha_{20} + \alpha_{12}\ PCAR + \alpha_{22}\ PFUEL + \alpha_{23}\ PCREW + \alpha_{24}\ PLOCO + \alpha_{25}\ PMNGT + \theta_{21}\ Y + \sigma_{21}\ S + \eta_{21}\ QK$$

CREWS:

$$XCREW = \alpha_{30} + \alpha_{13}\ PCAR + \alpha_{23}\ PFUEL + \alpha_{33}\ PCREW + \alpha_{43}\ PLOCO + \alpha_{53}\ PMNGT + \theta_{31}\ Y + \sigma_{31}\ S + \eta_{31}\ QK$$

LOCOS:

$$XLOCO = \alpha_{40} + \alpha_{14}\ PCAR + \alpha_{24}\ PFUEL + \alpha_{34}\ PCREW + \alpha_{44}\ PLOCO + \alpha_{54}\ PMNGT + \theta_{41}\ Y + \sigma_{41}\ S + \eta_{41}\ QK$$

NON-CREWS:

$$XMNGT = \alpha_{50} + \alpha_{15}\ PCAR + \alpha_{25}\ PFUEL + \alpha_{35}\ PCREW + \alpha_{45}\ PLOCO + \alpha_{55}\ PMNGT + \theta_{51}\ Y + \sigma_{51}\ S + \eta_{51}\ QK$$

where: XFUEL = Price of fuel • fuel purchased/cost

XCREW = Price of crew hour • crew hours purchased/cost

XLOCO = Price of loco hour • locomotive hours purchased/cost

XMNGT = Price of non-crew hour • non-crew hours purchased/cost

By purchased, we mean hours or amounts paid for. This is particularly important with respect to labor and locomotives since not all time paid for is used.

As will be observed from the cost function description, we have divided all variables by their means, i.e. an observation is divided by the mean of the observations before taking the logarithm. This is done mainly to protect the proprietary nature of the data. By so transforming the variables we only affect the intercept term, leaving the important coefficients undisturbed This way, actual costs for the railroad under study are only predictable by those with a proprietary interest while cost relationships are open to perusal by all. In view of this, we will not be publishing the variables means since they add nothing to understanding the cost functions, and only reveal proprietary information.

4.2 Estimation Results

The above equations were estimated as a system of seemingly unrelated equations [79] wherc we assumed an additive error structure. Since the factor share equations are derived by differentiation of the cost function, the error term in the cost function does not appear in the factor share equations. As in [15] we assume the disturbances are joint normal and estimate the system using a maximum likelihood technique and thus the results are invariant to which factor share equation is dropped. Table 2 provides the estimated cost functions, while Tables 3, 4, 5, and 6 provide the estimated factor share equations for fuel, crews, locomotives and non-crew labor respectively.

Table 2

Cost Function

VARIABLE	COEFFICIENT	ESTIMATE	STD. ERROR
-	α_0	.03997	.01123
PCAR	α_{10}	.31748	.00494
PFUEL	α_{20}	.04767	.00086
PCREW	α_{30}	.15185	.00130
PLOCO	α_{40}	.08354	.00084
PMNGT	α_{50}	.39945	.00300
QK	δ_{10}	-.92323	.13530
Y	β_{10}	.08939	.07851
S	γ_{10}	-.04843	.05306
$(PCAR)^2$	α_{11}	-.02637	.02267
PCAR • PFUEL	α_{12}	.00526	.00628
PCAR • PCREW	α_{13}	.00216	.00766
PCAR • PLOCO	α_{14}	.04086	.00765
PCAR • PMNGT	α_{15}	-.02167	.01535
$(PFUEL)^2$	α_{22}	,05928	.01022
PFUEL• PCREW	α_{23}	-.01716	.00835
PFUEL•PLOCO	α_{24}	-.02293	.00706
PFUEL• PMNGT	α_{25}	-.02422	.00133
$(PCREW)^2$	α_{33}	.10596	.01404
PCREW•PLOCO	α_{34}	-.01861	.00729
PCREW•PMNGT	α_{35}	-.07234	.01491
$(PLOCO)^2$	α_{44}	.03863	.00936
PLOCO•PMNGT	α_{45}	-.03794	.01019
$(PMNGT)^2$	α_{55}	.15617	.02461

Table 2 (continued)

VARIABLE	COEFFICIENT	ESTIMATE	STD. ERROR
$(Y)^2$	β_{11}	.23904	.49667
$(S)^2$	γ_{11}	- .07679	.14596
Y·S	τ_{11}	- .21683	.20978
PCAR · Y	θ_{11}	.024151	.04496
PFUEL·Y	θ_{21}	.00451	.00800
PCREW·Y	θ_{31}	.01064	.01191
PLOCO·Y	θ_{41}	- .01131	.00777
PMNGT·Y	θ_{51}	- .02800	.02758
PCAR · S	σ_{21}	.01480	.01839
PFUEL·S	σ_{22}	- .00463	.00324
PCREW·S	σ_{23}	- .00381	.00487
PLOCO·S	σ_{24}	.00367	.00319
PMNGT·S	σ_{25}	- .01004	.01120
$(QK)^2$	δ_{11}	-12.84350	3.36250
QK·Y	μ_{11}	.81675	.78897
QK·S	ϵ_{11}	- .00451	.28319
PCAR·QK	η_{11}	- .21474	.08022
PFUEL·QK	η_{21}	.02840	.01436
PCREW·QK	η_{31}	.05061	.02123
PLOCO·QK	η_{41}	.00755	.01399
PMNGT·QK	η_{51}	.12819	.04957

Table 3

Fuel Function

VARIABLE	COEFFICIENT	ESTIMATE	STD. ERROR
	α_{20}	.04767	.00086
PCAR	α_{12}	.00503	.00628
PFUEL	α_{22}	.05928	.01022
PCREW	α_{23}	-.01716	.00835
PLOCO	α_{24}	-.02293	.00756
PMNGT	α_{25}	-.02422	.01332
Y	θ_{21}	.00451	.00800
S	σ_{22}	-.00463	.00324
QK	η_{21}	.02840	.01436

Table 4

Crews Function

VARIABLE	COEFFICIENT	ESTIMATE	STD. ERROR
	α_{30}	.15185	.00128
PCAR	α_{13}	.00216	.00766
PFUEL	α_{23}	-.01716	.00835
PCREW	α_{33}	.10596	.01404
PLOCO	α_{43}	-.01861	.00729
PMNGT	α_{53}	-.07234	.01491
Y	θ_{31}	.01064	.01191
S	σ_{23}	-.00381	.00487
QK	η_{31}	.05061	.02123

Table 5

Locos Function

VARIABLE	COEFFICIENT	ESTIMATE	STD. ERROR
	α_{40}	.08354	.00084
PCAR	α_{14}	.04086	.00765
PFUEL	α_{24}	-.02293	.00706
PCREW	α_{34}	-.01861	.00729
PLOCO	α_{44}	.03863	.00936
PMNGT	α_{54}	-.03794	.01019
Y	θ_{41}	-.01131	.00777
S	σ_{24}	.00367	.00319
QK	η_{41}	.00755	.01399

Table 6

Non-Crews Function

VARIABLE	COEFFICIENT	ESTIMATE	STD. ERROR
	α_{50}	.39945	.00300
PCAR	α_{15}	-.02167	.01535
PFUEL	α_{25}	-.02422	.01332
PCREW	α_{35}	-.07234	.01491
PLOCO	α_{45}	-.03794	.01019
PMNGT	α_{55}	.15617	.02461
Y	θ_{51}	-.02800	.02758
S	σ_{25}	-.01004	.01120
QK	η_{51}	.12819	.04957

CHAPTER 5

RESULTS AND IMPLICATIONS

5.1. Introduction

This section reviews the estimation results presented in Chapter 4. Two major questions will be addressed.

1) Does engineering information add significantly to the explanatory power of the cost function? This is a test of the hybrid approach.

2) What implications can be drawn from the model? For example, what are the impacts of the various factor prices on cost?

5.2. On the Statistical Evaluation of Results

The tables in section 4.1 provide standard errors associated with the coefficient estimates. The standard error provides a simple measure of the quality of the coefficient estimate. While it has been traditional to report t-values (which are simply coefficient estimates divided by the standard error) in many regression analyses, such values can be very misleading in evaluating a model of the type at hand. This is because basic variables in the model appear in many terms in the form of composite variables. For example, speed occurs as itself, in a second-order term and in cross-products with all the other basic variables. Therefore to evaluate any of the basic variables, we must pose hypotheses on the appropriate vector of composite variables in the model; t-tests are inappropriate for this purpose. Instead, a vector-based test (the log-likelihood ratio test) will be used; it is described in section B.2.4.2.

5.3. A Test of the Hybrid Approach

As discussed in section 2.3, one would expect the coefficient of velocity to be negative. This is due to the relationship between congestion, short-run variable cost and velocity for a fixed configuration. This a priori knowledge of the sign allows us to perform a one-tailed test on speed, which appears to be marginally acceptable. If we perform a log-likelihood ratio test, as discussed in section 5.2 above, by setting all coefficients of speed-related terms to zero we find that the test statistic value is 21.49 with eight degrees of freedom. This causes us to reject the hypothesis that speed should not be in the model. Thus we conclude that the addition of engineering information adds significantly to the model.

5.4. Implications of the Model

In general the model exhibits the expected properties. The first-order price and flow terms are positive which implies that at the point of means (where one has the greatest confidence in the predictions) the partial derivative of cost with respect to a price, or flow, is positive. This is to be expected. Further, as one would expect, the fixed factor term has a negative coefficient. This is reasonable since we have estimated a short-run variable cost function and thus, no price on the fixed factor is included. Therefore improvements in the fixed factor should result in reductions in cost. For example, in the Cobb-Douglas case examined in section B.2.3.4 we see that the coefficient on the fixed factor is negative, again reflecting the lack of including a cost of making fixed-factor changes. While one cannot derive the same result for the translog, it is reassuring to see it come through so strongly.

Probably of more interest are the elasticities of cost with respect to factor prices. These are easily found at the point of means since they are the coefficients for the first-order price terms. For example, the elasticity of cost with respect to the price of cars is 0.317. Notice that cost is most responsive to non-crew labor wages, then car prices, then crew labor wages. Fuel and locomotives appear to have much smaller effects.

It is not difficult to understand why cars contribute heavily to costs: railroads, as regulated common carriers, must provide service to all who will pay the tariff. This translates into the requirement to have ready access to a number of cars of various types.

What, however, could be the reason for the disparity between non-crew labor and crew labor? A reasonable explanation is the following one. Since a crew is matched with a train (and not a specific number of cars), as crew costs go up, the firm has the option of running longer trains to amortize the crew cost. This is not true, however, of non-crew labor. Union restrictions, regulatory requirements for information (resulting in a significant amount of paperwork) and the fact that the firm under study is only partially computerized all probably contribute to the inability of the firm to substitute away from non-crew labor as its price rises. Thus, the relative effects of the two types of labor on cost is quite reasonable.

The coefficient associated with QK (i.e. $\hat{\delta}_{10} = -.92323$) is the elasticity of cost with respect to the quality of plant measure. If we hold configuration constant (and thus total mileage is fixed) then a one percent increase in the number of miles in the top FRA track category will result in almost a one percent drop in short-run variable costs. It should be stressed that since the short-run variable cost function does not include a price on QK, that the overall reduction in total costs is unclear. In other words, this elasticity neglects the cost of making the improvement to track.

The effect of such a change in plant can also be seen in the sign of the velocity term. Since $\hat{\gamma}_{10}$ is negative, then at the point of means this implies the improvements in speed would reduce costs. Since the track categories have effective speed limits associated with them (see section 3.4) this is a very reasonable result.

Finally, cost appears to be increasing in flow as seen by the fact that $\hat{\beta}_{10}$ (and most of the other terms involving Y) is positive. Thus, flow marginal cost is positive at the point of means. The cost elasticity of flow is approximately .1, indicating that at the point of means the cost function is relatively flat with respect to flow.

In terms of structure (e.g. separability, homotheticity, homogeneity) the joint test for separability and homotheticity (see section B.2.4.2) was performed. The test statistic value was 33.796 with ten degrees of freedom, and thus we reject the hypothesis. On the other hand, a review of the cross terms between output and prices and those between output and the fixed factor seems to lend some support to the notion that the function may be separable in inputs and outputs. Again, the t-values can be very misleading, and this is certainly not a test of the separability of the cost function; as is mentioned in section B.2.3.3 such test for the translog are complicated by the fact that the conditions for separability imply specific functional structure to the subaggregates thereby transforming the test into a joint test on separability <u>and</u> a specific functional form for the subaggregate. Of course, since the test for joint separability and homotheticity was rejected, it makes no sense to test homogeneity or unitary elasticities of substitution.

The rejection of the test is consistent with other work in this area (see, for example, [32]). It is also especially noteworthy that the second-order price terms (own and cross) are generally very strong, thereby providing further evidence that cost functions (and therefore production functions: see section

B.2.3.1 on the self-dual nature of Cobb-Douglas forms) that are Cobb-Douglas are overly restrictive. This is important since these have been very popular in analysis of transport cost and production.

What about density economies or diseconomies? From the discussion in section 2.3 above we see that Railroad X is probably suffering from diseconomies of density, at least during some parts of the year. Observation tends to confirm this. During parts of the year the yard becomes significantly congested. As to whether or not a configuration shift would really improve things is to be seen only by bringing in the cost of such a change (for example, improved yard facilities). To the degree that better operations procedures can be affected to smooth the congestion, the diseconomies will probably be reduced.

Our empirical results also motivate the following theoretical analysis of the relationship between density economies and maintenance policy, and the effects of regulation. We define a maintenance policy as that policy (level) of optimal maintenance activity given a fixed configuration and output level. Thus, since for any given configuration various levels of output will be consistent with different levels of maintenance activity, one can view the maintenance policy as giving rise to even "shorter-run" curves whose envelope is the fixed configuration curve (see Figure 12). For any given maintenance policy (which implies speed limits, among other things) density and speed again trade-off as congestion becomes significant. Note that regulation can act to provide incentives to not maintain in what would otherwise be an optimal fashion. Consider Figure 13, which shows a firm that is regulated to produce output on configuration B. Present output is at level $\bar{Z}$. The "appropriate" maintenance policy is the small U-shaped curve on the left-side

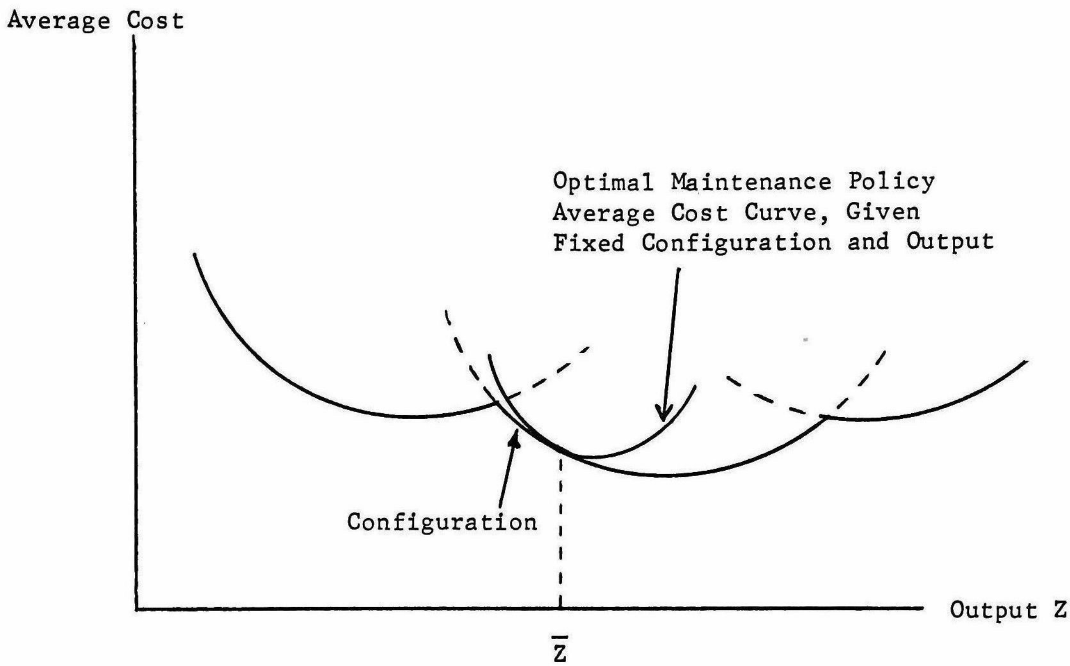


FIGURE 12 - CONFIGURATION AND MAINTENANCE POLICY

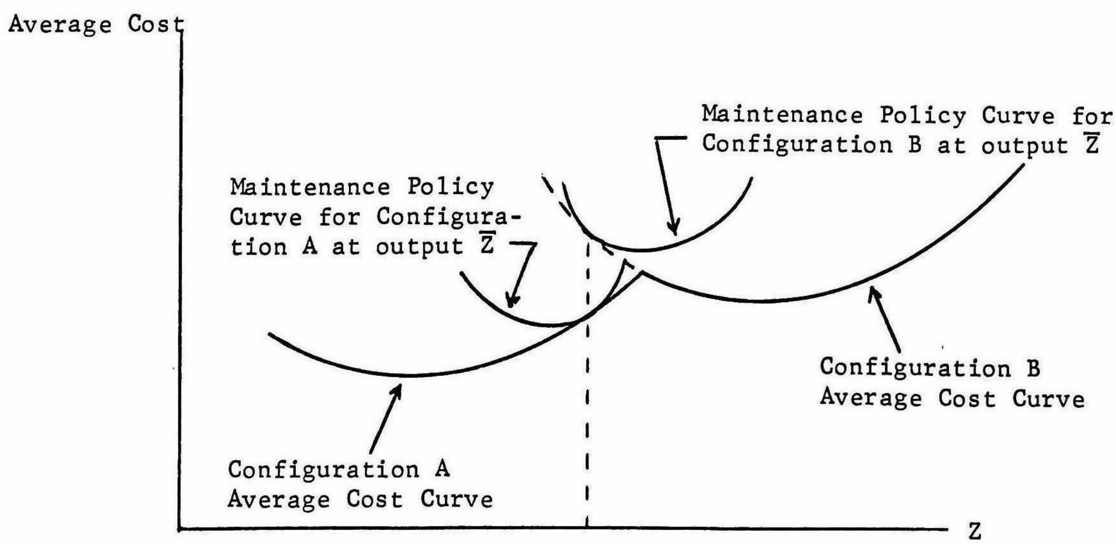


FIGURE 13 - DENSITY ECONOMIES AND MAINTENANCE POLICY

of the cost curve for configuration B. Clearly, the firm would be better off if it could shift to configuration A, which might involve some abandonment of service. Since it cannot, an alternative is to simply pursue the maintenance policy associated with configuration A: maintain some parts well and other parts not at all (as if they didn't exist). This would appear to be a policy of deferred maintenance, but in fact it is simply a result of the regulatory induced incentives.

CHAPTER 6

SUMMARY AND DIRECTIONS FOR FURTHER RESEARCH

6.1. Summary of First Year's Results

As demonstrated by the discussion in the previous chapter, economic and engineering principles can be brought together to provide a more complete cost model than either can produce by itself. This *hybrid* model incorporates both the comprehensive picture of the rail firm that an economic approach provides, and the depth of understanding of the production process that comes from detailed engineering analysis. The statistical test described in section 5.3, in which the hybrid model is compared to a model with the engineering-related variables omitted, provides strong empirical evidence of the value of the hybrid approach. Not only is this approach theoretically appealing, but it also produces a better explanatory model of costs from an empirical perspective.

In terms of theoretical development, the analysis included in Chapter 2 of the report has extended the traditional economic theory of cost and production to establish the appropriate role for engineering models in the study of cost. In the first year's work on this project, we have concentrated on using engineering models to predict one major service quality variable, average speed of shipment through the system. However, the approach we have developed is robust with respect to adding more service characteristics and network complexity. The procedure starts with a very general model of production which makes a minimum of economic assumptions (e.g. it makes no assumption as to whether or not there are returns-to-scale). This is important since we would like to examine (i.e. test) such economic attributes rather than assume them. Engineering models that reflect physical relationships among some of the input variables and some of the outputs are then added. These engineering models may reflect any such relationship which is physically meaningful to the production process. They are not limited only to describing speed. The engineering models increasingly restrict (and thereby further reveal) the model of production. Again, it should be noted that the restrictions will reflect physical realities and not economic assumptions that need to be tested. As more engineering relationships are added (reflecting network considerations or service characteristics) the economic attributes of the model become more and more refined, for the general production model becomes increasingly restricted by the engineering relationships and this, in turn, reveals more of the economic relationships.

The formulation of the hybrid cost model implies certain data needs for successful empirical analyses. We have been able to translate theoretical data needs specified by the model into data requirements than can be fulfilled by a firm using available information. Thus, the model allows us to specify ways of combining available firm data correctly to produce measures of cost that should be used in regulatory proceedings (e.g., "incremental" costs: see section 6.2). Further, because the model allows for multiple outputs, including service characteristics as well as volumes of commodities, marginal costs for particular commodities and services are computable.

It should also be noted that engineering considerations provide deeper insight into the nature of factors that contribute to scale economies (and diseconomies), and how size and density relate. This has been discussed in some detail in section 5.4, with particular emphasis on the interrelationships of traffic density, maintenance policy, and regulatory restrictions on network configuration.

The empirical work that has been done using railroad "X" as a case study has also produced interesting and valuable results. The short-run elasticities of cost with respect to various factor prices are quite illuminating. Short-run variable cost is most responsive to non-crew labor wages, then car prices and crew wages. Fuel and locomotives appear to have much smaller effects. The high elasticity of cost with respect to non-crew wages is an interesting (and somewhat surprising) result. It indicates the limited availability of opportunities for substitution of other factors for non-crew labor as non-crew wages rise.

The empirical work has also provided valuable insight with respect to appropriate structure of cost functions. As described in section 5.4, a joint test for separability and homotheticity (see section B.2.4.2) was performed, and these properties were rejected. The rejection of the test is consistent with other recent work in this area (see, for example, [32]). It is also especially noteworthy that the second-order price terms (own and cross) are generally very strong, thereby providing further evidence that cost functions (and therefore production functions) that are Cobb-Douglas are overly restrictive. This is important since these have traditionally been very popular in analysis of transport cost and production.

In summary, the first year's research on this project has provided a number of important theoretical and empirical results. The experience of this first phase of the research has also raised several important issues for further investigation. The focus of the second year's work, addressing some of these issues, is described in the following section.

6.2. Plans for the Second Year

Beginning with the Railroad Revitalization and Regulatory Reform (4R) Act of 1976, there was clear legislative recognition of the difficulties in using traditional railroad accounting data in regulatory proceedings. Such proceedings rely heavily on measuring the "cost" of providing service, either as a basis for setting rates, or more recently, as a basis for determining subsidies for the continued operation of lines which would otherwise be abandoned. However, the accounting procedures prescribed by the Interstate Commerce Commission (ICC) generally do not result in the collection of cost data in a form readily usable (in a manner consistent with economic theory) in such proceedings. The 4R Act included attempts to deal with some of these problems, using new terms such as "incremental cost" of providing service.

Except in rare instances, however, these new terms are not defined in the Act, but are left to the ICC to determine. Several of these definitional problems are being addressed in current proposed legislation. These terms will need to be defined and methods for actually applying them will have to be established. It is precisely on this point that we wish to focus in the second year of this research.

Specifically, we plan to address the question of defining and measuring "incremental costs." This term is closely associated with the economic concept of marginal cost, and reflects the change in cost incurred by a railroad as a result of changing either the amount of service provided or the nature (quality) of that service. The work done in this project during the first year provides an excellent basis for estimating short-run marginal costs. This is because the hybrid cost function technique that we have developed allows for multiple commodity types and service characteristics. Thus, short-run marginal costs with respect to the various components of the output vector can be found. With suitable extensions, the hybrid cost technique will provide useful guidance in developing procedures for implementing notions of "incremental costs" in regulatory and public policy settings, including estimation of long-run marginal costs. The three major areas in which our work to date requires extension concern: 1) more complete definition and development of the output vector in terms of commodities carried and service characteristics provided; 2) analysis of the costs of producing services on more complicated networks; and 3) appropriate measurement and pricing of fixed factor inputs, to allow construction of long-run cost functions.

First, the question of what is a rail firm's output and how it should be measured will be addressed. This will require us to examine problems of appropriate disaggregation of commodity flows to reflect both the nature of the goods carried and the network on which they move. This will also require an expansion of the service characteristics considered. During the first year of this work, we have concentrated on "speed of shipment" as the major service characteristic. However, there are obviously a number of other characteristics of importance such as transit time reliability and equipment availability.

Study of both transit time reliability and equipment availability require a model of network operations which produces estimates of the distribution of transit times by origin-destination pair, given an operating policy including train schedules, blocking, etc., and which reflects the degree to which equipment is available for provision to a shipper when requested. Previous work on service-differentiated demand models indicates that this is an important element of service quality to many shippers. A key aspect of the ability of a railroad to provide empty freight cars to shippers when requested is the effectiveness with which it distributes empty cars over its network. Thus, the second major area for model extension in the second year is the development of models of network operations which will allow us to focus on a broader range of service characteristics and associated costs.

The third major area for model extension in the second year concerns measurement and pricing of fixed factors. For the empirical work in the first year, a track quality index was constructed to measure the fixed factor representing quality of plant. Measures of size and condition of the railroad's network will be of importance in the second year also, and additional work will be done to examine this area thoroughly from both theoretical and empirical perspectives. The construction of appropriate prices for fixed factors is of particular importance if we are to be able to derive long-run cost functions from the estimated short-run functions. Because long-run functions are of central interest to government policy-makers and regulatory bodies, this is an important step to make.

Once the structural form of the hybrid cost model has been extended, it will be necessary to test the structure by acquiring data and estimating the model parameters statistically. As in the first year of the study, our approach will include working closely with a railroad as a case study. The nature of the extensions to the model to be explored in the second year necessitate working with a major railroad. A number of products will result:

1) first, an estimated model of a major rail firm's cost function and a comparison of it with our present results for a small rail firm;
2) a technique for defining and computing theoretically defendable incremental costs based on available data;
3) marginal cost functions that allow for multiple commodities, joint use of facilities by different services, and various service characteristics; and
4) finally, the resulting technique will provide necessary methodology for both the government (for policy analysis) and the railroads (for planning).

It should be emphasized that understanding costs of production is important both for policy analysis and regulatory review by the government and for planning purposes in rail firms, and that the procedures we are developing provide accurate, comprehensive cost analysis that can be conducted with readily available data. Questions of returns-to-scale, the relationship of output and input changes to changes in marginal and average costs, and the impact of various possible capital investments on the costs of providing service are problems that the technique can address and answer.

Bibliography

1. Allen, R.G.D. Mathematical Analysis for Economists, 2nd ed., Macmillan, London, 1938.

2. Baumol, W.J. "On the Proper Cost Tests for Natural Monopoly in a Multi-product Industry," American Economic Review, December 1977, pp. 809-822.

3. Beckmann, M., C.B. McGuire and C.B. Winsten. Studies in the Economics of Transportation, Yale University Press, New Haven, 1956.

4. Berndt, E. and L. Christensen. "The Internal Structure of Functional Relationships: Separability, Substitution, and Aggregation," Review of Economic Studies, 40, pp. 403-410, 1973.

5. Berndt, E. and L. Christensen. "Testing for the Existence of a Consistent Aggregate Index of Labor Inputs," American Economic Review, 64, pp. 391-404, 1974.

6. Blackorby, C., D. Primont and R. Russell. Duality, Separability, and Functional Structure: Theory and Economic Applications, North-Holland, New York, 1978.

7. Borts, G.H. "Production Relations in the Railway Industry," Econometrica, 20, January, 1952, pp. 71-79.

8. Borts, G.H. "The Estimation of Rail Cost Functions," Econometrica, 28 January, 1960, pp. 108-131.

9. Brown, R., D. Caves and L. Christensen. "Modelling the Structure of Production with a Joint Cost Function," University of Wisconsin, Social Systems Research Institute, Paper 7521, 1975.

10. Burgess, D.F. "Duality Theory and Pitfalls in the Specification of Technologies," Journal of Econometrics, 1975, pp, 105-121.

11. Car and Locomotive Cyclopedia of American Practice, Simmons-Boordman, Chicago, 1974.

12. Caves, D. and L. Christensen. "Modeling the Structure of Production in the U.S. Railroad Industry," University of Wisconsin (mimeo), 1976.

13. Caves, D., L. Christensen and J. Swanson. "Productivity in U.S. Railroads, 1951-1974," Economics Department, University of Wisconsin (mimeo), 1978.

14. Christensen, L., D. Jorgenson and L. Lau. "Transcendental Logarithmic Production Functions," Review of Economics and Statistics, 55, 1973, pp. 28-45.

15. Christensen, L. and W.H. Greene. "Economies of Scale in U.S. Electric Power Generation," Journal of Political Economy, 84, 1976, pp. 655-676.

16. Daughety, A. "Stochastic Production and Cost/Production Duality," Discussion Paper No. 356, Center for Mathematical Studies in Economics and Management Science, December 1978.

17. Davis, W.J. "Tractive Resistance of Electric Locomotives and Cars," General Electric Review, Vol. 29, 1926, pp. 685-708.

18. Davis, W.J. American Railway Engineering Association, 44: 190, pp. 670-679, 1943.

19. Denny, M. and M. Fuss. "The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates," American Economic Review, 67, 1977. pp. 404-418.

20. DeSalvo, J.S. "A Process Function for Rail Line Haul Operations," The Rand Corporation, p. 3729, 1967.

21. DeSalvo, J.S. and L.B. Lave. "A Statistical Engineering Approach to Estimation of Railway Cost Functions," The Rand Corporation, p. 3781, 1968.

22. Diewert, W. "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function," Journal of Political Economy, May/June 1971.

23. Diewert, W.E. "Applications of Duality Theory," in Frontiers of Quantitative Economics, Vol. II (M.D. Intrilligator and D.A. Kendrick eds.), North-Holland, New York, 1974.

24. Eads, G.C. "A Cost Function for the Local Service Airlines," unpublished Ph.D. dissertation, Yale University, 1968.

25. Eads, G., M. Nerlove and W. Raduchel. "A Long-Run Cost Function for the Local Service Airline Industry," Review of Economics and Statistics, August, 1969.

26. ElectroMotive Division, General Motors Corporation, Product Application Section, Trailing Load-Speed-Grade, Tables, 1000, 2000, 2500, 3000, for SW1000, GP38, GP35, SD40 Locomotives, LaGrange, EMD-GM, 1965.

27. English, G.W. "An Analytic Model for the Analysis of Single Track Railway Capacity," Institute of Guided Ground Transport, Queen's University at Kingston, Ontario, 1977.

28. Folk, J.F. Some Analyses of Railroad Data, MIT Department of Civil Engineering, Studies in Railroad Operations and Economics, Volume 6, June, 1972.

29. Freight Car Utilization and Railroad Reliability: Case Studies, Association of American Railroads Report R-283, October, 1977.

30. Friedlaender, A.F. The Dilemma of Freight Transport Regulation, Brookings Institution (Washington, D.C.), 1969.

31. Friedlaender, A.F. "Hedonic Costs and Economies of Scale in the Regulated Trucking Industry," in Motor Carrier Economic Regulation, National Research Council (Washington, D.C.), 1978.

32. Friedlaender, A.F. and R. Spady. "Equity, Efficiency and Resource Allocation in the Rail and Regulated Trucking Industries," CTS Report No. 79-4, Center for Transportation Studies, Massachusetts Institute of Technology, 1979.

33. Fuss, M. and D. McFadden, ed. Production Economics: A Dual Approach to Theory and Application, Volumes I and II, North-Holland, New York, 1978.

34. Gaver,D.P.,"Imbedded Markov Chain Analysis of a Waiting Line Process in Continuous Time," Annals of Mathematical Statistics,30:3,Sept.,1959, pp698-720.

35. Gentzel, R. Queuing Models for Analysis of Railroad Classification Yards, M.S. Thesis (in progress), The Transportation Center, Northwestern University, 1979.

36. Goldman, S.M. and H. Uzawa. "A Note on Separability in Demand Analysis," Econometrica, Vol. 32, No. 3, 1964, pp. 387-399.

37. Green, H.A.J. Aggregation in Economic Analysis, Princeton University Press, New Jersey, 1964.

38. Griliches, Z. "Cost Allocation in Railroad Regulation," Bell Journal of Economics and Management Science, 3, 1972, pp. 26-41.

39. Haight, F. Mathematical Theories of Traffic Flow, Academic Press, New York, 1963.

40. Hall, R.E. "The Specification of Technology with Several Kinds of Output, Journal of Political Economics, 31, 1973, pp. 878-892.

41. Harris, R.G. "The Economics of Traffic Density in the Rail Freight Industry," Bell Journal of Economics, Autumn, 1977, pp. 556-564.

42. Hasenkamp, G. Specification and Estimation of Multiple Output Production Functions, Springer-Verlag, New York, 1976.

43. Hay, W. Railroad Engineering, Vol. 1, John Wiley & Sons, New York, 1953.

44. Henderson, J.M. and R.E. Quandt. Microeconomic Theory, 2nd edition, McGraw-Hill, New York, 1971.

45. Hennes, R.G. and M. Ekse. Fundamentals of Transportation Engineering, McGraw-Hill, London, 1969, Chap. 25.

46. Houthakker, H. "Additive Preferences," Econometrica, 1960, Vol. 28, No. 2, pp. 244-257.

47. Jacobsen, S.E. "Production Correspondences," Econometrica, 38, 1970, pp. 754-771.

48. Keeler, T. "Railroad Cost Function: An Empirical Study," Unpublished B.A. thesis, Reed College, 1967.

49. Keeler, T. "Railroad Costs, Returns to Scale and Excess Capacity," Review of Economics and Statistics, May 1974.

50. Klein, L.R. "The Use of Cross Section Data in Econometrics with Application to a Study of Production of Railroad Services in the United States," National Bureau of Economic Research (mimeo), 1947.

51. Kulash, D.J. Routing and Scheduling in Public Transportation Systems, Ph.D. dissertation MIT Department of Civil Engineering, January, 1971.

52. Lau, L. "Duality and the Structure of the Utility Function," Journal of Economic Theory, 1969, Vol. 1, No. 4, pp. 374-396.

53. Lau, L. "Profit Functions of Technologies with Multiple Inputs and Outputs," The Review of Economics and Statistics, 1972, Vol. 54, No. 3, pp. 281-289.

54. Lau, L. "Comments on Diewart's 'Applications of Duality Theory' ", in M.D. Intriligators and D.A. Kendrick (eds.), Frontiers of Quantitative Economics, Vol. II, North-Holland, New York, 1974.

55. Lau, L. "A Characterization of the Normalized Restricted Profit Function," Journal of Economic Theory, 12, February 1976, pp. 131-163.

56. Leontief, W.W. "Introduction to a Theory of the Internal Structure of Functional Relationships," Econometrica, 15, 1947, pp. 361-373.

57. Mangasarian, O.L. Nonlinear Programming, McGraw-Hill, New York, 1969.

58. Marsden, J., D. Pingry and A. Whinston. "Engineering Foundations of Production Functions," Journal of Economic Theory, 1974, pp. 124-140.

59. Martland, C.D. and W.J. Rennicke. "Unit Costs and Capacity Relationships at Railroad Terminals," Proceedings of the Transportation Research Forum, XIX, 1978, pp. 136-144.

60. McFarland, H., "The Estimation of a Railroad Cost Function," Unpublished Ph.D. Thesis, Northwestern University, 1978.

61. Molinas, E.C. Poisson's Exponential Binomial Limit, Van Nostrand, Princeton, 1942.

62. Nerlove, M. "Returns to Scale in Electric Supply," in C. Christ (ed.), Measurement in Economics: Studies in Mathematical Economics and Econometrics, in Memory of Yehuda Grunfeld, Stanford University Press, Stanford, CA, 1963.

63. Osuna, E.E. and G.F. Newell. "Control Strategies for an Idealized Public Transportation System," Transportation Science, 6:1, 1972, pp. 52-72.

64. Panzar, J. and R. Willig. "Economies of Scale in Multi-Output Production," Quarterly Journal of Economics, 91, 1977, pp. 481-493.

65. Petersen, E.R. "Bulk Service Queues: With Applications to Train Assembly Times," School of Business Working Paper Series 71-2, Queen's University, Kingston, Ontario, August, 1971.

66. Petersen, E.R. "Over-the-Road Transit Time for a Single Track Railway," Transportation Science, 8:1, 1974, pp. 65-74.

67. Petersen, E.R. "Railyard Modeling: Parts I and II," Transportation Science, 11:1, 1977, pp. 37-59.

68. Powell, A. and F. Gruen. "The Constant Elasticity of Transformation Frontiers and Linear Supply System," International Economic Review, 1968, pp. 315-328.

69. Reebie Associates. Toward an Effective Demurrage System, prepared for U.S. Department of Transportation, Federal Railroad Administration, July, 1972.

70. Russell, R. "Functional Separability and Partial Elasticities of Substitution," Review of Economic Studies, 1975, pp. 79-86.

71. Saaty, T.L. Elements of Queuing Theory, McGraw-Hill, New York, 1961.

72. Scherer, F.M. Industrial Market Structure and Economic Performance, Rand McNally, Chicago, 1970.

73. Shephard, R. Theory of Cost and Production Functions, Princeton University Press, New Jersey, 1970.

74. Sono, M. "The Effect of Price Changes on the Demand and Supply of Separable Goods," International Economic Review, Vol. 2, September, 1961, pp. 239-271.

75. Spady, R.H. and A.F. Friedlaender. "Econometric Estimation of Cost Functions in the Transportation Industries," Center for Transportation Studies, M.I.T., September, 1976.

76. Spady, R.H. and A.F. Friedlaender. "A Hedonic Cost Function for the Regulated Trucking Industry," Bell Journal of Economics, Spring, 1978, pp.159-179.

77. Stigler, G.J. The Theory of Price, 3rd edition, Macmillan, New York, 1966.

78. Swanson, J.A. "Railroad Investment Demand, 1948-1965: A Neoclassical Model," Unpublished Ph.D. dissertation, University of Wisconsin, 1968.

79. Theil, H. Principles of Econometrics, John Wiley & Sons, New York, 1971.

80. Uzawa, H. "Duality Principles in the Theory of Cost and Production," International Economic Review, 5, 1964, pp. 216-220.

81. Varian, H. Microeconomic Analysis, W.W. Norton, New York, 1978.

82. Welding, P.O. "Time Series Analysis as Applied to Traffic Flow, Proceedings of the Second International Symposium on the Theory of Road Traffic Flow, OECD, London, 1963, pp. 60-72.

83. Welty, G. "Budget Planning 1979," Railway Age, Nov. 27, 1978, pp. 26-32.

84. Wright, W.B. "How Cars in Multiple Cut Cost," Railway Age, Jan. 4, 1960.

85. Yearbook of Railroad Facts, Association of American Railroads, Washington, D.C., 1978.

APPENDIX A

A MORE DISAGGREGATE COST FUNCTION

This model is essentially the same as the model in section 5.1 except for an expansion of the representation of flow. The flow vector has four parts:

Y1	Unit-train Coal	(STCC 11)
Y2	Low-value Bulks	(STCC 1-10, 11, 12-18, 20)
Y3	High-value Bulks	(STCC 21, 22, 26-29, 32, 40)
Y4	Manufactured	(STCC 19, 23-25, 30, 31, 33-39, 41-49)

The resulting model in the translog form with four factor share equations was estimated and is reported on the next few pages.

Table A-1

Cost Function

Variable	Coefficient	Estimate	STD. ERROR
	α_0	.02582	.01781
PCAR	α_{10}	.31857	.00498
PFUEL	α_{20}	.04779	.00107
PCREW	α_{30}	.15206	.00129
PLOCO	α_{40}	.08281	.00076
PMNGT	α_{50}	.39878	.00308
QK	δ_{10}	-.63609	.22069
Y1	β_{10}	.04510	.04299
Y2	β_{20}	-.05418	.12626
Y3	β_{30}	.04139	.09210
Y4	β_{40}	.15028	.07852
S	γ_{10}	.02412	.08172
$(PCAR)^2$	α_{11}	-.07265	.02836
PCAR• PFUEL	α_{12}	.01143	.00893
PCAR• PCREW	α_{13}	.01946	.00995
PCAR• PLOCO	α_{14}	.03221	.00894
PCAR• PMNGT	α_{15}	.09551	.01954
$(PFUEL)^2$	α_{22}	.06090	.01336
PFUEL• PCREW	α_{23}	-.02410	.00941
PFUEL• PLOCO	α_{24}	-.02188	.00760
PFUEL• PMNGT	α_{25}	-.02635	.01547
$(PCREW)^2$	α_{33}	.10297	.01515
PCREW •PLOCO	α_{34}	-.01999	.00795
PCREW •PMNGT	α_{35}	-.07835	.01547

Variable	Coefficient	Estimate	STD. ERROR
$(PLOCO)^2$	α_{44}	.04923	.00918
PLOCO•PMNGT	α_{45}	-.03957	.01001
$(PMNGT)^2$	α_{55}	.13472	.02592
$(Y1)^2$	β_{11}	.06039	.04627
Y1•Y2	β_{12}	-.46422	.38545
Y1•Y3	β_{13}	.02897	.28206
Y1•Y4	β_{14}	.09933	.19583
$(Y2)^2$	β_{22}	.77209	.74816
Y2•Y3	β_{23}	.20011	.48582
Y2•Y4	β_{24}	-.31962	.59987
$(Y3)^2$	β_{33}	-.02631	.17141
Y3•Y4	β_{34}	-.08702	.45818
$(Y4)^2$	β_{44}	.14835	.18496
$(S)^2$	γ_{11}	.07176	.10457
Y1•S	τ_{11}	-.02381	.11996
Y2•S	τ_{21}	.01726	.56547
Y3•S	τ_{31}	-.15968	.44404
Y4•S	τ_{41}	.07094	.25793
PCAR•Y1	θ_{11}	.02152	.01553
PCAR•Y2	θ_{12}	.03525	.05266
PCAR•Y3	θ_{13}	-.00362	.02659
PCAR•Y4	θ_{14}	-.02931	.03089
PFUEL•Y1	θ_{21}	-.00134	.00334
PFUEL•Y2	θ_{22}	-.00014	.01168
PFUEL•Y3	θ_{23}	-.00024	.00572
PFUEL•Y4	θ_{24}	.00662	.00708

Variable	Coefficient	Estimate	STD. ERROR
PCREW•Y1	θ_{31}	-.00148	.00401
PCREW•Y2	θ_{32}	-.00952	.01382
PCREW•Y3	θ_{33}	-.00061	.00686
PCREW•Y4	θ_{34}	.01533	.00827
PLOCO•Y1	θ_{41}	-.00288	.00238
PLOCO•Y2	θ_{42}	-.00810	.00843
PLOCO•Y3	θ_{43}	.00688	.00408
PLOCO•Y4	θ_{44}	-.01783	.00520
PMNGT•Y1	θ_{51}	-.01581	.00961
PMNGT•Y2	θ_{52}	-.01749	.03294
PMNGT•Y3	θ_{53}	-.00241	.01645
PMNGT•Y4	θ_{54}	.02519	.01940
PCAR•S	σ_{11}	.00645	.01847
PFUEL•S	σ_{21}	-.00308	.00400
PCREW•S	σ_{31}	-.00161	.00481
PLOCO•S	σ_{41}	.00139	.00285
PMNGT•S	σ_{51}	-.00315	.01145
$(QK)^2$	δ_{11}	-4.76311	3.14795
QK•Y1	μ_{11}	.52178	.50587
QK•Y2	μ_{21}	-1.10728	1.59404
QK•Y3	μ_{31}	-.15783	1.35530
QK•Y4	μ_{41}	.54583	.86997
QK•S	ϵ_{11}	.07494	.62061
PCAR•QK	η_{11}	-.16402	.08197
PFUEL•QK	η_{21}	.02452	.01789
PCREW•QK	η_{31}	.04047	.02134
PLOCO•QK	η_{41}	-.01672	.01282
PMNGT•QK	η_{51}	.11574	.05106

Table A-2

Fuel Function

Variable	Coefficient	Estimate	STD. ERROR
	α_{20}	.04779	.00107
PCAR	α_{12}	.01143	.00893
PFUEL	α_{22}	.06090	.01336
PCREW	α_{23}	-.02410	.00941
PLOCO	α_{24}	-.02188	.00760
PMNGT	α_{25}	-.02635	.01547
Y1	θ_{21}	-.00134	.00334
Y2	θ_{22}	-.00014	.01168
Y3	θ_{23}	-.00024	.00572
Y4	θ_{24}	-.00662	.00708
S	σ_{21}	-.00308	.00400
QK	η_{21}	-.02452	.01789

Table A-3

Crews Function

Variable	Coefficient	Efficient	STD. ERROR
	α_{30}	.15206	.00129
PCAR	α_{13}	.01946	.00995
PFUEL	α_{23}	-.02410	.00941
PCREW	α_{33}	.01030	.01515
PLOCO	α_{43}	-.01999	.00795
PMNGT	α_{53}	-.07835	.01547
Y1	θ_{31}	-.00148	.00401
Y2	θ_{32}	-.00952	.01382
Y3	θ_{33}	-.00061	.00686
Y4	θ_{34}	.01533	.00827
S	σ_{31}	-.00161	.00481
QK	η_{31}	.04047	.02134

Table A-4

Locos Function

Variable	Coefficient	Estimate	STD. ERROR
	α_{40}	.08281	.00076
PCAR	α_{14}	.03221	.00894
PFUEL	α_{24}	-.02188	.00760
PCREW	α_{34}	.01999	.00795
PLOCO	α_{44}	.04923	.00918
PMNGT	α_{54}	-.03957	.01001
Y1	θ_{41}	-.00288	.00238
Y2	θ_{42}	-.00810	.00843
Y3	θ_{43}	-.00688	.00408
Y4	θ_{44}	-.01783	.00520
S	σ_{41}	-.00139	.00285
QK	η_{41}	-.01672	.01282

Table A-5

Non-Crews Function

Variable	Coefficient	Estimate	STD. ERROR
	α_{50}	.39878	.00308
PCAR	α_{15}	.00955	.01954
PFUEL	α_{25}	-.02635	.01547
PCREW	α_{35}	-.07835	.01547
PLOCO	α_{45}	-.03957	.01001
PMNGT	α_{55}	.13471	.02592
Y1	θ_{51}	-.01581	.00961
Y2	θ_{52}	-.01749	.03294
Y3	θ_{53}	-.00241	.01940
Y4	θ_{54}	.02519	.00308
S	σ_{51}	-.00315	.01145
QK	η_{51}	.11574	.05106

APPENDIX B

PRODUCTION AND COST: THEORY AND EXAMPLES

B.1. Production

B.1.1 Definitions and Assumed Properties of Production and Transformation Functions

Let[1] $x = (x_1,\ldots,x_n)'$ be a non-negative n-vector of input levels (factors) used by the firm to produce a single output Z. Examples of inputs for a rail firm are fuel, various types of labor, locomotives, etc. A production function $f(x)$ is a mathematical model of the relationship between x and Z and thus:

$$Z = f(x) . \tag{B-1}$$

The above production function uses n inputs to produce one output; a classic example of an aggregate output measure for transport firms is total ton-miles of goods moved.

We define an isoquant of f to be the set of input levels that is just sufficient to produce a given output Z:

$$Q(Z) = \{x \mid Z = f(x)\} .$$

Thus if $\bar{x}$ is such that $f(\bar{x}) > \bar{Z}$ then $\bar{x} \notin Q(\bar{Z})$, i.e. the isoquants only reflect efficient production.

We assume[2] the following properties for $f(x)$:

1) $f(0) = 0$;
2) $f(x)$ is continuous with continuous first and second derivatives (unless explicitly stated otherwise) ;
3) if $x^1 \geq x$ then $f(x^1) \geq f(x)$;
4) $f(x)$ is quasiconcave [57], i.e. $f(\lambda x^1 + (1-\lambda)x^2) \geq \min [f(x^1), f(x^2)]$ for all $x^1, x^2 \geq 0$.

The first property states that positive production requires at least some positive inputs. The second condition imposes regularity on the function while the third condition means that more inputs will not result in less being produced. The fourth condition means that level sets[3] of f (i.e. combinations of x that provide at least a specified output) are convex sets. This in turn means that the isoquants are convex functions, i.e. they resemble Figure B-1a rather than Figure B-1b:

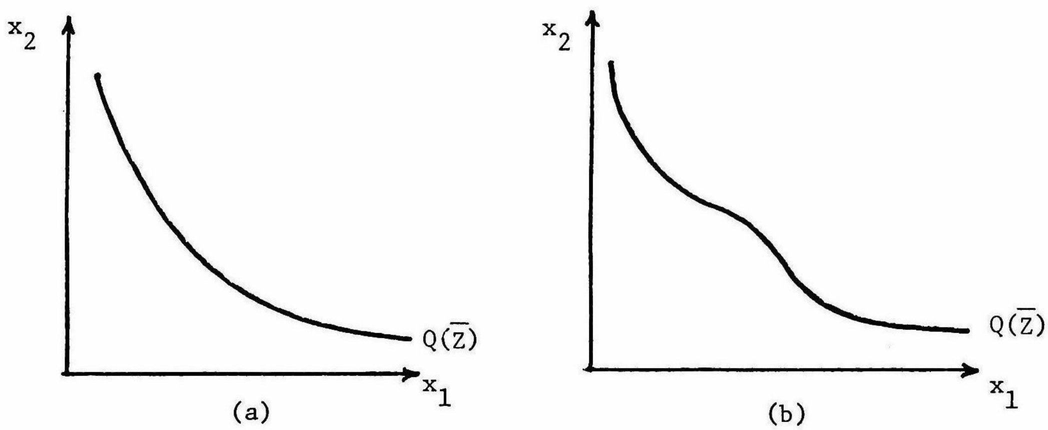


Figure B-1 ISOQUANTS

Many firms, and in particular transport firms, produce a vector of outputs rather than a single output. Transport firms, for example, move a variety of commodities to and from various geographical points. Furthermore, associated with the various commodity flows are characteristics of service such as speed of delivery, schedule unreliability, loss and damage, etc.

Let the firm's output vector be the non-negative m-vector of flows and characteristics $z = (z_1,...,z_m)'$. A <u>transformation function</u> T(z,x) is a mathematical model of the relationship between the input vector x and the output vector z. The vector z can be exactly produced from x if

$$T(z,x) = 0 \qquad \text{(B-2)}$$

which is the analogous statement to (B-1) above.[4] Typical conditions on T(z,x) are as follows:[5]

1) $T(z,x) \leq 0 \qquad \forall_{x,z}$;

2) $T(z,0) \leq 0 \Rightarrow z = 0$;

3) T(z,x) is continuous with continuous first and second derivatives;

4) $\nabla_z T(z,x) < 0, \ \nabla_x T(z,x) > 0$;

5) $V(z) = \{x \mid T(z,x) \leq 0\}$ is a closed, strictly convex set.

The first condition simply defines what we mean by producing z from x. This condition allows for both inefficient production (T(z,x) < 0) and efficient production (T(z,x) = 0). Condition (2) is analogous to condition (1) for production functions.[6] Condition (3) is the regularity condition analogous to condition (2) for production functions. Condition (4) which requires T to be decreasing in z and increasing in x is a stronger condition than condition (3) for production functions. Finally condition (5) is analogous to condition (4) for production functions: it will guarantee that a unique joint cost function exists (section B.2).

In what follows we examine various possible characteristics of production and transformation functions. The characteristics concern the way in which inputs combine to produce output(s). The main considerations in characterizing technology are as follows:

1. Does the technology exhibit economies (or diseconomies) of scale of production for various levels of output?

2. Under what conditions can a vector of outputs be aggregated into a scalar (e.g. ton-miles)?

3. Under what conditions can parts of the input vector be aggregated? For example, must we represent each and every type of labor used, car type used, etc. or can we use models that have aggregate labor and capital inputs.

This report addresses some of the above questions in detail. Others will be addressed more fully in the second year of the research.

B.1.2. Characterization of Production and Transformation Functions

Before proceeding to examine some of the characterizations of production and transformation functions we provide the following definition, which will be of use later in this section:

Definition. Let $H(u,v) = 0$ be continuous and differentiable with $\nabla H(u,v) \neq 0$. The marginal rate of technical substitution (MRTS) of v_i for v_j is

$$MRTS_{ij}(u,v) \equiv \frac{\partial H/\partial v_i}{\partial H/\partial v_j}$$

while the marginal rate of transformation (MRTr) of u_i for u_j is

$$MRTr_{ij}(u,v) \equiv \frac{\partial H/\partial u_i}{\partial H/\partial u_j}$$

Thus, in the case of the production function we will refer to $MRTS_{ij}(x) = f_i/f_j$ where $f_k = \partial f(x)/\partial x_k$ while in the case of the transformation function we may also be interested in $MRTr_{ij}(z,x) = (\partial T(z,x)/\partial z_i)/(\partial T(z,x)/\partial z_j)$. It is important to note that we have assumed that $\nabla H \neq 0$. While this is a stronger condition than condition (3) on production functions, most production functions satisfy this requirement. In general $\nabla f \neq 0$ will hold for most of the analyses; situations wherein this is not true will be noted.

B.1.2.1 Homogeneity and Almost Homogeneity

A production function f(x) is homogeneous of degree k (H.D.k) if

$$\lambda^k f(x) = f(\lambda x) \qquad \forall\, \lambda > 0$$

where λ is a scalar. The above condition states that multiplying all the inputs by a positive scalar multiplies the output by a power of the scalar. We observe the following classical categorization

1) $k > 1 \Rightarrow$ increasing returns-to-scale ;
2) $k = 1 \Rightarrow$ constant returns-to-scale ;
3) $k < 1 \Rightarrow$ decreasing returns-to-scale .

H.D.k functions satisfy[7] Euler's Theorem (see, e.g. [44]):

$$kf(x) = x' \cdot \nabla f(x).$$

Further it can be shown that the partial derivatives are H.D.(k-1). Thus we see that:

$$MRTS_{ij}(\lambda x) = \frac{f_i(\lambda x)}{f_j(\lambda x)} = \frac{\lambda^{k-1} f_i(x)}{\lambda^{k-1} f_j(x)} = \frac{f_i(x)}{f_j(x)} = MRTS_{ij}(x).$$

Thus the marginal rate of technical substitution is unaffected by changes in scale (i.e. it is H.D.0 in x). This assumes $\nabla f \neq 0$. Since we typically will take $\nabla f > 0$, then $MRTS_{ij}(x) > 0$. Production functions that have regions wherein $MRTS_{ij}(x) < 0$ are said to have non-economic regions since it will not generally be profitable to operate in such a region.

Homogeneity of degree k for a production function provides the intuition for the following definition of almost homogeneity for the transformation function, namely a transformation function is almost homogeneous of degrees k_1, k_2 and k_3 (AHD(k_1,k_2,k_3)) if and only if:

$$T(\lambda^{k_1} z, \lambda^{k_2} x) = \lambda^{k_3} T(z,x) \qquad \forall \lambda > 0 .$$

Lau has shown that such functions satisfy a modified Euler's Theorem [53] in: that T(z,x) is AHD(k_1,k_2,k_3) if and only if:

$$k_1 z' \cdot \nabla_z T(z,x) + k_2 x' \cdot \nabla_x T(z,x) = k_3 T(z,x) .$$

In general, since $T(z,x) = 0$ for efficient production, we will also refer to T(z,x) as AHD(k,1) where $k = k_1/k_3$ if it satisfies either of the above statements. It is also possible to show that $MRTS_{ij}(z,x)$ and $MRTr_{ij}(z,x)$ are independent of scale if T(z,x) is AHD(k_1,k_2,k_3).

B.1.2.2 Homotheticity

Homotheticity is a very important generalization of homogeneity. Many of the more popular production functions are homethetic, and homotheticity of the production function results in a very special structure for the cost function. Homotheticity was initially developed by Shephard [73].

A production function f(x) is homothetic if there exist functions d(u) and h(x) with u a scalar, d(u) monotonically non-decreasing and h(x) H.D.1 such that:

$$f(x) = d(h(x)) \qquad \forall x .$$

In other words if f(x) can be written as a rescaling of a H.D.1 function, it is homothetic. All homogeneous functions are homothetic. The reverse is not true; let $d(u) = e^u$ and $h(x) = x$ (x of size one). The resulting function is homothetic but not homogeneous of any degree.

If $\nabla f \neq 0$ then f is homothetic[8] if and only if $MRTS_{ij}(x)$ is H.D.0. in x $\forall$ i,j [52]. Thus independence of scale of the MRTS is a property of homothetic functions. Geometrically, this means that isoquants are radial expansions of the unit isoquant, i.e. Q(Z) can be geometrically constructed by passing rays from the origin through Q(1). This is shown in Figure B-2.

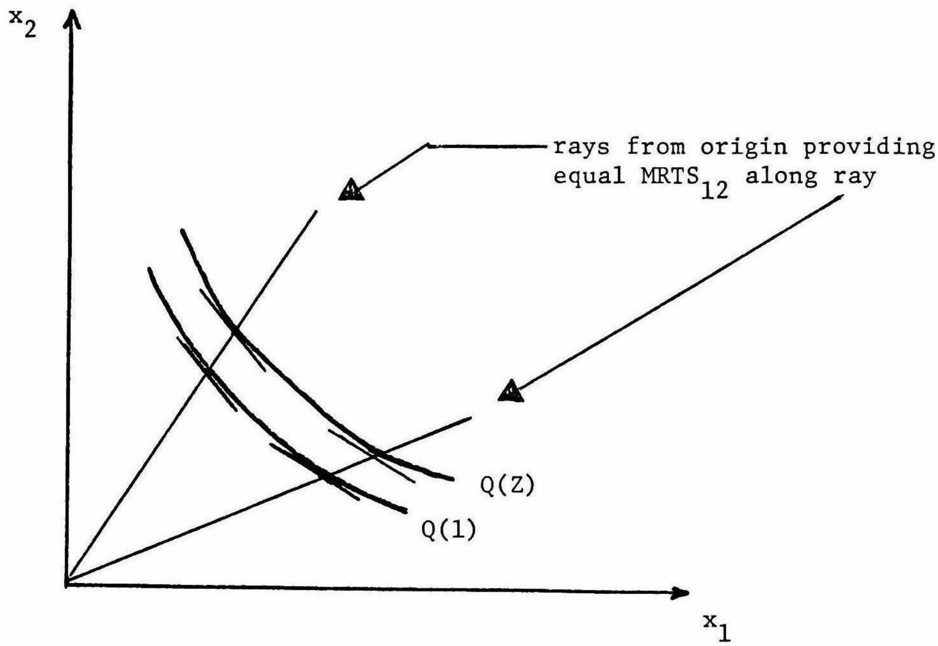


Figure B-2: HOMOTHETIC ISOQUANTS

This again illustrates the rescaling concept behind homotheticity.

This intuitive notion underlies the definitions put forward by Shephard [73, Ch. 10] and Jacobsen [47] and an alternative, more general notion in McFadden [33,Ch.I.1] (which is attributed to Hanoch). Shephard's definition is a straightforward extension of the single output definition to transformation functions that are input/output separable, i.e. we assume[9] $T(z,x) = g(z) - f(x)$. Further let $f(x)$ be homothetic and let $g(z)$ have properties (1), (2) and (3) of a production function with the added properties of: (4) quasiconvexity (i.e. $-g(z)$ is quasiconcave); (5) if $z' \geq z$ and $z' \neq z$ then $g(z') > g(z)$ and (6) as z becomes arbitrarily large, so does $g(z)$ (i.e. $g(z)$ unbounded for unbounded z). Notice that the function $g(\cdot)$ acts as an aggregation function on z; such a function may not exist. The basic notion of the definition is to place the homotheticity properties in $f(x)$ and use $g(z)$ as a surrogate output measure.

McFadden's definition, on the other hand, does not require input/output separability. A transformation will be <u>input-homothetic</u> if there exists a function $\alpha(\lambda,z)$, λ a positive scalar, with $\alpha(\lambda,z)$ increasing in λ and $\alpha(0,z) = 0$ such that:

$$V(z) = \alpha(||z||, z/||z||)\ V(z/||z||)$$

where $V(z) = \{x | T(z,x) \leq 0\}$ (the input requirements set of footnotes 3 and 4) and $||z||$ is a <u>norm</u> of z, e.g.:

$$||z|| = \left(\sum_{i=1}^{m} z_i^2\right)^{1/2} .$$

This definition is most easily understood by examining the single output case. Here $z = Z$, $||z|| = Z$ and therefore $z/||z|| = 1$. Then the definition reduces to:

$$V(Z) = \alpha(Z,1)V(1).$$

Thus $\alpha(Z,1)$ is the scaling effect on the unit isoquant represented by $V(1)$. In the multiple output case $z/||z||$ is a normalized output and $\alpha(||z||, z/||z||)$ acts as a scaling multiplier.

The two definitions will have somewhat different effects on the structure of cost functions to be discussed in section B.2. The two conditions are the same when $T(z,x)$ is separable in inputs and outputs, which is one type of separability to be discussed below.

B.1.2.3 Separability of the Production Function

The literature on separability (i.e. the ability to construct aggregate variables from disaggregate variables) is extensive; we will not attempt to review it in depth here. Instead we will provide a very basic overview of the area of separability which will be a primary focus for the second year's work.

Issues of functional structure were addressed by Leontief [56] and Sono [74]. An excellent overall reference is Blackorby, Primont and Russell [6]. Two questions addressed by the literature are as follows.

1) Under what conditions can one rewrite the function

$$Z = f(x_1,\ldots,x_n)$$

as

$$Z = f(g_1(x_1,\ldots,x_i),g_2(x_{i+1},\ldots,x_j),\ldots,g_\ell(x_k,\ldots,x_n))$$

or perhaps as

$$Z = f(\sum_i g_i(x_{j_i},\ldots,x_{k_i})) ;$$

in other words, form subaggregates (for example a labor variable to represent all different types of labor) of non-overlapping subsets of variables?

2) Under what conditions can one separate inputs from outputs in a transformation function, i.e. when can we write $T(z,x) = g(z) - f(x)$? Notice that both g and f act as aggregation functions with an aggregate input being just balanced by an aggregate output. Many technologies are apparently not separable this way (see, for example, [33, Ch. V.1], [10] and [13]).

Two types of separability have dominated the literature: weak and strong (see [6] for others). To define these, let P be a partition of the N indices of the a function h(x):

$$N = \{1,\ldots,n\}$$
$$\downarrow$$
$$\{N_1,\ldots,N_p\} \quad \text{Partition into p parts}$$

with 1) $N_i \cap N_j = \emptyset \qquad i \neq j \quad$ (mutually exclusive)

2) $N_1 \cup N_2 \cup \ldots \cup N_p = N \quad$ (exhaustive).

Thus this partitions the x-vector into p parts, i.e.:

$$(x_1,\ldots,x_n)$$
$$\downarrow$$
$$(x^1,\ldots,x^p)$$

in correspondence with the partitioning of the indices. Separability will be concerned with the effect of changes of a variable on the MRTS of other variables, i.e. we will examine when

$$\frac{\partial(h_i/h_j)}{\partial x_k} = 0 \qquad (\nabla h \neq 0)$$

where, as usual, the subscript on h refers to partial derivative. Now we can define strong (S) and weak (W) separability

$$\underline{\text{h is S}} \quad \text{if} \quad \frac{\partial(h_i/h_j)}{\partial x_k} = 0 \quad \begin{cases} \forall_i \in N_u \\ \forall_j \in N_v \\ \text{and } \forall_k \notin N_u \cup N_v \text{ ;} \end{cases}$$

$$\underline{\text{h is W}} \quad \text{if} \quad \frac{\partial(h_i/h_j)}{\partial x_k} = 0 \quad \begin{cases} \forall_{i,j} \in N_u \\ \forall_k \notin N_u \text{ .} \end{cases}$$

In words, h is strongly separable (S) in the partition P if when we pick variables from two parts of the partition (part N_u and part N_v) and compute their MRTS, it is independent of changes in variables <u>not</u> in either N_u or N_v. If this holds for <u>all</u> variables in all the parts of the partition then h is S. Weak separability doesn't require us to have i and j tested in different parts. In other words, weak separability tests each element of the partition against the elements of other subvectors in the partition.

Thus, for example, the following function

$$h(x) = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \ldots x_n^{\alpha_n}$$

is itself strongly separable and if we form two functions

$$h^1(x) = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \ldots x_m^{\alpha_m}$$

$$h^2(x) = x_{m+1}^{\alpha_{m+1}} \cdot x_{m+2}^{\alpha_{m+2}} \ldots x_n^{\alpha_n}$$

then h is strongly separable in the partition $\{N_1, N_2\}$ with $N_1 = \{1,\ldots,m\}$ and $N_2 = \{m+1,\ldots,n\}$.

Goldman and Uzawa [36], have related the S and W conditions to certain general functional forms. Berndt and Christensen [4] have (for homothetic production functions) related S and W conditions to constraints on elasticities of substitution. The elasticities of substitution attempts to measure the sensitivity of the optimal input factor mix to changes in the MRTS. For example, if production is a function of capital (K) and labor (L) alone, i.e.

$$Z = f(K,L)$$

then σ, the <u>elasticity of substitution</u>, is defined as [44]:

$$\sigma = - \left. \frac{f_K/f_L}{K/L} \cdot \frac{d(K/L)}{d(MRTS_{KL})} \right|_{Z \text{ fixed}} .$$

When the production function has more than two factors a number of possible measures can be constructed (see McFadden, Ch. IV.1 in [33]). If f(x) is homothetic then the Allen Elasticity of Substitution [1] (called AES) can be written as

$$\sigma_{ij} = \frac{\sum_{k=1}^{n} x_k f_k}{x_i x_j} \cdot \frac{|(\nabla^{2B} f)_{ij}|}{|\nabla^{2B} f|} ,$$

where $|\nabla^{2B} f|$ is the determinant of the bordered Hessian matrix (see footnote 5) and $|(\nabla^{2B}f)_{ij}|$ is the determinant of the i,j cofactor of $\nabla^{2B}f$. It will turn out that the σ_{ij} are computable from cost function information. Bendt and Christensen [4] relate restrictions on the σ_{ij} to issues of aggregation. This work is extended to non-homothetic production functions by Russell [70].

B.1.3. Examples of Production and Transformation Functions

In this section we will provide some examples of production functions, culminating with the most general forms currently in use.

B.1.3.1 Leontief Production

The Leontief (or fixed productions) production function is the following

$$Z = \min\left(\frac{x_1}{a_1}, \ldots, \frac{x_n}{a_n} \right)$$

where the $a_i > 0$. Thus inputs are used in fixed proportions (dictated by the a_i). Thus the isoquants (curves in the input space of constant output level) are corner or L-shaped as shown in Figure B-3. There is no substitution between factors: more of any factor will be wasted unless all factors are

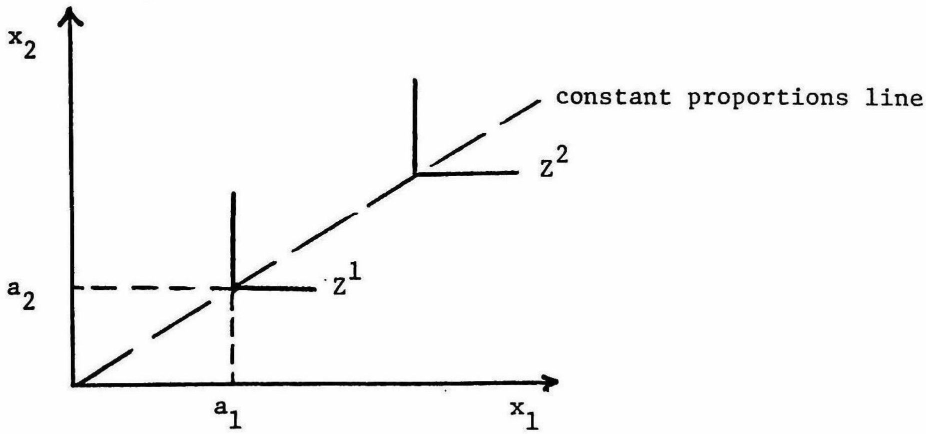


Figure B-3: LEONTIEF PRODUCTION

increased proportionately. This function is H.D.1 (i.e. constant returns-to-scale). Furthermore, if one views the above as a process and there are other processes available (i.e. other processes that entail different proportions) then there is a possibility of substitution between processes, as shown in Figure B-4.

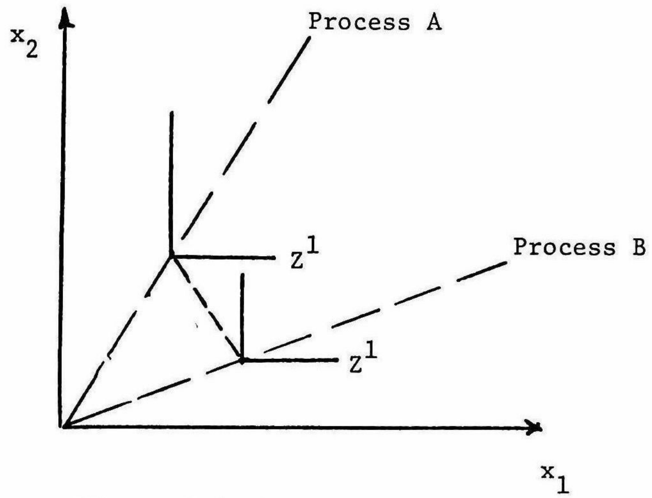


Figure B-4: SUBSTITUTION BETWEEN LEONTIEF PROCESSES

Thus this production function is not as limited as it seems at first glance. It is the ability to substitute between processes that has made this function so useful: linear programming models are based on Leontief production processes (each column in an L.P. being a fixed proportions production process).

It should be noted that this function is not differentiable. Thus, for this case we rely upon our original condition (3) for production functions.

B.1.3.2 Cobb-Douglas Production

This function has been extremely popular for a number of years. It is written[10] as follows:

$$Z = A \prod_{i=1}^{n} x_i^{\alpha_i} \qquad \sum \alpha_i = 1, \quad A > 0 . \qquad \text{(B-3)}$$

The Cobb-Douglas production function is H.D.1. A more general version is shown below (which is H.D.v):

$$Z = (A \prod_{i=1}^{n} x_i^{\alpha_i})^v \qquad \sum \alpha_i = 1,\ A > 0,\ v > 0 \qquad \text{(B-4)}$$

which is obviously homothetic. In this case any value of returns-to-scale is possible if (B-4) is estimated (see [62]). For what follows, we continue our analysis in the standard Cobb-Douglas, (B-3).

First, all factors are essential, i.e. if any one is zero then the function ascribes zero output to the process under study. This is not always a desirable result. Substitution between factors is possible (though complete substitution is not since all factors are essential). The elasticity of substitution, σ_{ij}, is constant and equal to one for all i,j pairs. Christensen and Greene [15] test for unitary elasticities of substitution by constraining certain coefficient estimates in their cost function estimation. We will return to this later.

B.1.3.3 Arrow-Chenery-Minhas-Solow CES Function and the CET/CES Transformation Function

Motivated by certain empirical evidence of the relationship between the log of value added per labor unit and the log of the wage rate, Arrow, Chenery, Minhas and Solow developed a general production function that was H.D.1, had a constant elasticity of substitution (not necessarily equal to one) and satisfied a simple model that explained (to some degree) the empirical results. The production function is called the constant elasticity of substitution production function (CES) and can be written as:[11]

$$Z = A\left[\sum_{i=1}^{n} \alpha_i x_i^{\rho}\right]^{1/\rho} \qquad \alpha_i > 0, \quad 1 > \rho$$

in which case $\sigma_{ij} = \sigma = \frac{1}{1-\rho}$ for all i,j. Special cases are the Leontief (when $\sigma \to 0$), the Cobb-Douglas ($\sigma = 1$), and the perfect substitute case ($\sigma \to +\infty$) where output is simply a weighted sum of inputs.

Powell and Gruen [68] were apparently the first to employ the CES function as a multiple output function to form the CET (constant elasticity of transformation) function. Joining the CET and CES functions, and assuming $T(z,x) = g(z) - f(x)$ we have:

$$T(z,x) = \left(\sum_{i=1}^{m} \delta_i z_i^{b}\right)^{1/b} - A\left(\sum_{i=1}^{n} \alpha_i x_i^{\rho}\right)^{k/\rho}$$

where k is the degree of homogeneity of the CES function (see note 11). Hasenkamp [42] has estimated such a function using cross-section data on U.S. railroads for 1929 and 1936.

B.1.3.4 Flexible Functional Forms: The Translog and the Generalized Leontief

Within the last ten years a number of reasonably general production functions have been developed. These are called flexible functional forms (see, e.g. [33, Ch. II.1], [6]). A general representation of such forms in the following:

$$\psi(f(x)) = \alpha_{00} + \sum_i \alpha_{i0}\phi_i(x_i) + \frac{1}{2}\sum_i \sum_j \alpha_{ij}\phi_i(x_i)\phi_j(x_j) \qquad \text{(B-5)}$$

where $f(x)$ is the production function, the α's are coefficients and $\psi(\cdot)$ and $\phi_i(\cdot)$ are suitable functions. To be more precise we have the following:

1) Transcendental Logarathmic Production Function[12] [14]

$$\psi(u) = \ell n(u)$$
$$\phi_i(u) = \ell n\ u$$

and thus we have:[13]

$$\ell n f(x) = \alpha_{00} + \sum_i \alpha_{i0} \ell n x_i + \frac{1}{2} \sum_i \sum_j \alpha_{ij} \ell n x_i \, \ell n x_j \; ; \qquad \text{(B-6)}$$

2) Generalized Leontief [22]

$$\psi(u) = u$$

$$\phi_i(u) = \sqrt{u}$$

yielding:

$$f(x) = \alpha_{00} + \sum_i \alpha_{i0} \sqrt{x_i} + \frac{1}{2} \sum_i \sum_j \alpha_{ij} \sqrt{x_i} \sqrt{x_j} \, . \qquad \text{(B-7)}$$

These functions can either be viewed as exact representations of technology or as approximations (second order) to an arbitrary technology. Lau [54] indicates that two notions of approximation have been used. McFadden [33, Ch. II.2] has viewed the flexible form as a second order approximation if first and second derivatives of the approximation are the same as the true function at the point of approximation. Christensen, et al [14] have viewed the flexible form as an approximation in the sense of a Taylor's series expansion.

In both views, the notion of an approximation is a local notion. While certain functional properties are globally inheritable by an approximation, a significant caution must be observed. Simply put, the factors that contribute to a good approximation and the factors that contribute to a good statistical estimation can be diametrically opposite to one-another. Approximations which are locally good depend on tightly packed data. On the other hand, good experimental design procedure usually calls for as great a dispersal of data points as is possible. This point is recognized in [33, Ch. II.1].

One of the major justifications for using the flexible forms is that many of the more standard production functions are sub-cases when restrictions are placed on the form. For example, in the translog (transcendental logarithmic) form, setting the second order coefficients to zero (i.e. $\alpha_{ij} = 0,\ i,j \geq 1$) yields the Cobb-Douglas case. The translog is also an approximation to the CES production function and others (see [14]). A more detailed review of such forms, their advantages and their failings is given in [33, Ch. II.1].

As an example of estimating and testing a model, consider the translog production function above. It is easy to show that the restrictions for the function to be H.D.1 are (assuming $\alpha_{ij} = \alpha_{ji}$, see note 12):

$$1)\quad \sum_i \alpha_{i0} = 1\ ;$$

$$2)\quad \sum_j \alpha_{ij} = 0 \qquad i = 1,\ldots,n\ .$$

To further restrict the form to be Cobb-Douglas (i.e. unit elasticities of substitution), one sets $\alpha_{ij} = 0$ for all i,j. Thus a procedure would be as follows.

1) Estimate the unrestricted function (with $\alpha_{ij} = \alpha_{ji}$).
2) Estimate the model with the H.D.1 restriction and test the new model against the unrestricted model (using a F-test or a likelihood ratio statistic).
3) If the restricted model can not be rejected, proceed to the next set of restrictions; otherwise stop.

B.2. Cost

B.2.1. Definition of the Cost Minimization Problem and Related Cost Functions: Long and Short Run Cost Functions

Let $p = (p_1,\ldots,p_n)'$ be an n-vector of given factor prices, i.e. the firm cannot affect p through individual firm actions. Moreover, we assume[14] $p > 0$. Finally we assume that the firm attempts to use the factors of production as efficiently as possible, i.e. for any specified output level, the firm chooses the input vector that produces the required output at minimum cost. Since cost is $p'x = \sum p_i x_i$ then the firms cost minimization problem (CMP) is as follows:

$$\text{(CMP)} \qquad \min_{x} \quad p' \cdot x \qquad \text{s.t.} \quad T(z,x) \leq 0$$

where p and z are given. Here we have written the problem for a cost minimization over a transformation function. We shall continue with this form with the understanding that the production function case is a subcase of (CMP).

The conditions on $T(z,x)$ guarantee that the solution to (CMP) exists and is unique: the objective function is linear and the set of x in the constraints is convex. If we vary z, holding p fixed a function is traced out relating minimal cost $C = \sum p_i x_i^*(z,p)$ (where $x_1^*(z,p)$ is the optimal solution to (CMP) for given z and p) to z and p. This is called the cost function (or long run cost function to indicate that all factors have been allowed to adjust to optimality) and is written:

$$C(z,p) \qquad \text{(B-8)}$$

The conditions on T(z,x) imply that the following characteristics of C(z,p) can be proved (see [73], [33, Ch. I.1], or [81]):

1) C(z,p) is H.D.1 in p, i.e. $C(z,\lambda p) = \lambda C(z,p) \ \forall \lambda > 0$;

2) C(z,p) is monotonic nondecreasing in p:

$$p' \geq p \Rightarrow C(z,p') \geq C(z,p) ;$$

3) C(z,p) is concave in p, i.e.

$$C(z, \delta p + (1-\delta)p') \geq \delta C(z,p) + (1-\delta)C(z,p')$$

for all δ such that $0 \leq \delta \leq 1$;

4) C(z,p) is continuous in p (for $p > 0$) .

The first condition reflects the obvious result that if all prices increase by the same proportion, so will the costs since the uniform price change will <u>not</u> affect the choice of the factor levels (since <u>relative</u> prices didn't change). The second condition is also straightforward: since C(z,p) represents minimal costs, one should not be able to reduce costs by <u>increasing</u> factor prices. The intuition for the third property takes more effort. Let (p,x^*) be the price and optimal quantity of inputs for some output level. This results in a cost C^*. Now if, say, just p_1 is increased slightly (to p_1'), then a slight reduction in x_1^* will have to be made, thereby not changing costs in proportion to the slight p_1 change. Figure B-5 illustrates the effect (see [33], [81]). Thus C(z,p) is concave. Continuity (property 4) follows from the concavity of C(z,p) (concave functions are continuous, except possibly at the boundary).

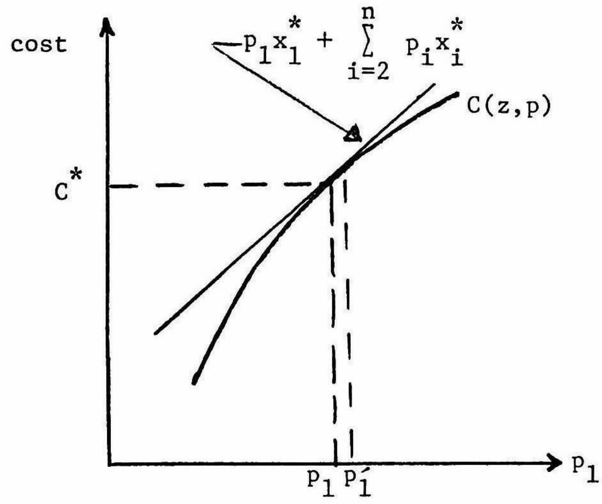


Figure B-5: COST FUNCTION CONCAVITY

An observation is in order. The analysis above and that which will follow rests on two important assumptions: (1) the firm faces fixed, known factor prices; (2) the firm minimizes costs. The assumptions cut two ways. On the one hand, we may develop cost functions for monopolists as well as perfect competitors: no issues about the market(s) for the output were raised. Moreover, as long as the entity being studied is trying to efficiently produce output we need not concern ourselves with problems of whether the firm is profit-maximizing or regulated to provide "socially optimal" output. However, it is important that the firm face reasonably competitive factor markets, something that may not be true for very large firms.

We can now define some standard related cost functions:

1) Marginal Cost: $MC_i(z,p) = \dfrac{\partial C(z,p)}{\partial z_i} \qquad i = 1,\ldots,m;$

2) For single output models

Average Cost: $AC(Z,p) = \dfrac{C(Z,p)}{Z} \qquad Z > 0.$

Using Shephard's lemma [73], [81] the optimal factor demand equations $x_i^*(z,p)$ are simply:

$$x_i^*(z,p) = \frac{\partial C(z,p)}{\partial p_i} \qquad i = 1,\ldots,n \ .$$

Some manipulation also shows that the elasticity of cost with respect to a factor price is:

$$\frac{\partial C(z,p)}{\partial p_i} \cdot \frac{p_i}{C(z,p)} = x_i^*(z,p) \ \frac{p_i}{C(z,p)} \qquad \text{(B-9)}$$

$$= \frac{p_i x_i^*(z,p)}{C(z,p)} \qquad \text{(B-10)}$$

which is simply the factor share, i.e. the percentage of cost spent on factor i. Notice that the left-hand-side of (B-9) can also be written:

$$\frac{\partial \log C(z,p)}{\partial \log p_i}$$

which will be especially useful in translog cost function studies (where log C(z,p) is expressed in terms of log z_i and log p_i).

If we restrict z to be a single output Z then a graph of a typical C(Z,p) can be drawn as shown in Figure (B-6). The cost function illustrated reflects economies of scale (increasing returns-to-scale) for outputs up to $\bar{Z}$ and diseconomies of scale for outputs greater than $\bar{Z}$. The result is a classical U-shaped average cost function with a minimum at $Z = \bar{Z}$. This will be the optimal size of the firm.

The above cost function represents the cost of producing output z given factor prices p assuming all factors are free to adjust their levels so as to minimize cost. This assumption is not always valid. Regulated common

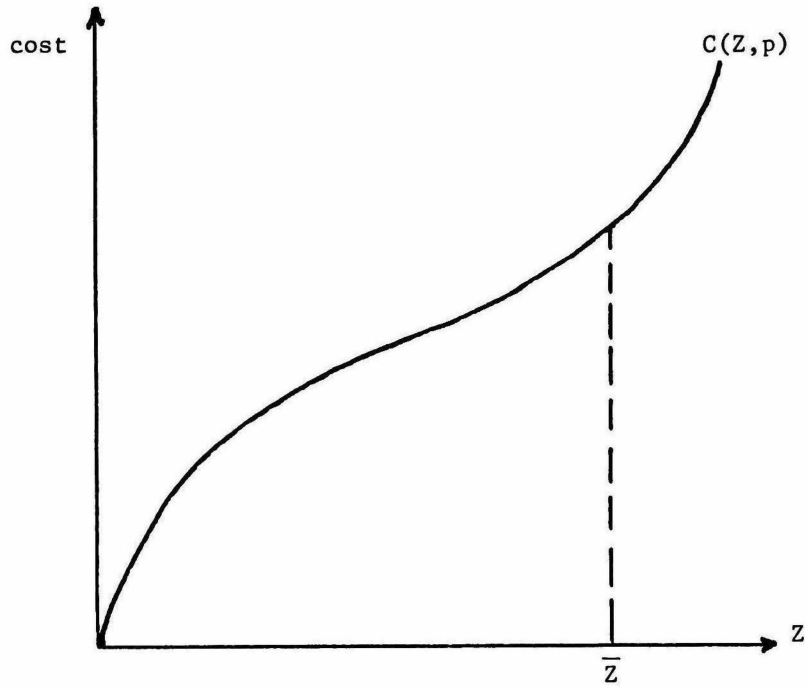


Figure B-6: COST FUNCTION

carriers often cannot adjust their capital stock through abandonment, for example, of service. There can be any number of reasons why, at least for a short period of time, a firm can not optimally adjust certain factors of production as it increases or decreases its output. This is an especially important issue when we try to estimate the firms C(z,p) function, since this means that some of the observations will lie on C(z,p) but some of them will lie <u>above</u> C(z,p). Notice that no observations could lie below C(z,p) by the definition of the function. Therefore if we attempt to pass a curve through a scatter of points we are doomed to overestimate the cost function. This was recognized by over a decade ago by Eads [24], Eads, Nerlove and Raduchel [25] and Keeler [48] and has been employed by a number of investigators since then (see [49], [41], [60], [32], [12]).

To formulate the short-run cost function, we partition x into two subvectors: x^v and x^f (for variable and fixed factors):

$$x = \begin{pmatrix} x^v \\ \\ x^f \end{pmatrix}$$

where x^v is of dimension $n_1 \geq 1$ and x^f is of dimension $n-n_1$. Now (CMP) becomes the short-run CMP (SRCMP):

(SRCMP) $$\min \quad p' \cdot \begin{pmatrix} x^v \\ x^f \end{pmatrix}$$

$$\text{s.t.} \quad T(z,x^v,x^f) \leq 0 \,.$$

Since the partition of x induces a similar partition on p and since $(p^f)'x^f$ is fixed then SRCMP becomes the short-run <u>variable</u> cost minimization problem

(SRVCMP) $$\min \quad (p^v)' \, x^v$$

$$\text{s.t.} \quad T(z,x^v,x^f) \leq 0 \,.$$

Again if we vary z the result is a cost function $C(z,p^v;x^f)$. Note that only p^v shows up in the cost function. The notation shows that the cost function is conditioned on the values of the fixed variables x^f. <u>Total short-run costs</u> are equal to <u>short-run variable cost</u> plus <u>short-run fixed costs</u>:

$$TC(z,p^v;\, x^f) = C(z,p^v;x^f) + (p^f)' \cdot x^f$$

Observe that short-run marginal costs $MC_i(z,p^v;x^f)$ can be calculated from either $TC(z,p^v;x^f)$ or $C(z,p^v;x^f)$. Thus

$$MC_i(z,p^v;x^f) = \frac{\partial C(z,p^v;x^f)}{\partial z_i} \qquad i = 1,\ldots,n \,.$$

If there is a single output, average cost is well defined and we have the definitions of <u>short-run average cost</u>, <u>short-run average variable cost</u> and <u>short-run average fixed cost</u>:

$$1)\quad AC(Z,p^v;x^f) = \frac{TC(Z,p^v;x^f)}{Z} \qquad Z > 0 \; ;$$

$$2)\quad AVC(Z,p^v;x^f) = \frac{C(Z,p^v;x^f)}{Z} \qquad Z > 0 \; ;$$

$$3)\quad AFC(Z,p^v;x^f) = \frac{(p^f)' \; x^f}{Z} \qquad Z > 0.$$

Again, Shephard's lemma yields the <u>short-run factor demand equations</u> and the <u>short-run factor share equations</u> for the variable factors (now with multiple outputs):

$$x_i^{v*}(z,p^v;x^f) = \frac{\partial C(z,p^v;x^f)}{\partial p_i^v} \qquad i = 1,\ldots,n_1$$

$$S_i^v(z,p^v;x^f) \equiv \frac{p_i^v x_i^{v*}(z,p^v;x^f)}{C(z,p^v;x^f)} = \frac{\partial \log C(z,p^v;x^f)}{\partial \log p_i^v} \qquad i = 1,\ldots,n_1$$

where $S_i^v(z,p^v;x^f)$ denotes the share of costs attributable to variable factor i.

Finally, the short-run functions provide the long-run function:

$$C(z,p) = \min_{x^f}(C(z,p^v;x^f) + (p^f)'\cdot x^f) \; . \qquad \text{(B-11)}$$

Thus, estimating the short-run variable cost function provides estimates of the short-run marginal cost functions, the factor demand and factor share equations. It should be noted that the estimated share equations from the variable cost function will not be the same as the estimated share equations from the total cost function. In fact the following relationship holds:

$$S_i^v(z,p^v;x^f) \frac{C(z,p^v;x^f)}{TC(z,p^v;x^f)} = S_i(z,p^v;x^f) \equiv \frac{\partial \log TC(z,p^v;x^f)}{\partial \log p_i^v}$$

$$i = 1,\ldots,n_1.$$

Thus, caution must be used in interpreting the estimated equations.

Furthermore, if p^f is known, the long-run cost function can be recovered by solving the optimization problem in (B-11) above. Thus by specifying a technology and a vector of prices, we can derive functional forms (in some cases explicitly as will be shown below) that can be estimated.

B.2.2. Cost Functions and Implied Technology

In the previous section we defined cost functions for technologies that were convex in their input factors (e.g. for production functions with isoquants as depicted in Figure B-1a). If the technologies are not convex in the sense shown in Figure B-1b then the cost function will be derived for the convexified technology. This is illustrated in Figure B-7.

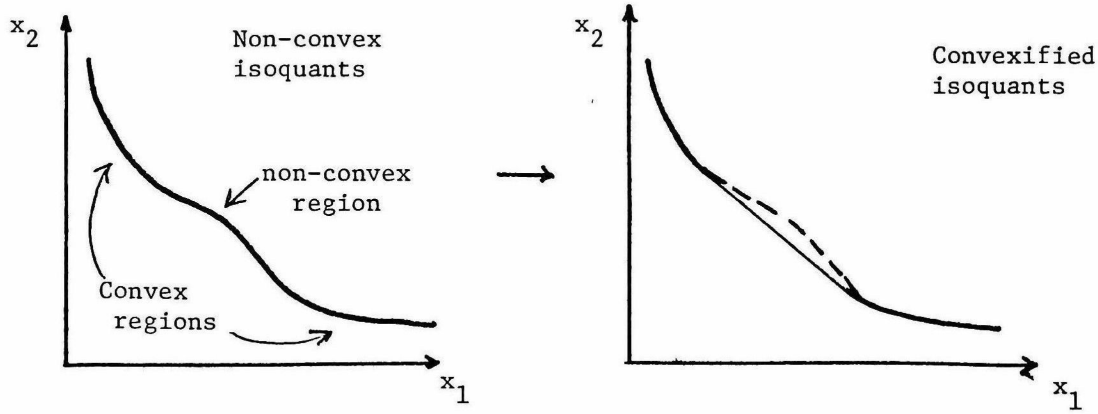


Figure B-7: CONVEXIFICATION OF NON-CONVEX ISOQUANTS

This means that all the technologies that have the same convexification have the same cost function. This is really not a problem, as McFadden [33, Ch. I.1] points out, since it can be shown that if the firm is facing given input factor prices and minimizing costs, then the firm never would choose an input mix that would place it in the non-convex region, i.e. it would act as if it worked with the convexification anyway. Thus, while a number of technologies can give rise to the same cost function, the convexified technology is all we need care about since the firm (if it obeys our assumptions on factor prices and cost minimization) would never be observed operating in the non-convex region anyway.

Now consider instead what information you could draw from a cost function $C(z,p)$. If you were given the function and told that it came from a cost minimizing firm that faced fixed prices, you could form the following set:

$$V^*(z) = \{x \mid p'x \geq C(z,p) \quad \text{for all } p > 0\}.$$

Geometrically, $V^*(z)$ is illustrated in Figure (B-8),

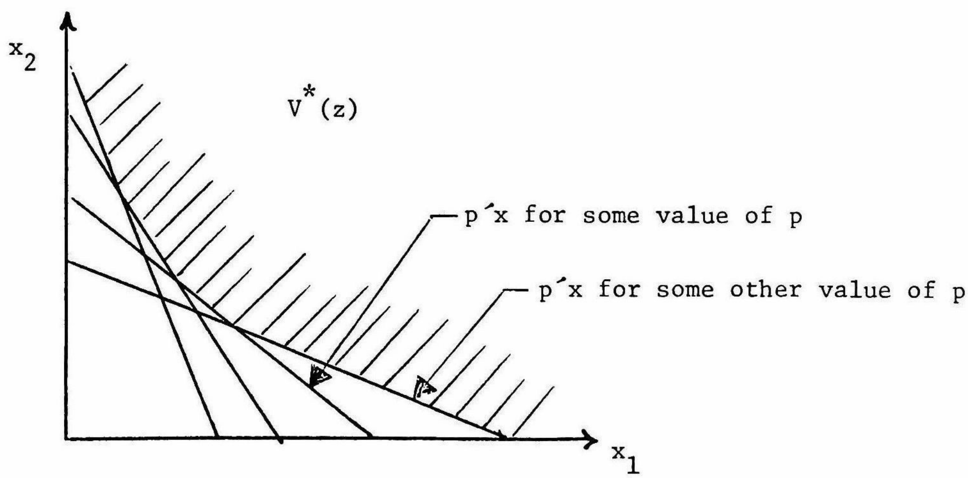


Figure B-8: CONSTRUCTING V*(z)

i.e. it is the set of points lying above all the straight lines. Thus, if we pick an output vector z and vary p and look at all the x that are northeast of the lines $p'x$, we have $V^*(z)$. There are an infinite number of such lines and the result is a curve that looks very much like an isoquant and the region to the northeast of it, as seen in Figure B-9 below.

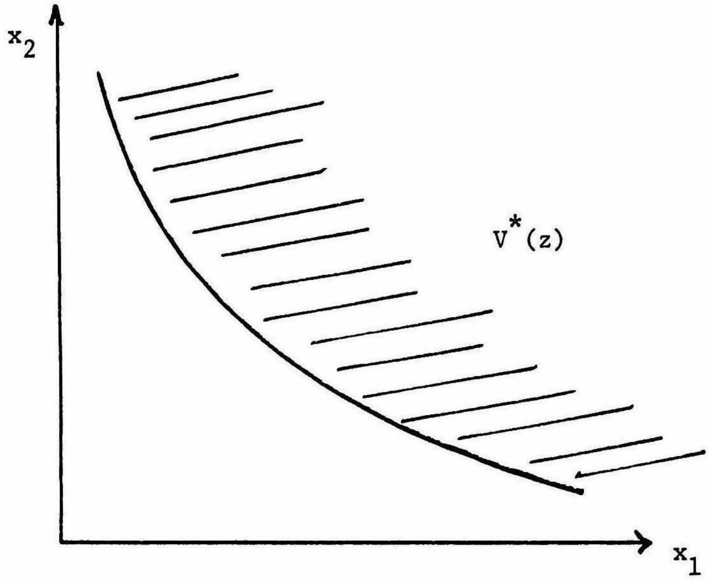


Figure B-9

It can be shown that $V^*(z)$ is always convex, irrespective of the technology that gave rise to $C(z,p)$. It also satisfies the properties that we require of a technology (see [81]). In fact we will call $V^*(z)$ the input requirements set (see footnotes 3 and 4) of our implied technology. We now have the following very important duality results (see, e.g. [81]).

1) Technology, Prices → Cost Function → Implied Technology .

2) If the original technology is convex in its inputs then the implied technology will be identical to it.

3) If the original technology is not convex in its inputs then the implied technology will be identical with the convexification of the original technology.

Therefore, a properly constructed cost function will provide all the information of interest about a technology (if the firm obeys our assumptions on fixed factor prices and cost minimization).

Put more practically, we can estimate either a production (or transformation) function or a cost function and get what we want to know about the underlying technology. We can use a cost function to inform us about the following.

1) Homotheticity .
2) Homogeneity .
3) Returns-to-scale .
4) Separability .

We will consider these in turn.

B.2.2.1 Homotheticity

A very useful and interesting result concerning the structure of the cost function occurs if we let the production function f(x) be homothetic, i.e.

$$Z = f(x) = d(h(x))$$

where d(•) is a monotonic increasing continuous function and h(•) is H.D.1. It can be shown [73] that if f(x) satisfies the above, then there is an inverse function to d (which we'll call s(•)) such that h(x) = s(Z). Now s(Z) is simply a scalar so we have the following result:

$$C(Z,p) = \min_x \{p'x \mid f(x) = Z\}$$

$$= \min_x \{p'x \mid h(x) = s(Z)\}$$

$$= \min_x \{p'x \mid h(\frac{x}{s(Z)}) = 1\}$$

$$= s(Z)\cdot \min_w \{p'w \mid h(w) = 1\} \qquad w = \frac{x}{s(Z)}$$

$$= s(Z)\cdot \ell(p)$$

To explain: the first line is a statement of (CMP), the second line the result of the transformation discussed above. In the third line we capitalize upon h(x) being H.D.1 and s(Z) being a scalar. Thus h(x) = s(Z) means h(x/s(Z)) = 1, i.e. if we divide every element of x by s(Z) then the output is 1. In the fourth line we change variables letting the vector w be the vector x with every element divided by s(Z). To maintain the equality we must multiply by s(Z). Finally in the fifth line we recognize that the minimum on line four will be purely a function of p (since w will be optimized out). Shephard has proved a more general version of the above (and its converse) and thus we have the following theorem.

f(x) homothetic $\Leftrightarrow$ C(Z,p) is multiplicatively separable in z and p, i.e.:

$$C(Z,p) = s(Z)\cdot \ell(p)$$

Note that since $C(Z,p)$ must be H.D.1 in prices p we know that $\ell(p)$ is H.D.1. Furthermore $s(0) = 0$, $s(Z) > 0$ if $Z > 0$, $s(Z)$ is continuous, etc. from the properties of the cost function.

This result can be extended to the transformation function case. Recall that Shephard assumed $T(z,x) = g(z) - f(x)$. In this case we have that $C(z,p) = g(z)\cdot\ell(p)$, i.e. again, $C(z,p)$ is multiplicatively separable if and only if the separable transformation function $T(z,x)$ is homothetic in x. Notice also that $g(z)$ is the aggregation function for the vector z.

If the transformation function is not separable then input homotheticity gains us somewhat less. McFadden [33, Ch. I.1] shows that in this case

$$C(z,p) = \alpha(||z||, z/||z||)C(z/||z||, p)$$

where $\alpha(\cdot,\cdot)$ is the scaling function discussed in section B.1.2.2 above and $||\cdot||$ is the norm function mentioned there also. What is important here is that tests for homotheticity that rely upon the multiplicative separability of $C(z,p)$ are actually testing separability <u>and</u> homotheticity together; a rejection may be a rejection of separability or homotheticity or both.

B.2.2.2 Homogeneity

Let $f(x)$ by H.D.k. Then in a manner similar to that in B.2.2.1 we see the following result.

$$\begin{aligned} C(Z,p) &= \min_x \{p'x \mid f(x) = Z\} \\ &= \min_x \{p'x \mid f(x/Z^{1/k}) = 1\} \\ &= Z^{1/k} \min_w \{p'w \mid f(w) = 1\} \qquad w = x/Z^{1/k} \\ &= Z^{1/k}\, \ell(p)\,. \end{aligned}$$

Thus, in particular, if f(x) is H.D.1 then $C(Z,p) = Z\,\ell(p)$ and vice-versa. In other words if C(Z,p) is linear in Z then f(x) is H.D.1.

It should be noted that from the above result we have that if f(x) is H.D.k then C(Z,p) is H.D.(1/k). Extending this to the multiple output case provides motivation for the following definition:

C(z,p) is output homogeneous of degree r (O.H.D.r) if $C(\lambda z,p) = \lambda^r C(z,p) \quad \forall \lambda > 0$.

It is with this definition in mind that we next consider economies of scale.

B.2.2.3 Economies of Scale

Baumol [2] has defined the notion of <u>decreasing average ray cost</u> for multiproduct firms. A firm has decreasing average ray costs if:

$$C(\lambda z,p) < \lambda C(z,p) \qquad \lambda > 1 .$$

For example, if we consider the single output case we have:

$$C(\lambda Z,p) < \lambda C(Z,p) \qquad \lambda > 1$$

$$\Downarrow$$

$$\frac{C(\lambda Z,p)}{\lambda Z} < \frac{C(Z,p)}{Z} \qquad \lambda > 1$$

which simply is the condition of declining average costs (i.e. returns-to- (or economies of) scale). Thus decreasing average ray costs should be associated with returns-to-scale, and they are (see Baumol [2]). Panzer and Willig [64] extend this notion to provide a measure of scale economies for multioutput firms. They show that the following measure captures returns-to-scale[15] in production:

$$S = C(z,p) / \sum_i z_i \frac{\partial C(z,p)}{\partial z_i} .$$

Notice that if C(z,p) is O.H.D.1 then S = 1 since by Euler's Theorem the numerator equals the denominator. Rearranging terms yields the following:

$$S = 1/\sum_i \frac{z_i}{C(z,p)} \frac{\partial C(z,p)}{\partial z_i}$$

$$= 1/\sum_i \frac{\partial \ell n C(z,p)}{\partial \ell n z_i} .$$

We shall see later that this function is particularly easy to calculate from a translog cost function. The measure above is the sum of the elasticities of cost with respect to output divided into one. The more inelastic the cost function is to output, the greater the returns-to-scale.

B.2.2.4 Separability

Analysis of the separability of production and transformation functions can be performed via the cost function. First, considering homothetic production functions Uzawa [80] has shown the following for the AES σ_{ij} (see section 2.1.2.3):

$$\sigma_{ij} = \frac{C(Z,p)C_{ij}(Z,p)}{C_i(Z,p)C_j(Z,p)} .$$

where $C_i(Z,p) = \partial C(Z,p)/\partial p_i$ and $C_{ij}(Z,p) = \partial^2 C(Z,p)/\partial p_i \partial p_j$.

Thus estimating the cost function provides estimates of the AES (the subscripts denote partial derivative with respect to price). Berndt and Christensen [4] show that weak separability ($i,j \in N_u$, $k \notin N_u$; see section B.1.2.3) implies $\sigma_{ik} = \sigma_{jk}$ ($i,j \in N_u$, $k \notin N_u$) which is true if and only if

$C_j(Z,p)C_{ik}(Z,p) - C_i(Z,p)C_{jk}(Z,p) = 0$. These conditions provide for weak separability of the cost function. Further Lau has shown that the cost function is weakly separable (strongly separable) with respect to a partition P in prices if and only if the production function is homothetically weakly separable (strongly separable) with respect to the partition P in inputs [33, Ch. I.3]. Again, this area is extensive; for further information see [6], [33, Ch. I.3], and [37] to name a few references.

B.2.3. Examples

As has been indicated above, there is a duality between production and cost: technology descriptions give rise to cost functions (when prices are incorporated) which give rise to implicit technologies.

In this section we provide cost functions for some of the production functions in section B.1. Furthermore we discuss some of the flexible functional form cost functions: the Generalized Leontief, the Hall function and the translog.

B.2.3.1 Cobb-Douglas Production and Cost

The Cobb-Douglas production function of section B.1.3.2 gives rise to the following cost function:

$$C(Z,p) = \left(A \prod_{i=1}^{n} \alpha_i^{\alpha_i}\right)^{-1} Z^{1/v} \prod_{i=1}^{n} p_i^{\alpha_i/v}$$

where v is the returns-to-scale in (B-4) in section B.1.3.2. Notice that C(Z,p) could be estimated by taking logarithms:

$$\ell n\, C(Z,p) = \alpha_0 + \sum_{i=1}^{n} \gamma_i\, \ell n\, p_i + \beta\, \ell n Z$$

where $\alpha_0 = \ell n(A \prod_{i=1}^{n} \alpha_i^{\alpha_i -1})$ $\gamma_i = \frac{\alpha_i}{v}$ and $\beta = 1/v$. For C(Z,p) to be H.D.1 in prices we would require the constraint $\sum \gamma_i = 1$.

B.2.3.2 CES Production and Cost

Referring to section B.1.3.3, the dual cost function would be as follows:

$$C(Z,p) = \frac{Z}{A} \left(\sum_{i=1}^{n} (p_i/\alpha_i)^{1-\sigma} \right)^{1/(1-\sigma)}$$

where $\sigma = \frac{1}{1-\rho}$ is the elasticity of substitution. Notice that as $\sigma \to 0$ (the Leontief case) we get a cost function that is simply a weighted sum of prices times the output level, i.e. for Leontief production

$$C(Z,p) = \frac{Z}{A} \sum_{i=1}^{n} (p_i/\alpha_i)$$

where the α_i correspond to the a_i of section B.1.3.1.

B.2.3.3 Flexible Functional Forms

The above cost functions are examples of what are known as self-dual technologies [46]. The coefficients of the production function appear in the cost function and vice versa and the dual functions are members of the same family. This is not in general the case with the flexible functional form. A transcendental logarithmic production function may not give rise to a translog cost function. The choice of which to use is thus a non-trivial one since it is possible that, for example, estimating a translog cost function and a translog production function could lead to different results (see Burgess [10]).

In what follows we will briefly describe three cost functions: the generalized Leontief (Diewert [22]), the generalized linear-generalized Leontief joint cost function (Hall [40]) and the translog (Christensen, Jorgensen and Lau [14]).

The generalized Leontief function resembles the production function of section B.1.3.4 above. It is as follows:

$$C(Z,p) = Z \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} \sqrt{p_i} \sqrt{p_j} \qquad \alpha_{ij} = \alpha_{ji} .$$

Notice that the cost function represents a Leontief production function if $\alpha_{ii} > 0$ and $\alpha_{ij} = 0$ for $i \neq j$. This is the source of its name. While this function is a second order approximation to any technology, it only admits one output.

The Hall function is an extension of the Diewert function for multiple outputs. It is:

$$C(Z,p) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k}^{m} \sum_{\ell}^{m} \alpha_{ijk\ell} \sqrt{z_k} \sqrt{z_\ell} \sqrt{p_i} \sqrt{p_j} \; .$$

Unfortunately, the Hall function assumes H.D.1 production and has a large number of parameters to estimate (n^2m^2).

Finally the translog is as follows:

$$\ell nC(z,p) = \alpha_0 + \sum_{i=1}^{m} \alpha_{i0} \ell n z_i + \sum_{j=1}^{n} \beta_{j0} \ell n\, p_j$$

$$+ \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_{ij} \ell n z_i \ell n z_j + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} \ell n p_i \ell n p_j$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{n} \gamma_{ij} \, \ell n z_i \, \ell n p_j$$

$$\alpha_{ij} = \alpha_{ji}, \; \beta_{ij} = \beta_{ji} \; \forall_{ij}.$$

Observe that if all the second order terms are zero, the translog reduces to the Cobb-Douglas. In fact, one could view the Cobb-Douglas as a first order approximation to a function (cost or production) and the translog as a second order approximation: both in logarithms.

A word is in order on factor demand equations. In general, most studies that estimate a cost function estimate it simultaneously with the factor demand equations. This provides increased efficiency in the estimation process. The factor demand equations for the Diewert form are quite simple:

$$X_i(Z,p) = Z \sum_{j=1}^{n} \alpha_{ij}(p_j/p_i)^{1/2} .$$

This is similarly true for the Hall function. On the other hand, the factor demand equations are not simple for the translog: they are non-linear in the parameters to be estimated. However, because the translog is expressed in logarithms the factor share equations are linear:

$$S_i(z,p) = \beta_{i0} + \sum_{j=1}^{n} \beta_{ij} \ln p_j + \sum_{j=1}^{m} \gamma_{ij} \ln z_j \quad .$$

Therefore in estimating the translog cost function we can append n-1 factor share equations (since the cost function makes the n^{th} equation).

In summary the principle advantages and disadvantages of the three flexible forms are as follows.

Generalized Leontief (Diewert [22]): Advantages - 1) Second order approximation to a cost function;
2) Low number of parameters to be estimated;

3) Convenient form for cost function and factor demand equations;
4) Allows easy test of non-substitution (Leontief) case;
5) Zero level of variables allowed;

Disadvantages-
1) Assume homogeneous production;
2) Assumes single output;
3) It is separability-inflexible (see below).

Generalized Linear-Generalized Leontief (Hall [40]):

Advantages-
1) Second-order approximation;
2) Multiple outputs;
3) Convenient form ;
4) Tests input/output separability;

Disadvantages-
1) Assumes constant returns-to-scale;
2) Large number of parameters to be estimated.

Translog (Christensen, Jorgensen and Lau [14]):

Advantages -

1) Second order approximation;
2) Reduces to popular forms (e.g. Cobb-Douglas, CES as limiting case);
3) Reasonable number of parameters;
4) Convenient form for estimating economies of scale;
5) Multiple outputs allowed;

Disadvantages -

1) Zero levels of variables not allowed (due to logarithms);
2) Factor demand equations non-linear in parameters (though factor share equations are not);
3) It is separability-inflexible.

Separability-inflexibility for the translog (see [6] and [19]) and the Diewert [6] means that imposing separability restrictions implicitly imposes significant structure on the aggregation functions themselves. Thus, not only is the separability of different variables being tested: the test of separability will also be testing a specific structural form for the aggregation functions. Thus rejection of the test may only reflect rejection of the forms, not the separability itself.

B.2.3.4 An Example of Finding Long-Run and Short-Run Cost Functions

To help tie-together some of the notions discussed in the sections above, this section presents a long- and short-run cost model. Assume a firm uses capital (x_1), labor (x_2) and fuel (x_3) to produce a single output (Z) following a Cobb-Douglas production function:

$$Z = A\, x_1^{\alpha_1}\, x_2^{\alpha_2}\, x_3^{\alpha_3} \;.$$

The long-run cost function C(Z,p) with $p = (p_1,p_2,p_3)'$ is found by solving (CMP):

$$\text{(CMP)} \qquad \text{minimize} \quad p_1x_1 + p_2x_2 + p_3x_3$$

$$\text{subject to} \quad Ax_1^{\alpha_1}x_2^{\alpha_2}x_3^{\alpha_3} = Z$$

which yields

$$C(Z,p) = \alpha_0 Z^{1/v}\, p_1^{\alpha_1/v}\, p_2^{\alpha_2/v}\, p_3^{\alpha_3/v}$$

where

$$v = \alpha_1 + \alpha_2 + \alpha_3$$

and

$$\alpha_0 = (A\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}\alpha_3^{\alpha_3})^{-1} \;.$$

Thus the factor demand equation for labor (x_2) is:

$$X_2(Z,p) = \frac{\partial C(Z,p)}{\partial p_2}$$

$$= \frac{\alpha_2}{v}\,(\alpha_0 Z^{1/v}\, p_1^{\alpha_1/v}\, p_2^{(\alpha_2/v)-1}\, p_3^{\alpha_3/v})$$

$$= \frac{\alpha_2}{v} \cdot \frac{C(z,p)}{p_2}$$

which means that the factor share equation for labor is

$$S_2(Z,p) \equiv \frac{p_2 X_2(Z,p)}{C(z,p)} = \frac{\alpha_2}{v}$$

which could also have been found by taking logarithms of the cost function and then computing $\partial \log C(z,p)/\partial \log p_2$.

The short-run cost function is found by fixing one or more of the variables. If we fix capital (x_1) at a given level $\bar{x}_1$ then (SRVCMP) becomes

$$\text{(SRVCMP)} \qquad \min \; p_2 x_2 + p_3 x_3$$

$$\text{s.t.} \quad A\bar{x}_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3} = Z$$

which yields the short-run variable cost function:

$$C(Z,p^v;\bar{x}_1) = \beta_0 Z^{1/u} p_2^{\alpha_2/u} p_3^{\alpha_3/u} \bar{x}_1^{-\alpha_1/u}$$

where

$$p^v = (p_2,p_3)'$$

$$\beta_0 = u(A\alpha_2^{\alpha_2} \alpha_3^{\alpha_3})^{1/u}$$

$$u = \alpha_2 + \alpha_3 .$$

It can be readily shown that we can derive the long-run cost function:

$$C(Z,p) = \min_{x_1} (C(Z,p;x_1) + p_1 x_1) .$$

The short-run factor demand function is:

$$X_2(Z,p^v;\bar{x}_1) = \frac{\alpha_2}{u} \frac{C(Z,p^v;\bar{x}_1)}{p_2} .$$

This means that the factor share function is as follows:

$$S_i^v(Z,p^v;\bar{x}_1) = \frac{\alpha_2}{u} .$$

Calculating returns-to-scale on the long-run function yields the following

$$S = \frac{C(Z,p)}{Z \cdot \frac{\partial C(Z,p)}{\partial Z}} = \frac{C(Z,p)}{\frac{1}{v} C(Z,p)} = v$$

which is the returns-to-scale parameter (see section B.1.3.2).

The system of equations to be estimated for the short-run variable cost function is as follows:

$$\log C(Z,p^v;\bar{x}_1) = \gamma_0 + \gamma_1 \log Z + \gamma_2 \log p_2 + \gamma_3 \log p_3 + \gamma_4 \log \bar{x}_1 + \varepsilon_1$$

$$\frac{p_2 x_2}{C(Z,p^v;\bar{x}_1)} = \gamma_2 + \varepsilon_2$$

where $\gamma_0, \gamma_1, \gamma_2$ and γ_3 are to be estimated, the ε_i are error terms and we have used the factor share equation for labor since there are only two factor share equations (labor and fuel, i.e. n=2) and thus n-1=1. The above system can be estimated as a seemingly unrelated equations system [79]. To enforce H.D.1 in p^v we would add the condition that $\gamma_2 + \gamma_3 = 1$. Since $1/\hat{\gamma}_1$ will estimate u, we could recover $\hat{\alpha}_2$ and $\hat{\alpha}_3$ from $\hat{\gamma}_2$ and $\hat{\gamma}_3$ (^ means estimated value).

B.2.4. The General Form of the Short-Run Variable Cost Function to be Estimated and the Associated Conditions to be Tested

In this report we present results of estimating a short-run variable cost function (see chapter 4). In this section we will provide the basic form for the cost model which will be made more specific later. We will also present the conditions on the model for homogeneity in prices, joint separability and homotheticity, homogeneity of degree k and unitary elasticities of substitution in inputs.

B.2.4.1 The Translog Short-Run Variable Cost Model

We have chosen to use the translog model for the reasons described above in section B.2.3.3. The model is as follows (ℓn means natural logarithm):

$$\begin{aligned}
\ell n\, C(z,p^v;x^f) = \alpha_0 &+ \sum_{i=1}^{m} \alpha_{i0} \ell n z_i + \sum_{i=1}^{n_1} \beta_{i0} \ell n p_i^v \\
&+ \sum_{i=1}^{n-n_1} \gamma_{i0} \ell n\, x_i^f \\
&+ \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_{ij} \ell n\, z_i\, \ell n\, z_j \\
&+ \frac{1}{2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \beta_{ij} \ell n\, p_i^v\, \ell n\, p_j^v \\
&+ \frac{1}{2} \sum_{i=1}^{n-n_1} \sum_{j=1}^{n-n_1} \gamma_{ij} \ell n\, x_i^f\, \ell n\, x_j^f \\
&+ \sum_{i=1}^{n_1} \sum_{j=1}^{m} \delta_{ij}^{pz}\, \ell n\, p_i^v\, \ell n\, z_j \\
&+ \sum_{i=1}^{m} \sum_{j=1}^{n-n_1} \delta_{ij}^{zx}\, \ell n\, z_i\, \ell n\, x_j^f \\
&+ \sum_{i=1}^{n_1} \sum_{j=1}^{n-n_1} \delta_{ij}^{px}\, \ell n\, p_i^v\, \ell n\, x_j^f
\end{aligned}$$

with $\alpha_{ij} = \alpha_{ji}$, $\beta_{ij} = \beta_{ji}$ and $\gamma_{ij} = \gamma_{ji}$. This results in a total of

$$\frac{m(m+3) + n(n+3)}{2} + mn + 1$$

parameters to be estimated.

The factor share equation for the i^{th} variable factor is as follows:

$$S_i^v(z,p^v;x^f) = \beta_{i0} + \sum_{j=1}^{n_1} \beta_{ij} \ell n\, p_j^v + \sum_{j=1}^{m} \delta_{ij}^{pz} \ell n\, z_j + \sum_{j=1}^{n-n_1} \delta_{ij}^{px} \ell n\, x_j^f \ .$$

B.2.4.2 Constraints

1) Homogeneity of degree 1 in p^v:

(a) $\sum_{i=1}^{n_1} \beta_{i0} = 1$

(b) $\sum_{i=1}^{n_1} \beta_{ij} = 0 \qquad j = 1,\ldots,n_1$

(c) $\sum_{i=1}^{n_1} \delta_{ij}^{pz} = 0 \qquad j = 1,\ldots,m$

(d) $\sum_{i=1}^{n_1} \delta_{ij}^{px} = 0 \qquad j = 1,\ldots,n-n_1$.

2) Separability of T(z,x) and Input Homotheticity

$$\delta_{ij}^{pz} = \delta_{ij}^{zx} = 0 \qquad \forall_{ij} \ .$$

3) Homogeneity of degree k (k given) in production $(\lambda^k Z = f(\lambda x))$ and almost homogeneity of degree (k,1) $(T(\lambda^k z,\lambda x) = 0)$:

$$\text{(a)} \quad k \sum_{i=1}^{m} \alpha_{i0} + \sum_{j=1}^{n_1} \gamma_{i0} = 1$$

$$\text{(b)} \quad k \sum_{i=1}^{m} \alpha_{ij} + \sum_{j=1}^{n-n_1} \delta_{ij}^{zx} = 0 \qquad i = 1,\dots,m$$

$$\text{(c)} \quad \sum_{i=1}^{n-n_1} \gamma_{ij} + k \sum_{i=1}^{m} \delta_{ij}^{zx} = 0 \qquad j = 1,\dots,n-n_1$$

$$\text{(d)} \quad k \sum_{j=1}^{m} \delta_{ij}^{pz} + \sum_{j=1}^{n-n_1} \delta_{ij}^{px} = 0 \qquad i = 1,\dots,n_1 .$$

4) Cobb-Douglas Production (i.e. unitary elasticities of input substitution):

$$\alpha_{ij} = \beta_{ij} = \gamma_{ij} = \delta_{ij}^{zx} = \delta_{ij}^{pz} = \delta_{ij}^{px} = 0 \qquad \forall_{i,j} .$$

The above are necessary and sufficient. Conditions similar to (1), (2) and (3) are discussed in detail in [75]. The standard procedure would be the following (see, e.g. [79]).

1) Estimate the unrestricted model, form the estimated covariance matrix, $\hat{\Omega}_u$.
2) Estimate the model subject to some restrictions and form the estimated covariance matrix $\hat{\Omega}_R$.
3) If the estimation procedure is a maximum likelihood procedure then form the following log likelihood ratio:

$$Q \, \ell n \, \frac{|\hat{\Omega}_R|}{|\hat{\Omega}_u|}$$

where Q is the number of observations. This statistic is asymptotically χ^2 distributed with degrees of freedom equal to the number of independent restrictions in R($|\cdot|$ means determinant).

4) Values of the statistic greater than a pre-set critical value on Type I error means rejection of the hypothesis implicit in the restrictions.

Notes for Chapter 2

1. Vectors are lower case letters, elements of vectors are lower case letters with subscripts, individual scalars (such as aggregate total output) are upper case letters (unless otherwise stated). Sets and matrices will also be upper case letters. All vectors are column vectors unless otherwise noted. A prime on a vector or matrix denotes transpose; superscripts on vectors are used to refer to different vectors.

2. Weaker properties are possible; see [73].

3. The level sets are sometimes referred to as input requirement sets:

   $V(Z) = \{x | f(x) \geq Z\}$ (see, e.g. [33, Ch. I.1], [81]

4. There are a number of important related concepts in the literature. Let $V(z) = \{x | T(z,x) \leq 0\}$, i.e. x can produce z as represented by $T(z,x) \leq 0$. Thus V(z) is the input requirements set analogous to note 3 above. The distance function D(z,x) is

$$D(z,x) = \max \{\lambda > 0 | \frac{1}{\lambda} x \in V(z)\} .$$

   D(z,x) is the amount by which a vector of inputs must be scaled down so as to just produce z. Thus, efficient production occurs when D(z,x) = 1 and therefore the relation between D(z,x) and T(z,x) is seen to be the following identity [40]:

$$T(z, \frac{1}{D(z,x)} x) = 0.$$

The concept of a distance function is also used in [73], [33, Ch. I.1], [33, Ch. II.1]; it first appeared in [73].

While the distance function is now becoming a more standard way of representing multiple output/multiple input production, we will continue to employ T(z,x) so as to readily address separability issues (see [40]).

5. ∇ means gradient, i.e. $\nabla g(x) = (\partial g/\partial x_1, \ldots, \partial g/\partial x_n)'$. ∇_v means gradient only with respect to the variables in the subscript. ∇^2 means the Hessian matrix of second derivatives, i.e.:

$$\nabla^2 g(x) = \begin{bmatrix} \partial^2 g(x)/\partial x_1^2 & \partial^2 g(x)/\partial x_1 \partial x_2 & \cdots & \partial^2 g(x)/\partial x_1 \partial x_n \\ \vdots & & & \vdots \\ \partial^2 g(x)/\partial x_n \partial x_1 & & \cdots & \partial^2 g(x)/\partial x_n^2 \end{bmatrix}.$$

Finally the bordered Hessian is denoted $\nabla^{2B} g(x)$:

$$\nabla^{2B} g(x) = \begin{bmatrix} 0 & \nabla g(x)' \\ \nabla g(x) & \nabla^2 g(x) \end{bmatrix}$$

which is an $(n+1) \times (n+1)$ matrix.

6. A slightly weaker condition that would allow making some outputs from other outputs is possible, but of limited interest in the current context. When this is of interest, one could respecify T(z,x) as an explicit production function in terms of one of the outputs (see [33, Ch. II.1]) or define a function in terms of a non-producible input (see [33, Ch. I.3]).

7. If $u = (u_1,...,u_\ell)'$ and $v = (v_1,...,v_\ell)'$ then $u \cdot v = \sum_{i=1}^{\ell} u_i v_i$ is called the dot (or inner) product of u and v. We shall employ the shorthand throughout the report.

8. Here homotheticity is, in fact, a slightly weakened version of the above. See[33, Ch. I.3], [33, Ch. III.3].

9. Here again, for convenience, we have taken slightly stronger properties than are in Shephard (see [73, p. 255]).

10. The capital Greek letter pi(Π) represents <u>product</u>, just as $\sum$ represents sum. Thus $\prod_{i=1}^{n} i = 1 \cdot 2 \cdot 3 ... n$.

11. This function is H.D.1. By taking a monotonic transformation $h(u) = u^k$ of f we could have

$$Z = h(f(x))$$

which would be H.D.k, allowing for increasing or decreasing returns.

12. ℓn means logarithm to the base e, (Naperian logarithms). It should be recalled that

$$\log(ab) = \log a + \log b$$
$$\log(a/b) = \log a - \log b$$
$$\log a^b = b \log a$$

13. In general, symmetry is enforced for the α_{ij}, i.e. $\alpha_{ij} = \alpha_{ji}$.

14. McFadden [33 ,Ch. I.1] considers weaker conditions, i.e. some prices that are zero (free factors).

15. The measure used here is referred to as $\hat{S}$ in [64].

☆U.S. GOVERNMENT PRINTING OFFICE: 1980 622-006/1075 1-3

REQUEST FOR FEEDBACK TO
The DOT Program Of University Research

DOT/RSPA/DPB-50/79/31 - "DEVELOPMENT OF HYBRID COST FUNCTIONS FROM ENGINEERING AND STATISTICAL TECHNIQUES: THE CASE OF RAIL"
DOT-OS-70061, Northwestern University

YES	NO	
☐	☐	Did you find the report useful for your particular needs? If so, how?
☐	☐	Did you find the research to be of high quality?
☐	☐	Were the results of the research communicated effectively by this report?
☐	☐	Do you think this report will be valuable to workers in the field of transportation represented by the subject area of the research?
☐	☐	Are there one or more areas of the report which need strengthening? Which areas?
☐	☐	Would you be interested in receiving further reports in this area of research? If so, fill out form on other side.

Please furnish in the space below any comments you may have concerning the report. We are particularly interested in further elaboration of the above questions.

Cut Out Along This Line

COMMENTS

Thank you for your cooperation. No postage necessary if mailed in the U.S.A.
FOLD ON TWO LINES, STAPLE AND MAIL.

RESEARCH FEEDBACK

Your comments, please . . .

This booklet was published by the DOT Program of University Research and is intended to serve as a reference source for transportation analysts, planners, and operators. Your comments on the other side of this form will be reviewed by the persons responsible for writing and publishing this material. Feedback is extremely important in improving the quality of research results, the transfer of research information, and the communication link between the researcher and the user.

FOLD ON TWO LINES, STAPLE AND MAIL.

Fold .... Fold

DEPARTMENT OF TRANSPORTATION
RESEARCH AND SPECIAL PROGRAMS Administration
WASHINGTON D.C. 20590

Official Business

PENALTY FOR PRIVATE USE, $300

POSTAGE AND FEES PAID
DEPARTMENT OF
TRANSPORTATION
DOT 513

U.S.MAIL

OFFICE OF UNIVERSITY RESEARCH (DPB-50)
Research and Special Programs Administration
U.S. Department of Transportation
400 Seventh Street, S.W.
Washington, D.C. 20590

Fold .... Fold

REQUEST FOR INFORMATION FROM THE UNIVERSITY RESEARCH PROGRAM

☐ Check here if you would like to be placed on the mail list for the University Research Program Solicitation Booklet (DT-63C)

IF YOU WISH TO BE ADDED TO THE MAIL LIST FOR FUTURE REPORTS, PLEASE FILL OUT THIS FORM.

Name ______________________________ Title ______________
Use Block Letters or Type

Department/Office/Room ______________________________

Organization ______________________________

Street Address ______________________________

City ______________________ State ____________ Zip ________

U.S. DEPARTMENT OF TRANSPORTATION
RESEARCH AND SPECIAL PROGRAMS ADMINISTRATION
WASHINGTON, D.C. 20590

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

POSTAGE AND FEES PAID
RESEARCH AND SPECIAL PROGRAMS
ADMINISTRATION
DOT 513

U.S.MAIL