Jan Owen Jansson

TRANSPORT SYSTEM OPTIMIZATION AND PRICING

AKADEMISK AVHANDLING
som för avläggande av ekonomie doktorsexamen vid Handelshögskolan i Stockholm
framläggs till offentlig granskning
måndagen den 8 september 1980, kl 10.15 i
sal Torsten å högskolan, Sveavägen 65
STOCKHOLM 1980

TRANSPORT SYSTEM OPTIMIZATION AND PRICING

EFI

The Economic Research Institute at the Stockholm School of Economics
Address: Sveavägen 65, Box 6501, S-113 83 Stockholm, tel 08-736 01 20

Basic Orientation

The Economic Research Institute (EFI), at the Stockholm School of Economics is devoted to the scientific study of problems in management science and economics. It provides research facilities for scholars belonging to these diciplines, training advanced students in scientific research. The studies to be carried out by the Institute are chosen on the basis of their expected scientific value and their relevance to the research programs outlined by the different sections of the Institute. The work of the Institute is thus in no way influenced by any political or economic interest group.

Research Divisions:

A Working-Life, Organizational and Personnel
B Accounting and Managerial Finance
C Managerial Economics
D Marketing . . . . . .
F Public Administration
G Economic Geography
I Applied Information Systems
P Economic Psychology –
S Macro economics

Independent Research Program:

Program for Participation and Organizational Development
Division for Applied Research in Business Administration

Additional information about research in progress and published reports is described in our project catalogue. The catalogue can be ordered directly from The Economic Research Institute, Box 6501, S-113 83 Stockholm, Sweden.

Jan Owen Jansson

TRANSPORT SYSTEM OPTIMIZATION AND PRICING

EFI THE ECONOMIC RESEARCH INSTITUTE
STOCKHOLM SCHOOL OF ECONOMICS

A Dissertation for the Doctor's Degree
in Economics

Stockholm School of Economics 1980

UDK 338.5:
656
ISBN 91-7258-121-2
LiberTryck Stockholm 011502

ACKNOWLEDGEMENTS

This study of "Transport System Optimization and Pricing" is an attempt to tie together a number of loose ends gathered from my own earlier work on pricing and investment problems in different transport industries. Thus its origin goes back to the end of the 1960s when I worked for the committees of inquiry into road costs and motor vehicle taxation (Vägkostnadsutredningen and Bilskatteutredningen), and the beginning of the 1970s when I was engaged in shipping and port problems, while attached to the ICHCA Head Office in London on a Johnson Foundation Scholarship.

I started to work on the present study in 1972, on the initiative of Professor Erik Dahmén. I am very grateful to Erik Dahmén who has been a constant source of encouragement in the ups and downs of my research. I am also grateful to Professor Karl-Göran Mäler for his penetrating criticism of the theoretical analysis, and to Professor Thomas Thorburn for his kind interest and expert advice on questions of sea transport.

I would like to thank two friends and collaborators in other projects most warmly, namely Dr Inger Rydén and Dr Dan Shneerson who have contributed both directly and indirectly to the present study. Chapter 9, on seaport capacity optimization and pricing, draws heavily on the work I have done together with Inger Rydén on Swedish seaports,[1] and this

[1] Jansson, J.O. & Rydén, I., Swedish Seaports - Economics and Policy. EFI, 1979.

chapter - as well as Chapters 7, 12 and 13 - have benefited greatly from extensive and fruitful collaboration with Dan Shneerson in the fields of liner shipping and port economics.

My work on bus transport modelling has benefited from discussions with Ingemar Lekteus, Brita Schwarz, and Jan-Erik Svensson, which I gratefully acknowledge.

I would also like to thank Professor Peter Bohm, Dr Nils Bruzelius, and Docent Alf Carling for reading and commenting on selected parts of the study.

Nancy Adler has done a great deal to correct and clarify the English of large parts of my original manuscripts. I have learnt a lot from our discussions.

I am very grateful to Ingrid Rosenvinge for diligently and competently typing the numerous versions of this work. Siv Johnsson has also been very helpful in various ways in the final stage of the production of the book.

Last but not least I would like to thank Börje Thunberg, head of the Traffic Division of the Swedish National Road and Traffic Institute, where I am working at present. Without complaint Börje Thunberg has put up with what must sometimes have been a most annoying amount of absent-mindedness on my part. I hope his tolerance will be rewarded.

This work has been financed by the Marcus and Marianne Wallenberg Foundation, whose support I acknowledge with my warmest thanks.

Linköping, March 1980

Jan Owen Jansson

CONTENTS

PART I

BASIC THEORY

1 INTRODUCTION
2 COMPLETION OF THE PRODUCTION FUNCTION FOR TRANSPORT SERVICES
3 A MODERN APPROACH TO OPTIMAL TRANSPORT PRICING
4 OPTIMAL PRICING OF SCHEDULED TRANSPORT SERVICES

1 INTRODUCTION

1.1 The problem and the purpose of the study

How should the pricing principles prescribed by welfare economic theory be translated into operational pricing policy for the transport services provided by railway and bus companies, airlines and shipping lines, or for the services provided by airports and seaports?

Although transport pricing has been prominent in applied welfare economics for a very long time, no consensus of opinions seems to be in sight. On the contrary, opinions differ as much as they ever did, both in the economics profession and outside it. In one important branch of the transport sector, however, great progress has been made, namely, in the theory of road pricing. The intensive work of economists on the question of road financing and motor traffic taxation during the 1960s produced results that have made a lasting impression on the general state of applied welfare economics.[1]

Apart from this, the discussion of railway and urban public transport pricing has got stuck in the endless and unproductive arguments about the long-run versus the short-run

[1] This work was "crowned" by Walters, A.A., The Economics of Road User Charges. World Bank Staff Occ. Papers, No. 5, 1968.

perspective, and when it comes to freight transport, economists have surprisingly little to say about such vital problems as the farreaching price-discrimination practiced in all modes of scheduled transport, or the allegedly unequal competition between "common carriers" and transport-for-hire or private transport.

There is an obvious need to develop a prescriptive price of theory and to operationalize the prescriptions for the different transport services apart from the road services. The aim should be at least to keep abreast with the development of the theory of road pricing.

The purpose of the present study is to contribute towards achieving this aim by suggesting a new approach to a prescriptive price theory for transport services (Part I), and by applying this approach to bus transport, cargo liner shipping, seaports and roads (Part II). The two first parts of this study are concerned with the "first-best". However, the results of the applications in Part II proved so disturbing with respect to the classical goal conflict between allocative efficiency and equity that it seemed essential to go into the sphere of the "second-best" as well (Part III).

A terminological comment is relevant already at this stage: in conventional economic analysis first-best optimal pricing is the same as marginal cost pricing. Unfortunately, this does not apply in the present study for reasons that will be made clear presently. The spirit of the pricing principle proposed here for transport services is certainly the same - it is a necessary condition for Pareto optimality - but the traditional designation "marginal cost pricing" is too misleading to be kept.

1.2 TECHNOLOGICAL AND ECONOMIC DIVISION OF THE TRANSPORT SECTOR

It would clearly be a good thing if a general price theory for the whole transport sector could be formulated. However, in view of the many important differences within the transport sector on both the demand and the supply sides, an *operational* general theory relevant to the whole transport sector is difficult to envisage.

Our first task is to divide the transport sector in a manner suitable to our present purpose.

From a technical point of view it is natural to distinguish in the first place between the different modes of transport. The nature of the "bearer" of the transport vehicles (water, road, rail, air, cable) plays the strategic role in this context. It determines more than anything else the method of propulsion and the other technical requirements of the transport vehicles, including their weight and size. For all modes of transport (except pipe-line), three principal "factors of transport production" are required: the transport vehicle carrying the goods or passengers, the bearer of the vehicles, and the "terminal"[1] for loading or unloading the vehicles. For the sake of brevity these three factors of transport production will be called *transport factors.*

[1] The quotation marks round "terminal" indicate that it is not a good term. There is, unfortunately, no substitute term. Terminal suggests the end-point of a transport. In a wider system context this is unfortunate. In the normal case a terminal is a point of exchange in the mode of transport. For example, James Bird has called his book about seaports, "Seaports and Seaport Terminals" to draw attention to the fact that the port function is normally the transfer of cargo between sea and land transport; but there are also seaports where the import cargo remains to be processed. In that case terminal is an adequate term. Hence, ports where oil refineries, mills etc are located on the immediate waterfront are called "seaport terminals". There is also another meaning of terminal. A bus terminal, for instance, can be both a point where a number of bus lines converge, and consequently, many riders board and alight, or simply a garage where the buses are kept when they are out of service. Aircraft hangars, car parks, etc are consequently "terminals" in the latter sense.

From an economic point of view it is natural to focus on different markets. The transport factors themselves - and the services provided by the transport factors, which will be referred to as *transport services* - are the objects that are bought and sold on the markets of the transport sector. An intermediate form of transaction, which is very important, is the hire of transport factors. As indicated in Table 1:2, it is appropriate to distinguish between the long-term and short-term hire of transport factors.

Table 1:1. Technological division of the transport sector

		The three principal factors of transport production		
MODES OF TRANSPORT	Shipping	Vehicles for carriage of goods and/or passengers	Bearer of moving vehicles	Loading/ unloading terminals
	Road transport			
	Railway			
	Air transport			
	Pipe line			
	Aerial cable way			

Table 1:2. Economic division of markets in the transport sector

		MARKETS FOR			
FACTORS OF TRANSPORT PRODUCTION	Vehicles	FACTOR ACQUISATION	LONG-TERM HIRE OF FACTORS ("TIME CHARTER")	SHORT-TERM HIRE OF FACTORS ("VOYAGE CHARTER")	FACTOR SERVICES
	Bearer of vehicles				
	Loading/ unloading "terminals"				

1.3 MARKETS FOR TRANSPORT FACTORS VERSUS MARKETS FOR TRANSPORT SERVICES

The producers of transport factors are practically never sellers of transport services. Various middlemen such as national road administrations, seaport authorities, or public transport companies act as buyers in the transport factor markets (besides "big" final users), and as sellers in the markets for transport services.

Most of the factors of the transport infrastructure category are acquired by various middlemen[1] in the first round, and then "portioned out" and resold in the form of services to the final users. In the case of transport vehicles, on the other hand, the final users of transport constitute the most important category of buyers on most of the factor markets as well. However, because of the great differences in the amount of transport services required by individual users, considerable differentiation of markets as well as of vehicle size has emerged. In freight transport markets the demand can range from, at one extreme, an oil company demanding billions of ton-miles of transport capacity per year to, at the other, the small importers of diverse manufactured goods, each demanding some hundred ton-miles per year, or even private persons sending small parcels once in a while. In passenger transport the most common unit of demand is of course the single individual, but there are also smaller or larger natural groups of individuals such as families, school-classes, troops, etc, as well as groups formed for the very purpose of travelling "in bulk".

[1] There are exceptions to this rule: private roads exist, and private terminals for freight loading and unloading are common in road haulage. Another example is the shipping line that hires a berth of a seaport for its own exclusive use.

The "small"[1] users who require too little transport to acquire their own transport vehicles can be divided into two categories: (1) those who are too small to achieve a reasonable rate of utilization *over time* if they acquire the transport vehicle in question themselves; they therefore prefer to hire the vehicle for the necessary period of time only; (2) those who are too small to achieve anything like full capacity utilization when they need transport and, moreover, their needs for transport are typically non-continuous. The latter user category generates the demand for scheduled transport services.

From a price theoretical point of view, a sharp dividing-line can be drawn between markets for factor hire and factor acquisition and markets for factor services. In the former, competitive conditions mainly prevail, particularly in the markets for transport factor hire. Tramp shipping, for example, is a rare object lesson in perfect competition. As regards markets for the acquisition of transport vehicles and transport factors of the transport-infrastructure category, the price formation is similar in character to the price formation in markets for other manufactured goods and for the produce of the construction industry respectively.

The present study is concerned exclusively with the markets for transport services. The reason for this is

[1] It is important to note that user "smallness" is a relative concept. What matters is the transport demand of an individual user in relation to the capacity of a factor of transport production of a viable size. Any kind of factor of transport production - a truck, a ship, an aircraft, even a road or a port - could in principle be made so small that its capacity would match the demand of quite small transport users. The point is that the cost per unit of transport output would be so enormously much greater that an individual solution is generally out of the question. An important exception springs immediately to mind: the private car has proved to be very competitive, to say the least, compared to various collective solutions even for relatively small users. But this is a very special case. The competitiveness of the private car is mainly explained by the fact that the driver is also the user; it is the prime example of do-it-yourself transport.

simply that there is no great need for a prescriptive price theory for the mainly competitive markets for transport factors.

The main reason for the very "special" nature of the markets for transport services is the generally extremely limited extent of each individual market. Given that a large number of people or freight shippers jointly make use of a particular transport factor, it cannot be left to the discretion of any individual user to decide where and when the transport is to take place.[1] In the case of common carriers, the destination and the time of departure has to be determined and announced in advance by the transport producer,

This is of fundamental importance to the delimitation of the different markets for transport services. Each particular service is defined by the geographical points between which transport is made possible, and even between adjacent pairs of points different services are not generally substitutable. This means that there is a literally infinite number of *potential* markets for transport services, but also that only a small minority of all potential markets can be served at all.

1.4 MARKET FORMS FOR TRANSPORT SERVICES

To underline this last point it is worth introducing a new member of the family of market forms. In the transport sector the most common market form is "*apoly*" (= a market where *no* seller can be sustained).

In the markets where the provision of transport services is viable, by far the most dominating market form is natural monopoly, so far as the transport infrastructure is concerned.

[1] In the case of the transport infrastructure, of course, the question of "destination" can only be asked before a particular factor for joint use is created. Once a road has been built, its location is given for ever.

By "natural monopoly" I mean here what I imagine most people have in mind when they use the term in this connection: it is "unnatural" - i.e. flagrantly uneconomic - to provide more than one identical piece of infrastructure between point A and point B; one four-lane road is much more economical than two two-lane roads, and a double-track railway is much more economical than two single-track railways on separate embankments, etc. On the other hand, on dense routes a road and a railway frequently run along the same stretch, and on some routes there is also a waterway, thus allowing for three different modes of transport. Opinions are divided regarding the rationality of such parallelism. A possible justification is that each mode of transport provides the best alternative for at least some transport users, and that this may compensate for the economies of scale that are missed as a result of the "product differentiation".

As regards scheduled transport services, market forms vary a great deal between the different transport industries. In the international airline industry, for example, oligopolistic markets are quite common due to a natural concentration of demand to the routes between administrative and commercial capitals. Many of these routes can sustain a number of airlines without apparent diseconomies.

Where the points of origin and/or destination of individual trips are more evenly dispersed over an area, which is characteristic of, for example, intraurban travel demand, then natural monopoly is typical of each individual service - bus line, or commuter train service.

Certain qualifications should immediately be made with respect to the natural monopoly positions enjoyed by many producers of scheduled transport services. First, for a greater or lesser proportion of the users a realistic alternative can be to hire or acquire a transport factor for their own exclusive use. Secondly, the "physical" ease of entry into markets for scheduled transport services makes the natural monopoly position of the incumbent much less secure than in the case of the transport infrastructure. Thirdly, there is a certain amount of competition between adjacent markets, in particular regarding short-distance transport services, which can be exemplified in the following way.

Consider the radial bus services in Circletown, which has its Central Business District right in the middle of the town. Suppose that there are 500 buses in regular service employed on 25 different routes between the outskirts and CBD. Each line has its own fairly well-defined market area, and no matter whether there is only one bus-line company or 25 bus-line companies in the city, the total market is most likely to be served by 25 natural monopolies. If two parallel lines are drawn by two different companies, complaints about "wasteful" or "cut-throat" competition would probably be made by the original incumbent; if the lines belong to the same company, they would probably ultimately be combined (see Figure 1:1).

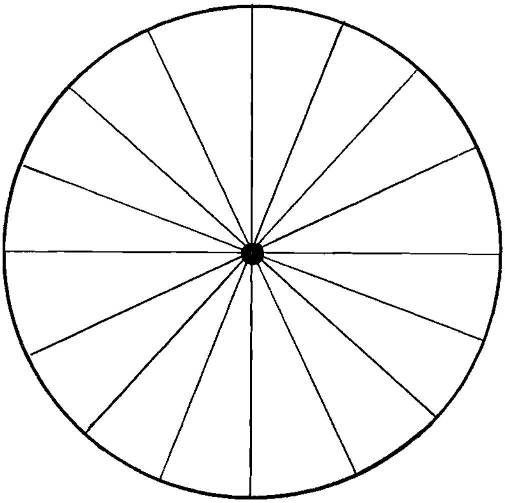


Figure 1:1. The radial bus services in Circletown

Even when the lines are equidistant, the natural monopoly position of each line is not absolute but "relative". It is clear that the market area of each bus line is inversely proportional to the total number of lines. If frequency of service is given priority over density of service, the 500 buses in Circletown could be concentrated to, let us say, no more than 10 different lines. In that case a much larger proportion of the total passengers on each line would be faithful to the line, regardless of relative bus fares. In the other extreme case, in which 500 buses are employed on perhaps 100 different routes, a large proportion of all passengers would waver between each adjacent pair of lines. Briefly, the cross-elasticity of the demands for adjacent services is strongly dependent upon the density of services, which in turn depends on the density of demand.

To protect themselves against "wasteful" competition and pirates "skimming the cream" off the market, urban bus companies (in Europe and the United States) have in many cases appropriated the sole right to run bus lines.[1]

A monopoly position can be used by the firm enjoying this position on the one hand to exploit customers by charging high prices and on the other to neglect the mandate of efficiency. In the case of public enterprises the former possibility is no problem. The latter possibility is a problem, and one which economists have considered. An interesting idea for mitigating the economic ill-effects of too quiet a life is to introduce "franchise bidding"[2] with a view to stimulating competition. It is basically the same idea as

[1] Coordination of services is also an important rationale for public transport monopolies.

[2] This procedure has been discussed with reference to various services in the literature of public utilities. See e.g. Demsetz, H., Why Regulate Utilities? Journal of Law and Economics, April 1968. Peacock, A.T. & Rowley, C.K., Welfare Economics and the Public Regulation of Natural Monopoly. Journal of Public Economics, 1972. Williamson, O.E., Franchise Bidding for Natural Monopolies - in General and With Respect to CATV. The Bell Journal of Economics, 1974.

inviting tenders from different contractors when a road, a bridge, or any large structure is to be built. In the case of scheduled transport services an additional advantage can be that a particular natural monopoly position need not be permanent; it can be reviewed at regular intervals, preventing the current incumbent from acquiring permanent tenure. In this field the idea could be realized either by a public authority selling the right to run a service along a specified route to the highest bidder, or by a public authority or private association inviting tenders from transport firms for the provision of certain specified services.[1] In the latter case the transport services are resold by the middleman to the final users.

However, the important point which must not be forgotten in discussing these qualifications of the natural monopoly positions enjoyed in many markets for scheduled transport services, is that any actual or potential competition does not necessarily eliminate "the decreasing-cost problem", i.e. the conflict that arises when optimal pricing yields insufficient revenue to cover the total costs of the services. This point is often lost in the literature of transport economics, to which we can now turn.

1.5 TWO MAIN ISSUES OF TRANSPORT POLICY AND ECONOMICS

In view of the rather dispersed pattern of market forms for different transport services markets, it is obvious that no one right answer can be expected to the problem of how to conduct the pricing and production of transport services. A large body of literature addressing these problems exists, but I will not attempt to summarize it in any detail here.

[1] See Bohm, P., et al., Transportpolitiken och samhällsekonomin. Liber Förlag, Stockholm, 1974, for an interesting discussion of this possibility.

However, the body of normative transport economics can perhaps be divided - admittedly very schematically - into two schools of thought, corresponding to the following two major issues of transport policy:

1. Freedom of market forces versus regulation of markets for transport services

2. Road and railway investment and finance

It is also worth noting that the two schools of thought in normative transport economics are concentrated on either side of the Atlantic.

The former issue has given the impetus to the development of a special branch of transport economics - the "economics of transport regulation".

As we have seen, the transport sector exhibits a very wide range of market forms. At one extreme perfect competitive conditions prevail, while at the other there are a great many local natural monopolies (let alone the innumerable "apolies"). There are also several intermediate market forms where "competition among a few" prevails.

Regulation in one form or another can be found in the whole range. Public regulation is exercised at both extremes, although with diametrically different rationales. One rationale is to protect transport *users* from monopoly abuse, and the other rationale is to protect transport *producers* from the evils of the alleged "cut-throat" competition caused by the ease of entry into some transport industries. In particular in the trucking industry, public interference to restrict competition has been a main issue since the 1930s.[1] To make matters even more confusing, a third rationale for transport regulation has now emerged: that nationalized

[1] For a recent Swedish account of this problem of the regulation of a basically competitive transport industry, see Kritz, L., Transportpolitiken och lastbilarna. IUI 1976.

transport undertakings - notably the state railways and airlines - need protection from increasingly competitive "charter" or private transport alternatives.

Finally, mention should be made of the widely practiced *self*-regulation[1] in the oligopolistic markets of the shipping-line and airline industries.

Economists have taken up an almost unanimous stand against all forms of regulation of transport markets, and particularly so in the United States.

On this side of the Atlantic the main concern of transport "political economy" is no longer the issue of transport regulation. To quote the seminal work "Modern Transport Economics" by the British transport economist J.M. Thomson:[2]

> ... transport is now treated as a branch of welfare economics in which market economics play a secondary role, rather than a branch of market economics with welfare implications. Pre-war economists were absorbed with questions of efficient competition within the road haulage; the underlying aim was to remove the deficiencies of the market in order to obtain the maximum advantage from private or public enterprise working under the freest possible conditions of competition. (Thomson, 1964, p. 12)

An important reason for this change of focus is, of course, that compared to pre-war conditions the road transport sector is much more dominating today.

The British literature on road pricing and related issues by Thomson himself, Walters, Beesley, and others has opened up a new ground.[3] Their work has been followed up in Sweden, for instance, in a number of studies on transport policy and economics, which have appeared in the last ten years. These include reports from two public inquiries into

[1] This term is often used in the industries concerned in preference to "price cartel" formation.

[2] Thomson, J.M., Modern Transport Economics. Penguin, 1964.

[3] A summary of the literature before 1970 is contained in Jansson, J.O., Prissättning av gatuutrymme. EFI, 1971.

the costs and prices of the services of roads and railways,[1,2,3] and a study by a team of economists of general transport policy and economics.[4] The main message of all of these is that pricing should be based on the social *short-run* marginal costs, and that social surplus maximization should be the investment criterion.

1.6 TWO IMPASSES

In the present context we will not restate either the case for deregulation of transport markets, or the case for short-run marginal cost pricing and social surplus maximization. Instead it will be argued that with regard to the present problem, both schools of thought seem to have landed in an impasse - although, it should perhaps be added, by no means the same one. In the remaining part of this introduction I will try to describe these impasses, and my thesis is that they provide the explanation of the relative underdevelopment of price theory for transport services.

1.6.1 Impasse 1: Denial of market failure due to confusion of two types of economies of scale and neglect of user costs

The basis of the argument against regulation is the alleged absence of economies of scale in transport.

In the literature on this subject the existence of economies of scale in the transport sector is denied both on a basis of various theoretical arguments and by reference to empirical studies. It thus becomes possible to argue that

[1] Vägtrafiken - kostnader och avgifter. SOU 1973:32 (VKU).

[2] Trafikpolitik - behov och möjligheter. SOU 1975:66 (TPUI).

[3] Trafikpolitik - kostnadsansvar och avgifter. SOU 1978:31 (TPUII).

[4] Bohm, P., et al., Transportpolitiken och samhällsekonomin. Liber Förlag, Stockholm, 1974.

there is no problem of market failure. The only problem is that the transport industries are too tightly regulated. The solution is, of course, to set competition free. A typical and authoritative treatise along these lines is "The economics of competition in the transportation industries" (1959) by John Meyer et al.[1] A later British work with much the same message is Foster's revised edition (1975) of "The transport problem".[2] Foster originally thought that there might be exceptions to the rule regarding non-existent economies of scale.

> Nevertheless as a generalization it can be maintained that a competitive solution will not be inefficient because of failure to realize economy of scale.
> (Foster, op. cit., p. 296)

In a postscript to the 1975 edition, Foster's main doubt about this generalization seems to have been removed in the light of:

> ... the arguments advanced by Dr Joy that railways are not decreasing costs industries if they pursue optimal policies on track ..... As a result it might be possible to price at marginal cost on railways, providing it is pursuing an optimal investment and disinvestment policy, without a deficit.
> (Foster, op. cit., p. 311)[3]

Two contributory causes of the one-sidedness of the argument against regulation has been (i) a confusion of economies of firm size with economies of density demand, and (ii) a general neglect of transport user costs.

Which concept of the economies of scale is the *relevant* one can vary very much, depending on the decisions or aspects of policy that are being considered. In the case of competitive markets, such as that commanded by the American trucking

[1] Meyer, J., et al.: The economies of competition in the transportation industries. Harvard University Press, 1959.

[2] Foster, D.: The transport problem. Revised edition. Croon Helm, London, 1975.

[3] The arguments referred to are to be found in Joy, S.: British Railways' track costs. Journal of Industrial Economics. Vol. 13, 1964. Reprinted in "Transport", ed. Munby, D., Penguin Modern Economics, 1968.

industry, for example, the most common question concerns *firm size*. Do small trucking firms perform as efficiently as large ones, or is it right to encourage the merger of smaller firms? In the case of markets for scheduled transport services, the same type of question *can* be relevant. But when it comes to the (dis)investment and/or pricing of a particular service, the relevant "scale" concept is *density of demand.*

There are several studies which suggest that economies of firm size are rare in railway transport, bus transport, or road haulage (see Chapter 4 for references). It may seem rather far-fetched to take this as a substantiation of the claims that all markets for transport services would work well if left alone, and that there is no serious conflict between optimal pricing and self-financing. Nevertheless, this is exactly what is often done.

There has been much less empirical study - in fact practically none - of the cost and output relationships that are relevant to these questions, in other words of how the costs of individual lines depend on the traffic volume. A growing awareness can be noticed in the literature, however, of the rather obvious fact that it is generally rather more costly to provide scheduled transport services on thin routes than on dense routes, although it should be added that the pricing policy that is then recommended runs counter to what will be suggested below as a result of the present analysis.[1]

The confusion concerning economies of scale would not have been so widespread, were it not for the deplorable neglect of the user costs. With the notable exceptions of Vickrey, Mohring and Turvey, transport economists have almost completely overlooked the highly relevant fact that (in the long run) the quality of transport services is positively

[1] See Harris, R.G., Economies of Traffic Density. The Bell Journal of Economics. Autumn, 1977.

correlated to the density of demand in practically all markets.

The following statement by Turvey and Mohring,[1] which actually refers to the problem of optimal bus fares, is equally relevant to the whole transport sector.

> The right approach is to escape the implicit notion that the only costs which are relevant to optimization are those of the bus operator. The time-costs of the passengers must be included too, and fares must be equated with marginal social costs.
> (Turvey & Mohring, 1975, p. 1)

This can serve as the motto for the present study. In the next chapter we will discuss how user inputs can best be taken into account in transport price theory.

In conclusion it should be pointed out that the wholehearted - and probably often quite justified - opposition to public regulations on the part of many - mostly American - economists, has diverted attention from the cases of real market failure and tended to belittle the decreasing-cost problem. And consequently, the need for a prescriptive price theory for transport services has not been felt very strongly.

1.6.2 Impasse 2: Denial of the relevance of the long run to prescriptive price theory

On the other hand, apart from Stewart Joy's study of railway track costs, in none of the aforementioned works on road and railway pricing and investment, has the question of increasing returns been investigated, except that Bohm, et al. comments on some results of the previously mentioned American studies of economies of scale in railway transport. Nevertheless, it seems to be taken for granted that SRMC-pricing of various transport services (with the exception of urban road services) would result in significant financial deficits;

[1] Turvey, R. & Mohring, H., Optimal Bus Fares. Journal of Transport Economics and Policy. September 1975.

otherwise all the attention paid to the second-best problem of reconciling SRMC-pricing with a budget constraint (in Bohm et al., as well as in VKU) would not be justified. But it is rather unsatisfactory that the existence of the problem under consideration is simply asserted rather than demonstrated.[1]

Generally speaking, this is rather typical of the SRMC-pricing school of thought. Making a case for SRMC-pricing was a difficult task, since it involved a break with the conventional wisdom of many economists and decision-makers or politicians, who tend traditionally to favour some variation or other of the principle of *long-run* marginal cost pricing. However, all this preoccupation with the question of which "run" is relevant to pricing has not been very productive.

A more productive line of development would be to formulate the theory in terms which would make it possible to examine what the financial result of optimal pricing would be.

In the discussion of road pricing the financial result for different links of the road network of charging "congestion tolls" has seldom been considered. To my knowledge the only economist who has provided an answer to this question is Herbert Mohring. As early as 1962 Mohring showed that, assuming that both the road user cost and road producer cost functions are linearly homogeneous, charging optimal congestion tolls and pursuing an optimal investment policy will result in exact coverage of the road capacity costs.[2]

[1] See Bohm, P., et al., pp. 86-88.

[2] Mohring, H. & Harwitz, M., Highway Benefits: An Analytical Framework. The Transportation Center at the Nortwestern University. Northwestern University Press, 1962, pp. 81-87.

Thus his conclusion, which has remained unchallenged ever since, was that there is no difference; the well-known theorem that the degree of economies of scale determines the financial result when MC-pricing is applied, holds true also in the case of congestion toll charging.

In the present study it will be shown that there is, in fact, a very considerable difference with respect to the financial result between optimal transport pricing and MC-pricing of goods where no user costs are involved. It will be demonstrated that when the scale-elasticity of output (E) differs from unity, the size of the financial deficit or surplus is *not* inversely proportional to the value of E. The financial result is much more sensitive to deviations (upwards or downwards) from the position of constant returns to scale.

The reason for the neglect of the question of the financial result of optimal pricing is probably that, in the heat of the short-run versus long-run debate, the supporters of SRMC-pricing made it a special point of their argument that "pricing has nothing to do with investment." Pricing decisions should be taken with a view to utilizing existing capacity optimally, and that is that.

This is going too far. Prescriptive price theory should not be confined - any more than descriptive price theory - to considering the short run only. In descriptive price theory a good balance exists between the analysis of price formation in the short run and price formation in the long run. The basic theory is that prices are set to maximize profits in the short run. However, it is also explained how the level of prices is settled in the long run - not by an alternative hypothesis about profit-maximizing behaviour, but by allowing for all the "market forces", on both the supply side and demand side, that only make their full impact in the long run. In practice the latter part of descriptive price theory is indispensible for predicting future price trends in a particular industry.

A similar balance should be struck also in prescriptive price theory. As in descriptive price theory, it is generally true that optimal pricing should not take sunk costs into account, or seek to incorporate the costs of planned investment in fixed plant. In prescriptive price theory, however, it is just as legitimate to ask what the level of prices is likely to be in the long run, on the assumption that an optimal pricing policy and an optimal investment policy are being pursued.[1]

1.7 TRANSPORT SYSTEM OPTIMIZATION AND PRICING

The main stimulus to the present study was my growing dissatisfaction with a transport price theory that goes no further than making a case for short-run marginal cost pricing. A theory should also be able to say something about what financial result can be expected from optimal pricing. To achieve this broader aim, the tempting short-cut of taking the whole supply side as given, which is typical of the extreme short-run approach to transport price theory, has to be resisted. Work must start instead with an investigation of the characteristics of the supply of transport services or, more exactly, of the characteristics of *efficient* transport systems. This provides the conditions for a richer theory of optimal pricing. And this is in fact the general attitude of the present study. It will be found that the following analysis does not completely live up to the high principles advocated. But an attempt is made to tackle both the problems of operational prescriptions for the pricing of bus transport, liner shipping, seaport

[1] A very special normative price theory would result if the latter assumption were not made. If a behavioural assumption of some sort is made to the effect that systematically inoptimal investments are undertaken, then strict SRMC-pricing cannot be recommended. The rather odd situation thus arises in which the pricing should also be designed with a view to coaxing investment decision-makers into behaving as good net social benefit maximizers.

and rural road services, and the problems of "transport system optimization" with a view to saying something more conclusive about the financial result of optimal pricing.

To achieve these two purposes it has proved necessary to adopt a new approach. The development of this approach will be the subject of the three following chapters.

2 COMPLETION OF THE PRODUCTION FUNCTION FOR TRANSPORT SERVICES

In the introductory chapter the need for a modern prescriptive price theory for transport services was indicated, and some possible reasons for the underdevelopment of existing theory suggested. The problem now is to propose a suitable design for such a theory.

The key ingredient in the approach suggested below is no new invention; it coincides with a well-known form of road investment appraisal, and it has also been incorporated in public transport systems analysis. This key ingredient is that the user costs are treated on a par with the producer costs.

From the theory of travel demand we can borrow the designation "generalized cost approach", which in that context implies that the user costs are monetarized and put on a par with all possible charges. In the present analysis of the supply side we will go a little further in the direction indicated, emphasizing also the concept of a generalized production function and including the user input as a principal factor of production.

2.1 THE TREATMENT OF USER INPUTS

I base my argument on a characteristic that applies generally to the production of transport services, namely that "user costs", i.e. the time and effort put in by the consumers of the services, constitute a large proportion of the total social cost. Under these circumstances much can be said for treating the user input as a factor of production alongside the "ordinary" factors provided by the producer of transport. This idea is generally applicable to services, as was first observed by Victor Fuchs.[1] The non-storability of services requires that production and consumption occur simultaneously. The service consumer necessarily becomes one of the factors of production; without the consumers' input no output of services could be achieved. A more widely known discussion along similar lines was started by Gary Becker.[2] His main point of departure was that *all* consumption - of storable as well as non-storable goods - takes time: the consumer has to take both a time constraint and a budget constraint into account in order to obtain maximum want satisfaction. Becker paid no special attention to the simultaneous nature of the production and consumption of non-storable goods. The fact that consumption is time-consuming is of profound importance to the choice of production *technique* in the case of non-storable goods only. Storability - which means that consumption can take place with a shorter or longer time lag after the completion of a particular product - makes it irrelevant to the consumers' problem of time allocation whether one hour or one week is needed to produce one unit of the product concerned.

In the present chapter I will try to justify an extension of the conventional concept of a production function to include the inputs provided by the consumers, in the case of services produced in the transport sector. An examination of

[1] Fuchs, V., The growing importance of the service industries. Occassional Paper 96. National Bureau of Economic Research. New York, 1965; and by the same author, The service economy. National Bureau of Economic Research. New York, 1968.

[2] Becker, G., The theory of the allocation time. The Economic Journal, 1965, Vol. 75. See also Burenstam Linder, S., Den rastlösa välfärdsmänniskan. Stockholm, 1969.

one type of transport service after another has confirmed my impression that if the transport users' input is only taken into account on the demand side, the production functions used are "incomplete". Hence, the title of this chapter. The first step in the argument is to illustrate the relative importance of transport user inputs.

2.2 THE IMPORTANCE OF USER COSTS

In the case of passenger transport, user time is by far the most important input in the production of transport services - the user cost share in the total costs of passenger transport is usually many times larger than the procucer cost share. An illuminating illustration of the general importance of travel time is provided by the charts in Figure 2:1, showing the time allocation of households in Stockholm. Between 6 am and 6 pm about 40 per cent of the time is spent at work or at school, and 10 per cent is spent on "movement" - travel between various activities. That is to say, the total time spent on transport is no less than one fifth of the total time occupied by work and preparation for work (studies). Bearing in mind that the total value of production in the transport industry (both passenger and freight) is no more than about 10 per cent of the total gross national product, it is clear that the user time input is the most important factor in the production of transport services. Everyday experience provides further evidence of this. In bus transport it is well known that the cost of labour (including clerical staff etc) is the dominating cost item in bus companies. However, on a crowded bus the time of the single driver does not constitute a very striking resource cost, compared to the time of all the passengers.

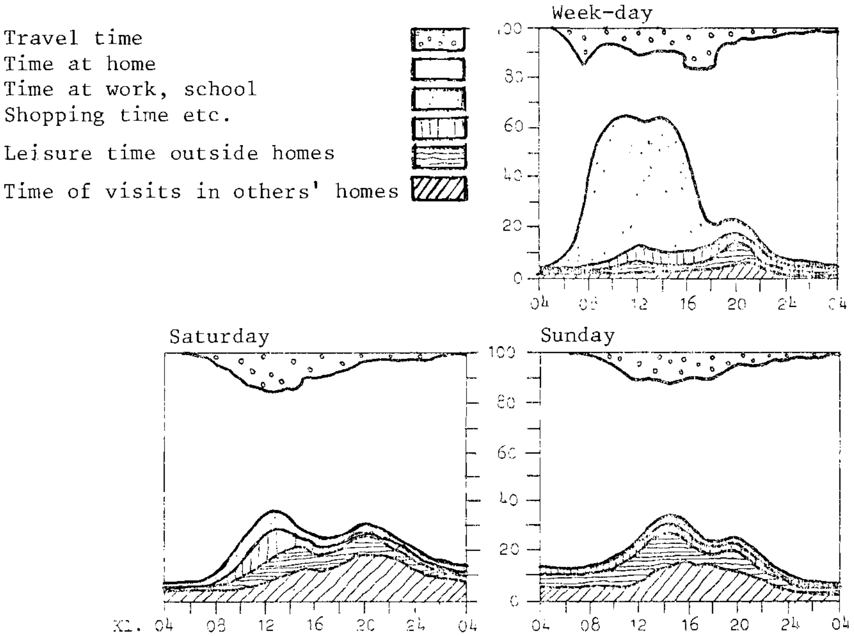


Figure 2:1. Time allocation of Stockholmers. Source: BFR:s arbetsgrupp för trafikforskning: Trafik och bebyggelse. Statens Råd för Byggnads forskning. January 1977.

When it comes to transport infrastructure services, the user cost share is even greater, since the time of the vehicles is then a user input. For example, on the main interurban roads the total cost to the road users (including the cost of travel time and accidents, as well as the cost of the vehicles) is something like ten times greater than the capital cost of the roads (see Chapter 10).

In the case of freight transport, the user cost share is naturally much less. On an aggregate level a relevant comparison would be to set the total time costs of all goods in transit, including transit storage, against the total producer costs of freight transport services. Again, the user cost share is much larger in the case of services supplied by the transport infrastructure. From the point of view of a port authority, for example, the costs of the ships being served appear as user costs alongside such items as interest costs on goods in transit, etc.

The large size of the user cost share is not sufficient reason in itself for putting the main emphasis on this factor in an economic analysis of the production of transport services. For example, in the production of wedding-rings, gold is probably the most costly input. However, to produce a good ring it may be that k grams of gold are required, no less and no more, regardless of the production technique used, and the most interesting questions of production economy concern the combination of skill and various tools and capital equipment. Thus the question is, does the input of user time into the production of transport services appear in a largely fixed proportion, as input materials often do in goods manufacturing? If this were so, it could be argued that the user inputs should be left out of account, or simply treated as a given implicit factor in transport production functions, despite the magnitude of user costs.

2.3 SUBSTITUTION BETWEEN PRODUCER AND USER INPUTS

In fact, there are considerable opportunities for substitution between producer inputs and user inputs in the production of transport services. In the short run - given the capacity of a transport facility - the main trade-off of user costs against producer costs is realized by the rate of capacity utilization as the balancing factor. The higher the rate of capacity utilization, the lower will be the capacity costs, but the higher will be the queuing costs and/or the congestion or overcrowding costs.[1]

[1] Congestion costs are connected with a prolongation of the travel time, while overcrowding increases the cost of a minute of travel time.

When the design of a transport facility can be regarded as variable, important substitution possibilities between producer and user inputs arise. Due to the simultaneous nature of the production and consumption of transport services, the production time - the service time or speed - is of vital importance to both producer costs and user costs. Other facility design variables which are significant determinants of both costs are "comfort" and the risk of accidents or damage to freight.

In Chapter 3 a general model of a transport facility will be presented; the model will focus on the trade-offs between producer and user costs resulting from variations in the rate of capacity utilization and the values of the facility design variables. This model forms a basis for the subsequent discussion of rules for optimal pricing.

The substitution possibilities between producer and user inputs are not exhausted by those just mentioned. However, to realize the substitution possibilities to the full, a *systems* approach to transport should be adopted.

The systems approach is required mainly because the cost of the "distribution" (as distinct from the "production") of transport services is usually quite an important item, and because the elasticity of substitution between the production and distribution of transport services is high. In the case of manufacturing industry, it is appropriate to divide activities into (1) the transport of input materials, (2) pure "production", which takes place at the production plants, and (3) "distribution". The latter can in turn be divided into selling activities and the actual transport of the final output. It is not always explicitly stated which of these activities a particular production function is meant to embrace. The most common assumption is probably that only pure production is included. This seems most appropriate when there is no substitution between pure production and transport, as when the location of the production plant is given. It is then practical to derive a separate "production function" and separate "transport functions" (for inputs and outputs). When the location of plants is not predetermined, there may be considerable opportunities for substitution between production and transport. The more significant the economies of scale in the pure production, and the higher the transport cost per product unit-mile in relation to the production cost per unit, the higher will be the elasticity of substitution between production and transport. In industries characterized by pronounced economies of scale in production, and a low value per m^3 of the output, an important choice has to be made between concentrating production to relatively few plants in order to exploit the economies of scale, or dispersing production to numerous plants in order to keep transport costs down.

For all transport services this choice is at the heart of the question of "system optimization". It may sound a little strange to speak of the "production" and "distribution" (let alone the "transport") of transport services, but it is nevertheless a logical way of applying the production and distribution dichotomy to various transport systems. In practically all transport systems individual door-to-door transports can be divided into one or more *feeder* transport stages and a *trunk-line* stage. From a theoretical point of view it is appropriate to compare the trunk-line transport to the production stage in manufacturing industry, and to regard the feeder transport as the "bridging of the space separating users and the production plant" - in other words, the distribution of transport services.

The "distribution cost" is a significant cost item in most transport systems. Hence the recent emphasis in freight transport on the whole door-to-door transport chain rather than on the trunk-line haul alone. In public transport, passengers discover every day that the cost of walking to and from stops and stations, and of waiting at stops and stations, is comparable to the cost of the transport service demanded.

When the financial result of optimal pricing is being considered, a complete system model is most relevant. Investment decisions have to take all possible "system effects" into account, and the consequences of doing this as regards the location, the number, the capacity and the general design of individual facilities in the system all have implications for the level of optimal prices.

In Chapter 4 this will be studied further with the help of a model of a system of scheduled transport services. It should perhaps be mentioned that this system model has not the degree of sophistication required for the investment problem. To handle all the degrees of freedom, the procedure best adapted to our purpose is to consider the effects on system design of changes in the *density* of demand in the system, on the assumption of homogeneity between individual system components. *On average*, the result should not differ too much from the heterogeneous transport systems to be found in real life.

2.4 INTRODUCING THE "PRICING-RELEVANT COST" ALONGSIDE THE MARGINAL COST

The theory and measurement of travel demand (including time valuation) based on the generalized cost approach has emerged as a useful speciality of transport economics. A corresponding specialization on the supply side could be very fruitful. The derivation of complete production functions for different modes of transport under different external conditions could become as important an aid to transport planning as studies of transport demand, provided that due care is exercised in applying average time values in view of the inevitable problem of the great variations in user time values.

Formally, the analysis of supply and demand according to the present approach will be similar to conventional supply and demand analyses, except in cne important respect which should be mentioned from the start to avoid misunderstanding. The main difference to bear in mind is that consumers play two roles in transport: the usual one of demanding the services, and another as a factor of production. Analytically, it is convenient to separate the two roles in the following way.

When the supply side is being considered, user inputs are treated in the same way as producer inputs. Given the factor prices, including the value of user waiting time, travel time, etc., an efficient factor combination will be one that results in the least total social costs in producing a particular output.[1] The "expansion path" is derived by tracing all efficient factor combinations, just as in ordinary production and cost theory. Note that when the efficient factor combinations are being determined, no account is taken of the role played by transport users on the demand side. That is one of the "tricks" of the approach. This neglect is made good when the demand side is brought into the picture. As in the ordinary case, knowledge of the expansion path is not a sufficient basis for choosing a particular factor combination (unless the demand is completely inelastic). The demand function also has to be traced. The double role of the transport users should now be allowed for, by taking into account that demand for the relevant service will be affected not only by prices but also by the other user costs. By monetarizing the latter costs and combining them in a "generalized cost", the well-known two-dimensional picture of supply (cost) and demand becomes applicable also to transport services. See Figure 2:2.

[1] One difficulty in the generalized cost approach - "weakness" is not the right word, because it is a general problem - is that the user input prices, i.e. the time values, may not remain constant at different levels of output. The same problem arises with regard to every factor of production where supply is not perfectly elastic. Where transport services are concerned, the same complication appears on the demand side too - which is probably why special attention has been paid to this problem in transport economics.

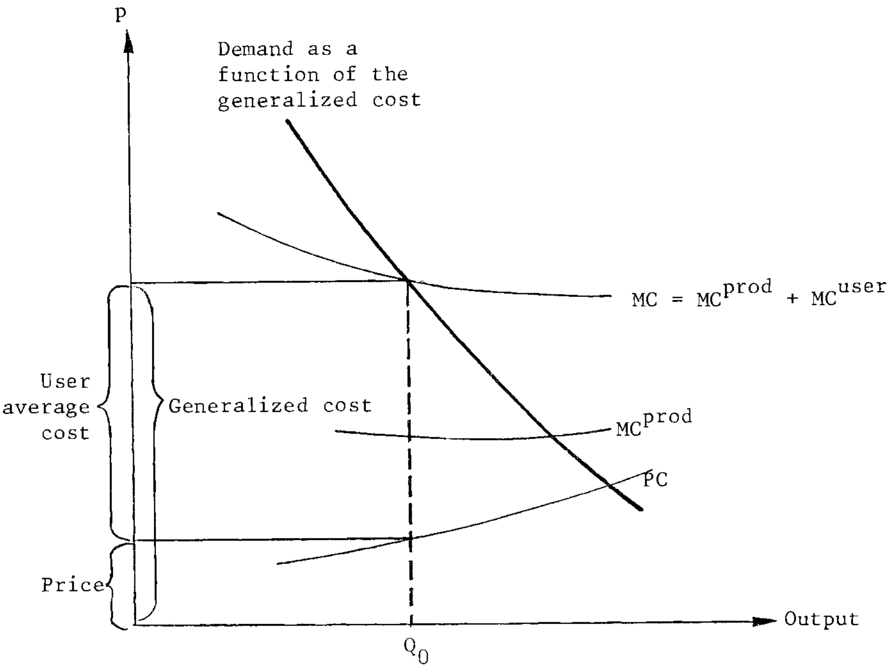


Figure 2:2. Marginal costs versus the pricing-relevant cost

The point of intersection of the demand curve and social marginal cost curve (MC in Figure 2:2) gives the optimal level of output, Q_O, just as in the ordinary case.

The optimal price is *not*, however, at the level of the point of intersection of the demand and MC. The generalized cost is at this level. The optimal price is only a smaller or larger fraction of the generalized cost - the remainder is constituted by the user average cost. If we define the social marginal cost - subsequently referred to simply as marginal cost and designated MC - as the sum of the producer marginal cost and the user marginal cost, it follows that optimal pricing does not mean that the price is equal to the social marginal cost,[1] and, unfortunately, the well-established term "social marginal cost pricing" cannot be used as a synonym for optimal pricing. This is an unavoidable disadvantage of the present approach. Instead we have to introduce a new cost concept, which we can call the "pricing-relevant cost" and denote PC. This cost assumes the role of the marginal cost of conventional analysis in that it is always equal to the optimal price. In Chapter 3 the pricing-relevant cost will be presented in detail. How is it defined? What shape does it assumes under different circumstances? How is it related to the conventional concepts of the long-run and short-run marginal costs of the producer of transport services? And how should the pricing-relevant cost be estimated for different transport services?

Before we embark on these questions we should first consider any possible alternative way of dealing with user costs.

2.5 ALTERNATIVE TREATMENT OF USER COSTS

2.5.1 External effects

One possible alternative approach to a price theory for transport services - given that user costs cannot simply be ignored - is to treat the user costs as "external effects". This was the original approach chosen by Pigou in his classic analysis of two alternative roads connecting two towns, one without and one with some traffic congestion.[2] And in later discussions of road pricing this tradition has persisted.

[1] This will be described in greater detail in the following chapter. Here the purpose is primarily to acquaint the reader with some basic terms and designations.

[2] Pigou, A.C., Economics of Welfare. London, 1920.

However, it is not really an alternative approach to the generalized cost approach. It is basically the same thing, but under a different name. In view of the fact that user costs often dominate over producer costs, it does not seem very appropriate to refer to the former as "negative external effects". It also seems inappropriate to define the system within such narrow limits that, for example, every individual user of a road plus the road space he requires defines a separate "system", so that any interference with fellow motorists becomes an external effect. Moreover it is rather difficult to adequately describe the effects of different transport-system designs on other user costs (apart from congestion and suchlike effects) with the help of this terminology. For instance, in the case of general cargo shipping, it does not sound right to describe as "positive external effects" the effect on feeder transport costs of increasing the number of ports-of-call or the effect on storage costs of increasing the sailing frequency.

I believe that a contributory cause of the underdevelopment of transport price theory outside the road sector is the traditional way of describing road user costs rather vaguely as external effects. This has made it more difficult to recognize that user inputs are major factors of production in all transport systems.[1]

2.5.2 Quality of service

Another approach could perhaps be to refrain from monetarizing the user inputs altogether, simply allowing for them in the definition of the quality of service. This alternative may appear to be the most natural approach from the point of

[1] The dearth of applications of welfare economics to other important service sectors like retailing is probably also due to insufficient understanding of the central role of user costs.

view of general demand theory: consumers' preferences for different qualities will be revealed on the market, and there is no need to impute different values to different products beforehand.

However, the basic philosophy of the present study is that this approach is inadequate to the formulation of a prescriptive price theory for transport services. This claim will not - and cannot - be proved at this stage. "The proof of the pudding is in the eating", and the purpose of this book is to demonstrate that useful results regarding optimal pricing can be obtained by following the approach I will elaborate below.

On the other hand, since we shall not return in Part I to the other main issue at stake - the issue of quality of service determination - it seems appropriate to examine in a little more detail the usefulness of the generalized cost approach as an analytical tool in the choice of quality of service.

The general proposition is that while the quality of ordinary goods (including transport factors) can be left to the market to deal with, the quality of transport services cannot rationally be chosen in any other way than by following the generalized cost approach.

In neo-classical micro-economic theory (as opposed to Lancaster's theory of consumer demand[1]), the "quality" of goods and services has always played a very insignificant role. There has been no need for the concept. Products of different quality are simply defined as different products, no matter how closely related they are. Given the basic assumption of constant returns to scale, it is quite reasonable to envisage that every conceivable product variation is

[1] Originally described in Lancaster, K.J., A New Approach to Consumer Theory. Journal of Political Economy, Vol. 74, 1966.

put to the test of the market. Then it is only a question of *quantity* determination, including the extreme of zero output - some products will not pass the test of the market.

This model is inapplicable when it comes to transport services. First, the extent of each particular market is often very limited and, secondly, the relevant output range is always the range of increasing returns. The normal case - particularly so far as the transport infrastructure is concerned - is that one (natural monopoly) enterprise only, supplying one service only in each time period, can be sustained in any one market.[1] There are certainly several duopoly markets and some oligopoly markets, and it should also be remembered that an individual enterprise can sometimes offer a limited choice of quality of service. One example is a bus route with two parallel bus lines, one is an express line and the other an ordinary line with 3-4 stops per kilometre; another example is the choice of 1st and 2nd class travel offered by railway companies and airlines. As a general rule, however, the "market solution" to the choice of quality of service is not feasible. Such a choice has to be made,[2] of course, and bearing in mind that transport services are "intermediate goods" - i.e. they are devoid of intrinsic utility and valued only to the extent that certain final

[1] This is not inconsistent with the fact that, for example, a shipping line can have a thousand entries on its tariff of freight rates. The explanation is twofold: on the one hand, different customers make different demands on capacity; on the other, price discrimination is widely practiced, so that different customers are charged different prices for the same service. Where different customers cause different costs, there is seldom a choice of price and quality for a particular customer; intrinsic properties of the commodities determine the relative costs.

[2] A tendency can be observed on the part of transport planners of seeking to avoid this choice by asking their principals (the politicians) to lay down "standards", or simply by relying on some sort of consensus within the profession regarding what should be considered to be a reasonable level of service of public transport, etc under different circumstances.

destinations can be attained - the generalized cost approach seems to be a reasonable choice. This does not mean that there will be no problems: various problems connected with estimating unit values for the user inputs remain to be solved, but an encouraging amount of progress has already been made on the theory and measurement of the value of time.[1]

Lastly, it should perhaps be mentioned in this connection that the generalized cost approach is by no means widely used in practice. As a digression some comments on this fact are made in the following last section of the present chapter.

2.6 DIGRESSION: THE GENERALIZED COST APPROACH TO DETERMINING QUALITY OF SERVICE IN THEORY AND PRACTICE

Generally speaking, the chief usefulness of the concept of a production function is that it provides in a compact term a continuum of efficient factor combinations for the production of any given level of output, without any conditions regarding the proportions or prices of factors. Conventional approaches to choosing a design for transport facilities, on the other hand, are constrained by various requirements regarding quality of service or norms regarding the way a particular facility should be designed. Such a deliberate imposition of constraints on the choice of factor combination has no equivalent in manufacturing industry, for example, and it seems to owe its occurrence as much to the analytical difficulties of handling many degrees of freedom with respect to qualities of service as to any real need for rigid standards. This inflexibility is particularly harmful in view of the fact that the transport sector is divisible into numerous individual markets. This means that very different volumes of output, and presumably also very different qualities

[1] See Bruzelius, N., Theory and Measurement of the Value of Time. Stockholms Universitet, 1978.

should really be produced at different production plants. Quite different factor combinations are likely to be appropriate where the density of demand is relatively high and where it is low, among other things as regards the relative importance of the user input. Under such circumstances the conventional approach tends to curb the technological imagination, suggesting too standardized a pattern of transport solutions. We need only recall, for instance, the notorious problems and inevitable arbitrariness of defining "minimum levels of quality" for the supply of transport services in sparsely populated regions. Another example of unnecessary inflexibility is provided by the "too short" short-lists prepared before making Cost Benefit Analyses of transport investment projects. A particular CBA is concerned usually for practical reasons, with evaluating a rather limited number of alternatives. The selection of the projects to be considered is normally a "pre-scientific" procedure, which, however, is often more important than the final choice between the short-listed projects. In road investment planning, for example, the potential use of the generalized system cost approach could be far greater than its actual use, if the aim of each individual road investment were set higher than at present. If the aim were to find the optimal design for each new road or road improvement, then thinking in terms of a road-service production function could be very fruitful. At present it is already decided in effect which type of road or road improvement is to be made, before a full CBA is carried out. All that remains to do is to optimize the timing of a given set of investment projects (which may mean postponing a project indefinitely).

Nevertheless, the main area of application of the generalized cost approach, so far as the supply side is concerned, is the road sector. The idea of calculating the social rate of return of road investment projects along these lines, and ranking them accordingly, seems to be applied in one way or another by most national road administrations.

In other areas of the transport sector the generalized cost approach is hardly used at all. Public transport appears to be a particularly promising area for its application. The reason why the generalized cost approach has not been incorporated into the planning routine of public transport undertakings, as it has been in the case of roads, cannot be that the relevant user costs are more difficult to monetarize than road-user time and accidents. What, then, can be the explanation?

Tradition - many public transport companies have a previous history as private enterprises - is probably the main reason why public transport companies have not followed the example of the national road administrations. Nor have external conditions been such as to encourage new thinking. Since the 1950s the relative demand for public transport has been on the decline, which has meant increasing financial difficulties for both bus transport and railway companies. Under these circumstances it is natural that all attention should be focused on the problem of meeting the imposed budget constraints. The general atmosphere has not been favourable to trying out any radical improvements in quality, such as are indicated by the generalized cost approach (see Chapter 5 for a discussion of this problem).

Conditions in the road sector have been just the opposite. The demand for road services is (i) generally very inelastic, and (ii) has been steadily growing for some decades. The first of these factors has tended to make road traffic the subject of quite heavy taxation in many countries - far above what is required for financing the roads. Nevertheless, road-builders have never had to fear that their roads would not be used. The road network has been continuously extended, which in generalized cost terms means that the cost to users of moving between given points has been steadily reduced. It is no wonder that finding methods for quantifying and setting a money value on improvements in road quality has appeared to be the most urgent task.

A more mundane and perhaps equally important reason why national road administrations have, in principle, accepted CBA in line with the generalized cost approach to investment appraisal, is simply that the alternative of Cost Revenue Analysis has been out of the question, since charging for road use does not come within the scope of the national road administration. In other words, if all roads in the national network were toll roads and a self-financing road system were introduced, it is not impossible that CRA would sooner or later replace the present principles of road investment appraisal.

3 A MODERN APPROACH TO OPTIMAL TRANSPORT PRICING

3.1 LONG-LIVED AND DEEP-GOING CONTROVERSY

Optimal pricing of transport services is one of the most controversial issues in the history of economic thought. From Dupuit until today there has been as much variety of opinions as in any of the famous controversies in economics. During the last century and at the beginning of this one, railway pricing was the main concern. The keen interest shown in the subject by leading economists such as Wicksell, Cassel, Pigou, Taussig, Hotelling and Dessus seems to have faded in more recent times. The reason for the less fervent interest of modern economists cannot be that agreement has been attained about the optimal pricing of scheduled transport services, but rather that matters have come to a deadlock. With the spectacular growth of road transport road traffic taxation has been a principal area of debate. In both areas suggestions for reforming pricing policy - made in sober earnest - still range from offering the services free of charge to self-financing full-cost pricing. Recently, second-best pricing problems have come up for discussion, but so long as the whole question of first-best pricing remains confused, it will be difficult to exploit any progress that is made. The present discussion is devoted to the tradtional problems of the first-best.

As is well known, two subjects of contention stand out with regard to all modes of transport:

- AC-pricing versus MC-pricing,

and, for those who have accepted the marginalistic view,

- LRMC-pricing versus SRMC-pricing.

The latter controversy has in my view been rather unproductive. It is rather curious that no one seems to have put this question in the century-long debate about transport pricing: from general micro-economic theory we learn that SRMC and LRMC are *equal* along the expansion path. Therefore, it should make no *systematic* difference whether the pricing is based on SRMC or LRMC. How is it then possible - as the LRMC-proponents claim, and the SRMC-proponents do not generally deny - that prices for transport services based on LRMC would be much higher than prices based on SRMC? The debate could never have remained so lively for so long if an answer to this question had been sought. The real cause of the controversy, it is argued here, is that the proponents of LRMC for transport services have systematically overestimated the optimal price, while the proponents of SRMC have often tended to underestimate the optimal price. The reason for this seems to be that *user costs* are not given due emphasis. They are either simply ignored - this is the principal sin of the LRMC-proponents - or misrepresented by a failure to couch them in terms of expected cost. It is generally necessary to abandon the deterministic view of the short-run costs, in order to deal appropriately with user costs.

The purpose of the following discussion is to pinpoint the fallacy of traditional approaches to optimal transport pricing, and to suggest a reformulation of the theory with a view to *operationalizing* the pricing prescription.

3.2 SCHEMATIC MODEL OF A TRANSPORT FACILITY

Our discussion will be assisted by a cost model of an unspecified transport facility. The simplest possible model for the purpose consists of (1) a production function, (2) the total costs of the transport producer and (3) the total costs of the users of the facility.

$$Q = \phi f(X) \tag{1}$$

$$TC^{prod} = g(X) \tag{2}$$

$$TC^{user} = Q \cdot h(\phi;X) \tag{3}$$

$f(X)$ = capacity

X = $X_1 \ldots X_i \ldots X_n$ = facility design vector

ϕ = rate of capacity utilization

Q = transport volume

Total output can be increased either by simply raising the rate of capacity utilization, ϕ, or by expanding the capacity $f(X)$ while maintaining ϕ at a given level. Capacity is determined by a number of design variables (including the manning requirement). The convention is adopted of measuring these design variables such that $\partial f/\partial X_i$ is always positive. The capacity of a road, for example, can be increased by making the road wider or straighter or less hilly. The width of the road, and the horizontal and vertical curvature expressed by the corresponding radii, are some of the design variables that are applicable in that case.

A common characteristic is that a design adjustment that increases capacity also effects the user cost per unit, given the rate of capacity utilization. Typically, the user cost, h, falls as X_i increases. It is difficult to find a good example of a purely capacity-augmenting design variable, let alone a design variable that raises capacity while also reducing quality. We thus postulate that $\partial h/\partial X_i \leq 0$. As regards total user costs: given all design variables, the normal relationship is that TC^{user} is roughly proportional to Q in the low range for the rate of the capacity utilization, i.e. AC^{user} remains more or less constant. However, as output approaches the capacity limit, AC^{user} generally increases quite sharply due to the emergence of congestion costs, queuing costs and so on, which means that the (collective) marginal cost of the users will increasingly exceed the average user cost.

For simplicity, all producer costs are assumed to be "capacity costs". These include also a good part of the maintenance costs. The use-dependent wear and tear of transport facilities is normally of a secondary order of magnitude. For instance, only a fraction of the total repair and maintenance costs of roads is considered to be caused by the traffic. There are even indications that roads actually *improve* by being used.[1] The facility in question can also be a bus line, airline, etc. In that case the use-dependent producer costs are possibly still more minute. The airline company costs of flying N aircraft full of passengers or N aircraft which are empty apart from the crew on a given route would differ very little. The problem of optimizing the maintenance policy is a complicated one. An attempt to include this side-issue in the present discussion would only confuse the main issue.

Now we shall look at the marginal cost picture. The marginal cost of the producer, MC^{prod}, can be generally written as the ratio of the total differential of the producer costs, dTC^{prod} to the total differential of output, dQ. In the present model this comes to:

[1] This is not to deny that heavy vehicles can cause severe damage to a road which is not built for heavy traffic; this problem is left out of consideration, however.

$$MC^{prod} = \frac{\sum_i dX_i \frac{\partial g}{\partial X_i}}{d\phi f(X) + \sum_i dX_i \phi \frac{\partial f}{\partial X_i}} \qquad (4)$$

It should be observed that the common terms of the "short run" and the "long run" have no exact self-explanatory definition. However, so far as most production of the kind that is conducted at fixed plants is concerned, this does not cause much confusion. The "putty and clay" metaphor pinpoints the key dichotomy: when long-run costs are being considered, it means that the factors of production are like putty, which is true only at the planning stage, i.e. before the production plant has been built. Short-run costs apply in the clay state: some problems of definition still exists, e.g. how the costs of different categories of salaried staff and labour are to be treated, but the main idea is clear enough.

In the present model the "short run" is accordingly defined so as to correspond to an assumption that all facility design variables are fixed. This means that in the short run $dX_i = 0$, and the short-run marginal producer cost is zero.

$$SRMC^{prod} = 0 \qquad (5)$$

In the long run all design variables are adjustable, and the whole of expression (4) defines the long-run marginal producer cost. Note, however, that $LRMC^{prod}$ is strictly undetermined as long as no further condition is presumed as regards the factor combination. Anyway, the common notion of $LRMC^{prod}$ is that this cost is at a much higher level than $SRMC^{prod}$ in the whole possible range, because in the long run all capacity costs are variable. $SRMC^{prod}$ is perhaps not usually thought to be zero, but certainly at a very low level - right up to the capacity limit. At that point $SRMC^{prod}$ ceases to be determinate according to the conventional way of looking at the matter. The possibility that the short-run and long-run marginal costs could coincide at any level of output is very remote.

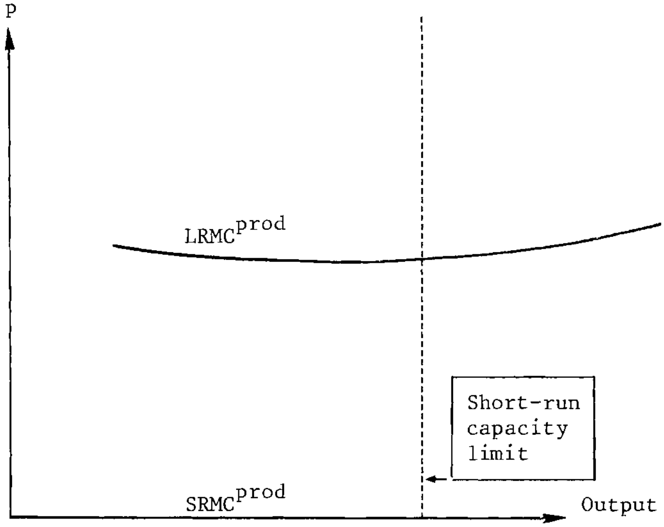


Figure 3:1. Typical positions of a transport producer's short-run and long-run marginal costs

3.3 GENERAL FORMULATION OF THE PRICING-RELEVANT COST

However, the important point is not which is chosen of $SRMC^{prod}$ or $LRMC^{prod}$ (or any "medium-run" producer marginal cost), as the basis for pricing - both are equally wrong - but that the optimal price should include a user cost item besides the producer marginal cost, or more exactly, the product of the transport volume Q and the change in AC^{user} caused by a unit increase in Q. This stands to reason. The consumer of an additional unit of transport should pay the cost caused to the transport producer as well as the cost caused to the original transport consumers. If the latter cost is zero, there should not of course be any addition to the producer marginal cost in the price, and *nota bene*, if the additional unit of transport consumption generates benefits rather than costs for the original consumer, it is equally obvious that a deduction from the producer marginal cost should be made to obtain the optimal price. A general formulation of the pricing-relevant cost is consequently:

$$PC = MC^{prod} + Q\,\frac{dAC^{user}}{dQ} \tag{6}$$

As far as the theory of road pricing is concerned, this formula has been well known and widely accepted for quite a long time. In this case the very slight road surface wear and tear caused by another car-kilometre corresponds to MC^{prod}, and the congestion costs imposed on the original traffic by an additional car-kilometre corresponds to the product of Q and dAC^{user}/dQ. The point of the present general formulation is that, if it is to be applicable throughout the transport sector, it is better not to presuppose that the extreme short-run version (assuming all facility design variables are fixed, as we are used to doing from the road pricing discussion) will always be the most relevant.

3.4 OVERESTIMATIONS AND UNDERESTIMATIONS OF THE PRICING-RELEVANT COST

The long-run versus short-run marginal cost controversy regarding transport pricing has no doubt its origin in the large apparent difference between $LRMC^{prod}$ and $SRMC^{prod}$. It is true that, typically, the supporters of the extreme short-run school of thought do not make very definite pronouncements about the level of optimal price: rather, they content themselves with pointing out that all other marginal costs are irrelevant to pricing. However, it has been more or less implied that a very low optimal price level follows from accepting the philosophy of "letting bygones be bygones" At least this implicit notion has been an inducement to the long-run marginal cost supporters' counter-attack. Large financial deficits have been regarded as unacceptable for both allocative and X-efficiency reasons, and several attempts have been made to show by theoretical arguments as well as by empirical studies of costs that, if prices are based on long-run rather than short-run marginal costs, the financial result would be much "better".

From a narrow producer cost point of view, it might seem that full recovery of the producer costs would ensue if long-run marginal cost pricing were applied, since in many studies the total producer costs have been found to be proportional or nearly proportional to the total transport volume.[1] This is true about cross-section studies of railway companies and bus companies as well as of time-series studies of aggregate road investment expenditure and road traffic volume. The latter type of evidence has seemingly supported the claim of the "development cost" argument regarding road finance,[2] that the long-run marginal cost pricing of road use might well balance the road budget.

The most important fact which is overlooked by the supporters of long-run marginal cost pricing is that - provided that the facility design can be adequately adjusted - the quality of service is normally higher, the larger the transport volume. Cases in point abound: scheduled transport services become more frequent, the larger the volume of traffic. Faster trains and more comfortable coaches are

[1] This is further discussed in Chapter 4, Section 4.4, where relevant references are given.

[2] A good example is Hiort, O.C., Kostnadsorienterte bilavgifter. Transportøkonomisk Institutt. Slemdal, 1964.

employed on railway trunk lines, while more or less obsolete rolling stock is relegated to branch lines. Roads are upgraded successively as the traffic increases etc, etc. Cross-section studies of the costs of railway transport, and of bus transport showing only moderate or no economies of scale in the *producer costs*, cannot be used as a basis for judging what the result of optimal pricing would be. It is quite obvious that the quality of service is not constant between low-volume and high-volume lines.

This does not mean that $\text{SRMC}^{\text{prod}}$ is a correct basis for pricing. In the short run, the quality of service decreases with increases in output due to rising congestion and/or queuing. To ignore these effects and to consider $\text{SRMC}^{\text{prod}}$ only is illegitimate for the same reason: the service in question is not the same at different levels of output. Whereas an *over*estimation of the optimal price will be the result of ignoring the user costs when the typical long-run marginal cost approach is adopted, an *under*estimation of the optimal price will result if pricing is based on $\text{SRMC}^{\text{prod}}$. In the theory of road pricing, a user-cost component was rightly included from the start.[1] Indeed it was the completely dominating component in cases where spillover effects on third parties are unimportant. However, with a few exceptions,[2] this crucial idea has not been followed up as it should have been in other areas of transport price theory. This is probably because a *deterministic* cost theory applies reasonably well to road pricing, while a deterministic view of user costs in other areas is very misleading.

[1] The classical reference is Walters, A.A., The theory and measurement of marginal private and social costs of highway congestion. Econometrica, 1961:4.

[2] Notably Turvey, R. & Mohring, H., Optimal bus fares. Journal of Transport Economics and Policy, September 1975.

In a deterministic setting characterized by a kinked short-run marginal cost curve, the optimal price will either be at the level of the constant $SRMC^{prod}$, that is nearly zero, or at a level determined by the intersection of the demand curve and the capacity limit. In this model the optimal price will exceed $SRMC^{prod}$ only when 100 per cent capacity utilization prevails. Since 100 per cent capacity utilization is rightly a rare occurence in transport, the uncritical use of a deterministic cost model can easily result in a systematic underestimation of the pricing-relevant cost of transport services. The famous controversy about which price "the passenger to Calais" should pay for a ticket, is one example of the pitfalls that exist when a deterministic view is assumed of a situation involving important stochastic elements.[1]

3.4.1 The degree of freedom in facility design makes no systematic difference to the pricing-relevant cost

With our cost model the common overestimation as well as underestimation of the pricing-relevant cost, and the futility of the whole long-run versus short-run marginal cost pricing controversy can be lucidly illustrated.

In the model the pricing-relevant cost, PC, is the sum of MC^{prod} according to (4) above, and a user cost component taking this shape:

$$Q\,\frac{dAC^{user}}{dQ} = Q\,\frac{d\phi\frac{\partial h}{\partial \phi} + \sum_i dX_i\,\frac{\partial h}{\partial X_i}}{d\phi f(X) + \sum_i dX_i \phi\frac{\partial f}{\partial X_i}} \qquad (7)$$

[1] This is discussed for example in Dessus, G., The general principles of rate-fixing in public utilities, International Economic Papers, No. 1, 1951 (also in Nelson, J.R. (ed.), Marginal cost pricing in practice, Prentice-Hall, 1964).

Adding (7) and (4) gives us:

$$PC = \frac{d\phi Q \frac{\partial h}{\partial \phi} + \sum_i dX_i \left(\frac{\partial g}{\partial X_i} + Q \frac{\partial h}{\partial X_i}\right)}{d\phi f(X) + \sum_i dX_i \phi \frac{\partial f}{\partial X_i}} \tag{8}$$

This expression for the pricing-relevant cost will be considered under the customary efficiency condition.

In the absence of spillover effects on third parties, the total social cost, TC, is equal to the sum of TC^{prod} and TC^{user} according to (2) and (3) above. The standard efficiency condition that TC should be as low as possible for each level of output is obtained by setting the derivatives of the following lagrangian equation equal to zero.

$$\Pi = TC^{prod} + TC^{user} + \lambda \left[Q - \phi f(X)\right] \tag{9}$$

The first-order conditions for a minimum are:

$$\frac{\partial \Pi}{\partial \phi} = Q \frac{\partial h}{\partial \phi} - \lambda f(X) = 0 \tag{10}$$

$$\frac{\partial \Pi}{\partial X_i} = \frac{\partial g}{\partial X_i} + Q \frac{\partial h}{\partial X_i} - \lambda \phi \frac{\partial f}{\partial X_i} = 0 \; (i=1...n) \tag{11}$$

$$\frac{\partial \Pi}{\partial \lambda} = Q - \phi f(X) = 0 \tag{12}$$

We can now use these equations for examining how PC above looks under the efficiency condition. Inserting $\lambda f(X)$ from (10) for $Q\partial h/\partial \phi$ in (8) and $\lambda \phi \partial f/\partial X_i$ from (11) for $\partial g/\partial X_i + Q\partial h/\partial X_i$ in (8) the following result is obtained:

$$PC = \frac{d\phi\lambda f(X) + \sum_i dX_i\lambda\phi \frac{\partial f}{\partial X_i}}{d\phi f(X) + \sum_i dX_i\phi \frac{\partial f}{\partial X_i}} \tag{13}$$

It is seen directly that in this new expression for PC, the two terms of the numerator are the same as the two terms of the denominator, except that each of the former include the lagrangian multiplier λ as a factor. The end result is consequently:

$$PC = \lambda \tag{14}$$

No matter which, or how many, of the design variables $X_i \ldots X_i \ldots X$ are fixed or variable, the pricing-relevant cost will come out the same, provided only that an efficient factor combination exists in the initial position. Setting an arbitrary number of $dX_1 \ldots dX_i \ldots dX$, and $d\phi$ equal to zero makes no difference to the result.

A completely different picture emerges when each of the producer cost and user cost components is looked at separately. Take first $Q \cdot dAC^{user}/dQ$.

Given that $\partial h/\partial X_i \leq 0$, while $\partial f/\partial X_i > 0$, it follows that the user cost component in PC decreases successively as more of the increments dX_i are given positive values. On the other hand, as $MC^{prod} = \lambda - QdAC^{user}/dQ$, it follows that the producer cost component in the optimal price increases as the degree of freedom in facility design grows.

In Figure 3:2 the offsetting relationship between the two components of PC is schematically illustrated. Along the horizontal axis the degree of freedom in facility design is measured from left to right. At the point of origin all design variables are fixed. At that level MC^{prod} is at its lowest and $Q \frac{dAC^{user}}{dQ}$ at its highest. To begin with, PC will be above MC^{prod}, and the user cost component is positive. When all design variables are fixed, the latter component equals the imposed congestion and/or queuing costs. The basic proposition here, however, is that sooner or later the user cost component of PC is likely to be negative, which will mean that PC is lower than MC^{prod}.

The principal moral is that the degree of freedom in facility design is not a very crucial factor when it comes to pricing, provided that an optimal investment policy is pursued. The important point to bear in mind is instead to be *consistent*. That is to say, if a particular degree of freedom in facility design is chosen for calculating the producer cost component in the pricing-relevant cost, the same degree of freedom has to be adopted also for calculating the user cost component.

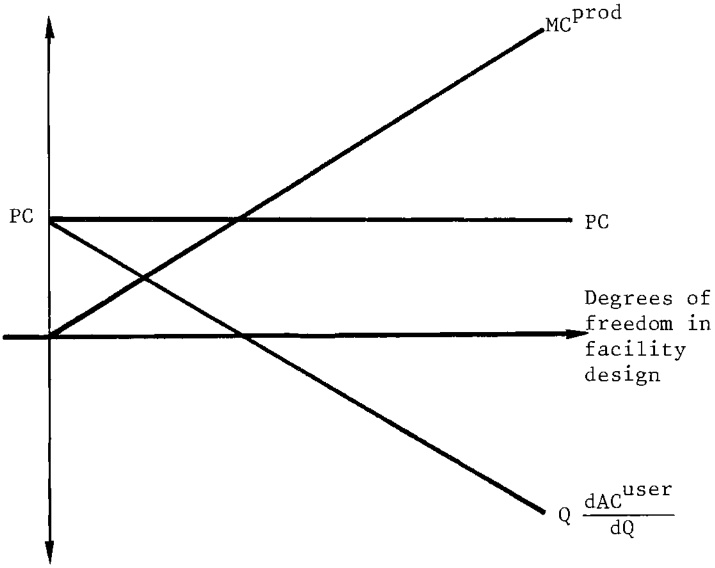

MC^{prod}
PC
PC
Degrees of freedom in facility design
$Q \frac{dAC^{user}}{dQ}$

Figure 3:2. Schematic illustration of the change in the proportion of the producer cost and user cost components of the optimal price, as the freedom of choice in facility design is increased.

3.5 A FUNDAMENTAL DICHOTOMY

Has the preceding generalization of the formula for the pricing-relevant cost, PC, any practical relevance to optimal transport pricing, apart from pinpointing the common overestimation and underestimation of the pricing-relevant costs? It will be argued here that when it comes to practical applications of the general theory of optimal pricing, the transport service sector resolves itself into two parts: in one the user cost component in the optimal price is positive, and in the other it is negative in the pricing-relevant "run". This is the same as saying that in one part of the sector, pricing based on short-run costs of the kind that is well established in the case of road services is relevant, and in another part a variation which is based on what can be called "medium-run" costs is relevant. The first of these parts includes services provided by:

- *Transport infrastructure*, i.e. roads, rail tracks, fairways, seaports, airports, and other "terminals",

and the other part includes services provided by:

- *Transport vehicles* using the roads etc.

The reaction to this may well be: "so what"? Was not the main point that it does not affect the pricing-relevant cost, which "run" is assumed to be applicable?

From a theoretical point of view, admittedly, it does not really matter, provided we can assume that the efficiency conditions are met. However, for the empirical work that is necessary if we are to apply the theory of optimal pricing, it makes a considerable difference whether the short-run or medium-run pricing-relevant costs are to be estimated. Above all, quite different user costs are of paramount importance in each of the two cases.

The costs to the user - freight as well as passengers - of coming into possession of services supplied by transport facilities can be categorized in the following way:

1. Costs of *going to/from* the facility. The service is not usually available on the door-step of each user. "Feeder transport" of varying length is normally required.
2. In the case of scheduled transport services, *waiting costs* (or equivalent) of varying size have to be incurred by users due to the discrete nature of departure times.
3. A special kind of waiting - conspicuous or latent *queuing* - is suffered when the facility is already fully utilized.
4. Last but of course not least are the costs of the time consumed by the *actual use* of the facility. As regards transport infrastructure, these costs consist of both user time and vehicle time as well as various supplementary inputs, whereas passengers' riding time or the interest cost on freight in transit are the only items as far as services supplied by transport vehicles are concerned.

The distinguishing characteristic of the production plants represented by individual pieces of transport infrastructure is that the plants in question are literally fixed (cannot be moved) once they have been established. In the short run - given the capacity as well as location of the transport infrastructure - the average user costs of categories (1) and (2) are fixed and can be ignored in any calculation of pricing-relevant costs. The primary determinant of the user costs of categories (3) and (4) is the rate of capacity utilization. These costs are therefore highly relevant to the pricing of services provided by the transport infrastructure.

Because of mobility of transport vehicles, the vehicle input is a variable factor of production in the pricing-relevant "run" in the case of transport vehicle services. In reality vehicle input is far more variable than tariffs of freight rates or passenger fares. For this reason the user costs which are primarily dependent on the rate of capacity utilization are less important in connection with price-making. This does not mean that knowledge about the queuing, congestion and crowding costs in public transport or scheduled freight transport is not worth striving for. On the contrary, such knowledge is necessary, strictly speaking, for deriving the pricing-relevant medium-run costs. The capacity adjustment should in principle be a trade-off of producer capacity costs against the relevant user costs, which include queuing and suchlike costs. At present the necessary knowledge of these user costs for such a trade-off is missing, and in practice it is common to resort simply to judgements of what is a reasonable "practical" limit of the capacity, or tolerable occupancy rate in the peak period. This is the same as approximating the true, continuously rising relationship between the user cost and the occupancy rate by a kinked cost function. It is in the nature of things that the kink of the approximate curve should not coincide with the maximum capacity. Depending on the variability in demand, the practical capacity should be set some way on this side of maximum capacity. Nor does this approximation mean that the well known-model illustrated in Figure 3:2 on page 54 becomes relevant for price-making. The main point concerning vehicle transport services is that there is no given limit of capacity in the pricing-relevant run.

In the following two sections we will discuss more in detail the problems that meet when it comes to deriving the pricing-relevant costs of services provided by different pieces of transport infrastructure, on the one hand, and by different transport vehicles, on the other.

3.6 EXPECTED SHORT-RUN COSTS

The main problem concerns the derivation of the short-run user costs. The problems of measuring the value of time in various occupations are well known. Here, in the first place, some conceptual problems of *expected* costs will be discussed.

A crucial difference exists between the expected and the deterministic short-run user cost in practically all relevant cases except roads. The short-run user costs of a "common facility", such as a road takes the form of congestion costs: the more cars there are in a given road network, the more obstructive each car will be to the movements of other cars. A short-run average user cost of the shape of (a) in Figure 3:3 can easily be envisaged, and is also clearly borne out by numerous empirical studies of the speed of flow of road traffic.

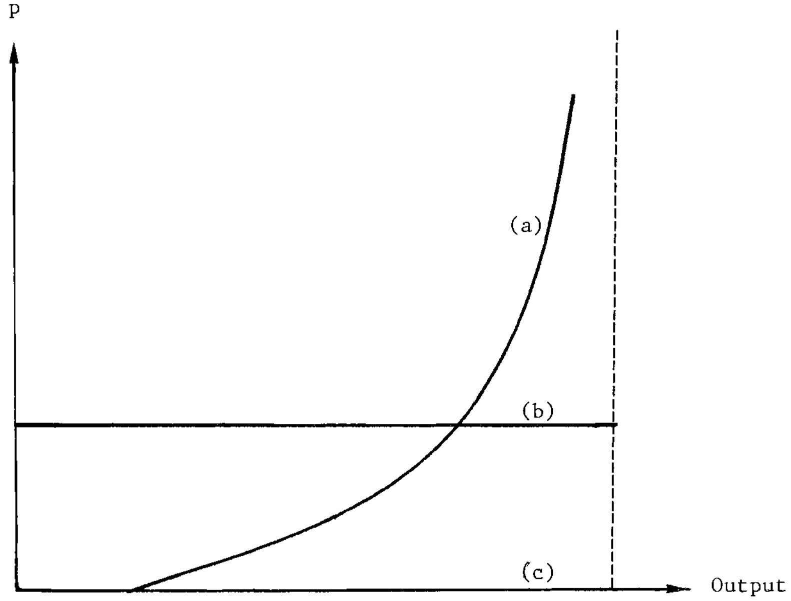


Figure 3:3. Standard shapes of the short-run average cost

Difficulties arise as soon as various "departmentalized facilities" come up for discussion. Where there are no congestion costs, because users do not interfere with each others' consumption of the services, a deterministic view of the short-run user costs easily leads to the assumption that these costs are constant (often zero) per unit right up to the capacity limit.

Our basic proposition here is that shape (a) is generally representative of the *expected* short-run user cost of transport services. Shape (b) is probably fairly common in goods manufacturing - it should be a good approximation of the short-run unit costs for material goods, in view of its importance in accounting, as well as in linear programming approaches to goods-production problems. Shape (c) is the result of a false analogy between goods production and service production. Services - non-storable goods - are "immaterial", and require no input material: hence the presumption that the unit costs are zero in the short-run, or equal to the tiny use-dependent cost of wear and tear. However, if the deterministic view of user costs is abandoned and the concept of expected costs introduced, another picture of the shape of the short-run costs will emerge. The expected short-run user cost as a function of the expected output will generally assume the shape of (a) in Figure 3:3 for departmentalized facilities as well, mainly because of the *risk of excess demand* in the face of a stochastic demand.

3.6.1 Failure of deterministic approaches to costs of non-storable goods

Let us start the discussion by a concrete example of which most people have daily experience - the supply of parking space on the street. The following quotation from Michael Thomson's seminal work "Modern Transport Economics"[1] poses the problem:

[1] Thomson, J.M., Modern Transport Economics. Penguin Modern Economic Texts, 1974.

> What is the *cost* of parking on the street? If the parked vehicle causes delay to moving traffic, a cost arises in the form of other people's time and operating expenses ... But if there is no such traffic effect and no other resource costs, why charge for parking? The answer is that, if parking space is in short supply, so that one person parking prevents another from doing so, there is an opportunity cost equal to the maximum price the other would pay for the opportunity to park ... Why does marginal cost pricing lead to a policy of charging what the traffic will bear for scarce parking place, but not for scarce road space, i.e. for moving traffic. The reason is that the cost of using parking space does not rise until the last space is occupied, i.e. the cost curve is kinked, whereas the use of road space creates social costs at a steadily growing rate as the volume of traffic rises towards the capacity of the road.
> (Thomson, 1974, p. 145)

That is to say, if the supply of parking space in a given area is S, and the workday demand is given by D (see Figure 3:4), the optimal price for parking space is, according to Thomson, the price which equates supply with demand. No short-run costs whatsoever are involved.

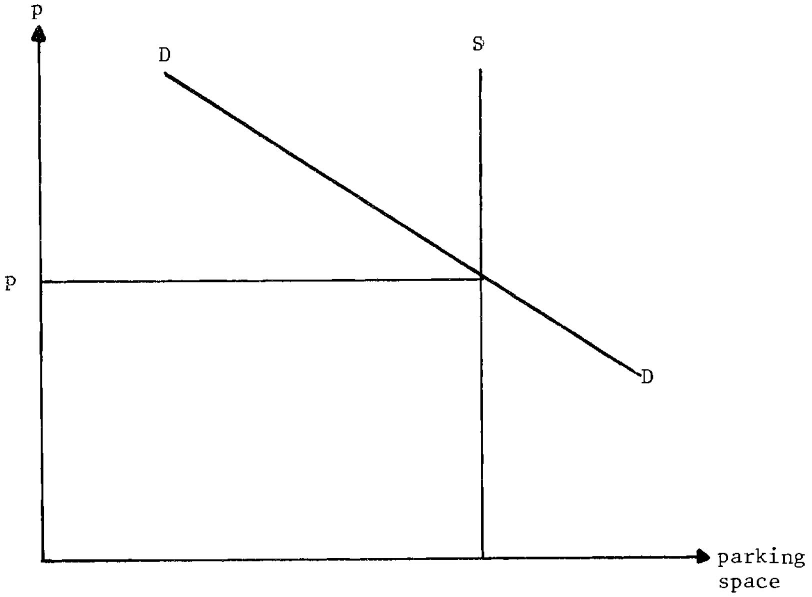


Figure 3:4. A deterministic view of optimal pricing of parking space

However, an important point is that the workday demand is highly unlikely to stay the same from one day to another. Given the price, the outcome will be that excess demand and excess supply situations alternate, and only occasionally will it happen that exactly S cars demand parking space. In situations of substantial excess supply it can be assumed that no user costs worth mentioning will arise. This is not so in situations of excess demand. Everyone who has ever driven into the Central Business District of any fairly large city has experienced the irritation and frustration of having to drive around looking for a place to park. In fact, the expected cost per parking space will rise sharply well before the expected demand for parking equals the supply of parking space. If the expected demand for parking space at a particular time of the day is, for instance, 75 per cent of the supply, it means that since the demand contains a more or less important stochastic element, it is bound to happen from time to time that some people fail to find a (legal) place to park, or have to park far away from their destination - perhaps outside the CBD - after a lengthy search for parking space.

The question is: what should be the optimal price, and the resulting occupancy rate, of the parking space in a particular area, given the realistic case of a stochastic demand for parking? It should be clear that it is inoptimal to aim at 100 per cent capacity utilization - which would be the right aim in a deterministic case, or if the occurence of excess demand costs nothing from a social point of view. In the latter unrealistic case the objective should be to maximize capacity utilization, and a rather low price - lower than p in the deterministic case (see Figure 3:4) - would probably be right. If the further unrealistic assumption is made that the available parking space could somehow always be allocated to those who are in greatest need of it without the aid of pricing, it is obvious that a zero price is optimal. Again, the reaction could well be: "so what"?

On the condition that (1) excess demand costs nothing, and (2) goods/services can be rationed as efficiently without prices as with them, then the price system has no allocative function to fulfil. The reason for mentioning this absurd way of arguing is that it does represent, albeit as something of a caricature, a line of thought on public utility pricing under risk which has appeared in the last ten years, inspired in the first place by Brown & Johnson (1969).[1] It was immediately refuted by Turvey (1970),[2] but has nevertheless flourished. Visscher (1973)[3] and others have rightly been unhappy with the implicit assumption that those who have the greatest willingness to pay can be found without resort to pricing. An even more important fallacy, in my view, is that the user costs of excess demand are ignored. The problem is how to handle these costs analytically.

Another approach that appears to side-step this problem, and which many economists impressed by the smooth working of currency markets, for example, are inclined to advocate, would be to arrange a kind of "auction" by means of which the market is continuously cleared. Concrete suggestions as to how such a system could work have been made with regard to long-distance air travel.[4] A crucial point about the efficiency of auctions, however, is that in order to make their considerable administrative costs worthwhile, the total turnover of each auction and of each individual transaction has to be quite large in terms of money, so that the

[1] Brown, G. & Johnson, B., Public Utility Pricing Under Risk. The American Economic Review, 1969. See also e.g. Andersen, P., Public Utility Pricing in the Case of Oscillating Demand. Swedish Journal of Economics, 1974.

[2] Turvey, R., Public Utility Pricing Under Risk: comment. The American Economic Review, June 1970.

[3] Visscher, M.L., Welfare-Maximizing Price and Output with Stochastic Demand: comment. The American Economic Review, March 1973.

[4] See e.g. Simon, J.L. & Visrabhanathy, G., Auction Solutions for Airline Overbooking. Journal of Transport Economics and Policy, September 1977.

relative pricing costs are low. These conditions, which are fulfilled on the stock and currency markets for example, are generally far from satisfied in the case of transport infrastructure or scheduled transport services which are particularly adapted to meet the needs of "small" transport users. Applied to short-distance public transport or parking space allocation, the system would be patently ridiculous. At least with the present methods of arranging auctions, there is no doubt that a continuous market-clearing price system for, let us say, parking space would generate very high pricing costs and yield only modest benefits. The only possible advantage of such extreme price flexibility would be that, on occasions of excess demand, the available parking space would be allocated according to willingness to pay rather than according to the principle of first-come-first-served. Note that the user costs of excess demand would not be eliminated as a result of introducing parking space auctions. Those who were outbidden would incur further costs of varying size, connected with finding a place to park elsewhere or, in the extreme case, going home again - and the total user costs of excess demand would be unlikely to be reduced. An auction system for parking space allocation, which would also eliminate the costs of excess demand, has to work in such a way that the allocation has already been made before the motorists demanding parking space have embarked on their trips. After the "Communication Revolution", which some people believe is coming, such auctions may be so cheap to arrange that they will become worthwhile. I doubt this, but there is no point in committing oneself more than necessary. At the present time and in the foreseeable future no system of continuous market-clearing pricing of public transport, of parking space, or of other low-value services is feasable.

3.6.2 Previous work on the derivation of expected short-run *producer* costs

The reason why economists have been slow in developing the theory of expected cost is most likely the excessive focus on storable goods in economic theory. For goods that can be stored at a relatively low cost, the more or less random short-term fluctuations in demand do not pose much of a problem. It will normally pay to maintain an even rate of output, irrespective of the actual time-profile of demand. Output (and input) stocks act as buffers between a constant rate of production and the variable rates of sales and the purchase of inputs. Under such conditions - which prevail in a major part of the storable goods production sector of an economy - uncertainty about demand is a problem in long-run production planning. The case of non-storable goods is, of course, a different matter.

The pioneer in the field of expected short-run cost derivation is Walters. At the CEMT symposium in Strasbourg on "Theory and practice in transport economics" Walters introduced a paper (called "Characteristics of demand and supply") by this observation:

> The non-storability of services implies that all demands must be met by production at the time required and all seats or spaces which are not filled are wasted. It is argued, however, that conventional cost analysis is not relevant to decision-making in the transport industries. Decisions are made about expected flows of traffic. Expected costs are the relevant magnitude for these decisions.
> (Walters, 1964)[1]

Walters has also touched upon the subject of expected cost in his famous study of road user charges, where he indicates a method of deriving expected short-run costs.[2]

[1] Walters, A.A., Characteristics of Demand and Supply. European Conference of Ministers of Transport: International symposium on theory and practice in transport economics. Strasbourg, October 8, 1964.

[2] Walters, A.A., The Economics of Road User Charges. World Bank Occasional Papers, No. 5, pp. 67-70.

Strangely enough, no one has followed up Walters' pioneering work. Independently of Walters, Rotschild & Stiglitz have considered the same problem of deriving expected short-run costs, although their interest is specifically geared to the "classical" question, whether uncertainty of demand in a profit-maximizing firm will result in a larger or a smaller optimal plant than the optimal plant in the "ordinary" riskless case.[1]

A summary of these various contributions to the issue of expected producer cost can be made as follows: the period of time up to the planning horizon (= planning period) of an enterprise is divided into a number of production periods. The production periods may be days, weeks or even years, depending on the particular circumstances. The price of the product canrot be changed during a production period. Orders come in more or less continuously during each period. All demand has to be met during (or at the end of) each such period. The *expected* demand per production period is constant up to the planning horizon, but subject to random fluctuations. The capital input is fixed. The fluctuations in demand have to be met by changes in the variable input(s) - for example, labour. An important although implicit assumption is that it is always possible to satisfy the demand of each production period, although at some cost.

The *ex post* Total Variable Cost of a production period is determined by the output, Q, that is demanded in the period concerned.

$$TVC = C(Q) \tag{15}$$

If Q were constant between production periods, there would be no difference between the ex post and ex ante TVC.

[1] Rotschild, M. & Stiglitz, J.E., Increasing Risk II: Its economic consequences. Journal of Economic Theory 3, 1971.

However, given the price, the quantity demanded (= output produced per production period) is a random variable assuming the values Q_1, $Q_2 \ldots$ Q_n with probabilities $f(Q_1)$, $f(Q_2) \ldots$ $f(Q_n)$. The expected value of Q is written:

$$E(Q) = \sum_{i=1}^{n} Q_i f(Q_i) \tag{16}$$

$$\text{where } \sum_{i=1}^{n} f(Q_1) = 1$$

As Q is a random variable, TVC is also a random variable. The expected value of TVC can be written:

$$E(TVC) = \sum_{i=1}^{n} C(Q_i) f(Q_i) \tag{17}$$

$$= E \left[C(Q)\right] \tag{18}$$

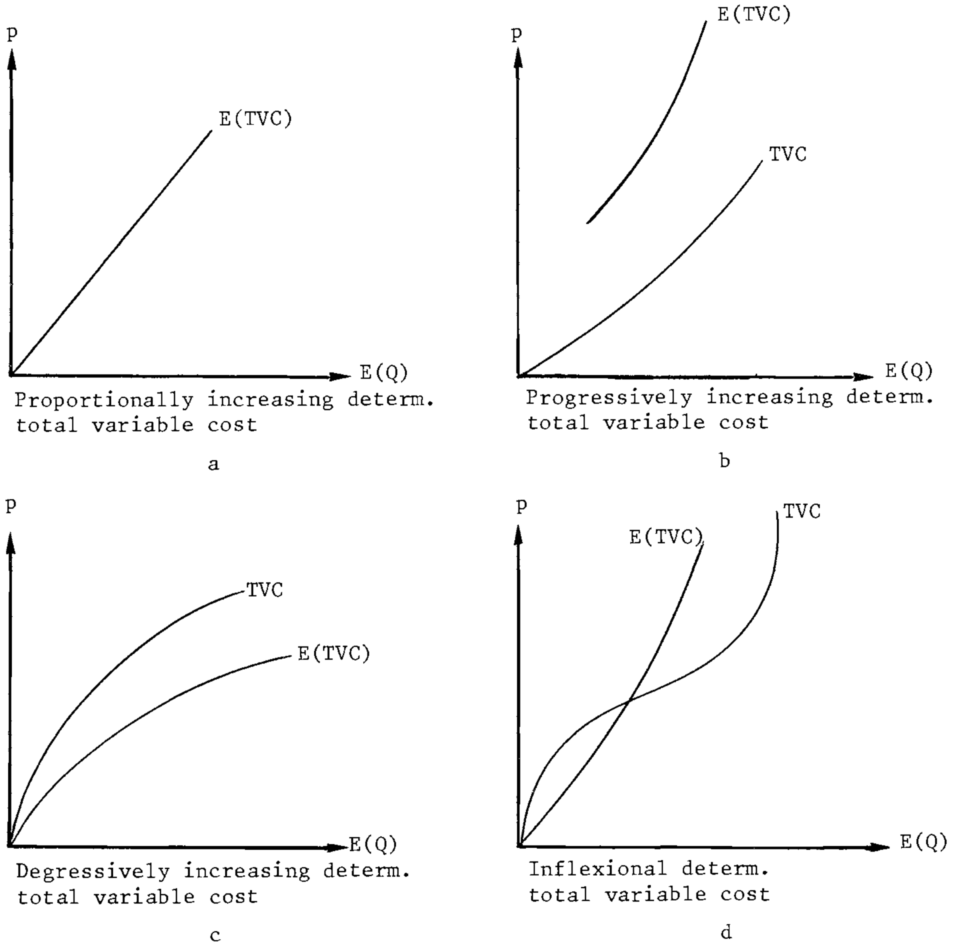
From equation (18) it can be seen that the expected value of a function of a random variable depends on the whole probability distribution of the random variable (in this case Q), and not just on one characteristic of the probability distribution, such as the mean.

Comparison between TVC and E(TVC) are therefore very difficult. However, it is possible to say, for any probability distribution of the output with the mean E(Q), whether the value of E(TVC) is higher or lower than the value of TVC for an output equal to E(Q).

Intuitively, it seems clear that if the riskless cost function C(Q) is linear, TVC and E(TVC) assume the same value regardless of the probability distribution of Q. This conclusion can be checked with the help of equation (17) above. In a linear case the right-hand sum is reduced to:

$$\text{constant} \cdot \sum_{i=1}^{n} Q_i f(Q_i) = \text{constant} \cdot E(Q) \tag{19}$$

The law of diminishing returns indicates a convex relationship between TVC and output. In such a case it is also intuitively clear that E(TVC) is higher than TVC for every value of E(Q) (as illustrated in Figure 3:5b below).



Proportionally increasing determ. total variable cost

a

Progressively increasing determ. total variable cost

b

Degressively increasing determ. total variable cost

c

Inflexional determ. total variable cost

d

Figure 3:5. Expected versus deterministic short-run total cost curves

Formally, we can support this conclusion by referring to "Jensen's inequality".[1]

If the relationship between TVC and output were concave in a deterministic setting, which is very unusual, I suppose, E(TVC) would be lower than TVC throughout. Walters has made the point that if the curvature of TVC changes from concavity to convexity (as it does in many economic textbooks), E(TVC) will be much less wavy.[2] This is illustrated in chart d of Figure 3:5.

A reservation has to be added with regard to charts b-d in Figure 3:5. It is not strictly correct to draw E(TVC) in a two-dimensional diagram, as if it were determined by E(Q) alone. The indicated relative position of E(TVC) and TVC holds good, irrespective of the shape of the probability distribution. The exact position of E(TVC), however, is not determined by a particular value of E(Q).

The conclusions about the importance of the curvature of the "ordinary" deterministic total variable cost curve to the shape of E(TVC) are interesting, but barely touch on the real problem, when we introduce risk into the cost analysis of various services. As far as departmentalized facilities are concerned, the main consequence of introducing a stochastic demand is that the short-run costs will rise because of the risk of *excess demand.* A primary problem of short-run operations is the possibility that capacity may occasionally be insufficient for the demand.

3.6.3 The crucial role of excess demand

A remarkable feature of both Walters' and Rotschild & Stiglitz's work is that they fail to take into account a short-run capacity limit. Walters uses short-run cost

[1] Rotschild, M. & Stiglitz, J.E., 1971, op. cit., p. 67.

[2] Walters, A.A., 1968, op. cit., pp. 68-69.

curves which are increasing but which do not reach infinity for finite levels of output. Rotschild & Stiglitz deal with a production function without limitational factors.

The reason why no capacity limit is taken into account is probably that it would cause serious analytical problems, so long as the analysis of expected cost is restricted to producer costs. The producer does not normally bear the brunt of the direct cost of excess demand. Indirectly the producer may become aware of all the costs because of a loss of revenue following the deterioration in the quality of service, which is bound to result when excess demand is a frequent occurrence.

In a context of net social benefit (rather than private profit) maximization these analytical problems look rather different.

We are back to the original question: given that the price of a particular service has to be set in advance, and that there are significant *user* costs connected with excess demand, how is the optimal price and output to be determined? In principle the nature of this problem is clear: it is a matter of trading off the costs of the risk of excess demand against the costs of excess supply (benefits foregone due to the underutilization of the available capacity.

Analytically, the most convenient approach is to derive a function for the user costs of excess demand expressed in monetary terms, and assume that demand depends on the "generalized" user cost rather than on the price alone.

The exact way of doing this depends on the particular circumstances of the service production in question. For many fully departmentalized facilities such as the berths of a seaport, or public transport where only sitting is allowed, etc., *queuing* in one form or another will occur. The queuing may be conspicuous or latent. Ships that cannot find suitable berths that are unoccupied are kept waiting, and short-distance public transport travellers who are denied a place will queue up at the stop or station, while the users of the less frequent long-distance services do not form a queue in the literal sense, but they suffer other inconveniences which are more difficult to measure, although they are certainly not negligible. By making use of queuing theory and related analytical tools, it is possible in most cases to get a good idea of the shape of the expected user cost functions. As is well known, there is a vast literature concerned with queuing theory and its applications (see Chapter 9). In the case of "semi-departmentalized" facilities such as underground commuter services where standing is allowed in the carriages, the inconvenience of *crowding* constitutes a special type of queuing cost. It can be regarded as queuing for a seat. The distinguishing characteristic of this cost is that it does not involve a waste of user time, but takes the form of an increasing disutility (cost) of travel time.

The greatest difficulty in monetarizing the user costs of excess demand arises in cases where these costs assume the shape of *frustration of not attaining possession* of the service concerned. In the case of parking in a particular area, the problem is manageable because a close substitute is available, i.e. parking elsewhere; the frustrated customers do not have to refrain from the service altogether. The only difference is that some user time is wasted on the extra search and a longer walk to the final destination. Let us conclude the discussion of pricing-relevant costs in the short run by exemplifying how, in principle, the expected short-run user costs might be derived.

3.6.4 Example of derivation of expected short-run user costs

Suppose that the total supply of parking space in a defined area is S, the quantity demanded is D, and output, in the shape of parked cars, is Q. The quantity demanded is a random variable, assuming the values $D_1 \ldots D_i \ldots D_k$, $D_l \ldots D_j \ldots D_n$, which are arranged in order of magnitude, with probabilities $f(d_1) \ldots f(D_n)$,

$$D_k \leq S, \text{ and } Q_i = D_i \tag{20}$$

$$D_l \geq S, \text{ and } Q_j = S \tag{21}$$

$$\sum_{i=1}^{k} f(D_i) + \sum_{j=l}^{n} f(D_j) = 1 \tag{22}$$

The expected demand E(D) is:

$$E(D) = \sum_{i=1}^{k} D_i f(D_i) + \sum_{j=1}^{n} D_j f(D_j) \tag{23}$$

As the demand is a random variable, so is the output Q. The expected output quantity E(Q) depends on D and S, and can be written:

$$E(Q) = \sum_{i=1}^{k} D_i f(D_i) + \sum_{j=l}^{n} Sf(D_j) \tag{24}$$

A high expected occupancy rate - a value of E(Q) close to S - will be bought to the price of frequent excess demand. This is clearly seen by writing the occupancy rate E(Q)/S.

$$\frac{E(Q)}{S} = \sum_{i=1}^{k} \frac{D_i}{S} f(D_i) + \sum_{j=l}^{n} f(D_j) \tag{25}$$

This expression is obviously always less than unity, as D_i/S is less than unity, and the sum of the probabilities $f(D_l) \ldots f(D_n)$ adds up to unity. For E(Q)/S to be close to unity it is required that the sum of $f(D_1)...f(D_n)$ assumes a high value at the expense of the sum of $f(D_1)...f(D_k)$. This occurs when the expected excess demand is great. Figure 3:6 below gives a typical shape of the relationship between E(Q) and E(D). Provided that the skewness, or the relative mass of the "tails" of the probability distribution of D do not change as E(D) goes up, further increases in expected output, E(Q) will be bought at a higher and higher price, i.e. result in a steadily increasing expected number of frustrated parkers.

The expected value of the excess demand E(ED) is defined:

$$E(ED) = \sum_{j=1}^{n} (D_j - S)\ f(D_j) \tag{26}$$

which is easily shown to be equal to the difference between the expected demand and the expected output.

$$E(ED) = E(D) - E(Q) \tag{27}$$

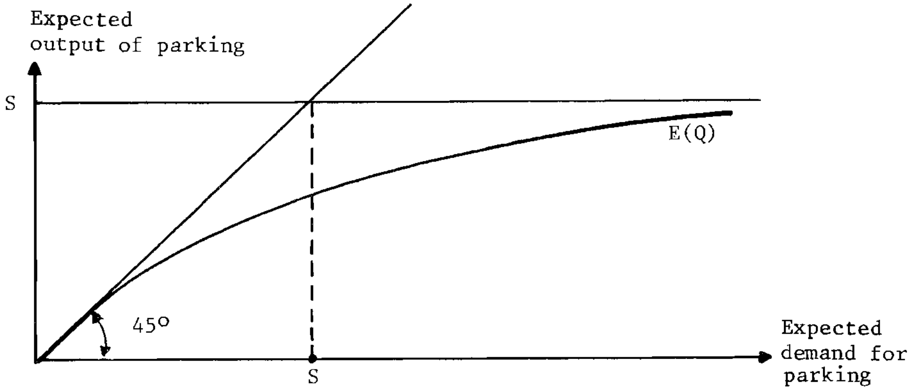


Figure 3:6. The expected output of parking as a function of the expected demand

Applying a unit cost c per frustrated parker, the expected user cost of parking can be written:

$$E(TC^{user}) = c \left[E(D)-E(Q)\right] \qquad (28)$$

The curve of Figure 3:7 giving the shape of the expected total user cost is obtained by taking the difference between E(D) and E(Q) in Figure 3:6 (the horizontal difference between the output curve and the 45°-line) for each value of E(Q).

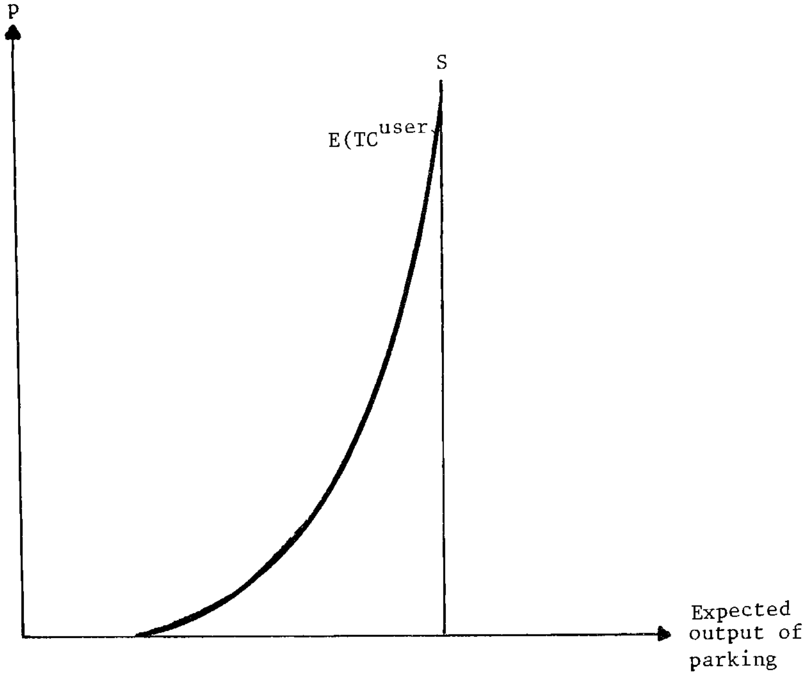


Figure 3:7. Expected total user cost of parking

In conclusion it can thus be seen that although there are no short-run producer costs, the introduction of a short-run capacity limit (which, it will be remembered, is absent from the diagrams in Figure 3:5a-d) gives rise to a cost which has to be taken into account in the pricing.

To find the optimal price, we naturally also have to know the demand function. It can be assumed that the expected demand is a function of the price P plus the expected user cost per would-be parker, $E(TC^{user})/E(D)=AC^{user}$. A statistical equilibrium occurs when the generalized cost $P+AC^{user}$, where AC^{user} is indirectly a function of P, is on a level that calls forth a demand consistent with this particular level of the generalized cost. The pricing relevant cost is, as was argued previously, equal to the product of the expected quantity of demand E(D), and the derivative of AC^{user} with respect to expected output E(Q). Note that it is the output that should be charged, not the demand.

$$PC = E(D) \frac{\partial AC^{user}}{\partial E(Q)} \qquad (29)$$

where

$$AC^{user} = \frac{c\,\left[E(D) - E(Q)\right]}{E(D)} \qquad (30)$$

From the above discussion it is clear that E(Q) is determined by E(D) and the supply of parking space, S.

$$E(Q) = f\,\left[E(D),\; S\right] \qquad (31)$$

Given the supply, S, the derivative of AC^{user} with respect to E(Q) comes to:

$$\frac{\partial AC^{user}}{\partial E(Q)} = \frac{cE(Q)}{\left[E(D)\right]^2}\;\frac{1}{\frac{\partial f}{\partial E(D)}} - \frac{c}{E(D)} \qquad (32)$$

After multiplying by E(D), the pricing-relevant cost is finally obtained:

$$PC = c\left[\frac{1}{\frac{\partial f}{\partial E(D)}\,\frac{E(D)}{E(Q)}} - 1\right] \qquad (33)$$

The pricing-relevant cost is equal to c multiplied by the inverse of the elasticity of E(Q) with respect to E(D), minus 1. To begin with, this elasticity (compare Figure 3:6) is equal to unity, i.e. in a situation of permanent excess supply, but it falls steadily with further increases in E(D). At its extreme it approaches zero, which makes PC very high. Such an extreme situation is, however, very unlikely to be an equilibrium.

To conclude: the content of the optimal price of parking space is of much the same kind as the content of the optimal price of driving a car on the roads. The pricing-relevant cost is the expected cost of a further space being occupied in terms of the time that has to be spent as a result by all other motorists demanding parking space.

3.7 A MEDIUM-RUN APPROACH TO THE PRICING OF SCHEDULED TRANSPORT SERVICES

The distinguishing characteristic of transport vehicle services in connection with price-making is that no given limit of the capacity should be assumed. Under this condition the approach to the pricing-relevant costs will be rather different from the one just discussed. Let us first briefly look at the formation of prices on the competitive markets for full-load transports to get the present issue of pricing of part-load transport services into perspective.

3.7.1 Short-run and medium-run supply of full-load transport

Thanks to the geographical mobility of transport vehicles, each particular market for full-vehicle-load transports - whether by land, sea or air - can on short notice be served by, in principle, any number of "production plants" (=vehicles), if only the reward is high enough. The short-run supply curve in a market for full-loads assumes the familiar rising shape of Figure 3:8 on the assumption that the opportunity costs of individual vehicles differ in varying degrees. Along the horizontal axis of Figure 3:8 the quantity of output is measured in terms of "the number of full-load transports between A and B during a particular period of time". When this period is "next week", an important fact is that some vehicles should happen to be at A or near A at the beginning of next week. For these vehicles the opportunity cost of making the voyage from A to B in the period concerned is comparatively low. Other vehicles at more distant places would require higher compensation to engage in transport between A and B next week. There may of course be other reasons for opportunity cost differences between vehicles; the point I want to make is simply that even in the "spot markets" for single-voyage charters the vehicles themselves, due to their geographical mobility, are to be regarded as variable factors of production. When it comes to time charters, i.e. the long-term hire of vehicles, the main difference is that the opportunity costs of the vehicles constituting the supply are roughly the same. The fact that different vehicles are scattered here and there at each particular moment and consequently will incur different costs for turning up at A, is now unimportant in relation to the large sum of money involved in a time-charter deal. Moreover, new buildings are a potential source of additional supply in the medium-run.

3.7.2 Supply of scheduled transport services

Let us now turn to the markets for part-load transports. Here our first observation is that, in the pure case where each consignment or each passenger requires only a small fraction of the total holding capacity of a vehicle, one type of service predominates in the various markets - *scheduled transport services.*

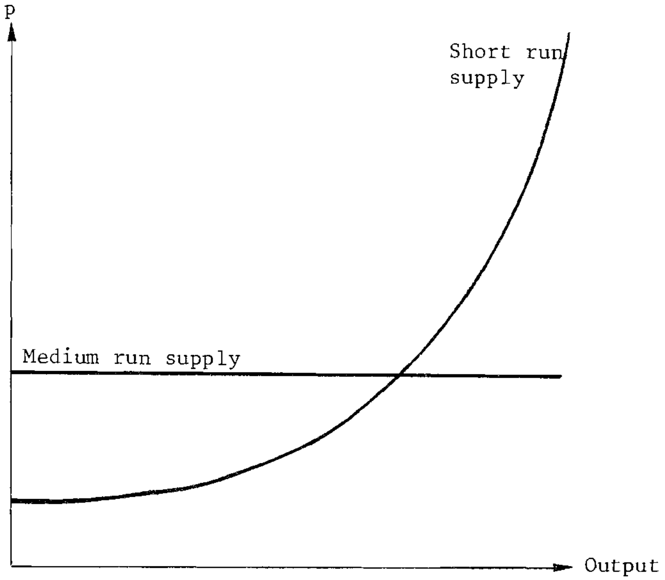


Figure 3:8. Pricing-relevant costs for full-load transports in the short run and in the medium run

A common characteristic of the provision of scheduled transpórt services is that it involves an explicit or implicit obligation on the part of the transport producer to maintain a regular service during an appreciable future period at a fixed price, regardless of what happens. A short-run supply curve of the shape depicted in Figure 3:8 is therefore rarely found in the case of scheduled transport services. The relevant opportunity cost of, for example, a ship for which engagement in a liner service is being contemplated, is not what the ship can earn in the tramp shipping spot market next week, but what can be earned by chartering out the ship for a year or more, or what it will fetch on the second-hand market for ships. In the "self-regulation" practiced in liner shipping through the conference system, rules are drawn up by the conference members whereby notice of a certain term must be given of any changes in freight rates. The period of notice generally lies between six months and a year. This is supposed to be part of the service provided to shippers. It is helpful in planning import and export business to know what the freight costs will be for some time to come. In the case of public passenger transport, the period of notice is still longer, due to the "administrative lag" and/or restrictions imposed by the regulatory authority concerned.

The pricing-relevant supply curve for scheduled transport services corresponds to the medium or long-run supply in the time charter market, as far as the conditions for calculating the vehicle opportunity cost is concerned. On the other hand, two differences exist which makes the shape of the supply (and pricing-relevant cost) of scheduled transport services quite different from that portrayed in Figure 3:8 above. The most apparent and least important difference concerns the disparate "least units" of supply and demand. When the "part-loads" are relatively small - e.g. a single passenger on a bus carrying, say, 60 passengers, or a typical general cargo consignment on a deep-sea liner - there may appear to be a problem of capacity indivisibility. However, on routes where a reasonably large number of vehicles are engaged in a regular service, it seems to me that this problem has been exaggerated out of all proportion.

3.7.3 The average cost of a marginal vehicle

Where a marginal unit of capacity constitutes a *relatively* large capacity addition, a problem of factor indivisibility can arise. What decides whether it is a reasonable approximation to treat the number of least capacity units as a continuous variable, is not whether 5, 25 or 75 passengers can be carried by an additional bus, or if 5 000, 25 000 or 75 000 tons of cargo can be carried by an additional liner ship; the crucial factor is rather the number of vehicles engaged on the route concerned. Treating this number as a continuous variable when calculating the pricing-relevant cost, is the same thing as approximating the marginal cost by the "average cost of the marginal plant". If there is only one plant serving a particular market, the average cost of an additional plant is certainly a poor marginal cost approximation in wide output intervals. If there had originally been 100 plants, it could have been a nearly perfect MC approximation.[1] There is obviously no magic limit in this respect below which it can be deemed illegitimate to disregard factor indivisibility. Suffice it so say that, compared to other industries for which the "average cost of the marginal plant" is a widely accepted approximation of the marginal cost, the production of transport services, where each vehicle corresponds to a separate plant, looks with a few exceptions like quite a suitable area of application. This opinion seems not, however, to be generally approved of in the transport economics profession. There is a number of conceivable reasons for this.

3.7.4 Problems in need of clarification

First, the idea that capacity is a variable input in the pricing-relevant run for scheduled transport services of freight and passengers, is not - strangely enough - generally accepted by leading economists.[2] However, operators of the services would be at a complete loss in the pricing-making, if they were to follow the recommendation that

[1] Compare Samuelson, P., The Monopolistic Revolution, in Kuenna, R.E. (ed.), Monopolistic Competition Theory: Studies in Impact. Wiley, 1967.

[2] Mohring, H., Optimization and Scale Economies in Urban Bus Transportation. American Economic Review, September 1972, and
Turvey, R. & Mohring, H., Optimal Bus Fares. Journal of Transport Economics and Policy, September 1975. They discuss pricing of bus services in a context of a given fleet of buses on a route. However, in a private letter Ralph Turvey has agreed that a medium-run approach is more practical.

prices should be based on the short-run marginal costs, which, as illustrated in Figure 3:9, are steeply rising as the capacity limit corresponding to each individual curve is approached: which short-run curve should be chosen as a match for the demand curve? My answer is that, since pricing and capacity adjustment should be carried out simultaneously, a medium-run curve which, however, takes an entirely different path than the short-run ones, is the natural match for the demand curve. The failure by economists to point this out has led price-makers - in their bewilderment - to resort to the only firm ground in sight - the basically constant level of the direct handling costs in the case of freight transport. This is particularly evident in freight transport by sea, where low-value cargo - thought to be unable to bear anything but very low charges - are accepted for freight rates well below the pricing-relevant costs. The received costing principle, unfortunately, is to assume that the pricing-relevant hauling cost of a part-load consignment is zero, and to set the charging floor at the level of the handling cost.[1] This problem is further discussed in Chapters 7 and 8.

Referring to Figure 3:9, the result is that the "wrong" - far too large - output is produced. As an interesting curiosity it can be mentioned that operators seldom make the mistake of equalizing capacity and expected demand. They are well aware of the fact that this would make the quality of service unacceptably low. A reasonable amount of reserve capacity is provided. Then a situation arises which seemingly justifies the chosen pricing policy; the ships usually sail with spare capacity: so was it not right to accept also low-paying cargo?

[1] In the literature this principle has been explicitly sanctioned by Gedda, S & Koch, A., Principer för fraktsättning vid sjötransport av containers. Skipsfartsøkonomisk Institutt, Bergen 1966, and Sturmey, S.G., Economics and International Liner Services. Journal of Transport Economics and Policy, May 1967.

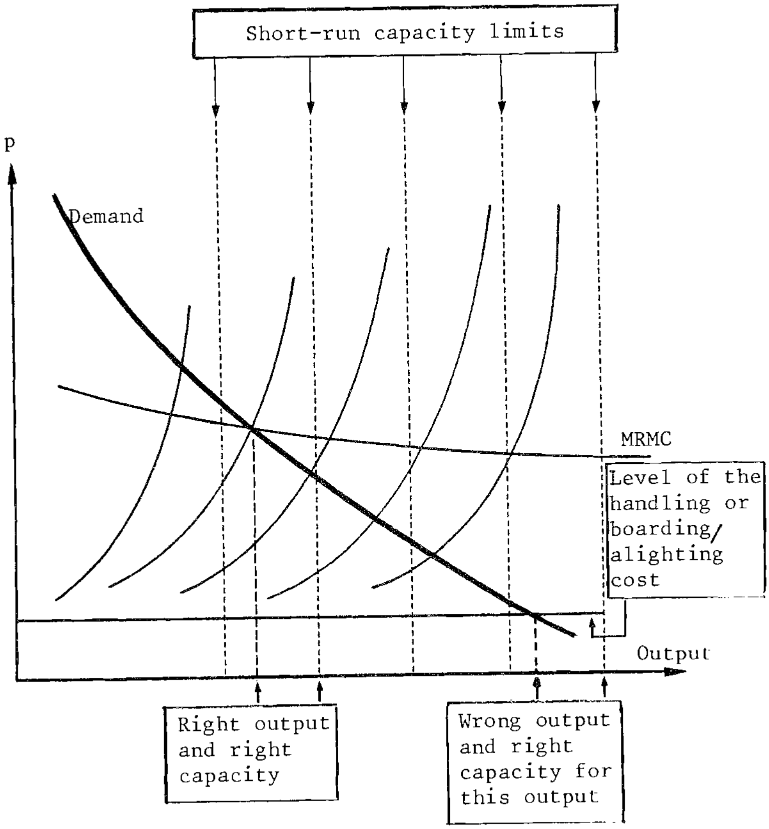


Figure 3:9. Illustrative example of short-run marginal cost curves for different capacities, and a corresponding medium-run marginal cost curve smoothed out by the average-cost-of-a-marginal-plant approximation

In the case of passenger transport, a similar fallacy is fairly common: one example is the idea that bus fares - even in peak-hours - should be based only on the costs imposed on others by the act of boarding and alighting.[1] In Chapter 6 the optimal level and structure of bus fares is extensively discussed. A classical area of capacity

[1] Mohring, H., 1972, op. cit.

limit disregard by short-run marginal cost pricing proponents is, of course, railway passenger transport. A peculiar kind of short-sightedness has been running for a century in this connection. Many advocates of free transit seems to make their claim from a position similar to that of a bystander who happens to be at the Gare du Nord just when the famous Passenger to Calais arrives to catch his train. (In passing, it can be noted that it is also typical that the observer is never at the station when passengers cannot catch the next train to Calais because it is full.) To avoid the bias of short-sightedness, it is helpful to bear in mind that railway transport production and pricing policy is concerned not with questions such as whether 557 or 558 passengers are to be transported from Paris to Calais by the train departing 17^{08} on the 6th of May this year, but with questions of which fare should be charged, and how much capacity should be provided in the afternoon peak-hours from Paris to Calais in the circumstances that the expected rate of demand E(D) is a certain function of the fare plus the user cost, and that a particular probability distribution applies to the demand for each given level of the generalized cost. By the latter approach the capacity indivisibilities - so apparent at a very short distance - disappear more or less.[1] I also think it would be a good thing in many cases to couch both supply and demand in terms of full-vehicle-loads. Few more important decisions require information in finer terms. Then it would be quite natural to focus on medium-run costs. It should be mentioned at the same time that in Chapter 9 about seaport optimization and pricing, some examples are given of situations where factor indivisibility is too outstanding to be ignored.

[1] Compare Walters' observation that with a stochastic demand factor indivisibilities tend to be smoothed out to a large extent. This point was made already in Walters, A.A., The Allocation of Joint Costs. American Economic Review, June 1960.

The position or shape of the medium-run marginal cost curve drawn in Figure 3:9 should not be taken very seriously. One interpretation of this curve which is positively wrong, however, is that it corresponds just to the marginal capacity cost of the transport producer.

When we move into the pricing-relevant medium-run, *other* kinds of user costs than queuing, congestion and crowding costs become highly relevant. The fact that the capacity costs of the producer is brought into the picture lead to the virtual "neutralization" of the user costs primarily dependent on the rate of capacity utlization. In their place, however, a category of user costs, which in the first place depends on the *number* of transport vehicles in the system, become non-neutral with respect to the volume of output. This is a very important aspect, deserving that full attention is paid to it. This will be done in the following chapter.

4 OPTIMAL PRICING OF SCHEDULED TRANSPORT SERVICES

In the present chapter we will discuss optimal pricing of scheduled transport services (which from now on is abbreviated to "sts") in a systems context. In the discussion certain general characteristics of sts are held up, which can be found in one form or another in all modes of transport - both for passengers or freight - and which are the decisive factors for the financial result of optimal pricing of sts.

A strategic limitation of the analysis should be mentioned at once. Vehicles producing sts are charged either a price or an accounting cost for the use of tracks and terminals. If the track and terminal costs appear as charges on the vehicles, it will be difficult to know where to draw the line in an analysis of optimal pricing of sts as soon as these charges can be expected to differ from the pricing-relevant costs of using the tracks and terminals concerned. The present analysis does not go into this problem. It is in keeping with the division of transport price theory suggested here to delegate the responsibility for achieving a total optimum, if one can put it like that, among experts in each of the three compartments - track, terminal, and sts price theory. As regards modes of transport such as railways, where the tracks and terminals are owned by the vehicle-owner, it could be argued that a discussion of optimal pricing of sts should also tackle the problems of the costing or internal pricing of the constituent track and terminal services. These problems will not be examined here. One advantage of our present division of transport price theory is that the particularly complicated pricing problems facing railway companies and other vertically integrated transport enterprises will be kept separate, so that theoretical advances in other areas can be put to maximum use. For example, the problem of rail track costing is more akin to the problem of road pricing, and the problem of railway terminal costing has more in common with the problem of port pricing than with the problem of costing and pricing train services.

4.1 BASIC CHARACTERISTICS OF SCHEDULED TRANSPORT SYSTEMS

A bird's eye view of any system of scheduled transport reveals a basic similarity: regardless of the mode of transport, whether the vehicles run on road, railway track, water, or in the air, the vehicles are moving to and fro along given routes, making halts at more or less regular intervals for loading/unloading at predetermined stops. However, for every mode of transport the size of the vehicles employed can vary within a very wide range. And, roughly speaking, there is a one-to-one trade-off between size and number of vehicles.

From an economic point of view the salient feature is that there are significant *economies of number of vehicles* serving a given geographical area, as well as equally important *economies of vehicle size.* The number economies are manifest in the *user costs*, while the size economies reveal themselves in the *producer costs.* The number economies are connected with the availability of sts and are normally reflected in both the frequency and the density of service.

Given the size and load factor of the vehicles serving a particular route, it is clear that the frequency of service will be roughly proportional to the demand on the route.

A symmetrical source of user cost degression is the positive correlation between the accessibility in space of sts and the density of demand. A study of maps showing the networks of railways, airlines, bus services or roads will reveal clearly that the more sparsely populated a region is, the more coarsely meshed the network will be. Similarly, the density of the shipping lines connecting two continents is positively correlated with the seaborne trade volume per mile of coastline.

Generally, the average feeder transport distance will be increasingly long, as the network becomes more coarsely meshed, or the density of sts less. Thus the costs falling to the users of scheduled transport services for transporting themselves or their shipments to and from the sts "stations", can be expected to be related to the density of demand in much the same way as the user costs stemming from the infrequency of services is related to the density of demand.

Thus, as far as transport policy is concerned, the main point is that the coexistence of significant economies of vehicle size and economies of vehicle number makes the sts sector a pronounced decreasing-cost industry, which means that optimal pricing will result in a financial deficit. This conclusion is particularly remarkable in view of the fact that the charter markets for sea, air and road transport, where basically the same sort of vehicles operate as those engaged in scheduled transport, are commonly regarded as some of the most perfectly competitive markets in existence.

The basis for our discussion here will be the following model of an "abstract mode" of sts. We will continue to take a bird's eye view so that we can grasp the common characteristics of sts without losing ourselves in particular details. The model is a straightforward development of the transport facility model in the previous chapter towards a sts-*system*.

4.2 MODEL OF A SYSTEM OF SCHEDULED TRANSPORT BY AN "ABSTRACT MODE"

The model seeks to describe the main effects on producer and user cost of different sts-system designs. The system is defined by a given geographical area, which is traversed in different directions by a smaller or larger number of transport vehicles plying fixed routes according to pre-announced time-tables. The aggregate output of sts per unit of time, Q, is given by the production function (1) below. A major simplification in the present analysis is that cyclical fluctuations in demand are disregarded. An alternative interpretation of the assumption of a homogeneous unit of output is that it represents a composite unit, including α units of peak output and β units of off-peak output.

The production function is supplemented by a function for total producer costs (2), and a function for total user costs (3).

$$Q = \phi NHV \tag{1}$$

where

$$H = H(X) \tag{1a}$$

and

$$V = V(X, Y, \phi) \tag{1b}$$

$$TC^{prod} = N \cdot Z \tag{2}$$

where

$$Z = g(X) + pY + c(X) \tag{2a}$$

$$TC^{user} = Q \cdot h(\phi, N, V) \tag{3}$$

Q	= transport volume
ϕ	= load factor or occupancy rate
N	= number of vehicles
H	= holding capcity of a vehicle
V	= overall speed of vehicles

X = vehcile design vector

Y = current input vector (fuel, cargo-handling labour, etc.)

Z = total producer costs per vehicle-hour

$g(X)$ = overheads, capital and crew costs per vehicle-hour

p = current input price

$c(X)$ = track and terminal charges

h = total user cost per journey (passenger trip or freight transport)

Total output is the product of the load factor/occupancy rate, the number, the hodling capacity, and the overall speed of the vehicles in the system. By overall speed is to be understood the ratio of the distance travelled per operating year to the total hours of operations, including all the time spent handling passengers or freight at terminals as well as hauling time. The holding capacity H, and the overall speed V are determined by a number of "primary" factors of production. These factors can be divided into two distinct classes: on the one hand there are a number of vehicle design variables, such as the hull dimensions of a ship and engine power, and on the other there are certain current inputs such as fuel. Holding capacity cannot be varied by varying the current inputs. Overall speed is a function of vehicle design, as well as the actual input of fuel and labour for freight handling at terminals. A third determinant of V is the load factor/occupancy rate, ϕ. As can be seen, ϕ appears both as an argument in the V-function, and as a scalar of the production function. Its latter role is self-explanatory: the transport capacity is given by NHV, and in order to produce transport output, the capacity provided has to be utilized by passengers or freight. Were capacity independent of the actual rate of utilization, the output would be strictly proportional to ϕ. This is not true in general, however. It is not that the load factor/occupancy rate affects the hauling speed very much (this effect is in fact almost negligible, at least so far as passenger transport is concerned); it is rather, in the first place, handling time that depends on the load factor. Given the average trip length, the more passengers carried by, let us say, a bus, the longer will be the time spent at stops, other things being equal. In freight transport this effect can be quite strong. This is particularly true as regards break-bulk cargo shipping, where the ships often have to spend more time in port than at sea (unless the load factor is abnormally low).

In the V-function, ϕ is consequently a factor which has a negative effect on output. The total effect of ϕ on Q is, of course, positive. The partial elasticity of Q with respect to ϕ equals 1 plus the partial elasticity of the overall speed with respect to ϕ.

$$E_{Q\phi} = 1 + E_{V\phi} \tag{4}$$

$E_{V\phi}$ is negative, but its absolute value is normally much less than unity.

The total factor costs of the producer of sts are written as the product of N and the cost per vehicle-hour, Z, which consists of the costs of the current inputs, pY, and the costs of all those factors which are fixed in the short run including "overheads", vehicle capital costs, crew costs, as well as track and terminal charges, which - given the type of vehicle - are assumed to be proportional to the length of the time of use of these facilities.

A common procedure in transport service costing is to make a sharp distinction between the traffic operation costs (including vehicle capital costs and crew costs) and those costs which are not directly associated with the traffic operations, i.e. the overhead costs. It is argued here that this sharp distinction is exaggerated, and when it comes to marginal cost calculation, can even be quite misleading. Overhead costs are, of course, a notorious problem in all costing, and they deserve a special discussion also in the present connection. However, in order not to break the flow of the main argument, this discussion will be postponed to the last section.

At this stage the only specification of the user cost function that is required is to note that ϕ, N and V, in the first place, are important determinants of this cost.

The load factor is the principal determinant of the queuing costs. In the case of passenger transport, a high occupancy rate can also cause crowding costs by adding to the cost of travel time.

Given the overall speed, the number of vehicles in the system, N, determines the accessibility of service in time and space. A given fleet of vehicles can be allocated to a large number of separate lines, maximizing the accessibility in space, or to a relatively small number of lines, making the frequency of services rather high. If the former course is chosen, the feeder transport costs will be low, but the actual or potential waiting-time cost will be high. If the latter course is chosen, the opposite will apply. An all-embracing name for the user costs caused by the necessarily imperfect accessibility of sts is "catchment costs".

The appearance of the third argument in the user cost function - the overall speed V - is explained both by the fact that the transport time cost is inversely proportional to V, and that, given the number of vehicles employed on a particular route, the frequency of service is proportional to V.

This concludes the presentation of the general model. In the following pages we will use the model in discussing optimal pricing of sts in greater detail than was possible in Chapter 3. On a basis of this discussion it will be possible to get a rough idea of what the financial result of optimal pricing would be for different categories of sts.

4.3 WHAT DOES THE OPTIMAL PRICE CONSIST OF IN TERMS OF COSTS?

As was pointed out in the previous chapter, optimal pricing of transport services as well as other services where the user cost is a non-negligible social cost component, implies that the pricing-relevant cost, PC, consists of two items - the cost effect on the producer, and the cost effect on the original users of the facility concerned of additional unit of output. The following expression is the most general formulation of the pricing-relevant cost.

$$PC = \frac{dTC^{prod}}{dQ} + Q\,\frac{dAC^{user}}{dQ} \qquad (5)$$

By this formulation no commitment is made regarding which factors of production should be assumed to be variable when it comes to pricing.

Mohring and Turvey discussed optimal bus fares in terms of short-run marginal costs.[1] With the help of the present model, we can now generalize their discussion, and a certain definitional vagueness can be clarified.

4.3.1 The pricing-relevant cost in the short run

Only the right-hand component of (5) is recognized as pricing-relevant in Mohring (1972) as well as in Turvey & Mohring (1975). That is to say, only costs imposed on fellow passengers are included in the optimal bus fares. In the present model this corresponds to assuming that all producer inputs are fixed, allowing only the occupancy rate ϕ to vary. Under this assumption the general expression for the pricing-relevant cost (5) takes this shape:

[1] Mohring, H., 1972, op. cit., and
Turvey, R. & Mohring, H., 1975, op. cit.

$$\mathrm{SRPC} = \frac{Qd\phi \left(\frac{\partial h}{\partial \phi} + \frac{\partial V}{\partial \phi}\,\frac{\partial h}{\partial V}\right)}{d\phi \left(NHV + \phi NH\,\frac{\partial V}{\partial \phi}\right)} = \phi\,\frac{\partial h}{\partial \phi} + \frac{\partial V}{\partial \phi}\,\frac{\partial h}{\partial V} \qquad (6)$$

Two separate imposed costs can be distinguished: one is the crowding and/or queuing cost caused by raising the occupancy rate, represented by the left-hand term, and the other is the imposed boarding/alighting cost, or handling-time cost in the case of freight. As can be seen from (6), the latter effect on the user cost of raising ϕ operates via the negative effect on V of an increase in ϕ.

The former effect appears in one form or another for all modes of sts. On buses or trains, where a considerable number of standing passengers are allowed, average comfort is reduced as soon as it becomes impossible to seat all passengers. As the number of standing passengers grows, the discomfort will be aggrevated; and when the passengers are packed like sardines, conditions are hardly endurable. In the "pure crowding case" it can be assumed, a priori, that only in very exceptional circumstances should buses or trains be so full that some passengers have to be left behind. (It is probably inoptimal to operate at such a high rate of holding-capacity utilization, that some passengers choose of their own free will to wait for the next bus/train in the hope that it will be less crowded.)

There is also a "pure queuing case", where there are no crowding costs. As far as freight transport is concerned, the pure queuing case obtains, provided only that the risk of damage does not increase as the load factor rises. In passenger transport it may be argued that the pure queuing case obtains, too, in all cases where no standing is allowed. However, although everybody has a seat, most travellers by air or railway, for example, probably consider their comfort to be impaired to a greater or lesser extent by a high occupancy rate. The imposed queuing cost appears as an increase in the expected actual or latent waiting time, while the imposed crowding cost appears as an increase in the cost of travel time.

A separate effect of raising ϕ is that, other things being equal, the travel time will increase. Given the transport capacity provided, more passengers or freight means that more time has to be spent on cargo handling or on the boarding/alighting of passengers. In Turvey & Mohring's discussion of optimal bus fares, the boarding/alighting time cost imposed on fellow passengers plays a major role, and in Mohring's earlier paper on urban bus transport it is assumed to be the *only* pricing-relevant cost.

From one point of view it may seem appropriate to reserve the epithet "short-run" for the costs we have just dealt with, which arise when all producer inputs are fixed, and only ϕ is a variable. This is in the closest analogy with the short-run costs of road use or the utilization of other fixed facilities. From another point of view this does not appear to be the most appropriate convention. From the point of view of the actual decision-making in a sts undertaking, the *fixing of the schedule* can be considered the most important dividing-line between questions of short-term operations and longer-term questions of policy. Some of the most typical short-term problems concern keeping to schedules in the face of varying condition, including random fluctuations in demand from one day to another - in other words, how to maintain a *reliable* service.

In this light Turvey & Mohring's short-run marginals cost concept appears as a half-measure only. They assume that total bus-hours are fixed, but not that a fixed schedule is adhered to. In their short-run cost model, an increase in demand will cause the original passengers to lose time because more time will be spent at stops. This must mean either that the bus line schedule is being adjusted in accordance with changes in the expected demand, or that the important desideratum of reliability of service is ignored. For the first alternative (where the schedule is assumed to be a variable), "short run" is not a very fitting designation. If the schedule can be adjusted, one might just as well assume vehicle input to be variable. To my mind the second alternative is an unreaslistic representation of the behaviour of most sts undertakings (including urban bus services).

The very idea of sts is that a published time-table is to be kept to, even under adverse circumstances. The main prerequisite for this is that the schedule is not too tight, but that a time loss incurred at one stage on the round trip can subsequently be made good by speeding up operations. With reference to expression (6) for PC, this means that in the short run the imposed boarding/alighting or handling-time cost is zero. As long as the schedule is fixed, this kind of costs caused by additional traffic will instead be reflected in various "schedule-keeping costs", which is an all-embracing name for a variety of minor cost items which appear when the actual demand is greater than usual.

In liner shipping, for example, when there is an unusually large amount of cargo which is difficult to handle at a particular sailing, the time lost in the handling operation will be regained by reducing both the transitional time and the cruising time of the sailing concerned. By working overtime in port, the ship's idle time can be substantially cut, and the time at sea can be reduced by steaming at maximum speed. This is obviously not done with impunity, and the tighter the schedule is, the higher will be the cost of saving a day in port or at sea. In passenger transport the main opportunity for making up time when boarding/alighting has taken unusually long, is to increase the running speed. In urban bus transport, for example, another frequently used method of making up lost time is to cut the transitional time - to the detriment of the comfort of passengers.

The time per stop can be reduced by abrupt stopping and starting, by rushing the boarding and paying, by ignoring the woman with the pram, by not waiting for the late-comer running to catch the bus, etc.

Conversely, when there is an unusually small amount of cargo, the ship does not normally steam ahead of schedule but keeps to its schedule by going slow and avoiding all overtime work. In passenger transport it is particularly important not to frustrate travellers by early departures.

In terms of the present model this means that an additional constraint should be imposed, namely that the overall speed, V, should be taken as given in the short-run. The existence of some current inputs, Y, then has to be recognized, since otherwise it would not be possible to maintain a fixed schedule in the face of random fluctuations in demand.

It is true that in practice delays occur in all modes of sts, which means that schedules are not made so flexible that it would always be possible to keep to them under any circumstances. It seems wrong in a general sts model, however, to take no account at all of the quality represented by reliability of service.

Assuming the current producer inputs, Y, to be variable, while a fixed schedule is maintained, which means that besides N and X, the overall speed, V, is also to be regarded as given in the model, we will get the following variant of the short-run pricing-relevant cost.

$$\text{SRPC}' = \frac{dYNp + Qd\phi\,\frac{\partial h}{\partial \phi}}{d\phi NH\bar{V}} \qquad (7)$$

The condition that $V(\bar{X}, Y,\phi) = \bar{V}$ means that

$$dY\,\frac{\partial V}{\partial Y} + d\phi\frac{\partial V}{\partial \phi} = 0 \qquad (8)$$

Inserting the resulting value for dY/dϕ in (7) gives the following result for SRPC':

$$\text{SRPC}' = \frac{E_{V\phi}}{E_{VY}} \frac{pY}{\frac{Q}{N}} + \phi \frac{\partial h}{\partial \phi} \tag{9}$$

where $E_{V\phi}$ is the elasticity of V with respect to ϕ, and E_{VY} is the elasticity of V with respect to Y. The ratio of pY to Q/N stands for the current input cost per unit of transport output. It corresponds to the average variable cost (AVC) of the sts producer. The whole left-hand term of (9) represents the schedule-keeping cost component of the pricing-relevant cost. It is consequently found that AVC multiplied by the ratio of $E_{V\phi}$ to E_{VY} plus the imposed crowding and/or queuing cost, $\phi \frac{\partial h}{\partial \phi}$ makes up the optimal price in the short run.

4.3.2 The pricing-relevant cost in the medium run

In actual practice, decisions about the pricing of sts are *not* taken in a short-run (= "fixed-schedule run") perspective. The tariff of freight rates or passenger fares of most sts undertakings is regarded as possessing a degree of "fixedness", at least comparable to that of the vehicle input. Price-making of sts is a "medium-run" concern. Let us see what shape the medium-run pricing-relevant cost assumes in the present model. If we write a completely general expression for PC - without any precondition as to which factors of production are fixed or variable - we get the following result:

$$\text{PC}= \frac{dN(Z+Q\frac{\partial h}{\partial N})+dX\left[N(\frac{\partial g}{\partial X} + \frac{\partial c}{\partial X})+Q\frac{\partial h}{\partial V}\ \frac{\partial V}{\partial X}\right]+ dY(Np+Q\frac{\partial h}{\partial V}\ \frac{\partial V}{\partial Y})+Qd\phi(\frac{\partial h}{\partial \phi} + \frac{\partial h}{\partial V}\ \frac{\partial V}{\partial \phi})}{dN\phi HV\ +dX\phi N(V\frac{\partial H}{\partial X} + H\frac{\partial V}{\partial X}) + dY\phi NH\frac{\partial V}{\partial Y} + d\phi(NHV+\phi NH\frac{\partial V}{\partial \phi})} \tag{10}$$

When it can be assumed that, as a general rule, the efficiency conditions are fulfilled, one property of the pricing-relevant cost in particular will be useful to a theoretical analysis of the contents of the optimal price, namely that PC is independent of which factors of production are assumed to be fixed. As can be seen from (10) above, a rather complicated expression for PC would result if non-zero values were assumed for all the differentials $d\phi$, dN, dX and dY. However, as was shown in the model in the previous chapter, and as is also demonstrated with the present system model in an appendix to this chapter, under the standard efficiency conditions, the pricing-relevant cost is the same, *no matter which factors are fixed or variable*. Even when all producer inputs are fixed, and only the load factor ϕ can be varied to produce different outputs, the result is the same, namely that the pricing-relevant cost is equal to the lagrangian multiplier introduced for the total cost minimization.

$$PC^{opt} = \lambda \tag{11}$$

A particularly useful formulation of the pricing-relevant cost can be obtained by allowing for the input of an additional unit of capacity - for instance, one more carriage per train or one more whole vehicle of the same design as existing ones - while letting the load factor/occupancy rate and/or current input per vehicle remain constant. Treating the number of vehicles engaged in sts production in the system as a continuous variable when calculating the pricing-relevant cost, is the same thing as approximating the marginal cost by the "average cost of the marginal plant".

This cost will look quite different in cases where each vehicle consists of a single unit, or where trains are made up of several carriages. In the second of these two cases, the relevant factor increment is really in the holding capacity per vehicle rather than in the number of vehicles. The two cases will be considered below, starting with the train case.

4.3.2.1 The train case

The holding capacity, H, has been written as a function of an unspecified vehicle design vector, X. Consider now a case where the holding capacity can be written as the product of the holding capacity of each individual carriage S(= "size") and the number of carriages per train L (= "length" of the train).

$$H = SL \tag{12}$$

The running speed will be negatively affected if the train is made longer, unless there is an increase in the current inputs Y (fuel and freight handling labour) to off-set this. It is both realistic and convenient to assume that such an increase is in fact introduced, because we can then assume that the overall speed, V, remains constant. The boarding/alighting time per train is generally speaking independent of the length of the train, since the boarding/alighting capacity (= number of doors) can be assumed to be proportional to the length. The same applies to the handling time in the case of a freight train. (It can be mentioned that in the latter case the loading/unloading of the railcars often takes place while the individual railcars are uncoupled from the engine, so that the hauling capacity of the engine is utilized to the maximum.)

Note that the chosen value of the factor increment dY has no effect on the value of PC (as long as the increment is small), provided that the efficiency conditions are fulfilled. Therefore, no further limitation is imposed by choosing a value of dY that just off-sets the negative effect on V of adding a further carriage. Capital costs, crew costs, and charges per vehicle-hour can be assumed in the train case to be equal to $\underline{a+bL+c(L)}$. That is to say, capital and crew costs are equal to a constant ($\underline{a}$) plus a component ($\underline{bL}$) that is proportional to the length of the train. In addition there are the "track user charges" on the train, $c(L)$, which are presumably a function of train length. Under these assumptions, the production function (13) and the total social costs (14) can be written:

$$Q = \phi NSLV(L, Y) \tag{13}$$

$$TC = N\left[a + bL + c(L) + pY\right] + Q \cdot h(\phi, N, V) \tag{14}$$

On the further condition that the overall speed remains constant, the medium-run pricing-relevant cost will be:

$$MRPC = \frac{dLN(b + \frac{\partial c}{\partial L}) + dY\ Np}{dL\phi NS\bar{V}} \tag{15}$$

The side-condition $V(L, Y) = \bar{V}$ gives us:

$$dL\ \frac{\partial V}{\partial L} + dY\ \frac{\partial V}{\partial Y} = 0 \tag{16}$$

Using (16) to eliminate dY from (15) gives us:

$$MRPC = \frac{b + \frac{\partial c}{\partial L} + p(-\frac{\partial V}{\partial L})/\frac{\partial V}{\partial Y}}{Q/NL} \tag{17}$$

The total revenue (TR) from optimal pricing is obtained by multiplying MRPC by Q:

$$TR = N\left(bL + cE_{cL} + pY\ \frac{-E_{VL}}{E_{VY}}\right) \tag{18}$$

where E_{cL} stands for the elasticity of the track user charge with respect to L, E_{VL} for the (partial) elasticity of the overall speed with respect to L, and E_{VY} for the (partial) elasticity of the overall speed with respect to Y.

Comparing the total revenue to the total producer costs N $\left[a+bL+c(L)+pY\right]$ shows that a contribution to the recovery of such capital costs and crew costs as are unaffected by the length of the train, aN, will be obtained if E_{cL} and/or the ratio $-E_{VL}/E_{VY}$ exceeds unity. If the track user charges were proportional to the train length, E_{CL} would be equal to unity. Nothing is known about this, since no empirical work on rail track congestion costs has to my knowledge been made. Given the engine, the value of $-E_{VL}/E_{VY}$ is initially well below unity. However, as the length of the trains is increased, sooner or later this ratio will probably exceed unity. It is quite conceivable that the continued addition of carriages will at some point result in a value of PC which equals AC^{prod}. The point, however, is that such a long train is unlikely ever to be consistent with the efficiency conditions. Well before the breakeven point is reached, the input of another whole train will probably be called for.

Thus, the result of allowing for additional carriages in the formula for the pricing-relevant cost is somewhat inconclusive. The following single vehicle case is in fact also relevant to the train case, since the optimal train length cannot be determined without taking account of the costs and benefits (of increasing the frequency of services) that would result from increasing the number of trains.

Lastly, a common misunderstanding of the marginal cost of railway transport services as it affects pricing should be mentioned: it is often thought that this cost will rise with the size of the necessary capacity addition - that, in other words, if the number of carriages in a train is given, the marginal cost of another passenger is practically zero; if it is allowed that another carriage may have to be added to the train, the marginal cost will be substantially higher; if a whole new train has to be put in, the marginal cost can be very high indeed, and if the central stations has to be rebuilt in order to accommodate more and longer trains ... etc.[1] The fallacy of this idea of a necessarily escalating marginal cost is that it ignores the user cost savings generated by each successive addition to capacity. Adding another carriage will reduce the users' queuing and crowding costs, and the input of another train will also save time spent waiting. Also, the rebuilding or modernization of railway stations can bring substantial benefits to the users. This, however, is really part of another subject, belonging properly to the theory of optimal pricing of transport terminal services.

4.3.2.2 The single vehicle case

Setting all the differentials dX, dY, and dϕ equal to zero in the general expression for PC according to (10), so that only dN assumes a non-zero value, a user cost effect becomes inevitable. Unlike the train case, the pricing-relevant cost will not this time be restricted to producer cost items only. In the train case it could be so restricted, since it was assumed that dY $\neq$ 0, so as to offset the negative effect on V of an increase in the length of the trains. It would, of course, be theoretically possible to do the same thing here, i.e. to offset the user cost effect of an increment, dN, by simultaneously considering the factor *decrements*, dX and dY, and/or an increment, dϕ, assigning values to these differentials such that the user cost stayed the same. However, this would be both unrealistic and inconvenient. The following attractively simple formulation of the pricing-relevant cost is much to be preferred.

$$\mathrm{MRPC} = \frac{Z + Q\,\frac{\partial h}{\partial N}}{\phi HV} = \frac{ZN}{Q} + N\,\frac{\partial h}{\partial N} \qquad (19)$$

[1] This line of argument is sometimes resorted to in attempts at resolving the paradox of the seemingly negligible cost of carrying "the passenger to Calais".

This result may appear too "simple". I believe that it is of fundamental importance for the issue at stake. Provided that the efficiency conditions are fulfilled, the optimal price of sts is equal to the transport producer costs per unit of output (passenger trip or freight transport), ZN/Q, plus the product of the number of vehicles, N, and the derivative of the average user cost with respect to N. A priori it is clear that the derivative $\partial h/\partial N$ is negative, which means that the optimal price will fall short of the average producer cost by an amount equal to the absolute value of $N\partial h/\partial N$.

4.4 THE FINANCIAL RESULT OF OPTIMAL PRICING

The standard method of predicting the financial result of optimal pricing is to find out whether there are economies of scale. As regards sts production, several studies have been made of the relationship between total *producer* costs and the "size" or turnover of individual sts firms.

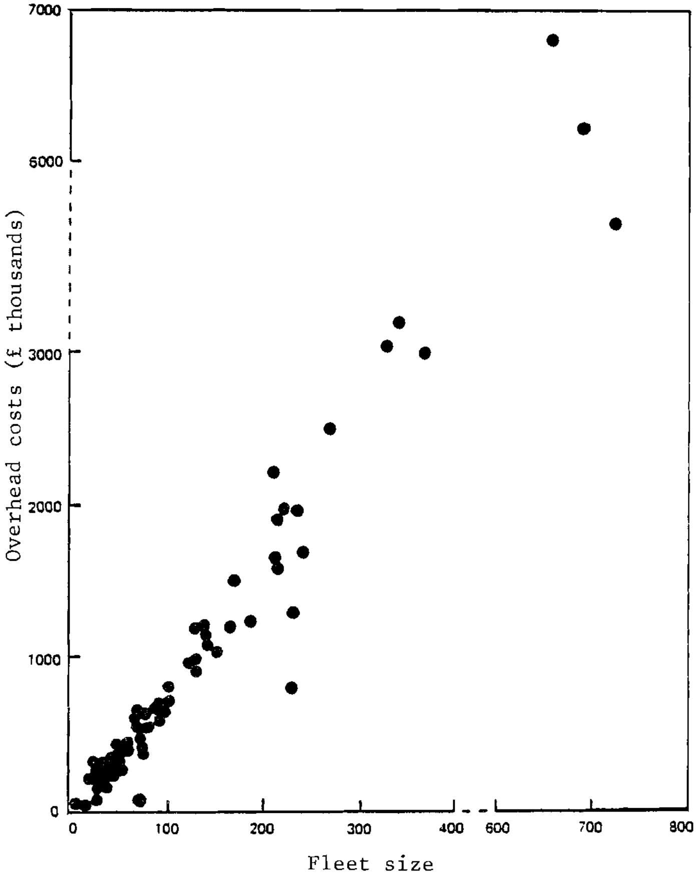
A typical point of departure has been to regard sts firms as markedly "multi-plant firms" on the grounds that each individual transport vehicle represents a production plant. Significant economies of firm size are to be found mainly in industries consisting of single-plant firms. Economies of plant size and economies of firm size are then synonymous. Multi-plant firms are presumably operating with plants of optimal size (at the time of the construction of the plants); in this case the economies of scale in production can be expected to be largely exhaused. Mergers of multi-plant firms can rarely be justified on grounds of production technology; instead, marketing considerations - or simply a wish to restrict competition - provide the rationale.

The pertinent question, consequently, is: what will happen to the vehicles used by a particular sts firm, following an increase in its market share? Studies of mergers of shipping lines, for example, have been rather inconclusive in this respect. A merger may be a means of facilitating the containerization of a particular trade route; in other cases, however, no particular development of existing production seems to result, and the same number of the same kind of ships continue to ply the route in question in much the same way as before. Or it may be possible to make savings in overhead costs by increasing the size of a firm. In sts undertakings the costs which are not directly associated with traffic operations - the costs of general management, administration, etc. - account for an appreciable portion of total costs. However, it is often discovered that any hope of saving on overheads as a result of an agglomeration is illusory. There is ample empirical evidence that, basically, overhead costs develop in proportion to the size of turnover of sts companies. To illustrate this, a chart of the relationship between overhead costs and fleet size for a number of bus companies in Britain is presented in Figure 4:1. The proportionality hypothesis is also borne out in liner shipping. A cross-section analysis of a large number of shipping lines undertaken by Ferguson et al. in 1960 showed that administrative costs amounted to about 10 per cent of gross revenue, regardless of the size of the fleet. A similar result was reported in an investigation of American shipping lines ten years later.[2] Cross-section studies of the *total* costs of (American) railroad companies and bus transport companies do not generally point to any

[1] Ferguson, A.R., et al., The Economic Value of the United States Merchant Marine. The Transportation Center, Northwestern University, 1961.

[2] Devanney III, J.W. et al., Conference rate-making and the West Coast of South America. Commodity Transportation and Economic Development Laboratory. MIT, 1972.

other direction. The evidence by and large supports the proportionality hypothesis.[1]



Source: APPTO Annual Summary of Accounts and Statistical Information, year ended 31st March 1973.

Figure 4:1. Relationship between overhead costs and fleet size for 63 municipal bus undertakings

[1] Borts, G., Increasing Returns in the Railway Industry. Journal of Political Economy, August 1954.
Borts, G., The Estimation of Rail Cost Functions. Econometrica, January 1960.
Friedlaender, A.F., The Dilemma of Freight Transport Regulation. The Brookings Institution, 1969.
(continued on next page)

But, and this is important, findings about the elasticity of the total cost of a transport firm with respect to the output of the firm cannot be used as a basis for assessing the likely financial result of optimal pricing of individual services. And, moreover, since user costs are disregarded, the cost picture is altogether inconclusive.[1]

In the present model it has been shown that a financial deficit will be the result of marginal cost pricing, whether or not total producer costs develop in proportion to the volume of transport. The only question is how large the deficit will be. As can be seen clearly from (19) above, the relationship between user cost and number of vehicles in the system is of crucial importance to the financial result of optimal pricing of sts. Let us look at this relationship a little more closely.

4.4.1 Vehicle-number economies in the user costs

The total user cost per unit of transport can be represented by the sum of (1) the catchment costs, (2) the queuing costs, (3) the transit time cost and, as far as freight is concerned,

(continued from the foregoing page)

Koshal, R.K., Economies of Scale in Bus Transport: Some Indian experience. Journal of Transport Economics and Policy, January 1970, and Koshal, R.K., Economies of Scale in Bus Transport: Some United States experience. Journal of Transport Economics and Policy, January 1972.
Lee, N. & Steedman, N., Economies of Scale in Bus Transport, Journal of Transport Economics and Policy, January 1970.
McDevitt, P.K., Returns to Scale and Technological Change in Urban Mass Transit. The Logistics and Transportation Review, No. 4, Vol. 12, 1976.
Meyer, J. et al., The Economics of Competition in the Transportation Industries. Harvard University Press, 1959.
Zvi Griliches, Cost Allocation in Railroad Regulation. The Bell Journal of Economics and Management Science. Spring 1972.

[1] Compare Vickrey, W., Returns to Scale in Transit: a comment. The Logistics and Transportation Review, Vol. 13, No. 1, 1977. On just four pages Vickrey makes an admirable summary of the theoretical and empirical problems in connection with the issue of economies of scale in public transport.

(4) the direct handling cost.[1] It can be assumed straight away that item (4) is independent of N. The relationship between items (1) and (2) and N can be better understood if we introduce the sts characteristics "density of service" (D), and "frequency of service" (F), and divide the catchment cost into (1.1) feeder transport cost, and (1.2) waiting cost. The concept of service density can be illustrated, for example, by the radial bus services of a centralized city (Figure 1:1), where D equals the number of the "spokes of the wheel" divided by the area of the city.

The feeder transport[2] costs falling to the users of sts are primarily a function of D, while the costs caused by the *in*frequency of service, including waiting time for passengers and storage time for freight, are obviously a function of F. Note the distinction made between waiting and queuing. Waiting occurs on account of infrequency of service, while queuing is caused by occasional excess demand. As for the latter cost, it is clear that, besides being a function of the load factor/occupancy rate, it also depends on F in much the same way as the waiting cost does. The more frequent a particular service, the less inconvenient will it be when excess demand does occasionally arise.

Provided that the routing of the services in the system adheres throughout to the shortest-distance principle,[3] it can be assumed that item (3) - the transit time cost - will be indpendent of N.

[1] That is, the efforts expended in loading the objects of transport on and off vehicles. These efforts are usually very small when it is a question of passengers, but they can be quite appreciable in the case of freight, which may incur high handling charges.

[2] Feeder transport includes all modes of catching the sts concerned, like walking or driving one's own car to or from bus stations, and hauling seaborne freight by road or rail to or from seaports.

[3] This is not necessarily the case for all categories of sts. In the case of freight transport, in particular, it is not unusual that the vehicles providing the scheduled transport services also provide a good deal of the feeder transport work, which means that there are some deviations, sometimes quite substantial, from the shortest-distance route pattern. Consequently, an additional cost-reducing effect of an increase in the density of demand may be that individual services will be increasingly in line with the shortest-distance principle.

To summarize, the total user cost per unit of transport can be written in this way:

$$AC^{user} = f(D) + w(F) + q(\phi, F) + t(\phi, V) + c_h \qquad (20)$$

$f(D)$ = feeder transport cost as a function of the density of service D

$w(F)$ = waiting cost as a function of the frequency of service F

$q(\phi, F)$ = queuing cost as a function of the load factor/occupancy rate ϕ, and F

$t(\phi, V)$ = transit time cost as a function of ϕ, and the overall speed V

c_h = handling charges

The point is, finally, that given the vehicle design and the load factor/occupancy rate, the number of vehicles in the system, N, is by and large proportional to the product of the density of service, D, and frequency of service, F.

$$N = \alpha DF \qquad (21)$$

The user cost term, $N \frac{\partial h}{\partial N}$ of the pricing-relevant cost according to (19) above can now be developed in the following way:

$$N \frac{\partial h}{\partial N} = D \frac{\partial f}{\partial D} + F \left(\frac{\partial f}{\partial F} + \frac{\partial q}{F}\right) \qquad (22)$$

Inserting this in (19), and multiplying by Q to get the total revenue from optimal pricing, we obtain:

$$TR = ZN + fQE_{fD} + wQE_{wF} + qQE_{qF} \qquad (23)$$

It can be seen immediately that a financial deficit would result, falling short of the total producer costs, ZN, by the sum of three products, each made up of a total user cost item, and the corresponding user cost elasticity with respect to D and F respectively.

$$E_{fD} = \frac{\partial f}{\partial D} \cdot \frac{D}{f} \tag{24a}$$

$$E_{wF} = \frac{\partial w}{\partial F} \cdot \frac{F}{w} \tag{24b}$$

$$E_{qF} = \frac{\partial q}{\partial F} \cdot \frac{F}{q} \tag{24c}$$

It should be pointed out that the first term of (22) may be inoperative under certain circumstances. It can happen that for one geographical reason or another, the density of services in the system is fixed even in the long run. For instance, the density of the shipping lines between two continents can be regarded in most realistic cases as having an upper limit, determined by the existence of suitable ports on either coast. In such a case the financial deficit of optimal pricing will be restricted to $Q(wE_{wF}+ qE_{qF})$.

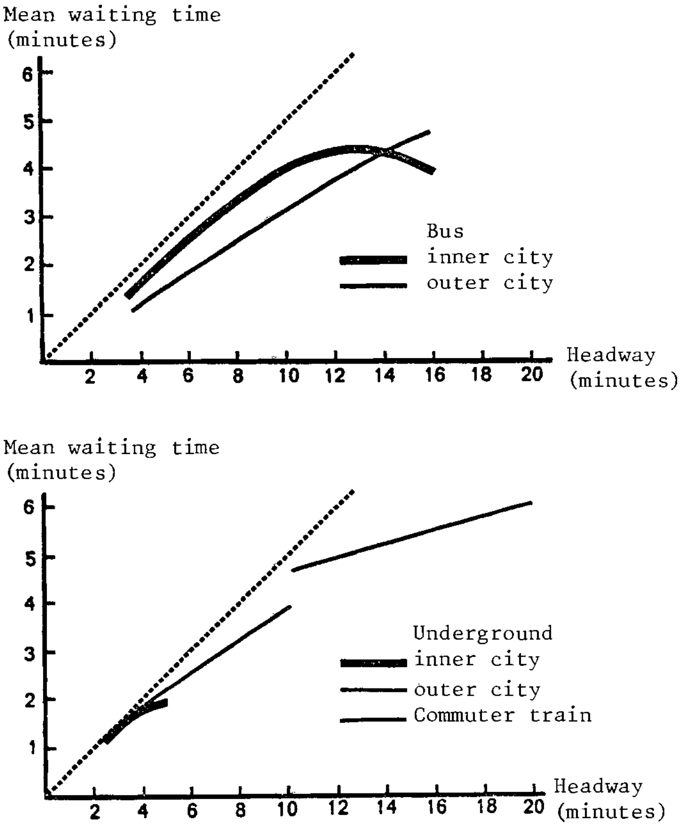
What can be said about the relative order of magnitude (relative to the producer cost) of the financial deficit? It is well known that the feeder transport cost constitutes a substantial proportion of the total system costs for practically all scheduled transport systems. This proportion certainly varies a great deal; it tends, for example, to be higher for local transport systems than for systems of long distance transport. Concerning the value of E_{fD},, it is helpful to develop this elasticity in terms of two other elasticities - the elasticity of the average feeder transport *distance*, d, with respect to D, and the elasticity of the feeder transport *cost* with respect to d. The average feeder transport cost per passenger or freight ton is directly determined by d, which in turn is determined by D.

$$E_{fD} = E_{dD} \cdot E_{fd} \tag{25}$$

If the points of origin and destination are uniformly spread in an area covered by a system of scheduled transport services, E_{dD} tends to equal minus one. The value of E_{fd} in the case of passenger transport is unity, so long as feeder transport is effected by walking or by private car. If puclic transport is used (i.e. the total trip includes both scheduled feeder transport and scheduled trunk line transport) E_{fd} will be less than unity only if feeder transport fares taper off with distance. For freight, feeder transport costs which diminish with distance are the rule rather than the exception. Thus, in brief, E_{fD} is typically equal to minus unity for passenger feeder transport and somewhat less than unity, absolutely speaking, (i.e. between -1 and 0) in the case of freight feeder transport.

The waiting cost w(F) and the queuing cost q(ϕ, F) are rather difficult to calculate, unless they take the form of actual waiting and queuing time at stops or stations. In urban public transport this is what happens; it is easily established that the total waiting cost to passengers accounts for an appreciable proportion of the total system costs in urban public transport, while the total queuing cost (rightly) constitute a marginal item. In an urban transport context a useful bench-mark value for the elasticity of the waiting cost with respect to the frequency of services, E_{wF}, is provided by the observation that if passengers arrive at random to stops or stations, the mean waiting time will be inversely proportional to the frequency of service (compare Figure 4:2 below). The same obviously goes for E_{qF}.

In the case of rural local bus services and interregional scheduled transports, on the other hand, users take the trouble to learn the time-tables, and the waiting cost due to infrequency of connections does not consist of actual waiting time at stops or stations. It may be that the waiting cost is still inversely proportional to the frequency of services; i.e. E_{wF} may still assume a value of minus



Source: TU 71: Trafikundersökningar i Stockholmsregionen hösten 1971, resultatrapport nr 1. Stockholms Läns Landsting, Trafiknämnden.

Figure 4:2. Mean waiting time for bus and train travellers in Stockholm

one (although the waiting cost is certainly less than it would be if passengers were arriving at random to the stops or stations). It is difficult, however, to give a simple rationale for this particular value of E_{wF} outside urban public transport systems.

So far as freight transport is concerned, some idea of the values of w(F) and E_{wF} can be obtained by an approach based on inventory theory. It can be shown, for instance, that under fairly general conditions the required level of safety stock tends to be inversely proportional to the square root of the frequency of service.[1]

Thus, to summarize, in the interval of values of E_{wF} between $-\frac{1}{2}$ and -1, a range of values towards the (absolutely) higher limit is likely to be applicable to passenger transport, and a range of values towards the (absolutely) lower limit is likely to be applicable to freight transport. Moreover, since the value of passengers' time is many times higher than the value of freight time, the absolute values of the total user costs fQ, wQ, and qQ are much greater in passenger transport than in freight transport. In comparison with producer costs this difference between passenger transport and freight transport user costs is presumably somewhat smaller. Nonetheless, the main conclusion is that coexistence of vehicle-size economies in the producer cost, and vehicle-number economies in the user costs can be predicted to make scheduled passenger transport a pronounced decreasing-cost industry, in the sense that optimal pricing will result in a relatively large financial deficit, regardless of the mode of transport. In the case of scheduled freight transport, the picture appears more diffuse. The total revenue from optimal pricing can always be predicted to fall short of the total producer costs. The relatively size of the financial deficit is likely to vary substantially between the sts for low-value and very high-value goods. On average, however, the character of a decreasing-cost industry is much less pronounced than it is in the case of passenger transport.

4.5 THE PROBLEM OF THE OVERHEAD COSTS

The conclusion that marginal cost pricing of sts would result in a substantial financial deficit does not probably come as a surprise to cost accountants brought up in the

[1] Baumol, W.J. & Vinod, H.D., An Inventory Theoretical Model of Freight Transport Demand. Management Science, March 1970.

modern "direct" or "marginal" costing school. If we accept that the vehicles are neither a common nor a fixed cost in the "run" that is relevant to pricing, the result of basing marginal cost pricing on direct costing principles would generally be a financial deficit equal to the total overhead costs.

It is true that the preceding analysis has led us to the conclusion that as a general rule, optimal pricing of sts will result in a financial deficit, which may be of the same order of magnitude as the total overhead costs but which may naturally be either smaller or substantially larger. However, the deficit is due to the number economies in the user costs. It seems appropriate to attempt to refute the alternative explanation, and show that the existence of overhead costs plays no part in the financial deficit of marginal cost pricing.

The problem of the overhead costs has been postponed to this last section. Now it is time to face this issue.

In the present model the total producer cost per hour is assumed to equal the product $N \cdot Z$, where Z represents the total cost per vehicle-hour, including vehicle capital and crew costs, fuel costs, and charges, as well as the overhead costs per vehicle-hour. By this assumption, the *incremental* producer cost per additional vehicle-hour equals the total average producer cost per vehicle-hour, that is Z.

$$IC^{prod} = Z \qquad (26)$$

This is in line with the above-mentioned evidence of proportionally between overhead costs and fleet size. However, one problem which has so far been avoided is that $TC^{prod} = Z \cdot N$ is really a long-run relationship. It can be argued that in the "medium-run" that is relevant to pricing, most overhead factors are fixed, and a different relationship between TC^{prod} and N should be used when discussing optimal pricing. The justification for the procedure adopted in the previous analysis is that although the incremental cost per additional vehicle Z is obtained from a relationship applicable in the long-run, Z still seems to be the best available approximation for the pricing-relevant medium-run value of the incremental cost per additional vehicle-hour. In the following discussion an attempt is made to substantiate the claim that it is a sounder approximation of $MRIC^{prod}$ than the average traffic operation cost per vehicle-hour, which follows from the application of modern costing principles.

In transport service costing it is common to distinguish the "cost of traffic operations" from other costs which are not directly associated with traffic operations and which are usually clumped together under the catch-all name of "overhead costs". The costs of traffic operations are assumed as a matter of course to be proportional to the capacity provided, e.g. to the total vehicle-hours of operation, while overheads are considered "unallocable", i.e. representing factors of production which cannot be allocated in accordance with any measure of capacity or output.

This may seem reasonable in itself. The blunder occurs when it comes to marginal or incremental cost calculations.

It may be helpful to find a name for the costing fallacy, which I shall attempt to put my finger on. In my view it boils down to a rather *disparaging treatment of the overhead costs*. For the sake of brevity, let total overhead costs be represented by the total costs of the salaried staff only, which can be designated $s \cdot M$ (s for "salary", and M for "man-hours"). Let us further denote the traffic operation cost per vehicle-hour, z. An expression of total producer costs that is typical of a costing approach assumes the following appearance:

$$TC^{prod} = sM + zN \qquad (27)$$

On the crucial assumption that z is independent of N, the incremental cost per additional vehicle-hour is apparently equal to z. It is claimed here that this underestimates the relevant incremental cost, and that $Z = z + sM/N$ is a closer approximation of the true value under normal conditions.

The basic problem of the costing approach is that all factors, the cost of which are considered to be "unallocable", are treated as if they were useless. As opposed to other factors of production, the overheads appear as pure "deadweight": they swell the cost side and do not seem to yield any benefits. This is illogical. The input M is bound to appear not only on the debit side but also either in the production function and/or as a cost-reducing factor in the function for the traffic operation cost per vehicle-hour, and/or in the user cost function. In the former capacity M would be a factor of production just like the others given in (1) on page 87. The achievements of at least some categories of administrative staff probably result in i higher output from any given fleet of vehicles. The contributions made by other categories of administrative staff are likely to appear as traffic operation cost *savings* or user cost *savings*. For example, an increase in the staff in garages and repair workshops could help to extend the economic life of the vehicles, and reinforcements in the OR-department could lead to improvements in the routing and scheduling of the vehicles,which could reduce feeder transport costs and the waiting time costs falling on the users. The most typical relationship, in my view, is that the traffic operation cost, z, is a function of M just as much as of various other production factors and design variables.

One of the main staff functions is after all to serve those in the operating line in various ways. In the absence of these services, line operators would perform their duties less efficiently.

If we incorporate the empirical finding that overhead costs tend to be proportional to fleet size, we find that the simplest way of writing the operation costs is the following:

$$N = z\ (\frac{N}{M}) \cdot N \qquad (28)$$

The higher the ratio of fleet size to salaried staff, the higher will be the traffic operation cost per vehicle-hour and vice versa. It can easily be shown that when neither N nor M, but only the ratio of N to M appears as a separate argument in the function for z, the optimal procedure will be to expand the overheads proportionately to the number of vehicles. Total producer costs are written:

$$TC^{prod} = sM + z\ (\frac{N}{M})N \qquad (29)$$

It is further assumed that M does not appear as an argument in either the production function or the user cost function but only in the z-function. Designating the ratio N/M=k, the efficiency condition for the input of M then assumes this simple shape:

$$s - k^2 \frac{\partial z}{\partial k} = 0 \qquad (30)$$

This means that M should be proportional to N, or that the efficient ratio of salaried staff to fleet size is constant and equal to k^{*}. Total differentiation of TC^{prod} gives us:

$$dTC^{prod} = dM(s-k^2 \frac{\partial z}{\partial k}) + dN(z+k \frac{\partial z}{\partial k}) \qquad (31)$$

The incremental producer cost per additional vehicle-hour equals the ratio of dTC^{prod} to dN.

$$IC^{prod} = \frac{dTC^{prod}}{dN} = \frac{dM}{dN}\,(s-k^2\,\frac{\partial z}{\partial k}) + z + k\,\frac{\partial z}{\partial k} \qquad (32)$$

In the medium run, dM=0 by assumption, and only the last two terms are operative.

$$MRIC^{prod} = z(k) + k\,\frac{\partial z}{\partial k} \qquad (33)$$

On the assumption that $\partial z/\partial k$ is positive throughout, it follows that the average traffic operation cost per vehicle-hour, z, is always below the medium-run incremental cost. In the long run dM is different from zero, but under the efficiency condition the first term of (32) comes to zero all the same. $MRIC^{prod} = LRIC^{prod}$ for $k = k^{*}$.

It remains to be demonstrated that these incremental costs also equal the total producer cost per vehicle-hour, Z. Along the expansion path, $dM/dN = 1/k^{*}$. If we multiply both terms within the bracket of (32) by $1/k^{*}$, we find that the second and fourth terms cancel each other out, and an alternative formulation of the long-run incremental cost emerges:

$$LRIC^{prod} = \frac{s}{k^{*}} + z(k^{*}) \qquad (34)$$

That this is equal to Z is clear since $Z = TC^{prod}/N$. Dividing TC^{prod} according to (29) by N gives us the same expression as that given for $LRIC^{prod}$ in (34).

The key point of the argument can also be illustrated diagrammatically as follows:

Suppose that the long-run total producer cost, including overheads, is proportional to the fleet size as illustrated in Figure 4:3. The medium-run total producer cost which is drawn on the assumption that a particular level of the overhead costs, sM_0 in the diagram, is fixed, takes the shape of the "true $MRTC^{prod}$" in Figure 4:3. The linearization of all cost and output relationships in cost accounting will lead to a differently shaped curve, or that given as the "false $MRTC^{prod}$" in Figure 4:3. The starting-point at the vertical axis, and the level at N_0, are the same for the false and the true $MRTC^{prod}$. The main difference will be increasingly pronounced, the further we look to the right of N_0. It cannot be true, as a general rule, that $MRTC^{prod}$ is below $LRTC^{prod}$ beyond N_0. $MRTC^{prod}$ has to be above $LRTC^{prod}$ in its whole range, except in the case of a fleet size equal to N_0, for which the medium-run fixed overhead costs sM_0 are optimal. It has to be a disadvantage that the overheads cannot be varied as the output varies, and this disadvantage will manifest itself in increases in the costs which are variable in the medium run. To assume a linear shape for $MRTC^{prod}$ even beyond N_0 is an (unconsciously) rather disparaging treatment of the resources representing the overhead costs. An undersized administration, meagre repair facilities, etc will certainly cause traffic operations to increase in cost in a way which could be avoided by adequately matching the overheads to the traffic volume.

This point about the shape of the curve of $MRTC^{prod}$ may seem rather too subtle in a total cost context. But it emerges as very important in a marginal cost context. In Figure 4:4 the systematic downward bias of marginal cost calculations, which is inherent in direct costing, is illustrated by portraying the slopes of the three total cost curves from Figure 4:3. The conclusion to be drawn from the argument in this last section can be summarized by referring to Figure 4:4.

Of the three curves shown, $LRIC^{prod}$ is the only one which is empirically based. An estimation of $MRIC^{prod}$ by statistical methods may prove an overwhelming task. In the absence of solid empirical evidence, we have to resort to theoretical reasoning. The point made here is that the disparaging treatment of overheads inherent in direct costing

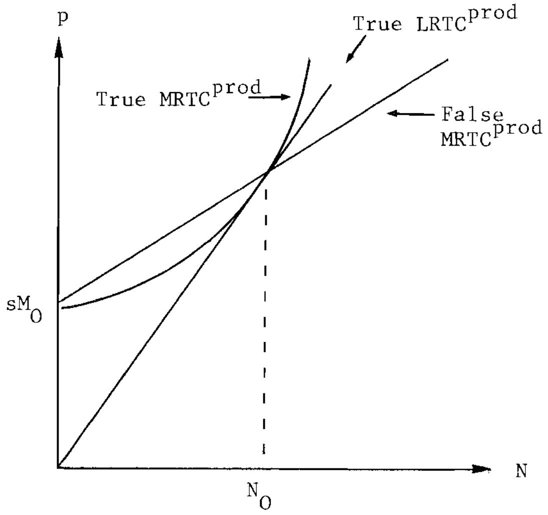


Figure 4:3. A diagrammatic representation of the "disparaging treatment" of the overheads inherent in direct costing

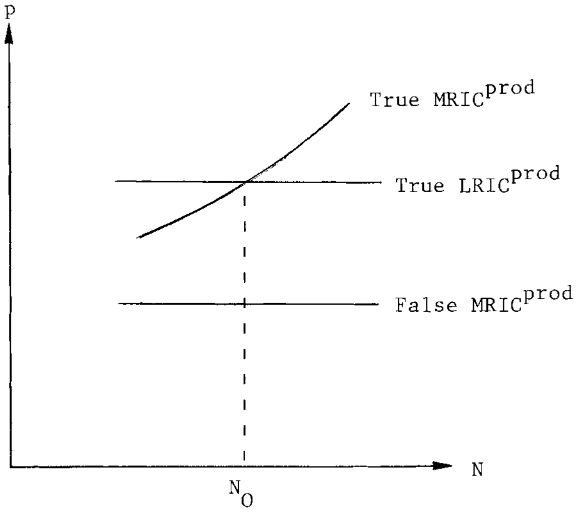


Figure 4:4. The result as regards the marginal cost calculation of the "disparaging treatment" of overheads

appears to lead to a systematic underrating of MRIC^{prod}. It seems preferable to assume that the input of overhead factors, too, can be chosen according to normal economic criteria, and thus that MRIC^{prod} = LRIC^{prod} under the standard efficiency conditions.

APPENDIX: EFFICIENCY CONDITIONS FOR THE DESIGN OF THE TRANSPORT SYSTEM IN THE MODEL

The efficiency conditions for the sts system design imply that the total system costs should be the lowest possible at each level of output. Forming a Lagrangian equation, the efficiency conditions are obtained as follows:

$$\Pi = TC^{prod} + TC^{user} + \lambda(Q-\phi NHV) \qquad \text{(A1)}$$

Taking the partial derivatives of Π with respect to N, X, Y, ϕ and λ gives:

$$\frac{\partial \Pi}{\partial N} = Z + Q\frac{\partial h}{\partial N} - \lambda\phi HV = 0 \qquad \text{(A2)}$$

$$\frac{\partial \Pi}{\partial X} = N\left(\frac{\partial g}{\partial X} + \frac{\partial c}{\partial X}\right) + Q\frac{\partial h}{\partial V}\frac{\partial V}{\partial X} - \lambda\phi N\left(V\frac{\partial H}{\partial X} + H\frac{\partial V}{\partial X}\right) = 0 \qquad \text{(A3)}$$

$$\frac{\partial \Pi}{\partial Y} = Np + Q\frac{\partial h}{\partial V}\frac{\partial V}{\partial Y} - \lambda\phi NH\frac{\partial V}{\partial Y} = 0 \qquad \text{(A4)}$$

$$\frac{\partial \Pi}{\partial \phi} = Q\left(\frac{\partial h}{\partial \phi} + \frac{\partial h}{\partial V}\frac{\partial V}{\partial \phi}\right) - \lambda\left(NHV + \phi NH\frac{\partial V}{\partial \phi}\right) = 0 \qquad \text{(A5)}$$

$$\frac{\partial \Pi}{\partial \lambda} = Q - \phi NHV = 0 \qquad \text{(A6)}$$

Using the efficiency conditions (A2)-(A5) to specify the general expression for the pricing-relevant cost, PC, according to (10) in the preceding analysis, it will be found that the denominator of (10) is equal to the numerator multiplied by $1/\lambda$. Thus, as shown earlier, under the efficiency conditions the optimal price is equal to the Lagrangian multiplier λ, *no matter which factors are fixed or variable.* Setting any number of the factor increments dN, dX, dY or dϕ equal to zero has the same effect on the numerator and on the denominator of (10).

PART II

APPLICATIONS TO URBAN BUS TRANSPORT, LINER SHIPPING, SEAPORTS AND ROADS

PART II

APPLICATIONS TO URBAN BUS TRANSPORT, LINER SHIPPING, SEAPORTS AND ROADS

"The proof of the pudding is in the eating." In this part the proposed analytical approach will be applied to urban bus transport, liner cargo shipping, seaports, and roads. These areas have been chosen so that the services provided by each of the three transport factors - transport vehicles, terminals, and tracks - should be represented. Moreover, in the case of transport vehicles, passenger transport and freight transport exhibit some conspicuous technical differences, which have motivated more detailed treatment of two modes of scheduled transport - urban bus transport and liner cargo shipping. Otherwise it has not been my purpose to cover every mode of transport. As argued in the introduction, an intermodal split of the transport industry is more productive from an economic point of view than the traditional, technologically biased division into modes of transport.

Nevertheless, in the present study I have aimed at width rather than depth. My purpose is to demonstrate the *wide* applicability of the approach. The models used are by no means as sophisticated as they would have to be for an analysis of a particular transport problem in a particular place. The following quotation from Robert Strotz well expresses my own methodological attitude.

> Due to the immense complexities of the problem a "prescientific" approach has to be adopted ... simple little stories, each of which highlights a particular though ubiquitous problem. From each of these we wish to draw a moral, a principle that ought not to be overlooked when a more complex situation is to be faced ... It is unfortunate, and it may seem self-

> depreciating to approach one's work in the manner described above. However, much of economic theory is of this sort. We construct funny little kites, each illustrating some basic principle of aerodynamics, but we don't expect any of these kites to really fly. This may still be good heuristics for the practical designer.
> (Strotz, R., 1963)[1]

The models are constructed so as to highlight the relationships of greatest importance to the issue at stake, while other characteristics and peculiarities thought to be of secondary importance are treated very schematically.

In the first four chapters of Part II scheduled transport services are discussed. In the last chapter in Part I - Chapter 4 - systems optimization and the optimal pricing of an "abstract mode" have been discussed in a bird's eye perspective.

In looking more closely at various systems of sts, different individual features inevitable catch the eye. The peakiness of demand in time and space is a salient feature of urban public transport. Chapters 5 and 6 examine the optimal frequency of service in peak and off-peak periods, and other characteristics of the design of an urban bus line, as well as the level and structure of optimal bus fares.

A major conclusion is that, due to the co-existence of substantial economies of bus size in producer costs and significant economies of vehicle number in user costs, urban bus transport enjoys very significant economies of scale in relation to the extent of the markets concerned.

Is this a characteristic also of modes of scheduled freight transport? In Chapter 7 a model of a liner cargo trade is constructed, with a view to examining the economies of scale in liner shipping. It is found that econo-

[1] Strotz, R., Urban Transportation Parables, p. 128, in Margolis, J. (ed.), The Public Economy of Urban Communities. Resources for the Future, 1965.

mies of scale do exist, but they are relatively insignificant, which means that liner shipping should offer strong competition to individual modes of sea transport even in thin trades, and that the conflict between optimal pricing and self-financing is generally of minor importance. In Chapter 8 it is argued that the pricing problems of scheduled freight transport are associated more with the *structure* of prices than with the price level. Tariffs of scheduled rail, air, and sea freight transport are notorious examples of high-degree price discrimination. It is shown that this results in far-reaching cross-subsidization between commodities moving in one and the same liner trade (intra-tariff), rather than between commodities moving in opposite directions, although this last is a more hotly disputed issue.

In Chapter 9 we move to the area of the transport infrastructure in a discussion of seaport-capacity optimization and pricing. One distinguishing characteristic of the transport infrastructure is, of course, the immobility of its constituent parts once they have been established. The location of ports, roads, etc is to a large extent historically given. In practice, therefore, "systems optimization" means the planning of those extensions of the transport infrastructure - necessarily relatively marginal - that are possible within the next 10, 20 or 30 years. The two chapters dealing with the services supplied by the transport infrastructure consist of partial analyses of a seaport in a given location and of a single link in the road network. As far as seaport problems are concerned, economists have hardly paid any attention to questions of investments or to pricing policy. It is argued here that various kinds of *queuing* models provide the appropriate tools for analysing seaport-capacity determination and optimal pricing, but that the application of operations analysis to seaports tends to forget that the throughput process is not a single-stage activity. A two-stage, multi-channel queuing model is built with a view to pinpointing the peculiar capacity problems, and clarifying some obscure issues concerning the structure of optimal port charges.

Our understanding of road investment and pricing has improved very considerably in the last two decades, to a large measure thanks to the work of economists. The prevailing school of thought emphasizes that pricing is a *separate* problem, which has nothing to do with road investment - in opposition to the "development cost" school of thought. In my view this point has been exaggerated. in Chapter 10 it is shown that by examining the investment problem, a simple explanation can be found for the paradoxical fact that optimal road-congestion tolls (outside urban areas) can be expected to cover only a fraction of the total capital cost of roads.

5 A SIMPLE BUS LINE MODEL FOR THE OPTIMIZATION OF SERVICE FREQUENCY AND BUS SIZE

As is well known, many countries have developed routines for road-investment appraisal based on the generalized cost approach, whereby the cost of road users' time and of car accidents, as well as car operating costs, are treated on an equal footing with road capital and maintenance costs. In other sections of the transport industry, for instance in bus transport which we shall be discussing here, the generalized cost approach to system design is not widely practised. This cannot be because the user costs are more difficult to monetize than road users' time and car accidents.

Bearing this in mind, it would be interesting to know whether public transport is generally efficient from a social point of view, i.e. whether the way present facilities are designed is roughly consistent with generalized cost minimization, or whether there is a systematic deviation from this efficiency condition. It is true that public transport companies are seldom unconstrained "net social-benefit maximizers". Nor are national road administrations - and just as the generalized cost approach has proved useful to them in their investment problems, so should it be equally useful to public transport undertakings *in deciding how facilities should be designed*.

A budget constraint may prevent the achievement of the optimal level of public transport output, but it should not stop each level of output from being produced at the minimum social cost.

The present analysis explores the consequences of including the user costs in designing urban bus services. Since the demand side is left out (that is, customers' reactions to the different prices and qualities are not dealt with), it will not be possible to determine what the optimal level of supply is in a particular situation. Nevertheless, by means of a simple bus line model it is possible to show that social cost minimization results in a pattern of service characteristics which is radically different from most present services, mainly in these respects: given the demand, more buses should be run, and the buses should be much smaller; in particular, the frequency of services should be substantially higher in off-peak; under certain conditions it is even optimal to run the same number of buses in off-peak hours as in peak. The most likely explanation of the "efficiency gap" found, is simply that the user costs are underestimated by operators. Consequently, it is crucial to the points made that the time values used are realistic. The values used in the present analysis have therefore been deliberately chosen on the conservative side, as compared to the findings of some recent econometric studies of travel demand.[1] The main point of departure here is the work of William Vickrey and Herbert Mohring, which can be summarized in the "square root formula" for service frequency optimization.[2]

[1] See e.g. Bruzelius, N., op. cit.

[2] Vickrey, W., Some Implications of Marginal Cost Pricing for Public Utilities. American Economic Review Proc. May 1955, Mohring, H., Optimization and Scale Economies in Urban Bus Transportation. American Economic Review. September 1972.

5.1 OPTIMAL FREQUENCY ACCORDING TO THE "SQUARE-ROOT FORMULA"

As a reaction against the "common sense" proposition of making the frequency of service proportional to total patronage, the so called square-root formula has been suggested by economists following a generalized cost approach. By trading off bus capacity costs against the user cost of waiting time, it is found that, given bus size S, the frequency of service F should be approximately proportional to the square root of the number of passengers carried on a bus line. Using a simple model of a bus line, the square-root formula is derived below.

Consider an urban bus line which starts and ends at a given point. It is of little consequence to the analysis whether it is a radial, diametrical, or circular route. The "production technique" is simply to make round trips by a certain number of buses, each holding a maximum of S passengers.

D = round trip distance of the route

R = total round trip *time*

N = number of buses on the route

F = frequency of service (bus flow per hour)

Z = total bus company cost per bus-hour

B = number of passengers boarding buses per hour

Q = average passenger flow per hour; $Q = B\frac{J}{R}$, where J is the average journey length

t = boarding and alighting time per passenger

T = running time plus total transitional time of the stops per round trip (transitional time is the time of stopping and starting)

c = value of riding time of passengers

v = value of waiting time of passengers

The total round trip time is the sum of T and the total time of boarding and alighting per round trip. The latter time is the product of t and the number of passengers per bus round trip, B/F.

$$R = T + t\frac{B}{F} \qquad (1)$$

The frequency of service F is equal to the product of the density of buses on the route, i.e. number of buses per km, N/D, and the overall speed, D/R.

$$F = \frac{F}{D} \cdot \frac{D}{T+t\frac{B}{F}} = \frac{N-tB}{T} \qquad (2)$$

The total cost components to be considered are (1) bus company costs, (2) passengers' waiting time, and (3) passengers' riding time.

The total bus company cost per hour is equal to the product of Z and N. Assuming that the mean passenger waiting time at stops is half the bus headway, the total waiting-time costs per hour is vB/2F, since the headway is the inverse of the frequency. The total riding time cost per hour for passengers is equal to the product of c, N, and the mean occupancy per bus, which is Q/F. The "balancing factor" in the trade-off between bus company costs and bus passenger costs, i.e. the control variable to be optimized, is the number of buses put on to the route, N. The total social costs considered, TC, can be written as a function of N.

$$TC = ZN + \frac{vBT}{2(N-tB)} + \frac{cNQT}{(N-tB)} \qquad (3)$$

For each given level of patronage of the bus line, the optimal number of buses on route N^{*} is found by setting the derivative of TC with respect to N equal to zero.

$$\frac{\partial TC}{\partial N} = Z - \frac{vBT}{2(N-tB)^2} - \frac{ctBQT}{(N-tB)^2} = 0 \qquad (4)$$

$$N^{*} - tB = \sqrt{\frac{BT(\frac{v}{2} + ctQ)}{Z}} \qquad (5)$$

The optimal frequency of service F^{*} is obtained by dividing N^{*}-tB by T (see (2) above).

$$F^{*} = \sqrt{\frac{B(\frac{v}{2} + ctQ)}{TZ}} \tag{6}$$

As can be seen, F^{*} will develop at a somewhat greater rate than only proportionally to the square root of the total patronage. However, for low to moderately high values of Q, the term v/2 will dominate over the second term within the bracket. When Q is comparatively high, a certain limitation of this simple analysis - namely the implicit assumption of the square-root formula that the total capacity of the buses on the route is always sufficient to meet the demand - will have more serious implications. Bus size, S, has to be introduced as a control variable as well as the number of buses, N. The typical sizes of the buses currently in use (compare Table 5:6 on page 151 for example) are so relatively large that it seems - from a social point of view - that the number of buses could be determined by taking the frequency of services only into account. The resulting total capacity seems almost always to be sufficient. The reason for this peculiarity is simply that existing buses are oversized from a social point of view. The matter of optimal bus size will be discussed later.

Another question concerns the optimal frequency of a given bus line at different times of day. It is clear that Z takes radically different values in peak and off-peak hours.

To be able to discuss the optimal frequency at different times of day, a more penetrating analysis of the structure of the costs of an urban bus company is required. This will be the subject of the following sections. We will then be in a position to extend the square-root formula to cover different peak and off-peak conditions.

5.2 THE STRUCTURE OF URBAN BUS COMPANY COSTS

The purpose of the cost analysis in the following two sections is to derive incremental costs of bus transport capacity in peak and off-peak hours. To achieve this, it is necessary first to make a more general dissection of the structure of the costs of urban bus companies.

It is rather difficult to discover the composition of the total costs of a bus company from the company's accounts. In Table 5:1 the composition of the total costs of 21 Swedish bus companies is presented in accordance with the cost categories used by the Association of Swedish Local Bus Service Operators.

It should be mentioned that the cost of "traffic" is practically synonymous with bus-crew costs, and that "workshops and garages" include the repair and maintenance of buses.

Table 5:1. The composition of the total costs of 21 Swedish urban bus companies in 1975 by "cost centres"

Cost centres	% of total cost
Administration	3.6
Traffic	41.8
Workshops and garages	13.0
Buildings	2.4
Insurance and taxes	3.9
Bus capital costs	20.9
Pensions	7.6
Fuel	6.8
	100

Source: Kollektivtrafik i tätort. Bilaga 1. SOU 1975:48.

A division into the two main categories of (1) traffic operation costs and (2) overhead costs is obtained by defining overheads as consisting of administration costs, pensions, and all capital costs excluding the capital cost of buses. The remaining costs - traffic operation costs - then represent about 80 per cent of the total.

The same figure is reached in a study of the costing of bus operations in Bradford 1972-1973,[1] if the so-called "variable overheads", which include the cleaning, lubricating, repairs and maintenance of vehicles are counted as traffic operation costs.

Crew costs represent about half the total traffic operation costs, and bus capital costs one quarter. The remaining quarter of the traffic operation costs is accounted for by fuel, repair and maintenance, insurance and taxes, and a few other minor items.

A common method of reallocating the total traffic operation costs of a bus company to different lines, is to start by allocating the costs according to the three following factors:

- Total bus-hours operated (H)
- Total bus-kilometres (M)
- Peak vehicle requirement (N)

A general assumption of bus transport costing is that the total traffic operation costs of a particular bus line can be reasonably well represented by the following linear relationship:

$$TC = aH + bM + cN \qquad (7)$$

[1] Travers Morgan, R., & Partners, Costing of Bus Operations, An interim report of the Bradford Bus Study. July 1974.

The main disagreement concerns the overhead costs. Can the bus company overheads be allocated at all between different bus lines, and if this is possible, should the overheads be apportioned according to the peak vehicle requirements of different lines, or should all the three coefficients a, b and c contain an overhead cost element? No definite solution to this problem exists, at least not at present.

When it comes to the further division of the costs of a bus line between peak and off-peak traffic, the direct bus running costs including fuel, lubricants, tyres, and wear and tear cause no particular problem. The running cost per bus-kilometre can be assumed to remain constant, regardless of the time of day a particular kilometre is driven. The remaining traffic operation costs consist of bus "standing costs" and crew costs.

The standing cost of a bus per unit of time includes the bulk of the bus capital cost and insurance,[1] garage costs, licence and other possible taxes. The capital cost of a bus once acquired should be seen as the opportunity cost of not using the bus on another route where the need for bus transport capacity is at its greatest. In the following analysis the convenient assumption will be made that this opportunity cost is equal - as it should be in the normal case - to the use-independent capital cost of a new bus.

The total bus standing costs are generally assumed to be proportional to the "peak vehicle requirements". This assumption is based on the fact that with respect to workday schedules, two principle categories of buses can be distinguished on an urban bus line:

[1] The economic life (in years) of a bus can be assumed to depend to some extent on the amount of use, i.e. bus-kilometres and/or bus-hours operated. Like insurance, part of the capital costs should therefore appear as an element of the running costs.

1. Buses in all-day service - "all-day buses"
2. Extra buses for peak service - "peak-only buses"

The total number of buses is determined by just this "peak vehicle requirement" in the sense that if the off-peak frequency of services is to be increased, no additional buses have to be acquired; a peak-only bus can simply be put into all-day service.

5.2.1 Crew costs

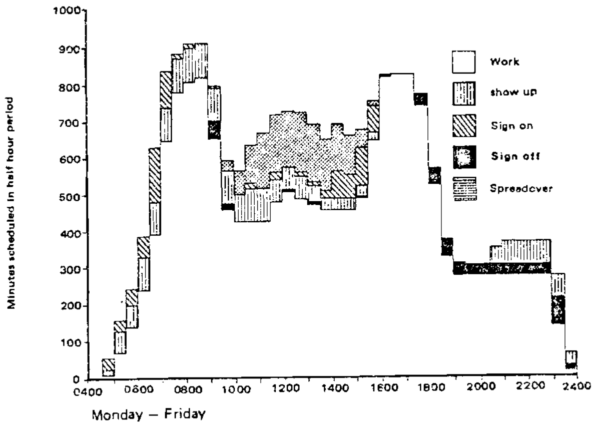
It is normally impossible to employ (and pay) drivers and conductors for the actual peak hours only, let alone just for the peak within the peak. In view of this, an additional problem is that the two diurnal peaks cannot be covered by one straight shift. Second-best solutions are to use double half-day working crews for extra peak services, or to employ "split shifts". In both cases the payable hours are at least twice the work-hours actually required.

In the normal case of demand subject to peak periods, crew scheduling and crew wage-rate differentiation with a view to minimizing crew costs are both very complicated. Actual systems of shiftwork, premium payments for early start, late finish, split shifts, etc are often too complex to be specified in detail in a standard bus operation costing framework. Various statistical costing methods has been used as a short-cut to a solution of this problem. Arthur Andersen & Co. have found that a satisfactory explanation of the total crew cost can be obtained from a linear relationship in which the total bus hours operated in the peak period and the off-peak period respectively, are explanatory variables.[1] It seems that the peak coefficient is about twice the off-peak coefficient.

[1] A method of bus route costing, developed by Andersen, A. & Co., Costing of Bus Operations. The proceedings of a symposium held at the Transport and Road Research Laboratory. Crowthorne, June 26 and 27, 1975.

In the study of the cost structure of bus operations in Bradford by R. Travers Morgan & Partners in 1972-1973, a detailed investigation of crew scheduling and wage costs was carried out (see diagram, Figure 5:1). The result of the study can be summarized in the following main conclusions. The basic all-day level of service is maintained by means of two consecutive straight shifts, while the extra peak service required is supplied by split-shift operations. As a result, the amount of idle time is about the same as the amount of effective working time (bus-hours operated) so far as split shifts are concerned, while for straight shifts the idle time is relatively insignificant. This accords with the aforementioned result that the peak coefficient is about twice the off-peak coefficient of crew hours.

The essential point, however, is that the addition of another bus in peak periods will cause an incremental crew cost equal to the cost of a complete split shift w', and the addition of another bus on all-day service will generate an incremental crew cost equal to 2w, where w is the total cost of a straight shift.



Source: Costing of bus operations. An interim report of the Bradford Bus Study by R. Travers Morgan & Partners, July 1976.

Figure 5:1. Bus-hours operated (unshaded area) and payable crew hours (total area)

5.3 INCREMENTAL COSTS OF PEAK AND OFF-PEAK BUSES

5.3.1 Costs to the bus company

On a basis of the preceding discussion, the incremental cost to a bus company of expanding capacity by the relevant "least unit of capacity" in off-peak and peak periods respectively, is obtained as follows.

Off-peak capacity expansion is effected by putting in another all-day bus making n round trips per day, $n-n_1$ in off-peak and n_1 in peak, while a peak-only bus making n_1 round trips is withdrawn.

Peak capacity expansion is obtained by putting in another peak-only bus, making n_1 round trips.

The incremental costs per day in the two cases, IC_0 and IC_1 are:

$$IC_0 = 2w - w' + r(n-n_1)D, \text{ provided that } N_0 < N_1 \tag{8a}$$

$$IC_1 = w' + s + rn_1 D, \text{ provided that } N_1 \geq N_0 \tag{8b}$$

r = running cost per bus kilometre

n = number of round trips per bus in all-day service

n_1 = number of round trips per peak-only bus

D = round trip distance

s = bus standing cost per day

w = crew cost of straight shift per day

w' = crew cost of split shift per day

N_0 = number of buses in service in off-peak periods

N_1 = number of buses in service in peak periods

It may seem far-fetched, but it will also be useful to consider a situation in which there are more buses in service in off-peak than in peak hours. This odd case would imply that some all-day buses stand idle during peak hours.

In this unlikely situation the incremental cost of another bus for off-peak service would be augmented both by the addition of the standing cost (s) and by the non-subtraction of w' (since there would be no peak-only bus to withdraw). The incremental cost of another bus for peak-only service would consist of the running costs rn_1D only.

$$IC_0' = 2w + s + r(n-n_1)D, \text{ provided that } N_0 \geq N_1 \tag{9a}$$

$$IC_1' = rn_1D, \text{ provided that } N_1 < N_0 \tag{9b}$$

The relative size of IC_0 and IC_1 is a key factor in finding the optimal frequency of service in peak and off-peak times. In the preceding section we saw that the following composition of the total costs of urban bus operations seems to be representative of both Swedish and British conditions:

Crew costs	4a
Bus standing costs	2a
Bus running costs	a or 2a
Overheads I	a or -
Overheads II	2a

Opinions differ about the treatment of some of the costs for repair and maintenance - about whether they should be treated as running costs or as a class of overhead costs.

The main problem in the calculation of the relative size of IC_0 and IC_1 is that the ratio of IC_0/IC_1 is rather sensitive to the way overhead costs are handled. A method for costing bus operations which has been much discussed recently was developed by R. Travers Morgan & Partners (RTM). In a report of the Bradford Bus Study this "marginal costing" approach was used with a view to finding the incremental costs per bus for peak and off-peak services respectively.[1]

[1] Travers Morgan, R. & Partners, Costing of Bus Operations, An interim report of the Bradford Bus Study. July 1974.

It involves the attribution of all overhead costs (overhead I and II) to the peak vehicle requirement. A different procedure has been adopted by the British National Bus Company (NBC), whereby first certain "semi-variable" costs are allocated per bus-hour and, secondly, the "fixed" overheads are divided into two parts: one attributable to the peak vehicle requirement and the other to bus-hours. A third method of incremental cost calculation would be simply to disregard overheads such as "general management" and other administrative costs which are particularly difficult to allocate, while sticking to the NBC way of treating the "semi-variable" costs.

Using these three methods of dealing with overhead costs, the relative sizes of IC_0 and IC_1 are calculated as shown in Table 5:2 below. It is assumed that the ratio of the two peak periods to the total off-peak period is 4:10, and that the crew cost per split shift is about 10 per cent higher than the crew cost per straight shift.

Table 5:2. Incremental costs of off-peak and peak buses, calculated according to three different principles of overhead costs allocation

Cost item	RTM		NBC		Disregard overheads	
	IC_0	IC_1	IC_0	IC_1	IC_0	IC_1
Crew cost	1.8a	2.2a	1.8a	2.2a	1.8a	2.2a
Standing cost	-	2.0a	-	2.0a	-	2.0a
Running cost	.7a	.3a	1.4a	.6a	1.4a	.6a
Overhead cost	-	3.0a	.9a	1.1a	-	-
Total incremental cost	2.5a	7.5a	4.1a	5.9a	3.2a	4.8a

As can be seen, the IC_0/IC_1 ratio comes to 1/3 according to the RTM method, while according to the NBC method it is over 2/3. If we disregard all overheads, we get an incremental cost ratio of 2/3.

A corresponding incremental cost calculation in which it is assumed that $N_0 > N_1$, gives us a IC'_0/IC'_1 ratio of 3/100 according to the RTM method and about 1/10 according to the other two.

5.3.2 Reduction in costs to passengers

The reduction in costs, or incremental benefit to a given number of passengers of putting in another bus is calculated by taking the derivative of the total waiting and riding time costs (see (3) above) of the passengers in peak and off-peak periods respectively, with respect to N_0 and N_1. The resulting incremental benefits, IB_0 and IB_1, are:

$$IB_0 = \frac{E_0B_0T_0(\frac{V}{2} + ctQ_0)}{(N_0 - tB_0)^2} = \frac{E_0B_0(\frac{V}{2} + ctQ_0)}{T_0F_0^2} \qquad (10a)$$

$$IB_1 = \frac{E_1B_1T_1(\frac{V}{2} + ctQ_1)}{(N_1 - tB_1)^2} = \frac{E_1B_1(\frac{V}{2} + ctQ_1)}{T_1F_1^2} \qquad (10b)$$

where

E_0 = extent of the off-peak periods per day
E_1 = " " " peak " " "
$E = E_0 + E_1$

By referring to the list of notations on page 129, the reader will find the remaining symbols self-explanatory.

The crucial point in the present context is that although the rate of peak demand is substantially higher than the rate of off-peak demand, the total number of passengers travelling in off-peak periods can be as large as, or even larger than, the total number of peak passengers. This tends to make the incremental benefits IB_0 and IB_1 of a comparable order of magnitude for each given frequency of service.

The proportion of total peak and off-peak travel naturally varies from one town to another, but some typical values can be given. The peak period per day, E_1, is typically about 4 hours - 2 hours in the morning and 2 hours in the afternoon. The "all-day" level of service may be maintained from, let us say, 6^{00} to 20^{00}. (After that a night service may be run, but this point can be kept separate from the present issue.) In that case $E_0 = 2.5E_1$.

The number of trips per off-peak hour, B_0, can be assumed to be from 1/3 to 1/2 of B_1 (remember that the directional imbalance is normally substantially greater in the peak period). Combining these two assumptions, we find that the ratio E_0B_0/E_1B_1 will range from 0.83 to 1.25. It seems, however, that the range above unity is more common than that below unity. In Bradford, for example, the total number of trips made in the four peak hours is 43 per cent of the total number of trips per day, according to the Bradford Bus Study.

It is clear that the value of the factor inside the brackets in the numerators of (10a) and (10b) is somewhat lower in off-peak than in peak hours, since $Q_0 < Q_1$. It can be shown, however, that this difference is of minor importance to the relative size of the whole value of the factor in the brackets. Not until the passenger flow is unrealistically dense on a particular route, will it make a significant difference. There are four parameters involved: the value of waiting time, v, the value of riding time, c, the boarding and alighting time per passenger, t, and the ratio of Q_0 to Q_1 which is denoted α. A sensitivity analysis is performed, giving "high" and "low" values to each of these parameters. Then the ratio of $(\frac{v}{2} + ct\alpha Q_1)/(\frac{v}{2} + ctQ_1)$ is calculated for different levels of the peak passenger flow in two extreme cases: in one case v and α are assigned the high values, while c and t are assigned the low, while in the other case it is the other way around. The parameter values used are:

	Low value	High value
v	£ 1	£ 2
c	£ 1/3	£ 2/3
t	2.15 sec.	4.25 sec.
α	1/3	1/2

It can be mentioned that the "low value" of t is applicable to two-man bus operation, and the "high value" of t is applicable to one-man operation.[1]

Table 5:3. Low and high extreme values of the ratio $(\frac{v}{2} + ctQ_0)/(\frac{v}{2} + ctQ_1)$

Q_1	Low value	High value
50	0.97	0.99
100	0.95	0.97
150	0.93	0.96
200	0.91	0.95
250	0.89	0.94
500	0.81	0.88
1 000	0.71	0.81

In conclusion, it thus seems that for one and the same frequency of service, the reduction in passenger costs from running another bus is more or less the same in peak and off-peak periods. In Figure 5:2 this condition is taken into account in that the two incremental benefit functions (IB_0 of N_0 and IB_1 of N_1) are given the same shape.

[1] See Cundill, M.A. & Watts, P.F., Bus Boarding and Alighting Times. TRRL Report LR 521. Crowthorne 1973, and
Quarmby, D.A., Effect of Alternative Fares Systems on Operational Efficiency: British experience, in Symposium on public transport fare structure; papers and discussion. TRRL Supplementary Report 37UC. Crowthorne, 1974.

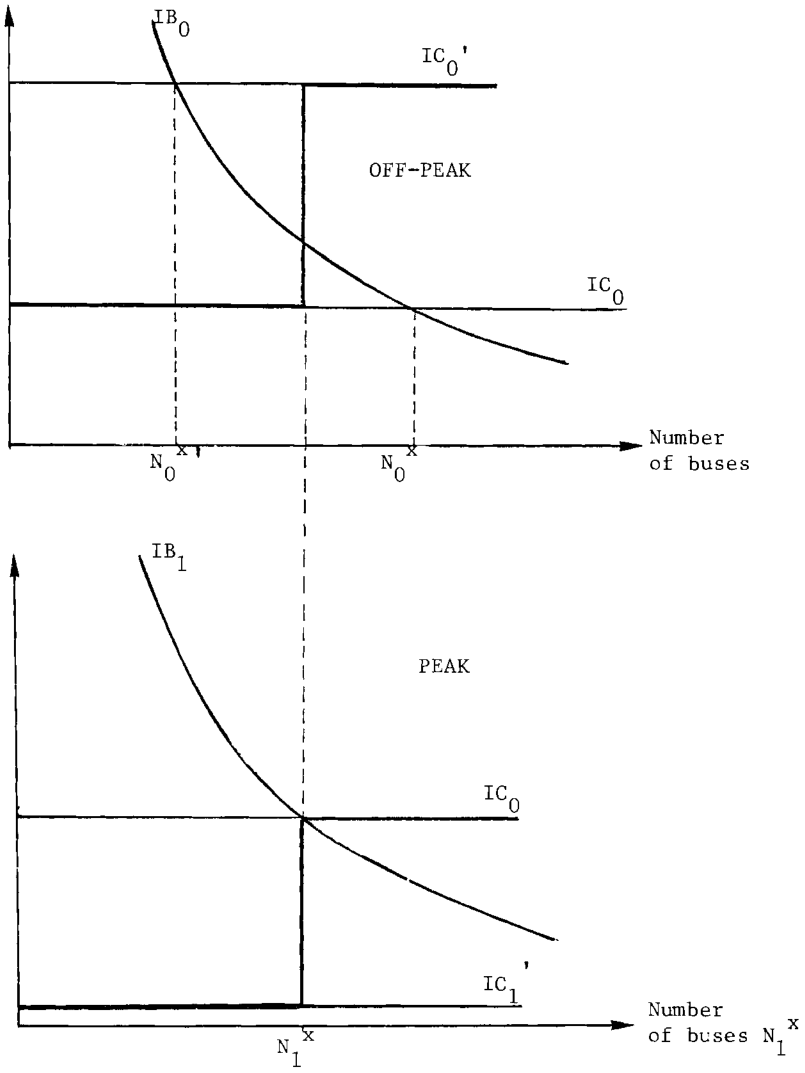


Figure 5:2. Incremental costs and benefits of another bus

5.4 THE CASE FOR RUNNING THE SAME NUMBER OF BUSES IN PEAK AND OFF-PEAK PERIODS

By balancing the incremental benefit and incremental cost of an additional bus in each period, we could presumably obtain the optimal number of buses. In the diagrams of Figure 5:2, IC_0 and IC_1 are put in, and it is assumed that the former is 2/3 of the latter. The point is now that the resulting "optimal" values for the number of buses in off-peak and peak service, N_0^* and N_1^*, do not fulfil the precondition for the relative level of IC_0 and IC_1, that is $N_0 < N_1$. It can be mentioned that if the RTM costing method were applied, the positive difference between N_0^* and N_1^* would be much larger still.

As expected, applying the value of the ratio IC_0'/IC_1' would not resolve the inconsistency. The result would be that $N_0^{*'}$ is much smaller than $N_1^{*'}$, but this is incompatible with the "perverse" precondition for the relative size of IC_0' and IC_1'.

These relationships suggest that a "corner solution" applies. Given the discontinuous character of the incremental costs, the optimality condition is not necessarily that equality of the incremental cost and benefit should apply, but that these inequalities should apply.

$$IC_0 \leq IB_0 \leq IC_0' \qquad (11a)$$

$$IC_1' \leq IB_1 \leq IC_1 \qquad (11b)$$

Under the conditions prevailing in the diagrams in Figure 5:2, i.e. that $N_0^* > N_1^*$ and $N_0^{*'} < N_1^{*'}$, the only alternative which is consistent with (11a) and (11b) is that $N_0^{opt} = N_1^{opt}$, i.e. that the same number of buses is run in peak and off-peak periods. Diagrammatically this result can be illustrated by writing IC_0 and IC_1 as "step-functions", where the discontinuity occurs when equality prevails between N_0 and N_1. The position of each of the corresponding incremental benefit functions is such that the point of intersection occurs at the vertical leg of the step.

Needless to say, exact identity of the peak and off-peak incremental benefit functions is not a necessary condition of the result that $N_0 = N_1$ in optimum. The crucial condition is apparently that $N_0 \geq N_1^*$, which in turn is determined by the relative value of the ratio in peak and off-peak periods of the total waiting-time costs at stops (of passengers on the bus as well as of passengers waiting for the bus) to the incremental cost of another bus.

The fact that the same number of buses should be run all day does not necessarily mean that the frequency of services also has to be the same. It would normally be possible to produce a slightly greater total mileage per hour by a given number of buses in off-peak than in peak periods. Given that $B_0 < B_1$, the frequency of services could be slightly higher in off-peak than in peak hours, thanks to the fact that fewer passengers are boarding/alighting per stop in off-peak periods. If in addition $T_0 < T_1$, this effect would be reinforced. It is not absolutely necessary to let this difference appear in the schedule. It is possible to maintain the same frequency in peak and off-peak by means of a more "relaxed" way of driving in off-peak periods. At such times the crew can afford to be patient with the old lady, can willingly help the mother with the pram, can avoid abrupt stopping and starting etc. The question is whether such an equalization of the schedule is desirable. This is a side-issue in the present connection. Let us only note that a schedule which is uniform all day is easier to remember. This fact is not unimportant. If moreover a really reliable service can be provided, then it is quite possible that the mean waiting time at stops can be reduced below the assumed "half the headway" value.

In conclusion it must be pointed out that $N_0^{opt} = N_1^{opt}$ is not a universally valid prescription. Parameter value constellations that would give us $N_0^* < N_1^*$ are certainly not inconceivable. For example, where the off-peak rate of demand is unusually low in comparison with the peak rate, the frequency of services ought probably to be higher in peak than in off-peak periods. However, a related point to the previous one is that in a situation where N_1^* exceeds N_0^* by a relatively small marging, which is what can be expected in all but the most exceptional cases, there is still a good case for equalizing the peak and off-peak service frequency. As has been mentioned, there are appreciable intrinsic advantages in such an equalization, which should not be sacrificed until N_1^* exceeds N_0^* by a really substantial amount. The simplicity afforded by a standard all-day schedule has considerable value in its own right.

5.5 A MODIFIED VERSION OF THE SQUARE-ROOT FORMULA

A new version of the square-root formula for the optimal frequency can be obtained by the following modified model: on the assumption that not only $N_0 = N_1$, but also that the total round trip time is the same in off-peak and peak periods, so that $F_0 = F_1$, the total social costs per day can be written as follows (compare (3) above):

$$TC = N \cdot IC + \frac{vEB}{2F} + \frac{cEQN}{F} \qquad (12)$$

The frequency of services that can be achieved by N buses is determined by the conditions regarding traffic congestion and the rate of boarding and alighting at peak hours.

$$F = \frac{N - tB_1}{T_1} \qquad (13)$$

Inserting this expression of F in (12), and taking the derivative of TC with respect to N, we obtain:

$$\frac{\partial TC}{\partial N} = IC - \frac{vEBT_1}{2(N-tB_1)^2} - \frac{ctEQB_1T_1}{(N-tB_1)^2} \qquad (14)$$

Setting this equal to zero, and solving for N (observing that $QB_1 = Q_1B$), we get:

$$(N-tB_1)^2 = \frac{EB(\frac{v}{2} + ctQ_1)T_1}{IC} \qquad (15)$$

The optimal frequency is then found by dividing by $T_1{}^2$, and taking the square root of the resulting expression.

$$F^{opt} = \sqrt{\frac{EB(\frac{v}{2} + ctQ_1)}{IC \cdot T_1}} \qquad (16a)$$

That is to say, when the diurnal peaks are taken into account, the original square-root formula (6) holds good, provided only that the demand per hour, B, is interpreted as a weighted average, i.e. $B = (E_0B_0 + E_1B_1)/E$, that the value of the peak flow, Q_1, is used in the formula, and that the incremental cost of an additional all-day bus IC divided by the total extent of a service day E is substituted for the previously used unspecified cost per hour, Z.

A useful variant of this formula can be obtained by expressing B in terms of the peak flow, Q_1, and eliminating the specific route characteristic, T_1. We can introduce the designation β for the ratio of the mean rate of passenger flow of the whole service day, Q, to the peak flow, Q_1.[1] The mean

[1] We have previously used α for the ratio of Q_0/Q_1. It consequently follows that

$$\beta = \frac{\alpha E_0 + E_1}{E}$$

number of passengers, B, can then be written as $\beta Q_1 D/J$. A more general route characteristic than T_1 is the ratio of T_1/D, that is the running time and transitional time *per kilometre*. Designating $T_1/D=h$, the optimal frequency can alternatively be written:

$$F^{opt} = \sqrt{\frac{\beta E Q_1 (\frac{v}{2} + ctQ_1)}{IC \cdot hJ}} \qquad (16b)$$

5.5.1 Example of service frequency optimization

It is interesting to consider a concrete example of the optimal frequency of a bus service. Assuming the route characteristics and "factor prices" given in Table 5:4, the previously derived formula for F^{opt} gives us the following pattern (compare Table 5:5 below):

Table 5:4. Parameter values used

Parameter	Description		Value used
h	Running and transitional time per km		2.8 min
J	Average journey length		3 km
α	Ratio of off-peak to peak passenger flow		0.4
E_0	Extent of off-peak periods		10 h
E_1	Extent of peak periods		4 h
t	Boarding/alighting time per passenger		4.25 sec
c	Value of riding time		50 p
v	Value of waiting time		150 p
IC	Incremental cost per work-day of all-day bus[a]	S=45	£ 56
	-"-	S=60	£ 61.50
	-"-	S=75	£ 67

[a] Source: Commercial Motor Journal. London, 1975.

Even for as moderate a flow as 50 passengers per peak hour, a frequency of services corresponding to a bus every ten minutes should be provided, and a bus every five minutes is right for a mean passenger flow of about 200 passengers per peak hour. A bus every third minute should be running when this figure is doubled. Much higher levels of demand are beyond the range that is interesting. When a frequency of service of 25-30 buses per hour is approached, further increases in the density of demand should be met under normal conditions mainly by increasing the density of bus lines (by making the bus line network more fine-meshed). At any rate, the values of the optimal frequency given in Table 5:5 are well above what is currently supplied by urban bus companies.

Table 5:5. "Optimum" vs. minimum frequency of service

Q_1	S = 45		S = 60		S = 75	
	F^{opt}	F^{min}	F^{opt}	F^{min}	F^{opt}	F^{min}
25	4.4	1.7	4.2	1.3	4.1	1
50	6.3	3.3	6	2.5	5.8	2
75	7.8	5	7.5	3.8	7.2	3
100	9.1	6.7	8.7	5	8.3	4
150	11.4	10	10.9	7.5	10.4	6
200	13.4	13.3	12.8	10	12.2	8
250	15.2	16.7	14.5	12.5	13.9	10
300	16.9	20	16.2	15	15.5	12
400	20.2	26.7	19.3	20	18.5	16
500	23.2	33.3	22	25	21.2	20
600	26.2	40	25	30	23.9	24

It can be observed that optimal frequency is relatively sensitive to bus size. In Table 5:5 three different sizes are compared. As can be seen, the differences in the values of F^{opt} for each level of passenger flow are quite insignificant.

The modified version of the square-root formula for finding optimal frequency assumes, like the original version, that the input of buses in peak as well as in off-peak periods can be determined solely on frequency-of-service grounds. When the most typical bus sizes currently in use for urban services are considered, it seems that this important condition generally holds within the interesting range of demand. We can compare, for example, the figures of F^{opt} with the corresponding figures for the minimum frequency 'of service required on pure capacity grounds.

The capacity requirement is allowed for by the condition that a certain mean occupancy rate must not be exceeded. The mean occupancy rate, ϕ, is defined as the ratio of the mean passenger flow to the product of bus size and service frequency.

Given the stipulated maximum value of the occupancy rate, ϕ_{max}, and the bus size, S, each capacity constraint defines a minimum permissible service frequency, F_0^{min}, and F_1^{min}.

$$F_0^{min} = \frac{Q_0}{\phi_{max} S} \qquad (17a)$$

$$F_1^{min} = \frac{Q_1}{\phi_{max} S} \qquad (17b)$$

What can be assumed about the value of ϕ_{max}? Two characteristics of radial bus lines in particular make the mean occupancy rate comparatively low even at the busiest times. These characteristics are (i) the tidal flow pattern of demand which gives rise to a marked directional imbalance in the short run, and (ii) the spatial peak. Only in the "critical section" on the main haul are the buses likely to be full or nearly full, and the length of the critical section is normally only a fraction of the length of the main haul. In Bradford, for instance, it was found that the mean occupancy rate in the four peak hours was only 0.31. For the present exemplification it is assumed that $\phi_{max} = 1/3$. Three bus sizes have been considered, namely buses carrying 45, 60 and 75 passengers. The first of these figures is rather low by current standards. The middle figure is a typical bus size for British double-deckers. So far as Swedish conditions in larger towns are concerned, this figure is on the low side, despite the exclusive use of one-man operations. An example can be provided by the bus company serving greater Stockholm (SL), where the bus fleet in 1975 was composed as shown in Table 5:6.

Table 5:6. Composition of the SL bus fleet

Bus holding capacity (no. of passengers)	Number of buses	%
11 - 20	5	0.3
21 - 30	0	0
31 - 40	4	0.2
41 - 50	0	0
51 - 60	8	0.5
61 - 70	151	9
71 - 80	1 187	71
81 - 90	215	13
91 - 100	0	0
101 - 110	34	2
121 - 130	71	4

As can be seen in Table 5:5, with buses carrying 45 passengers, the transport capacity resulting from optimizing the service frequency would be sufficient for a peak demand up to a passenger flow of about 200 passengers per hour. But with the typical Stockholm bus size of 75, the peak passenger flow has to rise to 600 before capacity rather than service frequency becomes the determining factor. The fact that no less than 24 buses an hour would be required to meet such a demand, indicates that this level of passenger flow represents an unrealistically dense bus route.

A reflection that suggests itself from these findings is that oversized buses would be unnecessarily wasteful. From a social point of view, optimizing the number of buses is not enough; a further improvement would be to adjust the bus size so that the capacity constraint becomes binding, whatever the level of peak flow happens to be.

It is true that the cost saving achieved by adjusting bus size so that the minimum permissible frequency and the optimal frequency are equal in the whole range, will not be very great. Crew cost is the one dominating cost item, and the cost of the driver will obviously not be influenced by the bus size.[1]

To get an idea of the size of the possible cost savings, let us examine the relationship between standing costs and running costs on the one hand and bus size on the other.

5.6 BUS COSTS AND BUS SIZE

The Commercial Motor Journal (London) regularly publishes tables of the standing costs and running costs of buses in the range from 12 to 86 seats. The data plotted in the two graphs below refer to 1975. As can be seen, a linear relationship seems to apply in both cases. The economies of bus size are particularly pronounced as regards running costs. The running costs per mile does not even double between the smallest and largest bus in the sample. It should be noted that equally significant economies of bus size would appear in the other traffic operation costs, if crew costs were added to the standing costs shown in Figure 5:3b.

[1] In the case of two-man operations, the inclusion of bus size as a variable may under certain circumstances indicate a changeover to one-man operations. In that case the elimination of the cost of the conductor does not, of course, constitute a net saving. Costs in the form of longer times at stops have to be deducted.

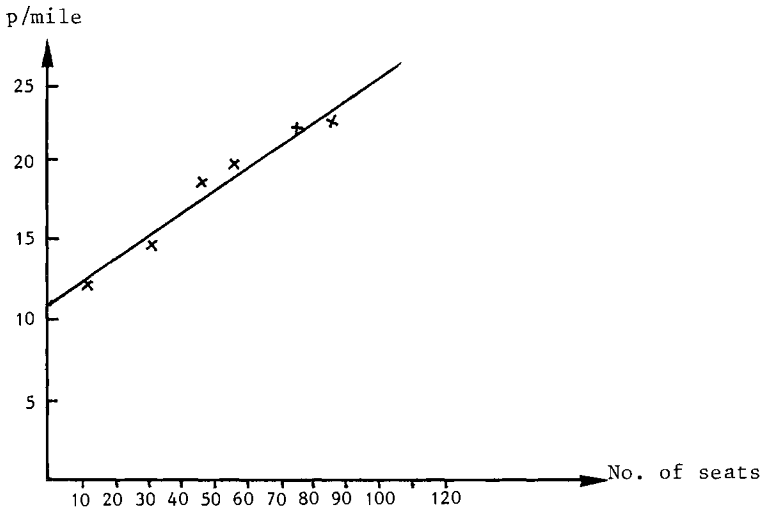


Figure 5:3a. Bus running cost

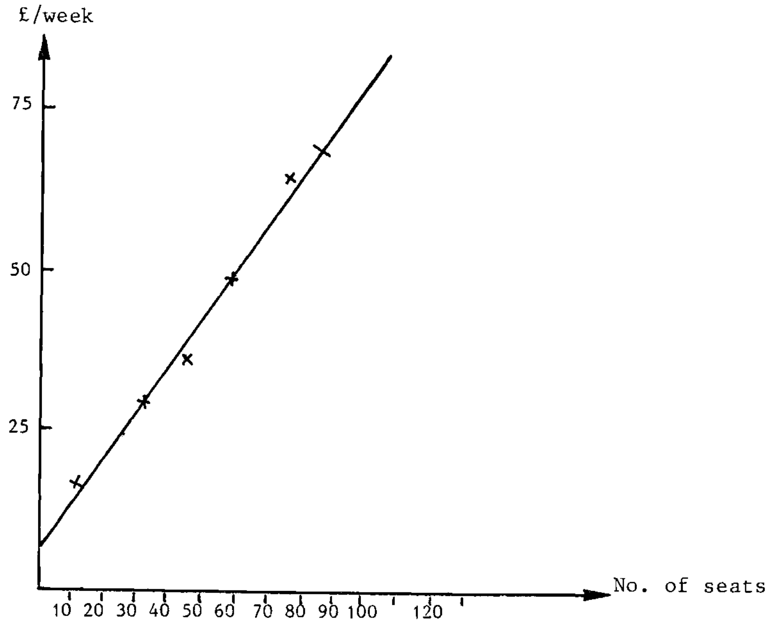


Figure 5:3b. Bus standing cost

By least squares regression, the following relationships were obtained:

$$\text{Running cost per mile (pence)} = 11 + .14S \quad (r^2 = .94) \qquad (18)$$

$$\text{Standing cost per week (£)} = 6.5 + .72S \quad (r^2 = .98) \qquad (19)$$

It thus appears that reducing the bus size by, let say, 50 per cent would save a little over 25 per cent of the running cost and standing cost per bus. Bearing in mind that the crew cost (of one-man operated buses) is not dependent on size, the cost saving from such a reduction in size would be only 12-15 per cent in relation to the total traffic operation costs. However, the adjustment of the holding capacity to the requirement is not a negligible source of efficiency improvement, in view of the present excessive bus sizes.

5.7 OPTIMAL BUS SIZE

When bus size, S, is included as a control variable, the formula for the optimal frequency (16) is no longer strictly relevant, since the incremental cost IC is a function of S. Moreover, the question is whether the case for running the same number of buses in peak and off-peak periods is still as strong, when buses of a more adequate size are being employed.

We will now extend the social cost analysis of a bus line by including bus size as well as the number of buses on the route as a control variable. This will make the analysis somewhat more complicated. Our earlier informal way of reasoning is no longer tenable, and a formal Kuhn-Tucker analysis will replace it.

The total social costs are assumed as before to include the total waiting and riding time costs of the passengers, and the total running, standing, and crew costs of the bus.

As the evidence speaks for a linear relationship between bus size and both the running cost per kilometre ($r = a_r + b_r S$), and the standing cost per unit of time ($s = a_s + b_s S$), it is possible to write the total incremental cost per bus in all-day service and peak-only service respectively, in the following abbreviated form:

$$IC = a + bS, \text{ and } IC_1 = a_1 + b_1 S \tag{20}$$

where

$$a = na_r D + a_s + 2w, \text{ and } b = nb_r D + b_s$$
$$a_1 = n_1 a_r D + a_s + w', \text{ and } b_1 = n_1 b_r D + b_s$$
$$a_0 = (n-n_1)a_r D + 2w - w', \text{ and } b_0 = (n-n_1)b_r D$$

The total social costs are then written:

$$TC = (a+bS)N_0 + (a_1+b_1S)(N_1-N_0) + \frac{\frac{v}{2} E_0B_0T_0 + cN_0E_0Q_0T_0}{N_0 - tB_0} + \tag{21a}$$
$$+ \frac{\frac{v}{2} E_1B_1T_1 + cN_1E_1Q_1T_1}{N_1 - tB_1}$$

As before the following inequality applies:

$$N_1 - N_0 \geq 0 \tag{21b}$$

On the other hand, now that bus size is a control variable, it is assumed that the capacity constraint is always binding in peak periods but may or may not be binding at off-peak periods.

$$\phi_{max} = \frac{Q_1T_1}{S(N_1-tB_1)} \tag{21c}$$

$$\phi_{max} \geq \frac{Q_0T_0}{S(N_0-tB_0)} \tag{21d}$$

As S is a function of N_1 (according to the peak capacity constraint), the total social cost can consequently be regarded as a function of N_0 and N_1 only.

The relevant lagrangian expression takes this shape:

$$\Pi = TC - \lambda(N_1-N_0) - \mu_0 \left[1 - \frac{Q_0T_0(N_1-tB_1)}{Q_1T_1(N_0-tB_0)}\right] \tag{22}$$

Given that both N_0 and N_1 are positive, the Kuhn-Tucker conditions for a minimum can be abbreviated as:

$$\frac{\partial\Pi}{\partial N_0} = a_0 + b_0S - \frac{E_0B_0(\frac{v}{2} + ctQ_0)T_0}{(N_0-tB_0)^2} + \lambda - \mu_0 \frac{Q_0T_0}{S(N_0-tB_0)^2} \tag{23}$$

$$\frac{\partial\Pi}{\partial N_1} = b_0N_0 \frac{\partial S}{\partial N_1} + a_1 + b_1S + b_1N_1 \frac{\partial S}{\partial N_1} - \tag{24}$$

$$- \frac{E_1B_1(\frac{v}{2} + ctQ_1)T_1}{(N_1-tB_1)^2} - \lambda + \mu_0 \frac{Q_0T_0}{S^2(N_0-tB_0)} \frac{\partial S}{\partial N_1}$$

$$\frac{\partial\Pi}{\partial\lambda} = N_1 - N_0 \geq 0 \tag{25}$$

$$\lambda \frac{\partial\Pi}{\partial\lambda} = 0 \tag{26}$$

$$\frac{\partial\Pi}{\partial\mu_0} = 1 - \frac{Q_0T_0(N_1-tB_1)}{Q_1T_1(N_0-tB_0)} \geq 0 \tag{27}$$

$$\mu_0 \frac{\partial \Pi}{\partial \mu_0} = 0 \tag{28}$$

Let us first examine the case of $N_0 < N_1$. In that case $\lambda = 0$. Two sub-cases are then distinguished, depending on whether or not the off-peak capacity constraint is binding. In the sub-case where the off-peak capacity constraint is non-binding, $\mu_0 = 0$. Unfortunately it is not possible in this sub-case to produce an explicit solution for the optimal design. However, a good idea of how the case for running the same number of buses in peak and off-peak periods stands, can be obtained in the following way. From (23) the optimal off-peak frequency can be written:

$$F_0^{\,2} = \frac{(N_0-tB_0)^2}{T_0^{\,2}} = \frac{E_0B_0(\frac{v}{2} + ctQ_0)}{(a_0+b_0S)T_0} \tag{29}$$

This repeats what has previously been shown, (when bus size was taken as given), namely that optimal frequency is found where the incremental cost (a_0+b_0S) is equal to the incremental benefit of an additional bus in off-peak periods. A corresponding expression for the peak period is obtained from (24), as $\partial S/\partial N = S/(N_1-tB_1)$.

$$F_1^{\,2} = \frac{(N_1-tB_1)^2}{T_1^{\,2}} = \frac{E_1B_1(\frac{v}{2} + ctQ_1)}{\left[a_1+b_1S - \frac{b_0N_0+ b_1N_1}{N_1- tB_1} S\right] T_1} \tag{30}$$

The right-hand term in the bracket of the denominator comes close to bS under all realistic conditions. The value of the denominator of (30) is consequently approximately equal to $(a_1-b_0S)T_1$. The optimal frequency of peak service is thus equal to the ratio of the total costs of waiting time and boarding/alighting in peak hours to a_1-b_0S (rather than a_1+b_1S which applies when S is given), while the optimal frequency of off-peak service is equal to the ratio of the total costs of waiting time and boarding/alighting time in off-peak to a_0+b_0S. This means that the previous inconsistency - that the "optimal" frequency should be higher in off-peak than in peak periods - is much less likely to arise. The treatment of the overhead costs is the crucial factor here. If the overheads are allocated per bus-hour or bus-kilometre, or are simply disregarded, the result under normal conditions will be that peak frequency should be somewhat higher than off-peak frequency. On the other hand, if the overheads are allocated per bus (in accordance with the RTM method) to boost a_1, under ordinary conditions the case for running the same number of buses all day would remain.

In the former case it would also have to be checked that the resulting optimal off-peak frequency is higher than the minimum permissible frequency. In other words, the ratio of F_0/F_1 must be higher than α (= Q_0/Q_1); otherwise the capacity constraint is binding even in off-peak periods, and the optimal values of the control variables have to be recalculated on the assumption that $\mu_0 > 0$. The squared ratio of the off-peak to the peak frequency is obtained by dividing (29) by (30). The result, (31), should exceed α^2 if the solution for the optimal off-peak frequency is not to violate the capacity constraint.

$$\frac{E_0B_0(\frac{v}{2} + ctQ_0)T_1(a_1-b_0S)}{E_1B_1(\frac{v}{2} + ctQ_1)T_0(a_0+b_0S)} \geq \alpha^2 \qquad (31)$$

It has been pointed out earlier that α seems to be in the range of 1/3 to 1/2. It appears safe to conclude that the above ratio is normally well over the upper limit of α^2, namely 1/4.

As has been shown, the case for equalizing the service frequency is much weaker when bus size is included as a control variable. Let us nevertheless consider the minority case, where it is still optimal to run the same number of buses in peak and off-peak periods. In this case explicit solutions are easily obtained for the control variables. The optimal bus size should be fairly representative also for the main case. In the main case it may be optimal to make N_0 10-30 per cent lower than N_1. But this should not make much difference as regards optimal bus size. The discussion below should thus be of more general interest than might at first appear, so far as the analysis of the determinants of optimal bus size is concerned.

Setting $N_0 = N_1 = N$, and summing (23) and (24) eliminates λ and gives us:

$$a - \frac{btB_1Q_1T_1}{\phi_{max}(N-tB_1)^2} - \frac{E_0B_0T_0(\frac{v}{2} + ctQ_0)}{(N-tB_0)^2} - \frac{E_1B_1T_1(\frac{v}{2} - ctQ_1)}{(N-tB_1)^2} = 0 \qquad (32)$$

Assuming as before that the frequency of service is equalized throughout the service day when $N_0=N_1$, the following expression is obtained for the optimal all-day frequency of service:

$$F^{opt} = \sqrt{\frac{EB(\frac{v}{2} + ctQ_1) + \frac{btB_1Q_1}{\phi max}}{aT_1}} \qquad (33)$$

Comparing this result with the modified square-root formula for the optimal service frequency (16a), which assumes a given bus size, two differences emerge. In the numerator of (33) the term btB_1Q_1/ϕ_{max} has been added, and in the denominator a has been substituted for IC. Both differences have the effect of raising F^{opt}, compared with the situation in which bus size is given. In most cases the addition to the numerator is not very significant. It has previously been shown that the right-hand term in the bracket, ctQ_1, constitutes only a fraction of the left-hand term, v/2, for moderate flows. A comparison of $EBctQ_1$ and btB_1Q_1/ϕ_{max} shows in turn that the latter term is only a fraction - about one third - of the former. The denominator has decreased since only the part of IC that is independent of size - that is, a - appears in the denominator (IC=a+bS). Thus, relaxing the assumption that S is given results not only in the elimination of all peak excess capacity but also in a higher service frequency at each level of demand (provided that the capacity constraint is non-binding in the original situation). If bus size is fairly high when S is given, the difference between IC and a, namely bS, is relatively significant.

The optimal bus size can be derived direct from (33) and the capacity constraint.

$$S^{opt} = \frac{Q_1}{\phi_{max}F^{opt}} = \frac{Q_1}{\phi_{max}}\sqrt{\frac{aT_1}{EB(\frac{v}{2} + ctQ_1) + \frac{btB_1Q_1}{\phi_{max}}}} \qquad (34a)$$

In a discussion of the determinants of optimal bus size, it is useful to rewrite (34a), expressing B and B_1 in terms of Q_1, journey length, J, and the inverse of the running speed which is assumed to be outside the control of the bus company. We then get:

$$S^{opt} = \frac{1}{\phi_{max}}\sqrt{\frac{ahJQ_1}{\frac{\beta Ev}{2} + tQ_1(\beta Ec + \frac{b}{\phi_{max}})}} \qquad (34b)$$

In view of the fact that the left-hand term completely dominates the denominator, it follows that S^{opt} is roughtly (but not quite) proportional to the square root of Q_1 in the interesting range of demand.[1] Examples of S^{opt} values at different levels of the peak passenger flow are given in Table 5:7 below. The strongest influence on S^{opt} comes, however, from ϕ_{max}. The more unbalanced a route, and the higher the spatial peak of the critical section rises above the mean level of flow, the lower has ϕ_{max} to be set and, as a result, the greater will the optimal bus size be. S^{opt} is nearly inversely proportional to ϕ_{max}. The reason for this relationship is basically that the frequency of service will be increasingly unimportant the fewer the passengers carried on the back-haul and, outside the critical section, on the main haul in relation to maximum flow. Under conditions of extreme spatial "peakiness" of demand, the capacity requirement can usefully be met mainly by increasing bus size. The same goes for the influence of β on S^{opt}. The smaller the number of passengers in off-peak periods compared with the peak flow (which determines the capacity requirement), the higher will the optimal bus size be. Journey length, J, plays a similar role. The farther each passenger travels on average, the smaller will the number of trips (B) be for a given level of passenger flow, (Q), and again the capacity requirement is met by bus size rather than frequency.

The influence of the "factor prices" involved, a, b, c, and v, are what one would expect. Optimal bus size is positively correlated to the size-independent bus cost component, a, and is negatively (although quite weakly) correlated to the proportionality constant, b, of the size-related bus cost component. An increase in the passenger time values, c and v, has the effect of reducing the optimal bus size (and of raising the service frequency to a corresponding degree). It can be noted that if crew wage costs and time values increase roughly parallel with one another in the future, the combined effect on S^{opt} will, other things being equal, be almost nil.

[1] As Q_1 goes to infinity, S^{opt} approaches an upper limit. Under the parameter values assumed in the previous discussion, the upper limit of S^{opt} comes to 95 passengers.

Table 5:7. Social cost comparison of three bus service designs

Q_1	Optimal design		Social costs per passenger		
	F^{opt}	S^{opt}	AC^{opt}	$AC^{opt}_{S=75}$	$AC^{\phi=\phi max}_{S=75}$
25	5	15	39.9	47.2	143.9
50	7.3	21	31.2	36	78.2
75	9	25	27.4	31.1	56.4
100	10.5	29	25.1	28.2	45.4
150	13.1	34	22.5	23.7	35.5
200	15.5	39	20.9	22.8	29
250	17.7	42	19.9	21.4	25.8
300	19.8	46	19.1	20.5	23.6
350	21.8	48	18.5	19.7	22
400	23.7	51	18.1	19.1	20.8
450	25.5	53	17.7	18.7	19.9
500	27.4	55	17.4	18.3	19.2
550	29.2	57	17.1	17.9	18.6
600	31	58	16.9	17.6	18.1

5.8 CONTRASTS WITH "CURRENT" PRACTICE

From the preceding account it is clear that a bus service designed to minimize the total social costs will look radically different from most existing bus service designs, principally in that (i) the frequency of services will be higher in general, and in off-peak periods in particular, and (ii) buses will be much smaller in general and on thin routes in particular.

It is interesting to examine the amount of the social cost savings that the different designs can yield. The above cost comparison includes three alternatives. On the one hand, the two following variants of total social cost minimization are considered.

1. The case just discussed,in which bus size is a control variable as well as the number of buses. For computational reasons, the "minority case" where $N_0 = N_1$ is the one considered. It is not generally optimal to equalize the frequency of services, when bus size is variable. However, the overestimation of the minimum social costs that can result is quite insignificant.
2. The case discussed earlier in which bus size is given, and the number of buses is the sole control variable.

It is not easy to choose an alternative for the comparison , which is representative of current practice, without tending towards caricature. A simple and clear objective has to be assumed, however, so that the calculations can be made. The following alternative representing "current practices" will be considered.

3. Bus size is given, and capacity - number of buses - is adjusted so that the occupancy rate will equal ϕ_{max} both in peak and off-peak periods.

The total social costs per passenger are calculated in each of these three cases.

If we let $N_0 = N_1$, and insert the values for F^{opt} and S^{opt} found in the preceding section into the total social cost expression, (21), we obtain the total social cost per day in optimum. Dividing this by EB gives us the social cost per passenger in optimum.

$$AC^{opt} = \frac{at}{\beta E} + \frac{bhJ}{E\phi_{max}} + chJ + 2\sqrt{\frac{ahJ\left(\frac{v}{2} + (c\beta + \frac{b}{E\phi_{max}})tQ_1\right)}{\beta EQ_1}} \tag{35}$$

The third and fourth terms are clearly dominating. The third term stands for the cost to the passenger of riding time proper. Given the running speed, it is clear that this cost is proportional to the journey's length, J. The fourth term is a conglomerate of the cost to the passengers of waiting time and stop time (on the bus), and bus company costs.

In the second case considered, the total social cost in optimum is obtained by inserting expression (16a) for the optimal frequency (given the bus size) into the total social cost expression (12). Dividing this by the total number of passengers per day gives us the following cost per passenger:

$$AC^{opt}_{given\ S} = \frac{tIC}{\beta E} + chJ + 2\sqrt{\frac{hJ(\frac{v}{2} + ctQ_1)IC}{\beta EQ_1}} \qquad (36)$$

In the third case considered, it is assumed that the bus size is given, and that the number of buses is adjustable so that the occupancy rate ϕ_{max} applies both in peak and off-peak periods. Under these assumptions the cost per passenger assumes this shape:

$$AC^{\phi=\phi_{max}}_{given\ S} = (t + \frac{hJ}{\phi_{max}S})\ \frac{\alpha IC_0 + IC_1}{\beta E} + chJ + ct\phi_{max}S + \frac{v\phi_{max}S}{2\beta Q_1} \qquad (37)$$

As can be seen from the three expressions of average cost, chJ is a term common to them all, that is to say, the cost of riding time proper is independent by assumption of the design of the bus service.

The same parameter values as used before have been inserted into (35), (36), and (37). In the last two cases, where bus size is given, the value of S is set = 75.

As expected, the cost difference is very striking in the low range of demand. The most important improvement on current practices that can be made is apparently to increase the frequency of services, especially in off-peak periods. However, the difference between $AC^{opt}_{S=75}$ and AC^{opt} is also appreciable in the case of low to moderate flows. The optimizing of bus size makes a non-negligible contribution to the total improvement that can be made on current practices.

If bus company costs only are taken into account, we get quite a different cost picture. Under the assumptions made in the third case, the bus company cost is constant throughout. The cost level represented by $PAC_{S=75}$ will be below both the other producer costs in the whole of the interesting range of demand. It is evident that an optimal bus service will be very costly in a *narrow* sense, i.e. if bus company costs only are taken into account. A change from current practice to an optimum situation would reduce bus user costs very considerably, at the expense, however, of an appreciable increase in the costs of the bus company.

This certainly poses a problem in view of the very strained financial situation of most urban bus companies. Who would dare to make the costly quality improvements indicated, and to raise bus fares to the extent required? That this should be done in a case where a budget constraint has to be taken into account, is the inescapable conclusion of the present analysis. A policy of quality improvement and fare increases in fact has the chance of *improving* the financial situation of bus companies, given that the present service design is inefficient. A change that implied the fulfilment of the efficiency condition would increase the total willingness to pay more than it would increase the total cost to the bus company.

A last reflection is that such daring on the part of bus companies would be a very good thing for the whole state of urban transport economics. It would put the time values used by transport economists to the acid test of transport service users' willingness to pay.

Table 5:8. Bus company (producer) costs per passenger

Q_1	PAC^{opt}	$PAC^{opt}_{S=75}$	$PAC^{\phi=\phi_{max}}_{S=75}$
25	17.9	20.6	5.6
50	13.5	15	"
75	11.6	12.5	"
100	10.5	11.1	"
150	9.2	9.4	"
200	8.5	8.4	"
250	7.9	7.7	"
300	7.6	7.2	"
350	7.3	6.9	"
400	7.1	6.6	"
450	7	6.3	"
500	6.8	6.1	"
550	6.7	6	"
600	6.5	5.8	"

APPENDIX: ONE-MAN VERSUS TWO-MAN OPERATIONS

So far a system of one-man operations has been assumed. It may seem to be a retrograde step even to raise the issue of OMO versus TMO. One of the main changes in urban bus operations in recent decades has been the changeover to OMO from TMO, which in countries such as Sweden was completed many years ago. Is this an unquestionably progress, from a social point of view? There is no clearcut answer to this question. There is one unknown factor involved, namely the prerequisite of OMO that any elaborate system of fare differentiation has to be abandoned to avoid unreasonably long boarding times. It is at least possible - although I do not think it likely myself - that, if they were known, the potential allocative benefits of fare differentiation would make a decisive difference to the OMO versus TMO issue. In spite of more or less flat fare rates, the boarding time for OMO buses still seems to be about two seconds longer than for TMO buses.[1] A secondary issue is: could this difference be an argument for TMO rather than OMO under any circumstances?

This issue can be usefully addressed with the aid of the present model for social cost minimization. It is surely more interesting to compare *optimal* OMO and TMO alternatives, than to compare the two systems on an "other things being equal" basis.

Let us first approach the issue in a more qualitative way to get a better understanding of the main factors involved. Later, the model will be used in calculating "OMO-regions" and "TMO-regions" in relation to certain critical factors.

[1] See Cundill & Watts, op. cit.

The advantage of having a conductor as well as a driver is, of course, that he relieves the driver from fare-collecting duty so that the corresponding time can be spent driving. This will reduce the ratio of time at stops to running time to the benefit of both the bus company and the passengers. The question is whether this benefit is worth the heavy price of the conductor's wage cost?

It would be patently ridiculous to have a conductor aboard a long-distance coach (London-Glasgow, for instance) taking up fares, while the bus is running. The majority of the passengers may travel the whole way from London to Glasgow, and the whole fare-collecting job takes perhaps no more than a quarter of an hour of the conductor's time - which would be a very poor rate of utilization of the conductor, assuming that he stays aboard for the whole journey. On the other hand, a guard aboard the London-Glasgow train, which may consist of some ten coaches, is by no means a ridiculous proposition. In this case the guard may be busy during the whole journey, collecting fares and checking tickets.

The point apparently is that for two-man operation to be an interesting alternative at all, it must be possible to achieve a reasonable rate of utilization of the conductor. Two equally important factors for the potential usefulness of a conductor are (i) the size of the total vehicle, which determines the total occupancy, and (ii) the length of journey per passenger, which determines the seat turnover. The ratio of these two factors gives us the number of boarding passengers per vehicle-mile. Since the occupancy rate, ϕ, is defined as the ratio of the passenger flow, Q, to the product of the bus size, S, abd the service frequency, F, and Q equals the product of the total number of boarding passenger of a bus line per hour, B, and the ratio of the length of the journey to the length of the route, J/R, we have:

$$\phi = \frac{Q}{SF} = \frac{BJ}{SFR} \tag{A1}$$

By rearranging the factors we find that the ratio of the occupancy to the journey length is equal to the number of boarding passengers per bus-mile.

$$\frac{\phi S}{J} = \frac{B}{FR} \tag{A2}$$

If this number is high enough to keep the conductor reasonably busy, two-man operation can be viable. Looking at the left-hand side of (A2), we can see that, as was indicated by the previous example, two-man operation is out of the question on a long-distance coach where J assumes a very high value. By reducing J and letting S (and ϕ) remain constant, we can see that sooner or later a situation will obviously be reached in which the number of passengers boarding per bus-mile will be sufficiently high to keep a conductor constantly busy. This is the most favourable situation for TMO as opposed to OMO. (Although it does not necessarily mean that TMO is preferable from a social point of view. This is another matter, concerning relative factor prices.) In certain cases in urban traffic, the mean journey length may be so short that a very high rate of conductor utilization could be obtained, provided that quite large buses were being used.

The main point that can be made in this connection on a basis of the present bus line model, where bus size is a control variable, is that S is strongly dependent on J. Given the passenger flow, a 10 per cent decrease in J will result in a 5 per cent reduction in S^{opt}, if - as we have found - S^{opt} is proportional to the square root of J. This means that the combination of circumstances, in which a high rate of utilization of a possible conductor can be attained, would be fairly rare when the design of urban bus services is determined by social cost minimization. It is quite conceivable that a conductor on a 75-seater double-decker carrying predominately short-distance passengers, pays his way in terms of social cost savings, given this bus size. But it is also quite conceivable that buses of this size are far too large from a social point of view in most places where they are currently used. (For example, halving the bus size and doubling the number of buses, while replacing TMO by OMO, is likely in many cases to reduce the total social costs.

With the aid of the model it is possible to exemplify the conditions under which TMO can still be superior to OMO.

First, it may be interesting to note that an assumption of two-man operation in the model will raise the optimal bus size in the whole passenger flow range. By inserting TMO-values for the crew cost and the boarding/alighting time in the model, we find that the optimal bus size is 15 to 30 per cent larger in the range 50 to 1 000 peak passengers per hour, assuming two-man operation.

The next step is to calculate the value of TC^{opt}/BE under TMO and OMO conditions, and to find the difference between them. Using subscripts 1 and 2 for OMO and TMO respectively, the cost difference per passenger assumes the following shape:

$$\frac{TC_1^{opt}}{BE} - \frac{TC_2^{opt}}{BE} = \frac{a_1 t_1 - a_2 t_2}{\beta E} + \qquad \text{(A3)}$$

$$+ \; 2\sqrt{\frac{vhJ}{E}}\left(\sqrt{\frac{a_1}{2Q} + \frac{a_1 t_1 (c+b')}{v}} - \sqrt{\frac{a_2}{2Q} + \frac{a_2 t_2 (c+b')}{v}}\right)$$

where

$$b' = \frac{b}{\phi \beta E} \qquad \text{(A4)}$$

Consider first the two products a_1t_1 and a_2t_2. If these two products are equal, or if a_1t_1 is smaller than a_2t_2, TMO will not be viable under any circumstances, because the first and the second terms of (A3) are both necessarily negative.

In other words: the whole issue would be settled from the outset, if the percentage increase in the size-independent traffic operation cost caused by adding a conductor were greater than the percentage reduction gained in boarding/alighting times. The parameter values assumed in the present study indicate, however, that $a_1t_1 > a_2t_2$. TMO cannot consequently be ruled out without some further inquiry.

Inserting the values previously used for all the other parameters, we find that the OMO-TMO cost difference is negative up to a peak flow level of a little more than 500 passengers per hour.

At this level social cost indifference obtains between maintaining a service frequency of 30 OMO buses per hour, each carrying about 50 passengers, and a service frequency of 24 TMO buses per hour, each carrying 62 passengers.

The sensitivity of this result to changes in the assumed parameter values has been tested. It turns out that the most critical value is constituted by the ratio (c+b')/v. So far it has been assumed that c=50, b'=13.5, and v=150 pence per hour. The value of waiting time has been found in several investigations to be at least three times greater than the values of riding time. It can be argued that this ratio is not quite relevant in the present context, because the "riding time" involved constitutes in fact *stop time*, i.e. passenger time consumed while the bus stands still for boarding/alighting. It is not unlikely that such time is perceived as much more tedious than riding time proper, and that the value of stop time is close to the value of time spent waiting for the bus. As is seen in Table A1, raising the value of c (that is, "c_{stop}" as distinct from "c_{ride}") to the level of v, would tip the balance in favour of TMO in quite a substantial range. The entries of Table A1 represent the critical values of the peak flow, implying that at flow levels below a given value, OMO is preferable, while TMO is preferable above the critical value. As well as the ratio c/v, the mean journey length, J, also varies. It is a bit surprising that this parameter did not prove more important to the issue at stake. This is because, as has been mentioned, the optimal bus size is strongly (positively) correlated with J.

Table A1. OMO versus TMO: critical values of the peak flow

J $\frac{c}{v}$	1	2	3	5	10	15
1/4	472	571	626	687	757	790
1/3	405	470	525	571	623	648
1/2	323	377	406	436	471	487
3/4	243	277	295	313	334	344
1	200	225	237	250	265	272

6 THE LEVEL AND STRUCTURE OF OPTIMAL BUS FARES

Urban public transport seems to be becoming increasingly heavily subsidized in many countries. Is this a problem? It is no doubt a financial problem for many towns and cities. The question is, is it a problem also from the limited point of view of resource allocation? It has often been argued that on second-best grounds, public transport prices should be below the social marginal cost, since this is the position as regards private car commuting. However, there does not seem to be a consensus of opinion as to whether or not such a harmonious state of affairs does actually exist, simply because the social cost of public transport are largely unknown. To my knowledge the only serious attempt to tackle this problem both theoretically and empirically was made by Herbert Mohring in 1972 in a paper on urban bus transport.[1] Mohring's paper and a later joint paper by Mohring and Ralph Turvey[2] provide the point of departure of the present chapter in which the purpose is to operationalize the theory of optimal bus fares and to calculate optimal fares under different conditions.

[1] Mohring, H., Optimization and Scale Economies in Urban Bus Transportation. American Economic Review, September 1972.

[2] Turvey, R. & Mohring, H. (1975), op. cit.

6.1 PREVIOUS WORK AND THE MAIN IDEA OF THE PRESENT APPROACH

In the 1972 paper Mohring calculates optimal bus fares by optimizing a bus line model. He finds that under greatly differing circumstances, the social marginal cost pricing of urban bus transport would require very substantial subsidization of bus companies. However, one omission in Mohring's bus line model throws too much doubt on the result to be acceptable: *The holding capacity of buses never enters the picture.* The social cost analysis is pursued as though bus holding capacity were infinite: crowding and queuing are disregarded. The bus fares calculated correspond only to the costs caused by the act of boarding and alighting by passengers. The main oddity to result from this disregard of the capacity constraint is that peak and off-peak fares appear to be almost the same, under normal conditions as regards the spacing of stops.[1] This is counter-intuitive in view of the oft-mentioned fact that the marginal buses are strictly required to carry only one load of peak passengers twice a day.

In the later joint paper (by Mohring together with Ralph Turvey) the queuing costs caused by a limited bus-holding capacity are recognized as being relevant to pricing. In this second paper there is no empirical analysis of the structure of optimal bus fares, and no comment on the validity of the empirical results of the 1972 article.

It is true that the queuing and crowding costs pose great empirical problems. For the time being the only solution seems to be to accept the current practice of fixing a "standard" with respect to queuing and crowding conditions in peak hours, i.e. to approximate a continuously rising unit cost function by a ⌋-formed one.

The present approach introduces a "maximum tolerable bus occupancy rate". In view of the stochastic nature of

[1] See e.g. Figure 5 on page 601 of Mohring (1972).

bus transport demand (and to a less extent supply), the practical capacity corresponding to the vertical leg of the synthetic cost function should be set some way on this side of maximum capacity. It seems irrelevant to the pricing issue to calculate bus capacity scarcity values, which would be the procedure if a short-run approach were chosen. It seems to me that bus capacity should be regarded as a variable factor of production, when it comes to determining bus fares. The number of buses put in on a particular route is normally adaptable on far shorter notice than the tariff of bus fares. The "medium-run" approach adopted in this study was not used by Mohring (1972) or Turvey & Mohring (1975), who instead discussed optimal bus fares in terms of the short-run marginal costs.

There is an inevitable *indivisibility problem* in the analysis of bus transport marginal cost, which is plain to see in the medium run but which may seem to be avoided in the short run. In their 1975 article Turvey & Mohring define their problem as being "to find out what marginal cost pricing of bus services consists of". They immediately point out "one apparently obvious answer which is wrong", namely, "that if a bus run costs X and carries Y passengers, and if all passengers travel the same distance, the marginal cost per passenger and hence the optimal fare is X/Y". Why is this a wrong answer? Because "this ignores the point that X is a joint cost", according to Turvey & Mohring. Later on the same point is made in this wording: "If we wish instead to look at costs when the system size is adjustable, i.e. to examine long-run rather than short-run marginal costs, we immediately meet with the difficulty that the system cannot be adjusted to provide for just one passenger more or less."[1]

[1] Turvey & Mohring, op. cit., p. 1 and pp. 3-4.

I do not think that the indivisibility problem is normally very serious in urban bus transport. The peak vehicle requirement on a main route is now typically in the range of 10-15 buses, and for an efficient service in respect of frequency and bus size, this range is likely to be substantially higher. In this situation one more bus will represent a sufficiently small addition to capacity to make the "average cost of the marginal plant" a reasonable approximation to the marginal cost.

In my view the reason why the "obvious answer" quoted is wrong, is not the indivisibility of the bus, but the fact that the answer ignores the benefits to the original passengers of the addition of one more bus. The addition of another bus to meet an increase in demand saves waiting time for the original passengers, and this benefit is generally very important in relation to the incremental cost. If it is taken into due account, a sound theory of optimal bus fares can be built on the basis of the average-cost-of-the-marginal-plant approximation.

The main idea is straightforward. When it comes to the actual calculation of the average cost of a marginal unit of capacity in urban bus transport, some problems arise. First, what is the relevant marginal unit of capacity, and what is its cost? It can be assumed that a complete round trip is the least unit of capacity. Bus round trips are far from homogeneous units in terms of incremental costs, however. The input of a bus round trip will generate very different costs for the bus company depending on the time of day when it is made. Further, it cannot be assumed a priori that a marginal bus put in to meet an increase in demand corresponding to the capacity provided by one round trip will in fact only make one round trip. Having committed a bus for service on a particular route, about half the total bus cost for the day is incurred as soon as the bus sets out on its first trip. So for how long should a marginal bus be kept in service, and how will the answer to this question affect the "average social cost of a marginal bus"?

6.2 PRICING-RELEVANT COSTS

The main concern of the present discussion is to discover the level of the optimal fare for the passengers who cause the "peak vehicle requirement", which should correspond to the average social cost of an additional bus required. But this is not the only cost that is relevant to the pricing of a bus service. In Table 6:1 nine cost items are designated, which could, in principle, be relevant in a pricing context.

Table 6:1. Pricing-relevant costs of

Period of time	Occupying space in the critical section	Boarding at the i^{th} stop	Alighting at the i^{th} stop
Off-peak	$\widehat{PC}_0$	$\overline{PC}^b_{0i}$	$\overline{PC}^a_{0i}$
The busiest round each peak period	$\widehat{PC}_{11}$	$\overline{PC}^b_{11i}$	$\overline{PC}^a_{11i}$
Remaining peak time	$\widehat{PC}_{12}$	$\overline{PC}^b_{12i}$	$\overline{PC}^a_{12i}$

The nine items result from the assumption of three distinguishable periods of time with respect to the rate of demand - the busiest peak round(s), the remaining peak period, and the off-peak periods - and from the fact that three "acts" on the part of passengers can cause pricing-relevant costs:

- Occupation of a seat or standing space on the bus in the critical section of the route
- Boarding
- Alighting

In addition, the pricing-relevant costs of boarding and alighting can vary a great deal within a given period of time, depending on the stop at which people get on or off the bus.

It is more important to note, however, that the demand for space on the bus has marked peaks, both in time and along the route. When it comes to pricing, it is not sufficient just to distinguish "peak" and "off-peak" levels of passenger flow. A common feature of urban bus routes is the existence of a distinct "peak within the peak". During a period of about a quarter to half an hour in the morning and again in the afternoon, the demand is at an appreciable higher level than during the remaining peak period. And further more, at a given time there will be variations in demand along a particular route. There are two main reasons for this. As regards radial services, the "tidal flow" of passengers causes a marked difference in demand between the main haul and the back haul. Added to which, spatial peaks exist in both directions. Particularly in the morning, the boarding rate is markedly higher than the alighting rate at the start of the run in towards the centre of the town, with the result that bus occupancy grows steadily. However, usually well before the terminal stop, the boarding rate is surpassed by the alighting rate, so that passenger flow reaches a maximum somewhere between the outskirts of the town and the central business district. A correspondingly peaked profile of bus occupancy appears during the afternoon peak in the opposite direction. The spatial peak on the main haul constitutes the "*critical section*" of the route. It is the passenger flow in the critical section in the "*busiest round*" that determines the total bus requirement.

The busiest round is the time interval between two successive departures of one and the same bus from a given stop in the critical section, when total passenger flow in the critical section has been at its maximum level for the day.

Bus fares should apply to rides between given stops at given times, and the pricing-relevant costs indicated in Table 6:1 should be combined accordingly (each fare is made up of three cost items) in the tariff of bus fares.

In practice a finely differentiated tariff of bus fares is out of the question. To keep fare collection costs down to a reasonable level, it is necessary to aim at a far-reaching standardization in fares. However, this should not be determined before the pricing-relevant costs have been calculated. Otherwise it is impossible to check whether a satisfactory balance has been struck between the allocative benefits and the costs of the price system.

The following analysis of the pricing-relevant costs is divided into two main parts. First, we will examine the pricing-relevant costs of occupying space on the bus, after which we will look at the pricing-relevant costs of boarding and alighting. For the sake of brevity, the terms "occupancy charges" and "b/a (boarding/alighting)-charges" will be used as appropriate. To obtain pure occupancy charges and pure b/a-charges, we will calculate the former as though an increase in the passenger flow in the critical section occurred without a corresponding increase in the rate of boarding and alighting, and vice versa in the case of the latter charges.

6.3 METHOD OF CALCULATING OCCUPANCY CHARGES

Whenever the capacity constraint is non-binding, it is obvious that the pricing-relevant cost of occupying space on the bus is zero. Increases in the passenger flow in the critical section will cause no costs in the alternative of retaining a given number of buses on the route. If in spite of this fact another bus is put in, it should mean that the incremental benefit of another bus is at least equal to the incremental cost, and the result as regards occupancy charge should still be zero. In cases where the capacity constraint is binding, another bus *has to* be put into service if demand goes up, and the average social costs of a marginal bus approximates to the occupancy charge.

This proxy consists of the difference between the incremental cost to the bus company and the incremental benefit to the original passengers of adding another bus, divided by the number of passengers which exhausts the additional capacity provided. The contents of the numerator and the denominator of this ratio will be discussed in turn below.

6.3.1 The numerator of the average social cost of a marginal bus

In the previous chapter two types of buses were identified in terms of use - all-day buses and peak-only buses. By using peak-only buses manned by a split shift, peak capacity can be raised above the basic all-day level of service. It is ruled out, a priori, that a bus should make only two round trips a day, i.e. operating during the busiest round only in each peak period. Under all realistic conditions the marginal cost of completing a full peak service is very small in relation to the marginal benefit. Under these conditions the cost per day of a marginal bus can assume three different values: the incremental cost of expanding off-peak capacity only, IC_0, is equal to the cost of putting in another all-day bus and withdrawing a peak-only bus; the incremental cost of expanding peak capacity, IC_1, is equal to the cost of another peak-only bus; and the incremental cost of increasing capacity throughout the whole day, IC, is consequently equal to $IC_0 + IC_1$. The incremental costs depend on the bus size, S. The following linear relationship has been found to apply:

$$\begin{aligned} IC_0 &= a_0 + b_0 S \\ IC_1 &= a_1 + b_1 S \\ IC &= a + b S \end{aligned} \tag{1}$$

where

$$a = a_0 + a_1, \text{ and } b = b_0 + b_1$$

The corresponding incremental benefit to the original passenger of expanding bus capacity were calculated in the previous chapter. It may be recalled that the incremental benefits were calculated on the assumption that total patronage is given. However, the results obtained are in fact directly applicable to the present problem, which is to calculate the occupancy charges separately. Let us consider a hypothetical situation in which the passenger *flow* has increased sufficiently to call for the addition of a bus, while the *number* of passengers stays the same. Such an outcome is possible, e.g. if the journey length is increased. This means that there are no offsetting boarding/alighting costs for the original passengers. Thus, the incremental benefit to the original passengers of putting in another bus following an increase in journey length only for passengers travelling in the busiest round is given by IB_1, as calculated in the previous chapter, if a peak-only bus is put in, and by IB if an all-day bus is put in.

In the (unlikely) event of full capacity utilization in off-peak hours, the addition of another all-day bus and withdrawal of a peak-only bus, following an increase in off-peak passenger flow in the critical section, would yield an incremental benefit equal to IB_0.

$$IB_0 = \frac{E_0 B_0 T_0 (\frac{v}{2} + ctQ_0)}{(N_0 - tB_0)^2} \tag{2}$$

$$IB_1 = \frac{E_1 B_1 T_{11} (\frac{v}{2} + ctQ_1)}{(N - tB_{11})^2}$$

$$IB = IB_0 + IB_1$$

The next question is: by what are we to divide the difference between the incremental cost and the incremental benefit?

6.3.2 The denominator of the average social cost of a marginal bus

So far as the peak is concerned, an increase in the passenger flow in the critical section requires an addition to capacity only if it occurs in the busiest round. It is possible that all peak rounds will be equally busy after the introduction of optimal bus fares. At present there is only one; the one that includes "the peak within the peak".

A common characteristic of bus routes is that the peak within the peak is (substantially) shorter than the round trip time, which means that each bus in the fleet contributes towards meeting the maximum peak within the peak demand at most once during each peak period (hence twice a day), and some buses will not pass the critical section at all during the peak within the peak.

If there is a marked peak within the peak, some concentration of buses is necessary. That is to say, the headway should vary in the busiest round in accordance with the expected variations in demand during this period. The total bus requirement, however, is determined by the passenger flow in the critical section during the whole period of a round trip. The point is that an addition to the passenger flow in the critical section, corresponding to another bus-load, requires an expansion of the bus fleet by another vehicle, no matter whether the increase in demand occurs in the peak within the peak or outside the peak within the peak, so long as it occurs during the busiest round.

Formally the peak vehicle requirement can be specified as follows:

$$\hat{Q}_{11}R_{11} \leq \hat{\phi}_{\max}\ SN \qquad (3)$$

$\hat{Q}_{11}$ = average rate of passenger flow in the critical section during the busiest round

R_{11} = round trip time in the busiest round

$\hat{\phi}_{max}$ = maximum tolerable occupancy rate in the critical section

S = bus holding capacity

N = number of buses on the route in the peak period (= the total bus fleet apart from reserve buses)

The peak demand as represented by the average boarding rate B_1 and the average passenger flow Q_1 is thus further divided into the higher rates B_{11} and Q_{11} that occur during the busiest round, and B_{12} and Q_{12} that occur during the remaining peak period. At present the latter are lower, but they can be equal to B_{11} and Q_{11} after the introduction of peak-load pricing. Spatially, the passenger flow in the busiest round is distinguished by its maximum value $\hat{Q}_{11}$, that occurs in the critical section, in addition to its mean value along the route.

The round trip time is equal to the ratio of the number of buses on the route to the frequency of service:

$$R_{11} = \frac{NT_{11}}{N - tB_{11}} \qquad (4)$$

Inserting this expression for R_{11} in (3), the following familiar formulation for the capacity constraint reappears: the maximum tolerable occupancy rate is equal to the ratio of the passenger flow to the product of bus size and service frequency in the busiest round, and equal to or greater than this ratio in the remaining peak period:

$$\hat{\phi}_{max} = \frac{\hat{Q}_{11}T_{11}}{S(N-tB_{11})} \geq \frac{\hat{Q}_{12}T_{12}}{S(N-tB_{12})} \qquad (5)$$

It should be observed, however, that the time taken by a round trip in the busiest round is the relevant period of time for assessing the capacity-determining demand, and hence the "peak vehicle requirement". This is particularly important to bear in mind when it comes to calculating the increase in passenger flow in the critical section, which should constitute the denominator of the expression for the average social cost of the marginal bus.

Total differentiation of (5) gives us the following expression:

$$d\hat{Q}_{11} \frac{T_{11}}{S(N-tB_{11})} + dB_{11} \frac{t\hat{Q}_{11}T_{11}}{S(N-tB_{11})^2} - \tag{6}$$

$$- dN_1 \frac{\hat{Q}_{11}T_1}{S(N-tB_{11})^2} = 0$$

It is clear that a continuum of value combinations for the increments dQ_{11} and dB_{11} can exhaust the capacity of one bus. At present, however, we are interested in the increase in $\hat{Q}_{11}$ which alone would call for another bus.

Setting $dN_1 = 1$, and $dB_{11} = 0$ gives us:

$$d\hat{Q}_{11}^{x} = \frac{\hat{Q}_{11}}{N-tB_{11}} = \frac{\hat{\phi}_{max}S}{T_{11}} \tag{7}$$

It should be observed that this increase in the *rate* of passenger flow can be sustained throughout the busiest round in each peak period by the input of another bus.

A question that presents itself is, how many "busiest rounds" are there each day? It is commonly assumed that the level of demand is always the same in the most busiest round in the morning, and in the busiest round in the afternoon. That makes two. It is quite possible, however, that *after optimal fares* are introduced, the rate of demand in the remaining peak periods will rise to the same level. This depends on the cross-elasticity of B_{11} and B_{12}, which is an unknown entity. The possibility that equalization of these demands may occur should be kept open. The relevant addition to total passenger flow, to use as denominators in calculating the pricing-relevant cost, is therefore equal to either twice or four times $R_{11} \cdot d\hat{Q}^{*}_{11}$. This product gives the maximum increase in the total passenger flow in the busiest rounds that can be accommodated by adding one more bus to the route.

$$2R_{11}dQ_{11}^{*} = \frac{2NT_{11}\hat{Q}_{11}}{(N-tB_{11})^2} \tag{8a}$$

$$4R_{1}dQ_{1}^{*} = \frac{4NT_{1}\hat{Q}_{1}}{(N-tB_{1})^2} \tag{8b}$$

As regards off-peak periods, it seems safe to assume that the rate of demand will not reach the level of the peak rate of demand as a result of peak-load pricing. The capacity constraint can be binding, however, even in off-peak hours, because of the possibility of differentiating peak and off-peak capacity. The applicable capacity constraint reads:

$$\hat{\phi}_{\max} \geq \frac{\hat{Q}_0 T_0}{S(N_0-tB_0)} \tag{9}$$

The main observation here is that this capacity constraint *should* rarely be binding. It is of course possible to envisage combinations of factor prices and other parameter values, which make the off-peak capacity constraint binding. Under such circumstances a rise in the level of off-peak demand will cause an additional capacity requirement. The distinguishing characteristic is that, since an increase in capacity takes the form of substituting an all-day bus for a peak-only bus, an increase in off-peak demand would never require an expansion of the bus fleet, but would simply result in a higher utilization of the vehicles of an existing fleet.

Analogously with the derivation of $d\hat{Q}_{11}^{*}$ above, the "capacity-exhausting" increment to off-peak passenger flow in the critical section is obtained as:

$$d\hat{Q}_{0}^{*} = \frac{\hat{Q}_{0}}{N_{0} - tB_{0}} = \frac{\hat{\phi}_{max}S}{T_{0}} \tag{10}$$

As it is assumed that the expected rate of demand is the same for all off-peak rounds, the total addition to passenger flow, which is relevant to a calculation of the off-peak occupancy charge, is equal to the product $E_{0}d\hat{Q}_{0}{}^{*}$, where E_{0} stands for the total extent of the off-peak periods.

This concludes the methodological discussion. In the next section the occupancy charges will be calculated in accordance with the principles laid down above. It could be expected, perhaps, that the results will be highly dependent on various characteristics of the route as regards both the demand and the supply of urban bus services. Interestingly enough, this is not how it turns out. It seems possible to draw some surprisingly general conclusions regarding the level and structure of optimal occupancy charges provided certain efficiency conditions are fulfilled regarding service design.

6.4 THE LEVEL AND STRUCTURE OF OPTIMAL OCCUPANCY CHARGES

To start with, let us write the occupancy charges for passengers traversing the critical section in the busiest round by inserting the values for IC_{1} and IB_{1} given by (1) and (2) above in the general formula for $\widehat{PC}_{11}$, assuming that only one busiest round occurs in each peak period.

$$\widehat{PC}_{11} = \frac{IC_1 - IB_1}{2R_{11} d\hat{Q}^{x}_{11}} = \frac{a_1 b_1 S - \dfrac{E_1 B_1 (\frac{v}{2} + ctQ_{11}) T_{11}}{(N - tB_{11})^2}}{\dfrac{2NT_{11}\hat{Q}_{11}}{(N - tB_{11})^2}} \qquad (11)$$

Eliminating S by using the capacity constraint (5), and splitting up (11) in three terms gives us:

$$\widehat{PC}_{11} = \frac{a_1 (N - tB_{11})^2}{2N\ T_{11} Q_{11}} + \frac{b_1 (N - tB_{11})}{2\widehat{\phi}_{max}\ N} - \frac{E_1 B_1 (\frac{v}{2} + ctQ_{11})}{2N\ Q_{11}} \qquad (12)$$

If the demand is given, the level of $\widehat{PC}_{11}$ is apparently highly dependent on the value of N, that is to say, the chosen combination of size and number of buses, in meeting the capacity requirement. Bearing in mind that tB_{11} normally constitutes a small fraction of N, it can be seen that the left-hand term of (12) is roughly proportional to N, the middle term is almost independent of N, while the right-hand negative term is inversely proportional to N. This means that $\widehat{PC}_{11}$ can be increased almost limitlessly by making N very large (and consequently S very small). Conversely, $\widehat{PC}_{11}$ will be very low - even negative - if a relatively small number of large buses are employed. The latter situation obtains in many towns and cities today. As argued in the preceding chapter, the salient feature of present designs of urban bus services is that too few and too large buses are used.

Expression (12) for $\widehat{PC}_{11}$ is thus a rather inconclusive result. Nothing else was to be expected. As long as no efficency conditions are imposed upon the choice of design variables, the pricing-relevant cost can assume any value. As an intermediate step towards design efficiency, let us regard N and only N as a control variable. In that case the situation that will most likely arise, following the "optimization" of the service in question, is one of excess capacity. Given the currently employed oversized buses, an optimal frequency of service would generally be accompanied by unnecessarily high capacity. And as has been pointed out several times, when the capacity constraint is non-binding, the optimal occupancy charge is zero.

The most interesting case to consider is that in which it is assumed that both the right number of buses and the right size of bus is chosen in each particular situation. To assume a truly optimal design seems to be the most logical approach in the present connection. Given that the will to charge optimal bus fares exists, misgivings about designing the service to minimize the social costs should not be entertained by the decision-makers. In the present analysis we allow for this assumption by reintroducing the efficiency conditions derived in our earlier Kuhn-Tacker analysis. In brief, these conditions can be written as follows:

$$IC_0 - IB_0 + \lambda - \mu_0 \frac{\hat{Q}_0 T_0 (N-tB_{11})}{\hat{Q}_{11} T_{11} (N-tB_0)^2} = 0 \qquad (13)$$

$$IC_1 - IB_1 - (b_0 N_0 + b_1 N) \frac{\hat{Q}_{11} T_{11}}{\hat{\phi}_{max} (N-tB_{11})^2} - \lambda + \qquad (14)$$

$$+ \mu_0 \frac{\hat{Q}_0 T_0}{\hat{Q}_{11} T_{11} (N_0 - tB_0)} = 0$$

$$\lambda = 0 \quad \text{when } N_0 < N_1 \qquad (15)$$

$$\lambda > 0 \quad \text{when } N_0 = N_1 \qquad (16)$$

$$\mu_0 = 0 \quad \text{when } \hat{\phi}_{max} > \frac{\hat{Q}_0 T_0 (N_1 - tB_{11})}{\hat{Q}_{11} T_{11} (N_0 - tB_0)} \qquad (17)$$

$$\mu_0 > 0 \quad \text{when } \hat{\phi}_{max} = \frac{\hat{Q}_0 T_0 (N-tB_{11})}{\hat{Q}_{11} T_{11} (N_0 - tB_0)} \qquad (18)$$

Six variants of $\widehat{PC}_{11}$ can be envisaged, depending, on the one hand, on whether the two rounds in each peak period will be equally busy, and on the other on whether the number of buses in the off-peak $N_0 = N$, on whether N_0 is less than N but greater than the minimum off-peak bus requirement, or on whether the capacity constraint is binding also in off-peak. Let us assume to begin with that only one "busiest round" occurs in each peak period. Three cases are then distinguishable. (i) If on a fairly thin route the rate of off-peak demand is comparatively high (relative to the peak rate of demand), it will probably be optimal to run the same number of buses in peak and off-peak hours. After the introduction of peak-load pricing, a levelling of the diurnal demand profile will probably occur, which will increase the likelihood that $N_0 = N$ at optimum. (ii) If a substantial difference between the rates of peak and off-peak demand prevails even under a new structure of bus fares, it may be right to use some peak-only buses, although the ratio of peak to off-peak capacity should normally be lower than the ratio of the peak rate to the off-peak rate of demand. (iii) On extremely dense routes it is possible that the capacity constraint should be binding in off-peak periods as well.

The pricing-relevant costs in these three cases - $\widehat{PC}^{i}_{11}$, $\widehat{PC}^{ii}_{11}$, and $\widehat{PC}^{iii}_{11}$ - are calculated below. The remaining occupancy charges which may or may not be different from zero will then be worked out. In case (i) where it is optimal to run the same number of buses in peak and off-peak hours, it is obvious that the capacity constraint is non-binding in off-peak periods, so that $\mu = 0$. When $N_0 = N$ it follows that $\lambda > 0$. However, by summing up (13) and (14) to obtain the difference IC - IB, λ is eliminated.

$$\widehat{PC}^{i}_{11} = \frac{(IC - IB)\ (N-tB_{11})^2}{2NT_{11}\hat{Q}_{11}} = \qquad (19)$$

$$= \frac{(b_0N_0 + b_1N)}{2\hat{\phi}_{max}\ N} = \frac{b}{2\hat{\phi}_{max}}$$

At optimum the occupancy charge in the busiest round is consequently equal to the proportionality constant, b, in the relationship between the cost per bus-day and S, divided by twice the maximum occupancy rate.

Although the model is exceedingly simple, this result seems to be robust. A check on its good sense is that it is consistent with the "total adaptation marginal cost" of an additional passenger traversing the critical section. If buses were made of putty, an additional passenger could be provided for by expanding the holding capacity of the bus concerned by $1/\hat{\phi}_{max}$ units. The total cost of this increase in capacity would be $b/\hat{\phi}_{max}$ and, assuming that the additional passenger uses the bus in both peaks, the optimal fare per trip would be $b/2\hat{\phi}_{max}$.

An attractive feature of (19) is that the pricing-relevant cost is nearly constant, regardless of the level of demand, value of time, etc. The chosen occupancy rate in the critical section $\hat{\phi}_{max}$ may vary somewhat, of course, depending on what the different bus companies accept as tolerable conditions as regards crowding and queuing.

Using the parameter values given in Table 5:4 and the estimated relationship between bus costs and bus size pertaining to British conditions in 1975, and assuming that $\hat{\phi}_{max} = 0.8$, the value of $\widehat{PC}^{i}_{11}$ will be 23 p.

The conclusion regarding $\widehat{PC}^{i}_{11}$ needs slight modification only when the two remaining cases are considered.

In case (ii) where it is optimal to run more buses at peak than at off-peak times, it follows that $\lambda = 0$. By assumption $\mu_0 = 0$, too. In this case the ratio of peak to off-peak frequency of service is not as low as the ratio of $\hat{Q}_0$ to $\hat{Q}_{11}$. Now $IC_1 - IB_1$ rather than $IC - IB$ has to be devided by $2R_{11}d\hat{Q}^{*}_{11}$, in order to obtain the pricing-relevant cost.

$$\widehat{PC}^{ii}_{11} = \frac{(IC_1 - IB_1)(N - tB_{11})^2}{2N\ T_{11}\ \hat{Q}_{11}} = \frac{(b_0 N_0 + b_1 N)}{2\hat{\phi}_{max}\ N} = \qquad (20)$$

$$= \frac{b_0 \frac{N_0}{N} + b_1}{2\hat{\phi}_{max}}$$

The value of $\hat{P}C^{ii}_{11}$ is apparently slightly less than the value of $\widehat{PC}^{i}_{11}$. For example, if 75 per cent of the buses are kept in service in off-peak hours, the value of $\widehat{PC}^{ii}_{11}$ is equal to 20 p (as compared to 23 p) under otherwise equal conditions.

In case (iii), finally, $\mu_0 > 0$ by assumption, which leads to the addition of a (negative) term that includes μ_0 (see (14) above).

$$\widehat{PC}^{iii}_{11} = \frac{b_0 \frac{N_0}{N} + b_1}{2\hat{\phi}_{max}} - \frac{\mu_0(N - tB_{11})}{2NT_{11}\ \hat{Q}_{11}} \qquad (21)$$

The value of μ_0 depends on how remote it is that the off-peak capacity constraint is non-binding. The role of this scarcity value is made more clear by considering the off-peak occupancy charge, $\widehat{PC}_0$. This charge is obtained by dividing the difference $IC_0 - IB_0$ according to (13) by the product of E_0 and $d\hat{Q}^{*}_0$ as given in (10) above.

$$\widehat{PC}_0 = \frac{(IC_0 - IB_0)\ (N_0 - tB_0)}{E_0\hat{Q}_0} = \frac{\mu_0}{E_0\hat{Q}_0} \qquad (22)$$

The total number of off-peak passengers traversing the critical section per day is $E_0\hat{Q}_0$. The total revenue from off-peak occupancy charges consequently amounts to μ_0. The total number of passengers traversing the critical section in the busiest round in each peak period is $2R_{11}\hat{Q}_{11}$. The total revenue from occupancy charges in the busiest round will consequently include a negative term = - μ_0.[1]

The total revenue from occupancy charges is thus the same in case (ii) and case (iii). When $\widehat{PC}_0 > 0$, the occupancy charge applying to the busiest round will be correspondingly lower than in the case in which the off-peak capacity constraint is non-binding.

$$\widehat{PC}_{11}^{iii} = \widehat{PC}_{11}^{ii} - \frac{E_0\hat{Q}_0}{2R_{11}\hat{Q}_{11}} \widehat{PC}_0 \tag{23}$$

Concerning $\widehat{PC}_{12}$ it is clear that if a zero occupancy charge does not raise the passenger flow $\hat{Q}_{12}$ to the level of $\hat{Q}_{11}$, a zero charge is appropriate. On the other hand, if a zero charge were to raise $\hat{Q}_{12}$ above $\hat{Q}_{11}$, a positive occupancy charge should apply also in the second round of each peak period. The aim should then be to levy such a combination of $\widehat{PC}_{11}$ and $\widehat{PC}_{12}$ that the level of demand is equalized throughout the peak period. Many such combinations of $\widehat{PC}_{11}$ and $\widehat{PC}_{12}$ exist. A further condition is, however, that the average peak occupancy charge should be equal to $b/4\hat{\phi}_{max}$.

$$\widehat{PC}_1 = \frac{\widehat{PC}_{11} + \widehat{PC}_{12}}{2} = \frac{b}{4\hat{\phi}_{max}} \tag{24}$$

On the assumption that $\hat{Q}_{12}$ will be equal to $\hat{Q}_{11}$ under optimal pricing,[2] the application of the formula for the occupancy charge of the busiest round(s) gives us a charge

[1] Remember that $R_{11} = \frac{NT_{11}}{N-tB_{11}}$

[2] It is obvious that $\hat{Q}_{11} < \hat{Q}_{12}$ is an anomaly. In that case $\hat{Q}_{12}$ determines the peak vehicle requirement, which implies that $\widehat{PC}_{12}$ should be $= b/2\hat{\phi}$, and $\widehat{PC}_{11} = 0$. Such prices, however, are inconsistent with the basic precondition that the intensity of demand is higher in the first round than in the second.

that is equal to $b/4\hat{\phi}_{max}$. If this level of charges were to prevail throughout the peak period, equality between $\hat{Q}_{11}$ and $\hat{Q}_{12}$ would not be obtained. It is necessary to find the charge differential which leads to equalization. If this is not achieved until $\widehat{PC}_{12} = 0$, a zero charge in the second peak round is appropriate in spite of the fact that the level of demand is the same as in the first round. If equalization is not achieved even for $\widehat{PC}_{12} = 0$, we are back where we were in the previous case. Without specifying the cross-elasticity in question, the peak occupancy charges cannot be determined more exactly than that:

$$\widehat{PC}_{11} = \frac{b}{2\hat{\phi}_{max}} - \widehat{PC}_{12} \tag{25}$$

and that

$$\frac{b}{4\hat{\phi}_{max}} < \widehat{PC}_{11} \leq \frac{b}{2\hat{\phi}_{max}} \tag{26}$$

$$0 \leq \widehat{PC}_{12} < \frac{b}{4\hat{\phi}_{max}}$$

6.4.1 The financial result of occupancy charges

The absolute level of the charges is not as interesting as the relative level. In other words: how would the revenue from optimal occupancy charges compare with the total traffic operation costs? These costs amount to $(a+bS)N$ or $aN+S(b_0N_0+b_1N)$. It is interesting to find that the total revenue from the occupancy charges is always equal to the size-proportional part of the total costs. The total revenue in the busiest round differs slightly, depending on whether $N_0 = N$ or $N_0 < N$, but so does the size-proportional traffic operation costs.

$$TR^{i} = \widehat{PC}^{i}_{11} \cdot 2R_{11}\hat{Q}_{11} = \frac{b}{2\hat{\phi}_{max}} \cdot \frac{2NT_{11}\hat{Q}_{11}}{(N-tB_{11})} = bSN \qquad (27a)$$

$$TR^{ii} = \frac{b_0 \frac{N_0}{N} + b_1}{2\hat{\phi}_{max}} \cdot \frac{2NT_{11}\hat{Q}_{11}}{(N-tB_{11})} = b_0SN_0 + b_1SN \qquad (27b)$$

It has previously been shown that when the off-peak capacity constraint is binding -case (iii) - the total revenue from peak and off-peak occupancy charges is the same as in case (ii). It has also been shown that when there are four rather than two "busiest rounds" per day, the average peak occupancy charge is reduced by half. Total revenue stays the same, however, since the number of passengers liable for an occupancy charge is twice as large. The financial contribution made by optimal occupancy charges is rather limited. At the very most - when really large buses are used - the size-proportional part constitutes a fourth of the total traffic operation costs. No contribution at all will be made towards the recovery of the size-independent part, aN, which includes among other things all crew costs.

6.5 METHOD OF CALCULATING BOARDING/ALIGHTING CHARGES

The remaining traffic operation costs after the revenue from occupancy charges has been deducted is thus equal to the part that is independent of size, aN. The question is, how far would optimal boarding/alighting charges contribute to the recovery of these costs?

An answer to this question can be obtained by calculating an average b/a-charge for each particular period of time. Taking an average in this way means, on the one hand, that no distinction is made between the two acts of boarding and alighting and, on the other, that differences in bus occupancy along the route are disregarded. In the following section, where a possible differentiation in fares between stops is discussed, a more detailed picture of the structure of optimal b/a-charges will be drawn.

To simplify the notations and to avoid too much repetition, it will be assumed throughout this section that the same number of buses, N, is in service all day. The exposition is also facilitated by generalizing the time division of the service day. The number of boarding passengers in the j^{th} hour is B_j. The total number of passengers carried per service day is consequently $\sum_j B_j$, where j = 1...E.

When the capacity constraint is non-binding, additional passengers would not cause any costs to the bus company worth mentioning, but the boarding/alighting costs are imposed on fellow passengers in the form of extra waiting and riding time (at stops). Given the number of buses on the route, the mean waiting time and the riding time will both be functions of the number of passengers boarding/alighting per hour.

Using the following notations, namely that AC_j^{wait} = the average cost of time spent waiting for the bus at stops per passenger in the j^{th} period and AC_j^{ride} = the average cost of riding time per passenger in the j^{th} period, then the pricing-relevant cost in question can be written in the following general form:

$$\overline{PC}_j = B_j \frac{\partial AC_j^{wait}}{\partial B_j} + B_j \frac{\partial AC_j^{ride}}{\partial B_j} \qquad (28)$$

Note that for this formulation to make sense, it has to be assumed that the schedule is a variable, i.e. costs are considered in the "medium run". In the short run - given the schedule - it is not necessarily true that an additional passenger boarding/alighting will impose any costs at all on the other passengers. The schedule should not be so tight that a marginal increase in demand would cause a delay. This does not mean that an increase in the boarding rate is free of cost in the short run. Normally, however, the additional cost should assume the form of various "schedule-keeping costs".

A given schedule will become too tight if the boarding rate goes up, and an adjustment in the schedule should be made, which means that extra riding time and waiting time costs for the passengers are substituted for the schedule-keeping costs temporarily borne by the bus service operator (as well as the passengers). For b/a-charges, too, the medium run rather than the short run is the "run" relevant to pricing. Moreover, the medium-run costs of additional boarding/alighting are much more easily calculable than the short-run schedule-keeping costs. The calculation of the boarding/alighting costs poses a tricky problem in cases where the capacity constraint is binding. Given the present formulation of the capacity constraint (6) or (9), the only way of meeting an increase in B_1 or B_0 at full capacity utilization is to put in another bus. In this case an indivisibility problem is inescapable. The size of a batch of additional riders that will absorb the whole capacity of another bus on the route, although none of them travel past the critical section, is very great indeed in relation to the original number of passengers. Setting $dQ_{11} = 0$ and $dN = 1$ in (6) above gives us the capacity-absorbing value of $dB_{11} = 1/t$, where t is given in hours.

In view of the fact that t ranges from about 1/1500 hour to 1/200 hour (in the case of an individual stopping service), there is no doubt that the average social cost of a marginal bus is a very rough approximation of the pricing-relevant cost in this case. However, as is shown in an appendix to this chapter, the result is quite reasonable. A dominating part of the pricing-relevant cost thus obtained corresponds to the imposed waiting and riding time costs according to (28) above, and a small remainder represents the admittedly somewhat suspect capacity cost of the boarding/alighting of another passenger. As could be expected, this remainder turns out to be equal to a certain fraction of the pricing-relevant cost of occupying space on the bus in the critical section, namely the fraction given by the ratio of the two capacity-absorbing increments $d\hat{Q}^{*}_{11}$ and dB^{*}_{11}. The b/a-charge applying to the busiest round can consequently be written:

$$\overline{PC}_{11} = \frac{d\hat{Q}^{*}_{11}}{dB^{*}_{11}} \widehat{PC}_{11} + B_{11} \frac{\partial AC^{ride}_{11}}{\partial B_{11}} + B_{11} \frac{\partial AC^{wait}_{11}}{\partial B_{11}} \qquad (29)$$

6.6 THE LEVEL AND STRUCTURE OF OPTIMAL BOARDING/ALIGHTING CHARGES

The crucial question for calculating the b/a-charges is: how far are the stop times of the buses dependent on the number of passengers boarding? So far it has been convenient and harmless to assume that limited changes in the ratio of the boarding rate to the frequency of the service will not change the number of stops per round trip. The time per passenger, t, stands for the "pure" boarding/alighting time, and the starting/stopping time (h per stop) has been assumed to be a part of T, i.e. the round trip time, except for the pure boarding/alighting time. Under this assumption it can be expected that the b/a-charges will generally be very low. This is characteristic of most "ordinary" bus services. In order to cover the whole range, we will also investigate the opposite extreme case of an individual stopping service.

If passengers set a relatively high value on savings in their walking time, the density of the stops in an ordinary bus service will probably be fairly high. Individual stopping services represent an interesting extreme case. The walking costs would be minimized if the passengers of a bus line were allowed to hail the bus and board anywhere along the route. The saving in walking costs would be counterbalanced by a substantial increase in the total stopping time per round trip. Compared to the ordinary case, the quality of service with regard to journey time would be rather poor, if the same bus size were used. To be viable, a bus line offering an individual stopping service has to use far smaller buses than is usual in the ordinary case. It is likely that the optimal system design would tend to be a "liner taxi service". By employing small buses/large cars, the stopping time per passenger is reduced (thanks to quicker acceleration and deceleration), and the number of passengers negatively affected by each stop will be correspondingly smaller, assuming the occupancy to be proportional to the bus size. Liner taxi services are provided in several cities in different countries. In Israel, for instance, it has been a popular alternative to the ordinary bus lines for quite a long time.

The round trip time, R, of the kind of "ordinary" bus service which has been discussed so far, consists of the three following components:

$$R = T^{run} + m \cdot h + t \cdot \frac{B}{F} \tag{30a}$$

where T^{run} is the pure running time, $m \cdot h$ is the stopping/starting time (m = number of stops on the route), and tB/F is the pure boarding/alighting time per round trip. We can conveniently write this as:

$$R = T + t\frac{B}{F} \tag{30b}$$

A liner taxi system is characterized by the fact that although a fixed route is adhered to, the passengers may get on and off whereever they like along the route. On the somewhat extreme assumption that everybody is always alone when boarding or alighting, the round trip time can be written:

$$R = T^{run} + 2(h+t)\,\frac{B}{F} \tag{31}$$

The results obtained in the following analysis of the "ordinary" bus service model are consequently directly applicable to a liner taxi service, simply by substituting h+t for t alone and T^{run} for T.

In the j^{th} period the average riding time cost per passenger is:

$$AC_j^{ride} = ck\ R_j = \frac{ckNT_j}{N-tB_j} \tag{32}$$

where k is the ratio of the average journey length to the round trip distance. Note also that $\frac{Q_j}{B_j} = k$. The cost of the increase in mean riding time caused by another passenger is obtained as:

$$\frac{\partial AC_j^{ride}}{\partial B_j} = \frac{ctk\ NT_j}{(N-tB_j)^2} \tag{33}$$

The total sum of the additional riding time costs caused by another passenger is equal to:

$$B_j\,\frac{\partial AC_j^{ride}}{\partial B_j} = \frac{ct\ Q_jNT_j}{(N-tB_j)^2} = \frac{ckNT_j}{N-tB_j} \cdot \frac{tB_j}{N-tB_j} = \tag{34}$$

$$= AC_j^{ride}\ \frac{R_j-T_j}{T_j}$$

Similarly, the imposed waiting time cost, relevant to pricing, is obtained as follows:

$$AC_j^{wait} = \frac{v\ T_j}{2(N-tB_j)} \tag{35}$$

$$\frac{\partial AC_j^{wait}}{\partial B_j} = \frac{vt\ T_j}{2(N-tB_j)^2} \tag{36}$$

$$B_j \frac{\partial AC_j^{wait}}{\partial B_j} = \frac{vt\ B_jT_j}{2(N-tB_j)^2} = \frac{v\ T_j}{2(N-tB_j)} \cdot \frac{tB_j}{N-tB_j} = \tag{37}$$

$$= AC_j^{wait} \frac{R_j-T_j}{T_j}$$

The b/a-charge is consequently equal to the sum of the average riding time cost and the average waiting time cost per passenger times the ratio $(R_j - T_j)/T_j$:

$$\overline{PC}_j = (AC_j^{ride} + AC_j^{wait}) \frac{R_j - T_j}{T_j} \tag{38}$$

As regards an ordinary bus service, the value of the ratio $(R_j - T_j)/T_j$ is no more than between 1/20 and 1/10, while it can be several times higher in the case of a liner taxi service.

6.6.1 The financial result of b/a-charges

By multiplying $\overline{PC}_j$ by B_j and summing over all j, and adding the extra charges on passengers boarding in the busiest round, the total revenue from b/a-charges is obtained as shown below. (For simplicity it is assumed that not only N but also F, R and T remain the same all day.)

$$\overline{TR} = (TC^{ride} + TC^{wait} + bSN) \frac{R-T}{T} \tag{39}$$

From expression (33), in Chapter 5, for the optimal frequency of service, the following relationship is easily derived:

$$F = \frac{TC^{wait} + TC^{ride}\frac{tB}{N} + tB_{11}bS}{aT} \qquad (40)$$

By multiplying both sides by $\frac{aTN}{N-tB_{11}}$ the following expression is obtained:

$$aN = TC^{wait}\frac{N}{N-tB_{11}} + TC^{ride}\frac{tB}{N-tB_{11}} + bSN\frac{tB_{11}}{N-tB_{11}} \qquad (41)$$

Deducting the total revenue according to (39) from (41) observing that $(R-T)/T = tB_{11}/(N-tB_{11})$, gives us:

$$aN - \overline{TR} = TC^{wait} \qquad (42)$$

The total revenue obtained from boarding/alighting charges would apparently fall short of the total size-independent cost by an amount equal to the total waiting time costs of the passengers. This normally represents a very large financial deficit. The size-independent part of the total traffic operation cost dominates over the size-proportional part; for a fleet of even rather large buses, aN is something like 3/4 of the total traffic operation costs. And as will be shown presently, the total waiting time cost is not much less than aN. The relative size of the financial deficit is given by:

$$\frac{aN - \overline{TR}}{aN} = \frac{TC^{wait}}{aN} = \frac{1}{1 + \frac{2ctQ}{v} + \frac{2t\beta\ IC}{vE}} \qquad (43)$$

where Q is a weighted average value for the passenger flow per hour, β is the ratio of B_{11}/B, and IC/E is the cost per service hour of an all-day bus. The right-hand term of the denominator is so small, relatively, that it can be ignored for our present purpose. On the assumption that c/v = 1/3, the following approximative expression for the relative deficit results:

$$\frac{aN - \overline{TR}}{aN} \approx \frac{1}{1 + \frac{2}{3}tQ} \qquad (44)$$

When both t and Q are relatively small, for instance, when t≈2 seconds and Q is less than 250 passengers per hour, the deficit will be some 90 per cent of aN. When t = 4.25 seconds (applicable to OMO buses where tickets are bought from the driver) and Q is as high as 1 000 passengers per hour, the deficit would be reduced to 56 per cent of aN. This is still a very substantial figure. Relative to the total traffic operation costs (a+bS)N, the deficit is about 50 per cent.

It can thus be concluded that ordinary bus services would require very heavy subsidization - in the range of 50-90 per cent of the total bus company costs - if optimal bus fares were charged.

A liner taxi service is quite a different matter. Replacing the single t of the formula (44) by 2(h+t), the financial result will be much less disastrous. Suppose that twice the sum of the stopping/starting time and the boarding/alighting time is 30 seconds - 15 seconds for picking up another passenger and 15 seconds for dropping him off - and supposing that all the other assumptions of the model are unchanged, the financial result of applying optimal liner taxi fares will be that 50 per cent of the traffic operation costs will already be covered at a density of demand (passenger flow) as low as 150 passengers per hour. And if the passenger flow per hour is, say, 500, about 80 per cent of the traffic operation costs will be covered. (The size-proportional costs will be covered by occupancy charges in an ordinary bus service as well as a liner taxi service.)

In view of the relatively small revenue arising from b/a-charges for an ordinary bus service, it can be expected that the absolute level of the b/a-charges will be very low. This is shown more clearly in the following alternative formulation of the b/a-charge, which is more revealing when it comes to considering the absolute level of $\overline{PC}_j$. Let us split (34) into these two following terms:

$$\frac{ctQ_jNT_j}{(N-tB_j)^2} = \frac{ctQ_jT_j}{N-tB_j} + \frac{ct^2B_jQ_jT_j}{(N-tB_j)^2} \qquad (45)$$

Substituting ϕ_jS for $Q_jT_j/(N-tB_j)$ in the first term of (45), and merging the second term of (45) with (37), gives us:

$$\overline{PC}_j = ct\phi_jS + \frac{tB_j\,(\frac{V}{2} + ctQ_j)T_j}{(N-tB_j)^2} \qquad (46)$$

The left-hand term corresponds to the *direct* imposed cost on the ϕ_jS passengers on the bus by the act of boarding/alighting. The right-hand term corresponds to the *indirect* imposed cost of riding time as well as waiting time arising from the increase in headway occurring as a result of additional boarding/alighting time. For the right-hand term a neat approximative value can be obtained by inserting the optimal value for the frequency of service F^{opt} according to (33) in Chapter 5. Disregarding the right-hand term of the numerator of the expression for F^{opt}, which rarely constitutes more than one or two per cent of the total value of the numerator, the following approximate expression for $\overline{PC}_j$ is finally obtained:

$$\overline{PC}_j \approx ct\phi_j S + \frac{at}{E} = t\,(c\phi_j S + \frac{a}{E}) \qquad (47)$$

The b/a-charge is consequently equal to the product of the boarding/alighting time, t, and the sum of the time cost per hour of the passengers on the bus and the size-independent bus cost per service hour. The rationale of the latter component is not that boarding/alighting makes a demand on capacity - by assumption, the capacity constraint is non-binding - but that this cost has to be incurred by the bus company in order to maintain the frequency of service (and avoid the indirect imposed cost).

Inserting the parameter values presented in the preceding chapter, the following conclusions can be drawn. When t is as low as about 2 seconds, which is applicable to TMO buses, $\overline{PC}_j$ will not exceed 1 p until bus occupancy is as high as 30 passengers. In the OMO case where boarding/alighting charges are levied in the form of tickets bought from the driver, t would be at least twice as high, and under certain circumstances the somewhat less trifling value of 2-3 p is indicated by the price formula. In the busiest round another penny should be added to the b/a-charge in each case.

In the case of a liner taxi service the b/a-charges should be much higher, although not quite as high as to make them proportional to the ratio 2(h+t)/t. The size of the optimal vehicle is much smaller when an individual stopping service is offered, which means that the value of the average occupancy $\phi_j S$ falls as t rises.

Finally, it can be added that a compromise between the ordinary case and the case of an individual stopping service can represent an improvement. A good many passengers would probably be no better off if they were to pay for the benefit of reducing the walking distance. Thus, several fairly widely separated fixed stops, where a concentration of starting-points and destinations can be expected, may be combined with the possibility of chosing one's stop. A differentiation in fares is then indicated, between those who prefer a bit of walking rather than paying a high fare and those who are prepared to pay for an individual stopping service.

6.6.2 Fare differentiation between stops

Two "extreme" cases that are opposites as regards the value of the boarding/alighting time have so far been considered: (1) The case of an ordinary bus service, where it can be assumed that limited changes in the boarding rate affects the pure boarding/alighting time only, (2) the case of an individual stopping service. In each case a differentiation of the b/a-charges between stops could be justified on grounds of variations in the bus occupancy along the route. However, so far as case (1) is concerned, such differentiations does not seem worth the additional fare collection costs, in view of the generally very low level of the pricing-relevant boarding/alighting costs.

Another ground for fare differentiation seems potentially more interesting, namely the differences in boarding/alighting rate that may exist between different stops. If the boarding/alighting pattern is fairly regular, implying that the number of passengers boarding and/or alighting per stop is relatively evenly distributed,[1] case (1) is probably a close approximation to the real case, and there is not much potential for a differentiation of fares between stops. Suppose on the other hand that one or more stops are used appreciably

[1] Needless to say, this does not mean that the occupancy rate will also be fairly even during a round trip.

less than the other stops on average. Given the stochastic nature of the demand, it may happen from time to time that the bus can by-pass a less frequented stop. This possibility brings a new pricing-relevant cost into the picture. Mohring (1972) has dealt with this aspect by assuming that the total boarding/alighting rate fluctuates at random around its mean value, which means that a Poisson distribution is applicable.

In Mohring's model the pricing-relevant cost in question is calculated per passenger trip, regardless of the starting-point or destination of a particular trip. Given the number of buses in service, the pricing-relevant cost considered arises because the probability that one or more stops can be by-passed decreases as the total demand increases, and hence the expected journey time increases. This averaging over all stops seems less interesting in view of the fact that, if there is a situation in which no boarding/alighting occurs at all, it is likely to be markedly "local", i.e. occurring at a limited number of less frequented stops, and relatively few passengers would be liable for an extra charge on this count. A highly diluted average charge on all passengers would be allocatively rather ineffective. It is therefore assumed in the present discussion that the stops on the route can be divided into two distinct groups.

1. The majority which can be characterized as "frequented stops", where a situation without any boarding/alighting at all is so rare that it can be ignored; these stops are interspersed by

2. a number of "less frequented stops" denoted 1...i...k, where it happens from time to time that no one wishes to board or alight.

The total stopping/starting time per round trip depends on the proportion of trips originating at or destinated for the infrequently used stops. The smaller this proportion, the greater will be the expected number of stops that can be by-passed.

A certain fare *differential* ($\Delta\overline{PC}_i$) between these two groups of stops can serve an allocative purpose: it would present users of the less frequented stops with a choice between a low fare plus a bit of extra walking and the higher fare charged for trips to and from the less frequented stops. Moreover, this kind of differentiation of fares can provide information that is useful in the difficult task of deciding the optimal number and location of stops.

On the other hand, it cannot be expected that the introduction of a possible addition to the fare for the b/a-charges of the less frequented stops would appreciably improve the financial result. It is in the nature of things that only a small minority of passengers would be liable for this extra charge.

To calculate the fare differential of the less frequented stops, it is naturally necessary to examine the demand along the route in more detail. The number of passengers using (boarding as well as alighting) the i^{th} less frequented stop per hour is designated X_i. For simplicity, the indication of the period of time is omitted.

The important point here is that this number is a random variable. Were this not the case, the marginal cost would be at the maximum level for $X_1 = 1$, and zero for all higher values of X_i.[1]

[1] Under such conditions the proponents of marginal cost pricing and the proponents of average cost pricing have traditionally been more than usually at logger-heads. The gap between the two schools of thought would be less wide in cases where "joint costs" are involved, if the proponent of marginal cost pricing took into account, as they should, that the marginal *expected* cost is the pricing-relevant cost. In connection with a related problem, this point was originally made by Walters, A.A., The Allocation of Joint Costs. American Economic Review, June 1960.

The expected stopping/starting time at the i^{th} stop per bus run is equal to the product of the stopping/starting time, h, and the probability that at least one passenger will board or alight.

$$h_i = h \cdot p(HX_i \geq 1) \tag{48}$$

where $H(= \frac{1}{F})$ is the headway between the buses on the route. The question is, how much additional stopping/starting time can be expected per hour following an increase in X_i by one passenger boarding or alighting per hour? The additional stopping/starting time is calculated by first expressing the expected stopping/starting time at the i^{th} stop per hour.

$$\frac{h_i}{H} = \frac{h}{H} p(HX_i \geq 1) \tag{49}$$

The partial derivative of h_i/H with respect to X_i gives the expected additional stopping/starting time.

$$\frac{\partial\left(\frac{h_i}{H}\right)}{\partial X_i} = \frac{\partial h_i}{\partial X_i} \frac{1}{H} = \frac{h}{H} \frac{\partial p}{\partial X_i} \tag{50}$$

Note that H is assumed to be independent of X_i. This assumption is justified when the following expression is used to represent the fare differential $\Delta\overline{PC}_i$.

$$\Delta\overline{PC}_i = \frac{h}{H} \frac{\partial p}{\partial X_i} \left(c\phi_i S + \frac{a}{E}\right) \tag{51}$$

The sum in the bracket gives the cost imposed on the passengers on the bus, *given the frequency of service*, plus the cost to the bus operator of maintaining a certain frequency. (Compare the discussion in connection with (47) above.)

By specifying the probability distribution of the number of passengers using the i^{th} stop, a better idea of the likely order of magnitude of $\Delta\overline{PC}_i$ can be obtained. The probability of X_i can be taken to be independent of the number of passengers using all other less frequented stops, since it is assumed that there is at least one frequented stop in between each pair of less frequented stops. It is then reasonable to assume that the probability that X_i passengers will use the i^{th} stop per hour is given by the Poisson distribution. In that case the probability that at least one passenger wants to use the i^{th} stop on each run is equal to $1 - e^{-HX}$i. The derivative of this probability with respect to X_i comes to:

$$\frac{\partial p}{\partial X_i} = He^{-HX_i} \tag{52}$$

The fare differential is then written:

$$\Delta\overline{PC}_i = he^{-HX_i}\left(c\phi_i S + \frac{a}{E}\right) \tag{53}$$

The shape of the function for $\Delta\overline{PC}_i$ is determined by the exponential factor. The main characteristic of $\Delta\overline{PC}_i$ is therefore that it falls sharply with increases in the expected number of passengers demanding the i^{th} stop. When HX_i is zero, i.e. when it is absolutely certain that a passenger getting on or off at the i^{th} stop is alone, e^{-HX_i} is unity. Already at $HX_i = 3$, for example, the value of e^{-HX_i} is down to 5/100, and when $HX_i = 6$, this factor assumes the insignificant value of 1/400. The shape of $\Delta\overline{PC}_i$ is shown in Figure 6:1. Putting figures on the parameters involved, a maximum value of $\Delta\overline{PC}_i$ in the range of 5-15 p is found. It can be assumed that the stopping/starting time, t is equal to 18 seconds, the cost of passenger time on the bus, c is equal to 50 p, and the bus cost per hour corresponding to the right hand term in the bracket 400 p. It is obvious that bus occupancy varies within wide limits. If it is 20 passengers, the maximum fare differential is 7 p. This value is doubled if the occupancy is 48 passengers.

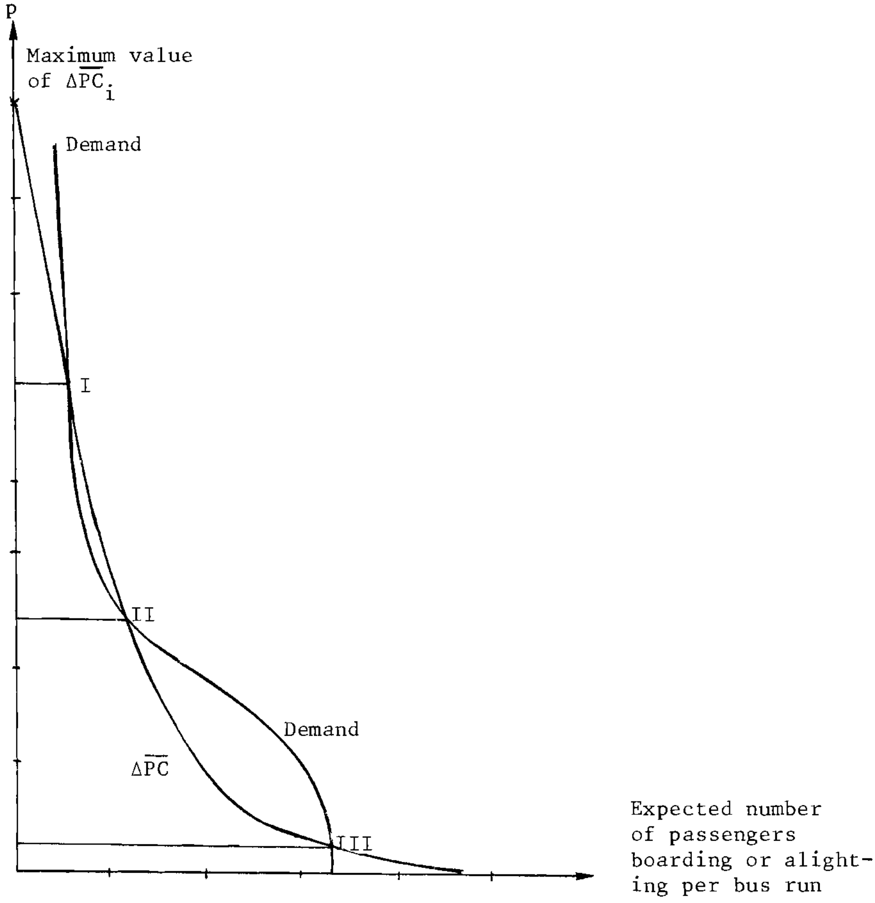


Figure 6:1. The shape of the $\overline{PC}_i$-function, and a hypothetical example of multiple points of intersection with the demand curve

In the present situation where there is no fare differentiation between stops, the only known point on the corresponding demand curve is its end-point on the horizontal quantity-axis. A lower limit for the optimal fare differential can consequently be given: it is safe to charge a fare at least equal to the vertical distance between the $\Delta\overline{PC}_i$-curve and the point on the quantity-axis representing the expected number of passengers at present demanding the i^{th} stop per bus run. If the demand is price-elastic, a considerably higher fare is of course conceivable. A "prohibitively high" fare may be optimal for particularly little frequented stops, which means either that the applicable fare will hardly ever be paid, and thus the stop almost never used by anyone, or that the stop is simply abolished.

As a final word of caution, it may be appropriate to point out that a sharply falling pricing-relevant cost curve poses certain problems for pricing in the normal case where the shape of the relevant demand curve is only vaguely known. Some idea of the shape of the demand curve can be obtained from the fact that for all potential users of the i^{th} stop, the relevant alternative is to use one of the two adjacent stops. The maximum differential anyone would be prepared to pay is equal to the product of the cost of walking and the extra walking distance incurred as a result of using the second-nearest stop. Both these factors vary greatly between individuals. Provided that the spread is unskewed as regards the cost of walking as well as the required extra walking distance, it can easily be shown that a demand curve of the shape illustrated in Figure 6:1 can be expected. The inflexion point of the curve corresponds to the product of the mean walking cost and the mean walking distance. The exact appearance of the curve should of course not be taken very seriously. The point I want to make by drawing this curve is simply that, given the shape of the $\Delta\overline{PC}_i$-function, we should be aware of the danger of arriving at a local net benefit maximum point, or worse, at a minimum point. For example, a rather high fare (corresponding to point I in Figure 6:1) may only leave those users of the i^{th} stop who put an extremely high value on the cost of walking, and/or who have to walk the maximum distance in order to get on or off the bus at the second-nearest stop. As is illustrated, this may be only a local net benefit maximum point. A somewhat lower fare differential (corresponding to point II in Figure 6:1) may also lead to a point of equilibrium situated on the $\Delta\overline{PC}_i$-curve, but it can represent a still worse situation, namely a net benefit minimum point.

The appropriate fare differential in the example corresponds to point III in Figure 6:1. However, taking the advantages of a flat fare (including a zero fare) into account, such a small fare differential does not seem worth the extra fare collection costs (or rather, the social surplus loss of not differentiating the fares is unlikely to be sufficiently great to justify a fare differential).

The financial effect of introducing fare differentials for using less frequented stops would be quite small. By definition, at present only very few passengers use the stops which would be liable to an extra charge, and it is not unlikely that almost prohibitively high fare differentials are optimal for some of the markedly less frequented stops.

6.7 SUMMARY AND CONCLUSION

We have discussed optimal pricing of urban bus services in a simple model of a bus line. The two key assumptions of the model which, however, are well borne out by empirical studies, are (i) that the mean waiting time of passengers is equal to half the bus headway, and (ii) that the bus and crew cost per service day is given by a relationship $a + bS$, where S is the bus size, and a and b are constants which can assume two different values, depending on whether the bus is in all-day or peak-only service.

Although the optimal design of bus lines as regards frequency of service, bus size, stop spacing, etc. is likely to vary greatly between different routes, certain general characteristics of the level and structure of optimal bus fares have been found. Needless to say, outside design optimum, "optimal" bus fares can be at almost any level. The basic philosophy of the present discussion, however, is that there is no real lasting "fixedness" in the design of a bus line, which can make it interesting to discuss "interim" tariffs of bus fares (i.e. "optimal bus fares on the way to a design optimum").

It is not usually possible to specify the optimal price level simply by analysing the cost structure. This can only be done if the pricing-relevant cost happens to be constant. A major conclusion of the present discussion of optimal bus fares is that the pricing-relevant costs are in fact constant, so far as the occupation of space on the bus is concerned.

The occupancy charge for traversing the "critical section" in the busiest rounds ($\widehat{PC}_{11}$) assumes the following shape:

$$\widehat{PC}_{11} = \frac{b}{2\hat{\phi}_{max}}$$

provided there is only one busiest round in each peak period. If there are two equally busy rounds in each peak period, the average peak occupancy charge can be described as:

$$\widehat{PC}_{1} = \frac{b}{4\hat{\phi}_{max}}$$

while the ratio of $\widehat{PC}_{11}/\widehat{PC}_{12}$ depends on the cross-elasticity of the demands of the two successive peak rounds. If the marginal bus is a peak-only bus, a slight modification of these two values should be made.

In all rounds in which the capacity constraint is non-binding, the optimal occupancy charge is zero.

Another pricing-relevant cost is caused by the boarding/alighting of passengers. The corresponding b/a-charges are critically dependent on the spacing between stops and the distribution of demand along the route.

We have distinguished three cases in this connection:

1. "The ordinary case" of a fixed number of stops where the distribution of the demands for individual stops is fairly even, so that it rarely happens that *no* one wishes to board or alight at a particular stop.
2. A variation of the ordinary case occurs when a fixed number of stops includes well-frequented stops interspersed by a few less frequented stops, which can sometimes be by-passed.
3. An individual stopping service.

In all cases the b/a-charges vary considerably both in time and space (along the route). The salient feature of the ordinary case, however, is the generally low level of these charges. Fare differentiation corresponding to the variations in the boarding/alighting costs, either requiring a conductor on the bus or leading to a doubling of the boarding/alighting times, is out of the question; the allocative benefits simply do not outweigh the fare collection costs. Only two alternatives seem worth considering: (i) either the b/a-charges are disregarded altogether, which means that except for the busiest rounds, the service is offered free of charge, or (ii) a flat fare is charged, amounting to 1-3p per trip outside the busiest rounds. An incentive to buy tickets at news-stands etc should be given. Regular travellers should be encouraged to buy a monthly "green card" at a price of, say £1, which entitles the bearer to unlimited travelling except across the critical section in the busiest rounds.

It can be argued - admittedly somewhat paternalistically - that a nominal fare to discourage travelling "for fun" or mischief is preferable to a zero fare, despite the fact that even the simplest flat fare system is not without cost.

Under certain circumstances it may be worthwhile to charge a fairly high extra fare to passengers using the much less frequented stops. Although the extra fare should be *almost* so high as to be prohibitive, it could serve an allocative purpose (as compared with cutting out the stop), provided passengers set very different values on their walking time.

Heavy subsidization of urban bus companies would be required if these fares were applied. In the ordinary case the financial deficit would equal the whole of those bus costs that are independent of size, i.e. aN (including all crew costs); if a nominal flat fare is charged outside the busiest rounds, the financial deficit would be slightly less. If some of the less frequented stops were liable to an extra charge, this would make very little difference financially. An individual stopping service, either partial or total, is a different matter. The total revenue from b/a-charges would make a substantial contribution to the recovery of the total costs of a liner taxi service.

This chapter started from the question of whether the present level of bus fares is generally above or below the pricing-relevant costs. It is well known that the operation of public transport is becoming an increasingly heavy burden on tax-payers in towns and cities. Nevertheless, the analysis we have made here suggests that present bus fares are inoptimally high, at least so far as off-peak periods are concerned. If off-peak fares were actually reduced to zero or almost zero, total revenues would fall by about half. This is one rough measure of the severity of the conflict between the goals of allocative efficiency and financial self-support in urban bus transport.

APPENDIX: DERIVATION OF THE PRICING-RELEVANT COST OF BOARDING/ALIGHTING IN THE BUSIEST ROUND

The distinguishing characteristic of $\overline{PC}_{11}$ is that the corresponding boarding/alighting makes demands on capacity. In the case of an optimized bus service, the capacity constraint should be binding in the busiest round, which means that additional boarding/alighting reduces the capacity to transport passengers across the critical section. If the boarding rate is rising by 1/t passengers per hour, while the passenger flow in the critical section remains unchanged, an additional bus has to be put in service in order to meet the capacity requirement. Suppose that it is decided to put on another all-day bus. The incremental cost in that case is equal to IC. The incremental benefit to the original passengers, however, is somewhat lower than IB. The passengers travelling outside the busiest round will enjoy the full benefit of an additional bus. The original passengers in the busiest round, on the other hand, will get no benefit at all and will even be subject to some disbenefit. The total additional boarding/alighting time caused is just sufficient to offset the input of another bus, as far as the frequency of service is concerned. There will thus be no savings in waiting time during the busiest round. Riding time will in fact be longer, because on an average more time will be spent at stops. The total increase in riding time for the original passengers in the busiest round is calculated as follows:

The average riding time per passenger can be expressed:

$$AC_{11}^{ride} = \frac{ckNT_{11}}{N-tB_{11}} \qquad (A.1)$$

On the assumption that the proportions between number of boarding passengers and passenger flow remain constant ($Q_{11}/B_{11} = k$), the total differential of the average riding time cost assumes the following shape:

$$dAC_{11}^{ride} = dB_{11} \frac{ctkNT_{11}}{(N-tB_{11})^2} - dN \frac{ctQ_{11}T_{11}}{(N-tB_{11})^2} \qquad (A.2)$$

Setting $dB_{11} = 1/t$ and $dN = 1$ gives:

$$dAC_{11}^{ride} = \frac{ckT_{11}}{N-tB_{11}} \qquad (A.3)$$

This has to be multiplied by the total number of passengers travelling in the busiest round in order to get the total increase in the stop-time cost.

$$R_{11}B_{11} \cdot dAC_{11}^{ride} = R_{11} \frac{cQ_{11}T_{11}}{N-tB_{11}} \qquad (A.4)$$

It is now clear that $\overline{IB}$, the incremental benefits to the original passengers from the addition of an all-day bus following an increase of R_{11}/t passengers in the busiest round (all travelling outside the critical section) comes to:

$$\overline{IB} = IB - \frac{2R_{11}B_{11}T_{11}(\frac{v}{2} + ctQ_{11})}{(N-tB_{11})^2} - 2R_{11} \frac{cQ_{11}T_{11}}{N-tB_{11}} \qquad (A.5)$$

Assuming that only one busiest round occurs in each peak period, the pricing-relevant cost of boarding/alighting in the busiest round can be defined as:

$$\overline{PC}_{11} = \frac{(IC-\overline{IB})t}{2R_{11}} \tag{A.6}$$

Inserting the above value of $\overline{IB}$, and recalling that $IC-IB = bNS/(N-tB_{11})$, and $R_{11} = NT/(N-tB_{11})$, we will find that $\overline{PC}_{11}$ comes out as:

$$\overline{PC}_{11} = \frac{tbS}{2T} + \frac{vtB_{11}T_{11}}{2(N-tB_{11})^2} + \frac{ctQ_{11}NT_{11}}{(N-tB_{11})^2} \tag{A.7}$$

Each of these three terms has a straightforward interpretation: the left-hand term should be compared to $\widehat{PC}_{11}$. The ratio of $\widehat{PC}_{11}$ to the left-hand term of $\overline{PC}_{11}$ comes to $t\phi_{max}$ S/T, which is equal to the ratio of dQ^*_{11}/dB^*_{11} as obtained from the capacity constraint (5) above. The left-hand term can consequently be interpreted as the pricing-relevant "capacity cost" of boarding/alighting in the busiest round. The middle term and the right-hand term represent the additional pricing-relevant costs of waiting time and riding time imposed on the original passengers travelling in the busiest round, as can be confirmed by a comparison with (34) and (37) above.

ALPHABETICAL LIST OF BASIC SYMBOLS

a	Size-independent bus cost
b	Proportionality constant of the size-proportional bus cost component
B	Number of passenger trips
c	Value of riding time
D	Round trip distance
E	Extent of diurnal service time period
F	Frequency of service
H	Headway between buses
IB	Incremental benefit to the original passenger of another bus
IC	Incremental cost of another bus
m	Number of stops on the route
n	Number of round trips per day of a bus in all-day service
N	Number of buses
p	Probability
PC	Pricing-relevant cost
Q	Passenger flow
r	Running cost per bus-km.
R	Round trip time
s	Bus standing cost
S	Bus size
t	Boarding/alighting time
T	Fixed component of the total round trip time
v	Value of waiting time
w	Crew wage rate
X_i	Number of boarding and alighting passengers at the i^{th} stop
ϕ	Bus occupancy rate

7 LINER SHIPPING COSTS IN THIN TRADES AND DENSE TRADES

The very significant economies associated with the volume of traffic in scheduled passenger transport mean, first of all, that individual modes of transport - in particular the private car - where there are practically no economies of scale, offer increasingly strong competition to public transport as the density of demand falls. It is generally very difficult to make public transport viable in sparsely populated areas. Secondly, it has been found that social optimal pricing of urban bus transport services would require heavy subsidization.

It is interesting to examine whether the same sort of "thin route" problems face scheduled freight transport. On a basis of the "abstract mode" model of Chapter 4, it was predicted that scheduled freight transport, too, is a decreasing-cost activity, although to a lesser extent than scheduled passenger transport. Two main reasons for this were identified:

1. In the case of freight, the *value of time* in relation to the cost of transport capacity is far smaller than the value of passenger time in relation to the cost of transport capacity.

2. Freight waiting costs due to infrequency of service mainly assume the form of *storage costs* and, unlike passengers' waiting time (at least in urban transport), the storage time necessary for freight going by liner

services is not nearly proportional to the headway (i.e. the interval between sailings). Instead, a "square-root law" seems to apply to this relationship.

Our purpose now is to examine how these general conditions find expression in a specific mode of transport. In this and the following chapter scheduled freight transport by sea will be discussed. In the present chapter a model of a liner shipping trade route is introduced, which will give us some idea of the economies of trade volume that obtain in this sector of the freight transport industry.

7.1 THE INTERNATIONAL SHIPPING-LINE INDUSTRY[1]

The market for liner shipping services include a few very dense routes and numerous thin routes. The bulk of the seaborne trade carried by liners moves between developed or industrial countries, with a marked concentration to a belt that includes Europe, North America and Japan. The flows of liner cargo between developed and developing countries are much thinner, and only minor quantities of general cargc move between developing countries. The question is whether there is a thin route problem in liner shipping, comparable to that in public passenger transport. To get a clear view of the competition facing liner shipping, it should be borne in mind that a large proportion of total general cargo moves in intercontinental trades and requires ocean shipping. In the case of intercontinental trade, land transport alternatives are with the notable exceptions of the Trans-Siberian Railway and the "land-bridge" across North America practically non-existent, and air cargo transport is competitive for a minor part of this trade only. The main competition comes

[1] For an economic-geographical description of liner shipping supply and demand, see Lawrence, S.A., International Sea Transport: the years ahead. Lexington Books, D.C. Heath & Co., 1972, and von Schirach-Szmigiel, Ch., Liner Shipping and General Cargo Transport. EFI, 1979.

from other forms of shipping. Keen rivalry between "common carriers" and individual transport is a salient feature of freight as well as passenger transport, and cargo shipping is no exception. Liner shipping has long been in competition with tramp shipping for shipments of materials such as wood pulp and copper, and various new forms of bulk services have recently been making inroads into traditional liner shipping business.

In view of this, it would be interesting to examine whether intercontinental liner shipping between Africa, South America, southern Asia, Australia, northern Europe, etc, where the density of the general cargo trade is relatively low, is at any great disadvantage as compared to liner shipping in the main trades of the world, or, in other words, whether significant economies of trade density exist in liner shipping. The existing form of organization in the liner shipping industry suggests that economies of scale will be less significant than in passenger transport. Unlike urban bus transport, for example, each particular market - "liner trade" - can normally accommodate more than one shipping line, and the main routes of the world are each plied by a considerable number of different shipping line.

On the other hand, this does not mean very much, in view of the fact that the liner shipping industry is largely cartelized. Almost every trade route in the world is covered by a separate coalition of shipping lines known as liner shipping "conferences" which fix freight rates and regulate - in different ways and to greatly differing degrees - the supply of shipping capacity on each particular route.

The coexistence of vehicle-size economies in the producer costs and vehicle-number economies in the user costs is the root cause of the inherent density-of-demand economies in scheduled transport services. Let us look at these two kinds of economies in liner shipping in turn.

7.2 MULTI-PORT CALLING AS A WAY OF MITIGATING THE THIN TRADE PROBLEM

For the two reasons mentioned at the outset, a low frequency of service is not such a serious disadvantage in freight transport as it is in passenger transport. A low *density* of service, on the other hand, is very disadvantageous in both cases. Even in deep-sea trades, land feeder transports (to/from the ports of loading/unloading) can cost as much as the sea transport itself, although the latter may be ten, twenty or more times as long as the former. It is a primary desideratum of users of liner shipping that the density of service be maintained at an adequate level even in very thin trades. The main routing technique for meeting this demand is multi-port calling. This technique can be successfully applied in freight transport on account of the relatively low value of transit time for goods as opposed to human beings. For instance, in urban bus transport the radial or diametrical bus lines can be characterized as "shuttle services", meaning that only minor deviations are made from a straight line between the starting-point and end-point of a route. As the density of demand decreases, a tendency to reduce the density of the bus lines can be observed, while maintaining the shortest-distance principle in the routing of each individual line. The reason why it is generally inoptimal to make any more substantial deviations from the shortest-distance route, is that the value of the riding time of commuters is relatively high. Suburban and rural bus lines are not always arranged as shuttle services, but can sometimes be designed in a criss-cross to cover a fairly wide area. Such services are no doubt perceived as "low quality" by the travellers, but the alternative of leaving some parts of a region without access to any service is no better.

In scheduled freight transport the possibility of increasing the cargo base of a line by substantial diversions is widely exploited. The value of time of the cargo is not usually the major restraint on the extent of these diversions; it is rather the value of the actual vehicle time (including crew time).

So far as transoceanic liner cargo services are concerned, the ships can call at up to ten ports in a range as wide as Vancouver to Long Beach or Hamburg to Le Havre, for example, before embarking upon the sea crossing.

The question is whether liner shipping is much more expensive in a thin trade than in a dense trade where a pattern of shuttle services between a single pair of ports can be viable. At a superficial glance, significant economies of trade density may seem obvious. In a thin trade it may be necessary to extend the cargo catchment area at each end to embrace almost a whole continent in order to maintain a reasonable frequency of sailings. This may involve very appreciable deviations from the least-shipping-cost pattern of the shuttle service.

However, as will be illustrated later, provided that the cargo concerned is not excessively valuable, the increasing length of the round voyage required for picking up part cargoes at a number of ports lying far apart is not as expensive as it first seems. A more important disadvantage is probably the inevitable transitional time (dead time before the start of cargo handling) incurred at every port of call.

The multi-port calling principle should not be carried to extremes. It is true that the density of service[1] can be maintained at any level, regardless of the number of shipping lines that can be supported in a particular trade, simply by increasing the ports-of-call per round voyage. However, this will be rather time-consuming for the shipowner in the end, and will of course also increase the mean transit time,

[1] The density of service is measured by the number of ports connected by liner shipping in a given trade.

while the benefit of increases in the frequency of service will diminish as the coastwise diversions are extended.

An alternative way of organizing liner shipping services is to introduce a system of sea feeder transport. This would make it unnecessary for the deep-sea liners to waste time on multi-port calling, since they could concentrate on the sea-crossing link. Ship size economies can be exploited to the maximum by using fairly large ships throughout for long-distance transport and smaller ships for supporting short-distance feeder transport.

Traditionally this type of division of labour has not been applied to liner shipping because of the double handling that would be required. For break-bulk cargo the handling costs (stevedoring charges plus the cost of the ship's time in port) are normally higher than the hauling costs: to incur handling costs twice over would, under practically any circumstances, be out of the question. In the wake of containerization, and with the spread of roll-on/roll-off ferry traffic to deep-sea routes, the conditions for sea feeder transport systems have indeed been introduced in some places. It is an open question, however, whether this will be the pattern for the future, or whether traditional multi-port calling, involving lengthy diversions at either end of the route, will continue to dominate. To organize a comprehensive sea feeder transport system - making schedules fit, etc - can be rather more complicated than running multi-port calling services.

The analysis in this chapter is restricted to a consideration of the economies of scale in liner shipping on the assumption that multi-port calling is the best way of securing a sufficient cargo base for liner services.

Besides the number of ports-of-call, the principal liner-service design variable is the size of the ship. The relationship between shipping costs and ship size will now be examined.

7.3 SHIPPING COSTS AND SHIP SIZE

The second prerequisite for economies of density of demand in scheduled transport services is the existence of economies of vehicle size. In this section some findings concerning the relationship between shipping costs and ship size will be presented. To put this into perspective, our earlier findings as regards the relationship between bus transport costs and bus size can be recalled.

The attractive simplicity of the bus line optimization model is due to the fact that the total bus cost per day can be described by this simple linear relationship:

$$a + bS_{bus}$$

The increase in bus size mainly takes the form of *lengthening* the bus and, as is well known, manning requirements are independent of bus size. Hence there is a linear relationship between cost per day and bus size, where the constant *a* dominates over the size-proportional element, bS. To obtain the relationship between cost per passenger and bus size, we have to take into account that the transport capacity of a bus is not quite proportional to bus size. There are pronounced diseconomies of bus size in the boarding/alighting operation. However, since this operation is relatively unimportant for buses in the usual size range, these diseconomies play a secondary role in the case of bus transport.

Neither of these basic facts is applicable to cargo ships.[1] For several reasons the relationship between cost per ship-day and ship size is better explained by an exponential function.

[1] The following conclusions regarding ship size and shipping costs are based on Jansson, J.O. and Shneerson, D., Economies of Scale of General Cargo Ships. The Review of Economics and Statistics, May 1978, which in turn is a summary of Chapter 3 in a forthcoming book "Liner Shipping Economics" by the same authors.

aS^b_{ship}

A ship grows in all three dimensions so to speak, and the relationship between capital cost and ship size seems to be a reflection of the geometric principle that the surface area of a vessel is proportional to the volume, raised to the power of 2/3. Manning requirements are not independent of the size of the ship (although the coming automation of shipping may make it so), but increases slowly with increases in S.

The elasticity with respect to ship size of the costs of the main inputs are given in Table 7:1. Factor costs are numbered in order of importance.

Table 7:1. Elasticity of factor costs with respect to ship size

Factor costs	Size-elasticity
Crew and other operating costs except fuel, f_1 (S) dollars per ship-day	.43
Capital costs, f_2 (S) dollars per ship-day	.66
Fuel costs, f_3 (S) dollars per ship-mile	.72
Interest cost of cargo in transit, f_4 (S) dollars per ship-day	1
Berth occupancy charges on the ship, f_5 (S) dollars per ship-day	1

The fourth factor cost item represents an input on the part of the shippers. Given the load factor, it is clear that the interest cost of the cargo is proportional to ship size. The fifth item represents certain inputs provided by the ports, the costs of which appear as charges payable by the shipowner. Port occupancy charges are traditionally levied to recoup the capital costs of port approaches, docks, etc. These charges are interesting in the present connection, because they are normally proportional to ship size (gross or net registered tonnage). But, in fact, it is doubtful whether there is a very close correspondence between these port charges and the real cost of the port resources used by ships. As the fifth factor cost item is fairly small in relation to the other factor costs, this problem will be ignored in the present analysis.

Total port charges also include stevedoring charges which, like some of the charges levied by the port authority, are paid per ton of different types of cargo. These charges are consequently invariant to ship size, and will appear in the model as a wholly "neutral" total cost item designated as "cargo cost".

To get a more complete picture of the economies of ship size, the size-elasticities of the factor costs should be set against the size-elasticity of shipping capacity. A fundamental point here is that total shipping capacity depends on both the "hauling capacity" (ton-miles per day), and the "handling capacity" (tons loaded/unloaded per day) of a ship. The former is the product of the holding capacity and the cruising speed. This product develops progressively, although only slightly, with increases in ship size, because speed tends to rise somewhat as ship size increases. The latter develops degressively, and markedly so, as ship size increases. A salient feature of liner shipping is that handling operations are relatively time-consuming for ships of practically all types and sizes. This is particularly marked in the case of break-bulk cargo shipping. Conventional liners generally spend at least half their time in port, and container ships which are specially designed to reduce the handling time nevertheless spend 20-30 per cent of their time there. The inherent diseconomies of vehicle size, so far as the handling operation is concerned, are therefore quite important in liner shipping. The "1/3-rule" identified by Thorburn (1960) in his pioneering study of sea transport,[1] postulating that the quantity of cargo loaded or unloaded per hour is proportional to the *length* of a ship of a given type, now appears to represent an upper limit for the development of handling capacity, as ship size increases.

A comparison of the size-elasticities in Tables 7:1 and 7:2 shows that ship-size economies prevail so far as the hauling operation is concerned. whereas ship-size diseconomies are the rule in the handling operation. In no case does the size-elasticity of the factor costs incurred at sea exceed the hauling capacity size-elasticity, while the opposite is true as regards the size-elasticity of the factor costs incurred in port and the size-elasticity of the handling capacity. Consequently, the hauling cost *per ton-mile* falls, and the handling cost *per ton* increases, with increases in ship size.

Table 7:2. Ship-size elasticities of shipping capacity

Capacity	Size-elasticity
Hauling capacity, S · V(S) ton-miles per day at sea	1 1/6
Handling capacity, L(S) tons loaded/unloaded per day in port	1/5 - 1/3

We can now proceed to use the above relationships between shipping cost and ship size in a total cost model of a liner trade, in which the least-cost service design is determined for different volumes of cargo.

[1] Thorburn, T., Supply and Demand of Water Transport. FFI, 1960.

7.4 A LINER TRADE MODEL

The purpose of the following model is to give us some idea of the economies of trade density in liner shipping. The main problems of the analysis concern the limitations of the relevant system in various dimensions.

The first question is to define the relevant markets, and this in turn will define the density of demand for a particular transport service. The model is adapted in the first place to intercontinental liner shipping. This implies, typically, that the trade between two opposite continental *coasts*, such as the Pacific coast of North America and East Asia (including Japan), or the west coast of Africa and the east coast of South America, constitutes the total shipping demand of a particular market. The density of demand is then defined as the trade volume per kilometre of coast. This affords an admittedly imperfect demarcation of the system, unless clearcut geographical boundaries happen to exist as they do for liner shipping between Australia and New Zealand, for example. In other cases some overlapping of markets is inevitable in reality, but there is no better alternative than to regard the different liner trades of the world as separate markets.

Another problem of system demarcation concerns the question of how far "upstream" the cost analysis should be taken. In a narrow sense shipping service can be defined as the carriage of cargo from the quay of port A to the quay of port B. It has been argued in the previous general discussion that both the feeder transport costs and the waiting costs should (as a general rule) be included in the cost analysis of scheduled transport systems. In the present case this would mean that all the costs of transport - *door-to-door* and storage costs at the premises of consignors and consignees - should be taken into account. However, a departure from this general rule will be made in the following model. The ports of loading and discharge for each individual consignment will be treated as given. Under this condition, inland feeder transport costs (to the port of loading and from the port of discharge) will be constant. It would then be inappropriate to take the inland transport costs into consideration, in any calculation of the elasticity of social costs with respect to trade volume. Adding a cost item which is independent of the trade volume *by assumption*, would prejudice the resulting elasticity. Only feeder transport operations undertaken by the liners themselves in the form of coastal diversions will be taken into account.

A third problem is that it cannot be assumed that one shipping line only serves each individual trade. However, in most trades the liner conferences act as a coordinating body. In the model it is assumed that the design of the shipping services fulfils the efficiency condition that each level of output is produced at the least total social costs. This is a necessary condition both for net social benefit maximization, and for total profit maximization. In other words: it is assumed that the liner conference/price cartel acts in accordance with the objective of maximizing the total profits of all members.

Given the above limitations and assumptions, the model is constructed as follows: consider the trade in general (liner) cargo between two continents separated by a sea; the problem is to design the liner shipping services in such a way as to minimize the total social costs of transporting - in the widest sense - the total volume of trade between a given number of ports - $\bar{m}_1$ at one end, and $\bar{m}_2$ at the other.[1]

[1] The number of ports to be included in the services, $\bar{m}_1$ and $\bar{m}_2$ will be treated as given in the present model. Experiments with a more comprehensive model have shown that these design variables are only slightly dependent on the density of trade. Fixing values for m_1 and m_2 will not thus have much effect on the result regarding the magnitude of the economics of scale. The main determinants of $\bar{m}_1$ and $\bar{m}_2$ are the inland transport costs, and the transitional time of ships per call. There is little dependence only between $\bar{m}_1$ and $\bar{m}_2$ and the other design variables. (continued on next page)

The setting is homogenized in that the same quantity of cargo is assumed to be generated between each pair of ports. The total demand for shipping consists of the trade flows in both directions. Very few trades are well balanced. It is therefore appropriate to allow for an export/import imbalance (in measurement tons), and designate:

$Q = X + M$ = total cargo flow

$X = Q/\mu$ = "export"[1]

$M = (1-1/\mu)\ Q$ = "import"

The measure of trade balance, μ, assumes values between 1 and 2. The value of the lower limit indicates that the ships go in ballast in one direction, and the value of the upper limit indicates that the trade is perfectly balanced.

In spite of the fact that the ports served by the conference liners are given, the itineraries of the ships can be fixed in widely differing ways. At one extreme each ship could call at every port on each round voyage, which would require the maximum coastal diversion at each end, but which would also give the highest possible frequency of sailings between each pair of ports (given the ship size). It is not necessary, of course, to call at every port on each round trip. Any number of ports-of-call between $\bar{m}_1 + \bar{m}_2$ and 2 per round trip is possible. However, the fewer the ports-of-call, the lower will be the frequency of sailings between each pair of ports.

(continued from foregoing page)
The choice regarding ports-of-call is thus restricted to the question of how many of the given ports should be visited at each sailing. Needless to say, the different shipping lines should not call at the same sub-set of ports, but should divide the total market between them so that the conference most conveniently covers the whole trade.

[1] The convention is adopted of calling the trade flow on the flat leg for "export".

The other trade-off is between size and number of ships: given the itinerary of the ships, as the ships increase in size, so will the shipping cost per ton drop and the shippers' storage costs rise. Consequently, the design variables to be optimized are the following:

n = number of round trips per year in the trade concerned
S = ship size
m_1 = number of ports-of-call at one end per round trip
m_2 = number of ports-of-call at the other end
F = frequency of sailings between each pair of ports

7.4.1 Total producer and user costs

The shipping costs directly borne by the shipowners are grouped into (i) those costs which are time-proportional and incurred all the time, i.e. crew and capital costs, (ii) fuel costs which are incurred only during hauling time, which is proportional to the sailing distance, and (iii) berth occupancy charges which are assumed to be time-proportional, and which are, of course, incurred only in port. The annual total costs of the first group are thus proportional to the product of the total number of round voyages, n, and the round voyage time, R; the total fuel costs are proportional to the total miles sailed by all ships, nD; and the total port charges are proportional to the total port time of all ships, which is equal to twice (loading and unloading) the total trade flow, Q, divided by the handling capacity, L, which - given the port productivity - is a function of the ship size.

The total costs of the shipping lines engaged in the trade can thus be written:

$$TC^{prod} = nR\left[f_1(S) + f_2(S)\right] + nDf_3(S) + \frac{2Q}{L(S)} f_4(S) \qquad (1)$$

where the round trip time

$$R = \frac{D}{V(S)} + (m_1 + m_2)h + \frac{2Q}{nL(S)} \tag{2}$$

The three components of the round trip time are from left to right, the total hauling time at sea, the total transitional time (h = transitional time per call), and the total time in port.

The round trip distance is made up of two sea-crossings and the coastal cruising at each end, the length of which depends on the number of ports-of-call. Assuming for simplicity that the ports at each end are equi-distant, the diversion necessary to call at another port is always equal to d. Designating the coast-to-coast distance D_0, we thus have:

$$D = 2D_0 + \left(m_1 + m_2 - 2\right)d \tag{3}$$

For the shippers the total costs appear in the form of liner freight rates, certain port charges, inland transport costs, insurance and interest costs on cargo in transit, and storage costs. None of these cost items can be disregarded because they are relatively insignificant. However, given that for each individual shipment, the port of loading and discharge is always the nearest port served by liner shipping, the inland transport costs do not vary with the design variables of the model and need not be considered. The direct cargo-handling costs, i.e. stevedoring charges, and port charges on cargo, are also constant. However, in this case the constancy represents an inherent characteristic of the costs concerned, and is not due to any limitation of the analysis; it would be a bit misleading to leave them out.

Bearing in mind that cargo-handling charges are paid for loading as well as unloading, the total direct handling costs are written:

Total cargo-handling costs = $2\bar{C}_0 \cdot Q$

The interest cost on cargo in transit is a size-dependent cost which can generally be written as the product of the total input of ship-days per year in the trade multiplied by the interest cost of an average cargo per day.

Total interest costs on cargo in transit = $nRf_4(S)$

The storage cost, finally, belongs to another category: this cost is a function of the frequency of sailings between each pair of ports, F. As in the case of cargo costs, it can be assumed that storage costs are incurred by both exporters and importers.

Total storage cost = $2C_1(F) \cdot Q$

Summing up the costs borne by the shippers, we get the total user costs:

$$TC^{user} = 2Q\left[\bar{C}_0 + C_1(F)\right] + nRf_4(S) \qquad (4)$$

Apart from the level of freight rates and cargo costs, the questions of greatest concern to the shippers are apparently the frequency of sailings between each particular pair of ports and the transit time. In a given trade, the average frequency of sailings, F, is determined by the ship size and the number of ports-of-call per round trip. The total number of ports at each end is $\bar{m}_1$ and $\bar{m}_2$, respectively. This means that there are $\bar{m}_1\bar{m}_2$ *pairs* of ports to be served. On each round trip, m_1m_2 pairs enjoy service. The average frequency of sailings is therefore equal to the total number of round trips per annum, n, multiplied by the ratio of m_1m_2 to $\bar{m}_1\bar{m}_2$.

$$F = n \frac{m_1 m_2}{\bar{m}_1 \bar{m}_2} = \frac{X m_1 m_2}{\phi S \bar{m}_1 \bar{m}_2} \qquad (5)$$

where ϕ is the mean load factor (occupancy rate) on the export leg, and thus ϕS is the practical holding capacity. Adding total producer costs according to (1) and total user costs according to (4), and dividing by Q, gives us the following expression for the total social cost per ton:

$$AC = 2\bar{C}_0 + 2C_1(F) + \frac{nR}{Q}\left[f_1(S) + f_2(S) + f_4(S)\right] + \qquad (6)$$

$$+ \frac{nD}{Q} f_3(S) + \frac{2}{L(S)} f_5(S)$$

Inserting $X/\phi S$ for n, and the aforementioned expression (2) for R, and rearranging the terms, yields the following result:

$$AC = 2\bar{C}_0 + 2C_1(F) \quad + \frac{\left[f_1(S) + f_2(S) + f_3(S) \cdot V(S) + F_4(S)\right]}{\mu \ \phi \ S \ V \ (S)} +$$

$$+ \frac{(m_1 + m_2)h \left[f_1(S) + f_2(S) + f_4(S)\right]}{\mu \ \phi \ S} + \qquad (7)$$

$$+ \frac{2 \left[f_1(S) + f_2(S) + f_4(S) + f_5(S)\right]}{L \ (S)}$$

The five terms of (7) represent from left to right: the cargo-handling cost per ton, the storage cost per ton, the hauling cost per ton, the transitional cost per ton, and the indirect handling cost (i.e. the cost of ship's time in port) per ton.

7.4.2 Simplifying approximation

To allow quantification of the cost relationships involved, certain approximations have to be made regarding the combined size-elasticities of the costs per ton. As has been mentioned, all the relationships between factor cost and ship size, as well as those between hauling and handling capacity and ship size, are satisfactorily explained by exponential functions. The relevant size-elasticities are given in Tables 7:1 and 7:2 above. An interesting peculiarity of these values, and one which is very useful to the present analysis, is that the combined size-elasticity of the hauling cost per ton, and the combined size-elasticity of the transitional cost per ton, are close to - 1/2, while the combined size-elasticity of the handling cost per ton is nearly 1/2.

Given a wage rate for seamen on a level with that applying in the United States and the Scandinavian countries, and assuming that the average value of the cargo is low to middling, the weighted average size-elasticity of the factor costs incurred per day at sea comes to about .65. The weighted average size-elasticity of the factor costs per day of transitional time is only about .5 (since no fuel cost worth mentioning is incurred). The corresponding elasticity of the costs in port is almost .7 on the assumption that the port charges are proportional to both time and tonnage. Subtracting the size-elasticity of the hauling capacity from the first-mentioned figure, the size-elasticity of the holding capacity (i.e. unity) from the second figure, and a value in the low range of the given span of handling capacity size-elasticities from the third figure, the stated result is obtained. If this is to provide a good approximation, it is essential that the crew cost should be the most important item and that the cargo interest cost should be a far less important factor cost item. Where there is a relatively low wage level, and where the value of the cargo is notably high, the two former combined size-elasticities are absolutely less than, and the latter is somewhat greater than, 1/2.

This approximation is particularly helpful in view of the fact that a "square-root law" seems to apply also to the relationship between storage cost and frequency of service.[1] Given that $F = nm_1m_2/\bar{m}_1\bar{m}_2$, the complete approximate version of expression (7) for the social cost per ton can then be written as follows:

$$AC = 2\bar{C}_0 + 2\alpha_1\sqrt{\frac{\bar{m}_1\bar{m}_2\phi S}{m_1m_2\ X}} + \frac{\alpha_2\ D}{\mu\phi\sqrt{S}} + \frac{\alpha_3(m_1+m_2)h}{\mu\phi\sqrt{S}} + 2\alpha_4\sqrt{S} \qquad (8)$$

where $D = 2D_0 + (m_1 + m_2 - 2)d$

The proportionality constants, α_1, α_2, α_3 and α_4 are conglomerates of the proportionality constants of the factor cost functions (which in turn are determined by the relevant factor prices), the proportionality constants of the hauling and handling capacity functions and certain route characteristics such as the cargo-handling productivity of the ports. The most interesting route characteristics to consider explicitly are the total main-haul cargo volume, X, the mean load factor on the main haul, ϕ, the directional cargo balance, μ, the route distance (across the sea), D_0, the required coastal cruising distance per additional port-

[1] See Baumol, W.J. & Vinod, H.D., An Inventory Theoretical Model of Freight Transport Demand. Management Science, March 1970. Making somewhat different assumptions applicable to liner cargo shipping, it can be shown (Jansson, J.O. & Shneerson, D., Liner Shipping Economics, op. cit.) that the safety stock requirements of shippers in the Poisson case is not exactly proportional to the square root of the interval between sailings, but comes close to this in the range where the interval between sailings is not very great, and provided that the transit time is of moderate length.

of-call, d, the transitional time per call, h, and the total number of ports served by liner shipping at either end, $\bar{m}_1$ and $\bar{m}_2$.

The nature of the economic balance involved is very clear from this version of AC. Ship size, S, is the primary balancing factor. There are apparently very significant economies of ship size in the hauling cost as well in the transitional cost,[1] whereas there are equally important diseconomies of ship size in both the indirect handling cost and the storage cost. The second balancing factor is the extent of the diversion made at either end of each sailing, represented by the number of ports-of-call, m_1 and m_2. The coastal cruising distance is proportional to $m_1 + m_2 - 2$, and the transitional time per round trip is proportional to $m_1 + m_2$, while the storage cost is inversely proportional to the square root of $m_1 m_2$.

7.4.3 Optimal ship size, multi-port diversion, and frequency of sailings

The optimal economic balance is obtained by minimizing AC with respect to S, m_1 and m_2. Taking the partial derivatives concerned, and equating to zero gives us:

$$\frac{\partial AC}{\partial S} = \alpha_1 \sqrt{\frac{\bar{m}_1 \bar{m}_2}{m_1 m_2 \; XS}} - \frac{\alpha_2 \; (D_0 - d)}{\mu \phi S \sqrt{S}} - \frac{(m_1 + m_2) \; (\alpha_2 d + \alpha_3 h)}{2 \mu \phi S \sqrt{S}} + \frac{\alpha_4}{\sqrt{S}} = 0 \qquad (9)$$

[1] At first it may seem to run counter to intuition to postulate that lengthy diversions involving many ports-of-call favour large ships. Suppose that for every ship one day is lost in transitional time per call. One would surely expect this to be very expensive for the very large ships. But we have to remember that, counted per ton of cargo, it will be increasingly expensive, the smaller the ship.

$$\frac{\partial AC}{\partial m_1} = -\frac{\alpha_1}{m_1}\sqrt{\frac{\bar{m}_1\bar{m}_2S}{m_1m_2X}} + \frac{\alpha_2 d + \alpha_3 h}{\mu\Phi\sqrt{S}} = 0 \qquad (10)$$

$$\frac{\partial AC}{\partial m_2} = -\frac{\alpha_1}{m_2}\sqrt{\frac{\bar{m}_1\bar{m}_2S}{m_1m_2X}} + \frac{\alpha_2 d + \alpha_3 h}{\mu\phi\sqrt{S}} = 0 \qquad (11)$$

The optimal ship size S^{*} is solved by multiplying (9) by $2\sqrt{S}$, (10) by $m_1/\sqrt{S}$, and (11) by $m_2/\sqrt{S}$, and then adding the resulting equations.

$$S^{*} = \frac{\alpha_2\,(D_0-d)}{\alpha_4\,\mu\phi} \qquad (12)$$

The optimal ship size is proportional to the coast-to-coast distance minus "one diversion". A well-known characteristic of the structure of the world liner fleet is that short-sea liners are far smaller than deep-sea liners, and that in the latter category the ships operating on the North Atlantic, for example, are considerably smaller than ships used on routes between antipodes.

It is surprising that only the trade balance, μ, and not the volume of trade, Q, appears as a determinant of optimal ship size. The more balanced the trade, the smaller will be the optimal ship. This can be explained as follows: given the export volume, X, which determines the total shipping capacity, we will find that the greater the volume of imports, M, the higher will be the incremental benefit of an additional sailing. This means that the ship size and number combination should be increasingly "number-intensive". On the other hand, if X and M increase parallel to one another, there will be no effect on S^{*}. In an indirect way the trade volume can be said to influence S^{*}: namely, as Q increases, the optimal round trip distance decreases without causing any reduction in ship size.

From (10) and (11) it is clear, first, that $m_1 = m_2$ at optimum. Symmetrical itineraries including the same number of ports-to-call at each end are optimal. The optimal number of ports-to-call on each coast, $m^* = m_1^* = m_2^*$, can be calculated by inserting the value found for the optimal ship size in (10) or (11):

$$m^* = \left(\frac{\alpha_1 \alpha_2 \; (D_0 - d)}{\alpha_4 \; (\alpha_2 d + \alpha_3 h)}\right)^{1/2} \left(\frac{\bar{m}_1 \bar{m}_2}{X}\right)^{1/4} \tag{13}$$

The most important determinant of m^* is the distance between adjacent ports, d. The greater this distance, the fewer ports will be visited on each round trip.

It is an interesting fact that the sea-crossing distance, too, is an important determinant of m^*. This relationship also holds in reality. A common feature of liner trades is that, given the volume of the trade, the greater the coast-to-coast distance, the more extensive will be the service range. The explanation is roughly that a diversion of a given length, $\bar{d}$, results in an increase in the mean transit time and makes a further demand on shipping capacity, which will increase *relatively* as the coast-to-coast distance, D_0, diminishes.

The optimal frequency of sailings, F^*, is obtained from (12) and (13).

$$F^* = \frac{\mu \alpha_1}{\alpha_2 d + \alpha_3 h} \sqrt{\frac{X}{\bar{m}_1 \bar{m}_2}} \tag{14}$$

The optimal sailing frequency is proportional to the square root of the total export volume. As the trade density increases, the frequency of sailings does not increase commensurately; instead, the advantage of a greater trade volume is partly exploited by reducing the multi-port calling diversions.

7.4.4 Economies of trade density and the financial result of optimal pricing

By inserting the values for S^* and m^* in (8) above, we obtain the total cost per ton at optimum. It is an interesting coincidence that the total cost per ton proves to consist of, apart from the direct handling cost, two pairs of equal entities which can be identified as the storage cost ($2C_1$) and the "diversion cost" ($2C_3$ = the cost of coastal cruising and transitional time) on the one hand, and the sea-crossing cost ($2C_2$) and the indirect handling cost ($2C_4$) on the other. (Note that each cost is incurred twice for each shipment.)

$$AC^* = 2\bar{C}_0 + 2C_1 + 2C_2 + 2C_3 + 2C_4,$$

where

$$C_1^* = C_3^* = \sqrt{\alpha_1(\alpha_2 d + \alpha_3 h)}\left(\frac{\bar{m}_1\bar{m}_2}{\mu\phi Q}\right)^{1/4} \qquad (15)$$

and

$$C_2^* = C_4^* = \sqrt{\frac{\alpha_2\alpha_4\ (D_0-d)}{\mu\phi}} \qquad (16)$$

Looking at (15), we can see that Q is a determinant of the storage and diversion costs. As was previously found, the frequency of sailings increases with increases in Q, but not fully in proportion to the increase in Q. A parallel effect is that the coastal diversions will be less extensive.

Looking at (16), we can observe that the trade volume is not a determinant of the sea-crossing and indirect handling costs. The reason, as we have seen, is that the optimal ship size is independent of the trade volume. On the other hand, the trade balance, μ, has the effect of reducing these costs. So far as C_2^* is concerned, a better trade balance does reduce costs, despite the fact that the optimal ship size falls with increases in μ. The economy of a higher overall load factor, however, is a stronger force. It is also interesting to note that the elasticity of C_2^* with respect to the coast-to-coast distance, D_0, is as little as about 1/2. A fourfold increase in the route distance will only double the sea-crossing cost.

The total cost per ton can now be written:

$$AC^* = 2\bar{C}_0 + 4\sqrt{\alpha_1(\alpha_2 d + \alpha_3 h)}\left(\frac{\bar{m}_1\bar{m}_2}{\mu\phi Q}\right)^{1/4} + \qquad (17)$$
$$+ 4\sqrt{\frac{\alpha_2\alpha_4\ (D_0-d)}{\mu\phi}}$$

The extent of the economies of scale can be measured by the elasticity of AC^* with respect to Q. The first and third terms of AC^* are apparently independent of Q, while the elasticity of the second term with respect to Q is equal to - 1/4. The overall elasticity thus depends on the ratio of the second term to the total cost per ton. This ratio falls as Q increases. Consequently, the elasticity of AC^* with respect to Q starts at the minimum level of - 1/4, and gradually approaches zero as Q increases.

$$E = \frac{\partial AC^*}{\partial Q} \cdot \frac{Q}{AC^*} = -\frac{1}{4} \cdot \frac{C_1^* + C_3^*}{\bar{C}_0 + C_1^* + C_2^* + C_3^* + C_4^*} \qquad (18)$$

A more revealing expression for this elasticity can be obtained with the help of the equalities $C_1^* = C_3^*$ and $C_2^* = C_4^*$, by considering the elasticity in terms of the direct handling cost, the sea-crossing cost, and the diversion cost.

$$E = -\frac{1}{4}\;\frac{C_3^*}{\bar{C}_0 + C_2^* + C_3^*} \tag{19}$$

A value of E almost as low as - 1/4 is clearly outside the realistic range. It would mean the diversion cost completely dominating the sea-crossing cost and the direct handling cost, which is never the case in reality. We can get a good idea of a realistic range of values for E by taking levels for the individual cost items which are typical of existing liner services. As has been mentioned, the service range at each end can embrace the coastline of a whole continent. Nevertheless, it is very unusual for the coastal diversion to be as long as the sea-crossing leg of a round trip, because the really extensive service ranges are to be found only in connection with long-distance trade routes. It can also be shown that greater diversions (in relation to the coast-to-coast distance) are inefficient under real-life conditions. The increase in service frequency which will be attained is insufficient compensation for the additional shipping capacity required and the extending of the transit time.

In a short-sea trade the direct handling cost (stevedoring charges) is the dominating item. This means that neither C_2^* nor C_3^* is of an order of magnitude comparable to C_0^*. In this case, the value of E barely differs from zero. In a deep-sea trade the sea-crossing cost and indirect handling cost ($C_2^* = C_4^*$) can be on the level of the direct handling cost. A safe lower limit for the value of E under these conditions is obtained by setting $\bar{C}_0 = C_2^* = C_3^*$. This gives us E = - 1/12.

The point-elasticity does not tell us much about the difference between the costs of liner shipping in a thin trade and the costs in a trade where the density is ten or twenty times greater or even more. The aforementioned lower limit of E was obtained by equalizing the sea-crossing cost and diversion cost. Letting this represent the extreme thin trade case and $Q = \infty$, $C_1^* = C_3^* = 0$ represent the extreme dense trade case, the maximum ratio that can be expected of the cost of a thin route to the cost of a dense route (all factors except trade density being constant) appears to be about 5:3.

Needless to say, the factor prices contained in α_1, α_2, α_3, and α_4 are not the same all over the world. In reality we can therefore find many examples of higher costs in dense trades than in thin trades. Above all, the direct handling cost C_0 varies greatly from port to port. Cargo handling is particularly expensive in American and Scandinavian ports. On the other hand, the indirect handling cost C_4 is sometimes very high in many ports in low-wage countries, due to long queues. But these inter-trade differences are irrelevant in our present context. There is very little competition between liner shipping of different trades. Where it is relevant to compare costs, however, is between different forms of shipping in one and the same trade. Our purpose here has been to examine whether a low density of trade is likely to put liner shipping at any appreciable disadvantage in competition with tramp shipping and other possible forms of bulk services in which costs are commonly held to be largely independent of the trade density.

The conclusion is that a thin-route problem does exist in liner shipping, too, but that its gravity is nothing in comparison to the thin-route problem in public passenger transport.

A different but related question concerns the gravity of the conflict between allocative efficiency and equity in liner shipping, given the inherent decreasing-cost character of this mode of transport. In other words: what level of freight rates can be expected in relation to the shipowners' average cost as a result of the application of the principle of (social) optimal pricing to liner shipping?

In Chapter 4 above, "Optimal pricing of scheduled transport services", it was found that the financial result will be a deficit, which is equal to the total waiting costs (storage cost in the case of freight) multiplied by the absolute value of the elasticity of these costs with respect to the frequency of service, E_{WF}, plus the total feeder transport cost multiplied by the absolute value of the elasticity of the feeder transport cost with respect to the density of service. In the present model the density of service has been held constant, for which reason this last elasticity will be equal to zero. (Observe that this limitation does not reduce the financial deficit; on the contrary, it slightly increases the deficit since the fixing of the density of service leads to very much larger waiting costs in thin trades.) It has been assumed in the model that $E_{WF} = -\frac{1}{2}$. Given this assumption, the financial deficit can be expected to amount to half the total storage costs, if freight rates have been set according to the pricing principles described in previous chapters.

This prediction is confirmed: basing the freight rates on "the average social cost of the marginal ship" in the present model, and observing the conditions of design efficiency, we will find that the total freight revenue falls short of the total shipping costs (= shipowners' costs) by half the diversion cost. In other words, the pricing-relevant cost per ton of cargo, PC, is equal to the sum of the shipowners' costs per ton, except for the diversion cost at one end.

$$PC = 2(C_2^{*} + C_4^{*}) + C_3^{*} \qquad (20)$$

The optimal freight rate would fall short of the shipping cost per ton by about 1/5 at most. In relatively dense trades this fraction would be considerably smaller. For example, if the total diversion cost at optimum is 1/3 of the sea-crossing cost, the discrepancy between the shipping cost and the optimal freight rate would be 7 per cent only of the former.

In discussions of liner shipping economics and policy the question of the optimal level of freight rates has not been much to the fore; nor should it be. Full cost pricing and optimal pricing would not after all produce very different results. What has been to the fore so much the more is the optimal *structure* of freight rates. This is connected with the marked multi-product character of liner shipping services, which has so far been disregarded. We can now turn to this matter.

8 INTRA-TARIFF CROSS SUBSIDIZATION IN LINER SHIPPING[1)]

The salient feature of the pricing of scheduled freight transport services is the markedly discriminatory structure of freight rates. This is equally true of land transport, air transport, and sea transport. The subjects to be discussed in this chapter are pricing practices in liner shipping, and the "self-regulation" which is a prerequisite of price discrimination.

Given that a number of shipping companies serve each particular trade, it would be impossible to maintain a markedly discriminatory structure of freight rates without price cartels being formed. Consequently, together with the international airline industry, the liner shipping industry stands out in the world economy as being completely cartelized. The shipowners' main arguments in defence of the conference system is that it facilitates the "co-ordination" of services and stabilizes freight rates. As has been mentioned, a much weightier rationale of the liner conferences is that they make it possible to increase total freight revenue by means of the large-scale confiscation of the consumers' surplus.

[1] This chapter is based on an article of the same title written some time ago. Although the empirical material is quite old now, it seems that the problem discussed is equally relevant today.

8.1 THE PROBLEM

A very puzzling feature is that this does not generally help to make liner shipping very profitable. On the contrary, the liner shipping industry has long been below average as regards profitability. My explanation of this paradox is the traditional practice of charging freight rates for many commodities that are well below the average shipping costs.

On the other hand, the conventional wisdom of shipping economists and operators alike is that due to the existence of a large portion of *common costs*, the appropriate charging floor is right down at the level of the direct handling costs. This view is challenged here: the question is, however, whether the "cross-subsidization" in liner shipping between high-value and low-value commodities can be justified on grounds of falling average social costs. The previous analysis indicates that, apart from exceptionally thin trades, the economies of scale in liner shipping are so relatively modest that any downward deviation in freight rates from the average shipping costs exceeding 10 per cent is unlikely to be justified; and in dense trades this figure should be considerably lower. As will be shown below, in the case of low-rated cargo to and from the United States, downward deviations from average costs are far greater than this.

The governing principle of liner conference rate-making is to "charge what the traffic will bear", which in practice means that every commodity (rather than shipper) is assigned an individual freight rate in accordance with a (mostly intuitive) estimate of the relative rate elasticity.

It is usually considered that the art of charging what different commodities can bear consists of finding the "ceiling" for the freight rate of each commodity. Various ingenious rules-of-thumb for achieving this have been worked out by conferences. A generally neglected aspect, on the other hand, is the establishing of a proper "charging floor".

This is bound up with cost, and liner shipping cost accounting is peculiarly ill adapted to cost allocation by commodities.

In the following discussion the costs of liner shipping, which should be relevant to rate-making are analysed. A comparison of freight rates and costs are made, which demonstrates a pattern of cross-subsidization. It is based on material submitted in the congressional hearings on "Discriminatory Ocean Freight Rates and the Balance of Payments". A brief introduction to that *inter*-tariff issue will serve to put the present *intra*-tariff problem into perspective.

The establishment of intra-tariff cross-subsidization poses the problem of why liner conferences - apparently against their own interests - persist in rating a large part of the traffic below marginal costs. Professor Sturmey has put forward one explanation for this seemingly irrational behaviour: "The idea of deterrent pricing appears to be acceptable as a long-run explanation of conference behaviour and, in terms of the long-run survival and profitablity of member lines, is rational".[1] By keeping the rates of commodities which are particularly exposed to tramp/charter competition rather unattractive, the conference prevents those outsiders from obtaining a foothold and subsequently making inroads into the really profitable part of the liner business.

The second half of this chapter seeks to add some features to that rate-making picture. After all, it is odd that individual conference members continue without demur to carry a lot of unremunerative commodities just to deter potential competitors, when liner companies' profits are modest anyway.

[1] For a representative and well argued paper, see Sturmey, S.G., Economics and Liner Services. Journal of Transport Economics and Policy, May 1967.

It is suggested that to a large extent this can be explained simply by the fact that *from the standpoint of the individual shipping line* hardly any low-rated cargo appears unprofitable. The conflicting conclusions about cross-subsidization from the standpoint of the whole conference on the one hand, and from the standpoint of the individual conference member on the other, have an obvious policy angle. This is dicussed in the concluding section of this chapter.

8.2 US JOINT ECONOMIC COMMITTEE HEARINGS

United States shipping legislation exempts the cargo liner shipping industry from the operation of the Anti-Trust Acts, and allows the formation of conferences to fix freight rates. But agreements between carriers on freight rates are to be allowed only after approval by the government agency responsible for administering shipping policy - since 1961 the Federal Maritime Commission (FMC).

In the early 1960s the decline in American steel exports, together with some information about inbound and outbound freight rates for steel products, induced Senator Douglas, Chairman of the Joint Economic Committee, in hearings before the Committee on 20 June 1963, to ask the Chairman of FMC why FMC had not done anything to remedy "the apparent freight rate discrimination towards American exporters". This started several investigations of export and import freight rates, and during the following years a large amount of research by various bodies was directed towards proving - and disproving - the allegation that US exporters were victims of favours shown to foreign exporters.[1]

[1] It became clear in the course of the hearings that the existence of allegedly discriminating price cartels throughout the liner shipping industry did not mean that liner companies themselves were unduly favoured. A "Comparative Financial Analysis of American Industry", prepared by Standard & Poor's Corp., which was used as evidence, placed the shipping industry close to the bottom of a list of industries named in order of profitability, regardless of how relative profitability was measured.

In the absence of a consistent theory of liner shipping costs, all attempts to arrive at a definite conclusion about discrimination were doomed to failure. Very generally, unlawful price discrimination occurs when "likes are not treated alike". The crucial question is, obviously, to establish what constitutes "likes". The economist naturally looks at the pertinent marginal cost. This is not necessarily the natural thing for the lawyer. Less subtle criteria, such as material and physical characteristics of a good, are required for a legal condemnation of specific rate-making behaviour. The only really clear-cut case of unlawful freight rate discrimination therefore occurs when exporter X of a certain commodity is charged a higher rate than other exporters of the same commodity on the same route.

Things are much less straightforward in the case of commodities moving in opposite directions. Overwhelming problems arose in the hearings when the question under discussion was whether or not a *general* disparity of import and export freight rates existed.[1] As the comparison embraced both different liner services and a great many different commodities, it is not surprising that nothing came out of it. In the second place, the less ambitious task of scrutinising certain *individual* rates was embarked on. The FMC examiners tried to single out those rates which could be judged to be "so unreasonably high or low as to be detrimental to the commerce of the United States". No more precise criterion was given, and both elements - whether a particular rate is unreasonably high or low, and whether the unreasonableness of the rate causes detriment to commerce - involved great difficulties of interpretation. In the end, the only inbound and outbound rate comparison to which sufficient weight could be attached to warrant action was between the freight rates for *identical* commodities moving in opposite directions. With a few excep-

[1] For a fuller discussion of the issue at stake, see Bennathan, E. and Walters, A.A, The Economics of Ocean Freight Rates. Praeger Special Studies in International Economics and Development, 1969.

tions, such as "books", however, a particular commodity moves in any quantity in one direction only. When both an inbound and an outbound rate exist, the possible rate disparity is normally to the disadvantage of the direction in which the smaller quantity moves. Therefore, it is not difficult to understand that only trifling entries fell within the examiners' cognisance. The tangible result of many years' investigations was that seven not very important outbound rates of the North Atlantic United Kingdom Conference were disallowed as being detrimental to American export trade.

In the course of the hearings a lot of material of interest for *intra*-tariff aspects of freight rates was published. On the one hand, several tariff extracts were produced, and on the other, a cost allocation study, by commodities, of major US inbound and outbound liner trade routes was prepared by a Washington firm of accountants, Ernst & Ernst.[1] On this evidence, which failed to establish inter-tariff freight rate discrimination on any scale, it appears that far-reaching intra-tariff cross-subsidization was prevalent.

The Ernst & Ernst study is a conventional "fully distributed cost" exercise. In the following three sections I will try to justify the view that - contrary to what most authorities hold - the "charging floor" should be close to the fully distributed cost in the main liner trades.

8.3 THE "COMMON COST" ILLUSION

The main point to bear in mind is that what should constitute the proper charging floor on a particular route, is to be judged from the standpoint of the whole conference.

The individual shipping line considers the *voyage* as its basic production unit. This is reflected in shipping

[1] Ernst & Ernst, Selected Commodity Unit Costs for Oceanborne Shipments via Common Carriers (Berth Liner). Undated report (but pertaining to 1964) to the US Department of Commerce.

cost accounting. The Ernst & Ernst cost allocation does not represent standard accounting practice in liner shipping. Cost accounting is in practice normally organized on the basis of "voyage accounts".[1] The total revenue of a round voyage is compared with the cost of the voyage to show the voyage profit. No attempt is made to split up this cost between individual consignments over and above the allocation of "cargo costs", which include stevedoring charges, commissions and claims. This has probably contributed to the predominant view that for rate-making purposes fuel costs and time-proportional costs, either at sea or in port, cannot be allocated between commodities. This leaves only the direct cargo costs to set the "charging floor". If this agreed with the true cost picture, a widely dispersed freight rate structure in accordance with varying demand elasticities would not qualify as *prima facie* evidence of cross-subsidization.

It is true that in the *short run* a large portion of the cost of a liner service is fixed, and fixed costs cannot be allocated either between units of different commodities or between units of the same commodity. Short-run average variable cost (SRAVC) is obviously much lower than the average total cost, when the proportion of fixed cost is high. A very popular fallacy is to argue that this implies that short-run marginal cost (SRMC) is very low too: a "rationale" for setting the charging floor low. (In cost accounting thinking, SRAVC and SRMC are frequently confused, as economists have been pointing out for a long time without much success.) The truth, of course, in the case where the proportion of fixed cost is high, is that SRMC is very low when the degree of capacity utilization is low, and very high when the carrying capacity of a fleet of liners is strained.

[1] Gross, R.O., Some Financial Aspects of Shipping Conferences. Journal of Transport Economics and Policy, May 1971.

Bearing this variability of SRMC in mind, one would not be surprised to find some very low freight rates, *if* the market period in liner shipping were as short as the market period in (e.g.) tramp shipping. If liner freight rates could be separately negotiated prior to every (scheduled) voyage, then at least at-sea costs could be regarded as fixed during the "market period" and common to all cargo lifted on a particular voyage. When total demand for space in such a case happens to fall short of the ship's capacity on a particular sailing, the most rate-elastic commodities should be accepted for freight rates just aboove direct "cargo costs". The profit-maximizing "temporal monopolist" would endeavour to charge what the immediate traffic would bear, and would discriminate against the inelastic demand. On occasions when demand exceeded the ship's capacity, the voyage freight rates would be similarly dispersed according to demand elasticities, but the charging floor would be set higher, because the opportunity cost of capacity would be above zero.

This picture of rate-making serves only as a contrast to reality. The very notion of a tariff implies stable rates for fairly long periods. The *raison d'etre* of liner conferences is to avoid short-term bargaining and frequently fluctuating rates, by means of long-term agreements (among shipping lines) to quote freight rates on every occasion in accordance with a fixed tariff, regardless of short-term demand fluctuations. Under these circumstances the relevant cost picture for rate-making looks quite different. The range of vision must be widened from the single voyage of an individual ship to all sailings within the market period, and to the operations of the entire fleet plying on the route concerned. Adjustments of the fleet, such as adding or withdrawing ships, are practical possibilities within the market period.

The accountant easily accepts that most costs become variable *in the long run.* However, even in the setting of a relatively long market period, the view that at-sea costs cannot be meaningfully allocated by commodities is often maintained, not by claiming that a large proportion of costs are fixed and cannot therefore be allocated, but by arguing that at-sea costs are a common cost, because the production factor is indivisible. It is true that a ship - especially a big container ship - is a rather "lumpy" bit of a production factor. The relevant time unit may, however, be short. Ship capacity is not geographically fixed like industrial plant. In some areas where many trades are more or less interwoven, several alternative itineraries may be practical possibilities in the short run. In this case the capacity provided by one ship on a particular round voyage is a more adequate definition of the smallest capacity unit. Indivisibility is furthermore a relative concept. The withdrawal of even one whole ship from a fleet of, say, 25 ships is not a very substantial reduction of the carrying capacity in percentage terms. It seems perfectly legitimate to compare the freight revenue earned per year by the 4 per cent least-paying commodities (moving on the "fat leg") with the annual avoidable cost of the marginal ship, in order to ascertain whether all commodities pay their way.[1] Moreover, one must not forget that important output adjustments can be made within the market period, even when the fleet size is given. The annual carrying capacity of a ship is not fixed in any way. By speeding up the sea voyage and by using more stevedoring manpower per call (including overtime) an increasing quantity of cargo can be carried in a year, although naturally only at steeply rising short-run marginal cost.

This is not to deny that production factor indivisibilities can be of lasting importance. There is a "thin trade problem", but this is a problem not of common but of *decreasing* cost.

[1] "The average cost of the marginal ship" is an easily understood, practical proxy for the pricing-relevant cost of shipping a cargo ton. The marginal ship is either the oldest, least efficient ship on the route, or a new ship. Provided, however, that the second-hand ship market is operating, the total opportunity cost per ton carried should be roughly the same for all ships of a particular line.

The important point is, however, that in terms of cargo tonnage the thin trade sector constitutes a small fraction of the liner shipping industry. If we draw the line where the volume of trade is sufficient to support two sailings per week, it seems that less than 15 per cent of world liner trade is carried on thin trade routes. The containerization of liner shipping may change this proportion.

The fact that a container ship normally replaces three or four conventional liners is important in the context of economies of scale. So far containerization is concentrated to the densest routes of the world trade. It is likely, however, that a tendency for ship size to grow faster than the liner business will spread to almost every route in the future, thus accentuating the thin trade problem of decreasing cost.

8.4 PRINCIPLES OF COST ALLOCATION

Conventional (break-bulk) liner shipping is a truly multiproduct undertaking. The physical heterogenety of general cargo in intrinsic qualities and package types makes a detailed allocation of costs quite complicated.[1] However, if the reasoning of the preceding section is accepted, the following rough and ready allocation principles should follow logically. With a few exceptions, which will be mentioned below, the principles are the same as those applied in Ernst & Ernst's cost allocation study. The total cost of a liner service is divided into three parts: (1) port costs, (2) at-sea costs, and (3) administrative costs.

[1] In passing it can be noted that the very intricacy of costing has probably been a contributary cause of the respectability of charging what the traffic will bear. The appearance of containers in general cargo shipping has borne out the "psychological" importance of non-identity of break-bulk commodities. Now that "all boxes look alike", the continued wide dispersal of commodity freight rates seems to many to be increasingly irrelevant.

8.4.1 Port costs

Port costs are further divided into direct handling cost and indirect handling cost. The former is mainly made up of stevedoring charges, which are payable per ton loaded or unloaded of each item. The indirect handling costs are less immediately allocable: however, a quite simple and sound allocation principle is available. All costs - apart from the direct handling cost - incurred while in port, such as crew costs, vessel capital costs and port dues, are by and large proportional to time. The time spent by a ship in a particular port is *on average* commensurate to the quantities handled of different articles. The relative manoeverability of goods - the time required to (un)load one ton of commodity X as compared with commodity Y - is consequently the appropriate basis for the allocation of the indirect handling costs. The argument often advanced against this principle - namely that, since only a fraction of a ship's time in port is effectively spent in cargo handling, only a corresponding fraction of the whole indirect port cost could be meaningfully allocated to different articles - is refuted by the observation that the idle time and working time of a ship in port are more or less proportional. The number and duration of coffee-breaks and other stoppages are related to the effective time of loading and unloading, in the same way that nights in port are related to workdays in port.[1]

Ernst & Ernst did not allocate the indirect handling costs in this way. They argued, just as the shipping industry often does, that it would be too much trouble for a somewhat doubtful advantage. Their unit cost estimates are therefore too little dispersed.

[1] It should be noted that the considerable time from entering a port to the start of cargo handling, and from the end of cargo handling to actual departure, is fixed per call, or at least does not vary with the quantity of cargo handled, and should therefore properly be counted as at-sea time.

8.4.2 At-sea costs

The at-sea costs, or capacity costs as they are sometimes called, present allocation problems in two "dimensions". First, the total round voyage at-sea costs have to be split up between the two (or more) legs of the voyage; secondly, it is necessary to determine the proper unit of measurement for apportioning the ship's capacity between different commodities.

The dominant route pattern of the liner services studied by Ernst & Ernst consisted of two-leg round voyages. The balance of trade is obviously the crucial factor for the allocation of the at-sea costs between the outbound and inbound legs. Ernst & Ernst adopted the convention of making this allocation in proportion to the volumes carried.

When the imbalance is marked, the economic approach to cost allocation is to make the "fat" leg bear all the at-sea costs. When the gap between export and import trade volumes is narrower, some at-sea costs may be attributed to cargo on the "thin" leg, on account of the *possibility* that during the next planning period demand may be such that the fat and thin legs are reversed. Generally, a proper treatment of this problem should be in terms of probability distributions of the demand for shipping in each direction and marginal *expected* at-sea costs, as suggested by Walters.[1]

Ernst & Ernst calculated all the unit costs on a "measurement ton" basis (including broken stowage), because "it was found that cargo liners generally achieve substantially higher utilization of available space-capacity than of cargo deadweight capacity" (p. 40). This was no doubt an adequate procedure in that case, i.e. in the case of definite "measurement trades". The generally correct method is that, in principle, *both* the weight *and* the measurement of a particular article should influence the charge, in direct analogy

[1] Walters, A.A., 1960, op. cit.

with Walters' approach to the allocation of costs between the inbound and the outbound. The composition of cargo - like the total quantity of cargo - is not constant from one sailing to another. Sometimes a ship in a particular trade may sail "full" but not "down", and sometimes it is the other way round. The greater the probability of the former situation, the more importance should be attached to cargo measurement in the capacity (at-sea) cost allocation.[1]

In addition, some high density cargo may add stability to the ship when bottom loaded, and hence may well carry a *negative* weight-proportional capcity cost component in a measurement trade.

8.4.3 Administrative costs

The allocation of administrative costs or "overheads" is a general problem of cost accounting. Suggested solutions range from treating overheads as strictly fixed (and common to all sections of the business) even in the longest run, to treating them as roughly proportional to the volume of trade. The particular conditions of liner shipping speak fairly convincingly for the latter view. The costs which are not directly associated with the operation of ships or with cargo handling amount to 10-15 per cent of the total costs, according to Ernst & Ernst; they include the cost of overall management, selling and advertising costs, and miscellaneous items like certain local taxes. On the assump-

[1] Note that an allocation of capacity costs on a "freight ton" basis is always incorrect. The freight ton is the most common cargo quantity unit in liner shipping. It is a weight ton up to a stowage factor of 40 cu.ft./2240 lbs; for higher stowage factors a freight ton is a measurement ton. The correct capacity cost of a particular consignment is the sum of the products of X and its total weight and Y and its total volume, where X is the weight unit cost (= 0 in a measurement trade) and Y is the measurement unit cost. This applies irrespective of stowage factor of the consignment. When allocating capacity costs per freight ton, on the other hand, the volume of a consignment would be of no consequence if its stowage factor were less than 40; if its stowage factor exceeded 40 the weight of the consignment would not affect the capacity cost assigned to it.

tion that the total volume of cargo is approximately proportional to the number of customers, and possibly also the number of tariff entries, the simple procedure adopted by Ernst & Ernst of allocating administrative costs on a per-ton basis seems *a priori* reasonable for rate-making purposes. As mentioned in Chapter 4, it has been established that the overhead cost share is by and large constant in liner shipping, regardless of the fleet size.

8.5 COMPARISON OF FREIGHT RATES AND COSTS

8.5.1 Freight rates

The freight rate material presented below is an adaptation of tariff extracts prepared by the American Steamship Traffic Executives Committee and the Federal Maritime Commission. These extracts were compiled independently of the Ernst & Ernst study of cost allocation. It is also important to note that they were not meant to illustrate *intra*-tariff rate structures, but to make inter-tariff rate comparisons. The ASTEC figures were submitted as evidence in support of the view that American exports were in fact *not* unfavourably discriminated by liner conferences. To demonstrate this, freight rates for the principal export commodities on several main routes were presented together with the freight rates of comparable (if not exactly identical) import commodities. This comparison came out favourable for the export rates, as could be expected, but it was of course criticized on the grounds that the odd inbound commodities naturally carried higher rates, and that these rates, were often anyway to be characterized as "paper rates". It is unusual for the same commodity to flow in both directions at the same time; but tariffs do contain all sorts of odd rates - the fossilisation, as it were, of past trade patterns. The ASTEC export rates, however, are well suited to the present purpose.

Tables 8:1-8:4 give US outbound rates; Table 8:5 gives the rates of the principle commodities moving from Japan to US Atlantic ports. These were originally produced by the Federal Maritime Commission in the course of the controversy. By applying the ASTEC method of export and import rate comparison, but in reverse order, FMC wanted to show the inherent bias of such a method.

To achieve comparability with the Ernst & Ernst unit cost data, it has been necessary to convert all rates into measurement ton rates. The ASTEC and FMC rate-book extracts adhere to the "multiple-unit" arrangement, which constitutes the standard rate quotation system for liner tariffs. The conversion factors required are the stowage factors of commodities going under weight ton rates, or rates levied per "weight or measurement ton (W/M), whichever gives the highest revenue".[1] As pertinent stowage factors are seldom given in the original sources, information available in the standard stowage factor manuals must do. This no doubt impairs the accuracy of the conversions. In the nature of things, the stowage factor of a commodity varies - sometimes substantially - from one trade to another. Difficulties of identification are also unavoidable. Several freight rates had to be dropped because of identification problems. But the remaining rates serve the purpose of illustrating the order of magnitude of freight rate dispersal in a conference tariff.

[1] The option "W/M" was introduced because, by variations (e.g.) in the packing, a commodity may from time to time be presented in varying shapes. To protect against consequent loss of space in the hold, the option is reserved to charge the commodity on the basis of measurement instead of by weight.

Table 8:1. Freight rates and costs for liner cargo from US Pacific coast ports to Japan in 1964 (US dollars per measurement ton)

Principal commodities	*Freight Rates (net)*	*Unit Costs of ship under*	
		US flag	*Japanese flag*
Waste paper for pulping	14	26–32	21–28
Flour in bags	16	28–34	24–31
Cotton, raw, high density	17	26–32	21–28
Infusiorial earth	21	28–34	24–31
Dried fruit	22	29–35	23–30
Oil and grease, lubricating	22	28–34	24–31
Borax	23	28–34	24–31
Hides, green	24	26–32	21–28
Flax, tow waste	25	26–32	21–28
Scrap metal (non-ferrous)	26	36–42	32–39
Stainless steel scrap	27	36–42	32–39
Resin, synthetic	30	28–34	24–31
Aluminium ingots	41	36–42	32–39
Coffee, roasted	44	29–35	23–30
Machinery and parts	48	28–34	22–29
Copper	52	36–42	32–39
Aeroplane parts	55	28–34	22–29
Household goods	63	29–35	25–32
Lead ingots	64	36–42	32–39

Table 8:2. Freight rates and costs for liner cargo from US North Atlantic ports to Belgium and Holland in 1964 (US dollars per measurement ton)

Principal commodities	*Freight Rates (net)*	*Unit costs of ship under*	
		US flag	*Norwegian flag*
Tyres and tubes, rubber	9	26–28	24–27
Scrap, aluminium	10	35–37	34–37
Blocks foam glass	11	26–28	24–27
Road building equipment, packed	13	23–25	21–24
Tobacco, unmanufactured, hogsheads	13	24–26	22–25
Automobile parts	13	23–25	21–24
Jukeboxes, automatic record players	13	23–25	21–24
Rosin, resin, synthetic	14	24–26	22–25
Latex, packed	15	24–26	23–26
Machinery, textile	17	23–25	21–24
Machines, air conditioning, H. H., etc.	18	23–25	21–24
Machinery, mill, steel, metal, etc.	19	23–25	21–24
Cigarettes	23	24–26	23–26
Automobiles, used, unpacked	27	22–24	19–22
Machines, metal-working and parts	28	23–25	21–24
Chemicals	30	24–26	22–25
Office appliances	37	23–25	21–24
Typewriters and parts	43	24–26	23–26
Film, Kodak, not for cine-kodaks	48	26–28	24–27

Table 8:3. Freight rates and costs for liner cargo from US Atlantic ports to South American east coast ports in 1964 (Us dollars per measurement ton)

Principal commodities	*Freight Rates (net)*	*Unit costs of ship under*	
		US flag	*Argentine flag*[1]
Wood pulp	16	31-37	36
Wheat flour and semolina	17	32-38	37
Naval stores, gums and resins	18	31-37	36
Autos, trucks, buses and parts	24	31-37	37
Manmade fibres	24	31-37	36
Dried milk	27	33-39	37
Lubricating oils and greases	30	31-37	36
Glass and glass products	31	33-39	38
Railway locomotives, cars and parts	37	31-37	37
Construction and conveying machinery and parts	48	31-37	37
Other flour and grain preparations	51	32-38	37
Vegetables and preparations	56	33-39	38
Tobacco, manufactured	61	32-38	37

[1] For Argentine operators this was a route with fairly well balanced trade; "fat" and "thin" legs cannot be clearly distinguished.

Table 8:4. Freight rates and costs for liner cargo from US Atlantic ports to Japan in 1964 (US dollars per measurement ton)

Principal commodities	*Freight Rates (net)*	*Unit costs of ship under*	
		US flag[1]	*Japanese flag*[1]
Wheat flour	17	30	28
Rosin and sizing	18	29	27
Carbon black	19	30	28
Cotton, raw, high density	19	29	27
Scrap metal, brass	23	37	34
Shells, mussel	24	30	28
Lubricating oil and grease, packed	24	30	28
Synthetic resin	31	29	27
Tobacco, unmanufactured	34	29	27
Petroleum solvents	38	29	27
Autos, unboxed	42	28	28
Machinery, n.o.s.	52	29	27
Aeroplanes and parts	58	29	27
Cargo, n.o.s.	71	31	29
Tinplate	81	37	34

[1] For both US and Japanese operators this was a route with balanced trade.

Table 8:5. Freight rates and costs for liner cargo from Japan to US Atlantic ports in 1964 (US dollars per measurement ton)

Principal commodities	*Freight Rates (net)*	*Unit Costs of ship under*	
		US flag[1]	*Japanese flag*[1]
Barbed wire	18	39	36
Novelties: ash trays, picture frames, etc	20	32	30
Radios	20	33	30
Christmas ornaments	20	32	30
Wallboards, plywood, and veneer	23	39	36
Rugs and carpets	23	33	31
Toys and games	25	32	30
Canned fruits	28	33	31
Canned meat	28	33	31
Doors, furniture, k.d.	29	39	36
Cotton goods and yarn	31	32	30
Hardware, shovels	32	39	36
Ramies	35	32	30
Machinery and parts	36	30	29
Electric motors	37	30	29
Linen	39	32	30
Iron or steel rods	46	39	36

[1] For both US and Japanese operators this was a route with balanced trade.

The rates are given net. Loyalty rebates of 15 per cent are paid on tariff rates to shippers who have signed exclusive patronage contracts. Practically all cargo is carried under this condition, so the rates of the ASTEC and FMC tariff extracts have been reduced by 15 per cent. All figures are rounded to the nearest dollar.

8.5.2 Costs

The idea of the Ernst & Ernst cost allocation study was to calculate full costs per measurement ton for as representative a selection as possible of commodities moving on each route. By picking out the main package types and distinguishing between low, medium and high density commodities when appropriate, eight standard items were obtained, the unit costs of which should cover fairly well those of the principal commodities moving on each route.

A liner conference trade normally connects several ports at both ends. Ernst & Ernst therefore calculated unit costs for every applicable pair of ports. As the conference freight rate for a commodity is the same regardless of the port of origin or destination, the relevant cost figure for the comparison is the weighted average of the unit costs given by Ernst & Ernst for every pair of ports.

Unit costs have been estimated for the predominant ship type on each route of both the United States registry and the most competitive operator under a foreign flag. For American ships the costs are given with, as well as without, subsidy. The figures including subsidies are relevant to our present purpose.

Where a cost range is given in the tables, the lower figure accords with the Ernst & Ernst method of cost allocation, while the higher figure is obtained by attributing all at-sea costs to the fat leg in each case.

The freight rates of the five routes covered by the tables cannot be claimed as statistically representative samples in a strict sense; nor is the matching of rates and costs perfect, owing to the comparatively small number of standard commodities selected in the Ernst & Ernst study. This may have resulted in too narrow a band of unit cost figures. The value of the data is rather that they provide a concrete illustration of the extent of intra-tariff cross-subsidization. After all, the fact that cross-subsidization must be a general feature of conference rate-making can be deducted from other sources. A number of recent freight rate studies[1] lend indirect but strong support to this conclusion - and also indicate that intra-tariff cross subsidization is not restricted to trade routes to and from

[1] See Jansson, J.O. & Shneerson, D., Liner Shipping Economics, op. cit.

the United States. These studies have served to substantiate the common belief that the value of goods is a principle determinant of freight rates.[1] On the constant-returns-to-scale assumption, the very fact that value-of-service pricing is applied is indicative of cross-subsidization, unless huge monopoly profits are being made by liner companies. The only way of reconciling the fact that freight rates are differentiated according to the value of goods with the known and certified fact that liner shipping is not particularly profitable,[2] is that the profits made on high-rated cargo, for which a demand-based element is added to marginal cost, are offset by losses on a comparable amount of traffic rated well below marginal cost.

8.6 WHY DOES CROSS-SUBSIDIZATION SURVIVE?

The demonstration of intra-tariff cross-subsidization raises the question as to *why* liner conferences maintain such tariffs. I will use the term "low-rated cargo" to mean cargo that moves at rates below the charging floor as previously defined. In addition to the "deterrent pricing" aspect, there is one reason why low-rated cargo constitutes a large part of total liner business, and it is important to understand this before the problem of cross-subsidization can be tackled.

The crux of the matter is that *from the standpoint of the individual shipping line low-rated cargo is not normally unprofitable.* From the collective standpoint of all the shipping lines serving a particular trade, on the other hand, the conclusion that low-rated cargo is loss-making is valid.

[1] The FMC, for instance, defines "an unreasonable rate as one which does not conform the rate-making factors of cost, value of service, or other transportation conditions". By the term "value of service" is implied "more than anything else the value of the commodity being shipped".

[2] This is clearly the case of British liner shipping as evidenced by the report of the Rochdale Committee of Inquiry into Shipping. London, H.M.S.O. 1970.

And this view has implications for the efficiency of resource allocation: if the individual viewpoint dominates over the collective viewpoint, misallocation of resources will result. There are two reasons why individual shipowners may have different ideas about low-rated cargo, and both have to do with the quality of service - spare capacity and frequency of sailings.

8.6.1 Differentiation in quality of service

The natural point of departure is to ask: are the rates of conference tariffs adhered to in practice? Given the cross-subsidization pattern, one would expect member lines in competition for the best rated cargo to grant secret rebates, which would tend to level out disparities in freight rates and costs. It is in the nature of things that little is known about this. Looking at the lower end of the scale, "surcharges" on top of the official rates are certainly not levied by individual lines, but there is another method of partly offsetting losses on low-rated cargo. The quality of service is differentiated to reduce the gap between costs and rates.

This can be done in the following way: the individual shipowner can be regarded in his daily operations as a price-taker, but he is not necessarily always obliged to "clear the market", as it were. Short-term more or less random fluctuations in demand for shipping space make it unreasonable to require that cargo should *never* be rejected. The idea of liner shipping is certainly to guarantee that space will be offered regularly, at pre-determined dates; but, as this guarantee cannot be regarded as absolute, some room for manoeuvre in the selection of cargo is left with the individual shipowner. A bookings policy which always leaves the least-paying cargo behind on the (rare) occasions when demand exceeds supply, is optimal from the point of view of a particular line. It is not strictly consistent with the "common carrier" image of liner shipping, but it is a natural way to adjust to a rigid freight rate structure that is out of line with the cost structure.

It can be added that it would be desirable for shippers to be offered an explicit choice of rate/quality combinations. Quality differences would reflect the relative shipping cost; urgent commodities cause higher capacity costs than commodities which can sometimes wait, thereby helping to raise load factors in slack periods. A number of "classes" can thus be envisaged, which on strict cost grounds would involve some degree of freight rate differentiation. A low stand-by rate, for example, which should include hardly any capacity cost element, could be charged to shippers who were prepared to let their cargo always take the end position in the queue.

So-called supplementary or filling cargo is dealt with in that way by liner operators, but this class of cargo moves at "open rates", which are not fixed by the conference. Supplementary cargo is lifted on an *ad hoc* basis in keen competition with tramps. It is well known that the rates often barely cover the handling charges involved, and liners show interest in the tramp market only on sailings when bookings of ordinary liner cargo happen to fall short of carrying capacity.[1]

If all liner cargo proper - which is far more important than supplementary cargo - went at the same rate, there would be no incentive to treat individual shipments differently. Shipping lines are not supposed to give preferential treatment to anybody, but it is not too difficult to understand that an individual shipowner, faced with the widely dispersed freight rate structure of the tariff, may try

[1] Abrahamsson, B.J., A Model of Liner Price Setting. Journal of Transport Economics and Policy, September 1968. Abrahamsson emphasises the relationship between the liner and tramp markets.

to carry out his own quality differentiation. When the average load factor is high, it is tempting to give preference to coveted cargo, for example by declining low-paying cargo in anticipation of more rewarding offers.

Such a policy cannot be followed too persistently. The shipping line practicing it would become unpopular not only with maltreated customers but also with the other conference members, since it will be the next ship (of another line) that has to lift the rejected low-paying cargo.

8.6.2 Low-rated traffic and frequency of sailings

I have not so far answered the question why the low rates were considered acceptable in the first place. The member lines themselves have fixed the rates at which various commodities are to be carried. A more fundamental aspect of the conference system comes into the picture when we try to answer this question. The heart of the matter is the *jointness of capacity provided and frequency of sailings.*

A major objective of a shipping line is to get as large a relative share as possible of the traffic that can bear high rates. A way of achieving this is, paradoxically enough, to carry low-rated traffic as well. This is manifest in the shipowners' canvassing for new business.

Consider a shipowner who is thinking of putting another ship (of the optimal size) on a certain route. His problem is to muster sufficient cargo to fill the ship. In this situation the total cost of the ship appears genuinely common to all its future cargoes. Every consignment that makes a contribution towards covering the common cost (i.e. gives revenue exceeding the direct cargo cost) interests the shipowner in this situation.

Bearing this in mind, it is understandable that the shipowner, as a conference member, is not inclined to argue that cargo rated below the full cost floor should not be accepted. Other members of the conference are, or will sooner or later be, in a similar position. This means that in general the sum of low-rated "part cargoes" for all lines would easily be sufficient to occupy a whole ship.

A variation on this theme is the common situation where a bigger ship of lower cost per ton is made profitable by a complement of low-rated commodities. The position is now that the shipowner does not regard the size of the comtemplated new ship as given, but realizes that technical developments and/or a change in relative factor prices have made profitable a bigger ship than the one now in operation - on condition, of course, that the new ship will be fully employed. Now the cost of the ship is not "common" to all its future cargoes; but the marginal cost of ship-building per unit of carrying capacity is a decreasing function of ship size, so long as economies of ship size exist. The charging floor for the required new business will be below the full unit cost of the new ship.

Again, whereas this is perfectly rational from the standpoint of shipowner X, the fact that all conference members are reasoning in the same way, results in rates being fixed too low for a large part of the total traffic on each route.

From the point of view of the conference, one way of checking whether or not cross-subsidization is going on is to compare the full opportunity cost of the marginal ship of the conference fleet with the revenue from a corresponding amount of marginal traffic, regardless of how this traffic is at present distributed between the member lines.

Even an individual line would find as a rule that the sum of low-rated cargo on all sailings during one year exceeds the annual carrying capacity of at least one ship. The dilemma is, however, that for one line to refuse to take the low-rated cargo and to withdraw one ship is not likely to cure the anomaly, because some high-rated traffic may easily be lost through the consequent reduced frequency of sailings.

This dilemma is even clearer when there is too much tonnage on a route - the extreme case of low-rated occupancy of a ship's hold is empty hold space.

8.6.3 Excess capacity - A symptom of the dilemma

Excess capacity - as distinct from spare capacity - has plagued the liner shipping industry for a long time (not only in recent years as a consequence of the container revolution). This cannot probably simply be ascribed to accidental "bad luck". The market mechanism is not well adapted to getting rid of excess capacity. Each individual line has but a weak incentive to reduce its carrying capacity *in isolation*, because in the peculiar order of things "demand follows supply" when viewed from the standpoint of the individual line.

Shipowner X, for example, employs four ships on a route, but each one sails half-empty every voyage. At first sight, the obvious thing for shipowner X to do, in order to secure full capacity utilization, is to take two of his ships out of service. The trouble is, however, that business might well drop at more or less the same rate as offered capacity, and the remaining two ships would still sail half empty.

Unless rate concessions are made to the shippers originally catered for, a contracting line has great difficulty in keeping its customers and maintaining its market share so that an appreciably higher load factor can be achieved. Shippers would have to postpone some shipments or despatch them earlier to adjust to the sparser schedule of shipowner X, and why should they do something for nothing?

It might have been possible to maintain the original sailings frequency while reducing capacity by half, if four ships half as big had been substituted for the original ones. Ignoring the practical difficulties in implementing such far-reaching changes in fleet composition, serious economic disadvantages are associated with this line of action when the existing ships are of "the right size" for the route (distance) concerned. The shipping cost per ton of a short-sea liner on a deep-sea route would be substantially higher than the cost of an adequately sized ship.

This "jointness" of capacity provided and quality of service is peculiar to regular transport services. A goods manufacturer can reduce production capacity without fear of loosing cutomers as a result.

It is interesting to compare Stewart Joy's account of the problems of the British railways at the beginning of this century; this is another example of the "demand follows supply" predicament. Then there were about 100 railway companies. Excess capacity was a main problem, but

> No individual railway company was prepared to reduce capacity (and therefore cost) for fear of not being followed by its competitors. In such a cost-variation impasse, revenue maximisation equalled profit maximisation (or loss minimisation). Of course, it can now be seen that for the railway *industry*, if not for the railway company, capacity profitably have been reduced.
> (Joy, S., p. 236)[1]

In liner shipping the dilemma is reflected in the natural inclination of conference members (traditionally not approved of by the US Maritime Administration) to go beyond rate-making to "pooling" sailings, cargo or freight revenue. Whichever form they take, pooling arrangements are well adapted to facilitating appropriate adjustments of supply. The container consortia appearing in some trades are a continuation of this trend.

But this may lead to another dilemma. "Open conferences", with agreements on freight rates only, have admittedly not been successful in preventing such evils as excess capacity. The remedy - agreements also covering the supply of carrying capacity - considerably enhances the monopoly

[1] Joy, S., Pricing and Investment in Railway Freight Services. Journal of Transport Economics and Policy, September 1970.

power of liner conferences, and may lead to the emergence of countervailing power in the form of increased activity by shippers councils or of outright governmental interference. Thus, matters may go from bad to worse. At the negotiating table, where the relative discipline of the market place is completely absent, nationalistic short-sightedness may well aggravate the misallocation of resources.

8.7 CONCLUSION

The traditional organization of international liner shipping is under great pressure, both from rapid technical change and from the new ambitions of emerging shipping nations. Yet the rate-making practices of liner conferences have remained unaltered for almost a century. It remains to be seen how far this will continue to be true if and when an international "code of conduct for liner conferences", now in the melting-pot, replaces the system of conference self-regulation.

It seems both unrealistic and undesirable to fight for the preservation of the old conference system. However, in order to prevent a complete politicizing of international liner shipping, liner companies should offer an alternative to the traditional system of complete self-regulation. I think the elimination of cross-subsidization - written into a code of conduct of liner conferences - should be a main ingredient in a viable reorganization of the liner shipping industry.

The crucial question is whether the capacity-regulation dilemma can be resolved - without making each liner trade a complete monopoly. I think this would be possible, if a cost-based structure of freight rates were introduced. The root cause of the dilemma is the existence of "high-rated" and "low-rated" commodities. Wasteful service competition (including excess capacity) is in great measure the result of rivalry for the high-rated commodities. One reflection of this is that no individual line dares to stop carrying low-rated commodities for fear of losing its share of the high-rated traffic. If price discrimination were abandoned, so that all traffic were equally rewarding, the incentive to sustain losses on one part of the business would be absent, and other excesses in service competition could be eliminated.

9 SEAPORT CAPACITY OPTIMIZATION AND PRICING

Seaports have barely been considered hitherto, either theoretically or in practice, as suitable subjects for a pricing system geared to the principle of net social-benefit maximization. Very few economic treatises on pricing in seaports has appeared.[1] In none of these, in my view, the peciliar characteristics of seaports have been duly allowed for. The discussion has been so general that no very definite commitment has really been made.

On the other hand, specific and advanced models have been produced with a view to optimizing seaport capacity from a social point of view within the framework of seaport Operations Analysis. This OA work has been mainly concerned with the development of *queuing* models of seaports.[2]

In view of this, a natural and important task seems to be to develop port pricing theory along the same lines as - and to keep abreast with - the OA work on capacity optimization, making use of the queuing models that have been evolved. The present study has therefore been divided into two main parts. In the first part a queuing model for the expansion of seaport capacity is developed with a view to de-

[1] Notably Walters, A.A., Marginal Cost Pricing in Ports. The Logistics and Transportation Review, Vol. 12, No. 3 (1976). A more conventional "engineering economy" approach is represented by Heggie, I., Charging for Facilities. Journal of Transport Economics and Policy. Jan. 1974.

[2] For some references, see footnote on page 279.

riving expansion paths for the capacity to load/unload ships and the capacity to accommodate cargo in transit storage. In the second part this model is used for calculating optimal charges on ships and cargoes, and for examining the likely financial result of optimal charging. As will be shown, the question of port pricing - which has become bogged down in rather unfruitful speculations about how best to exact the "cost responsibility" of different port users or how the "value of service" can best be reflected in the structure of port charges - appears in a new light when a queuing model approach is adopted. From a more general theoretical point of view, some interesting problems come to light when - as is mandatory in light of the stochastic nature of short-term demand - the analysis is couched in explicit terms of *expected* cost. Under this condition the question of the differentiation of port charges proves very involved. The pronounced multi-product character of port services has led operations analysts to treat the service time of ships as a random variable, in the same way as the arrival rate. Naturally, in so far as the variability in service time is caused by differences in ships and cargoes (as distinct from differences in external conditions like the weather), a port pricing theory must not sidestep the problem in the same way, but must face the fact that a major potential benefit of a properly differentiated tariff of port charges is that the distribution of service time can be influenced so as to reduce the variability, which in turn makes possible a higher rate of capacity utilization without any increase in queuing costs.

I make no claim in the present study to have found a definitive solution to the last-mentioned problem. But I have, I believe, been able to bring into focus an interesting new area for theoretical and empirical research.

9.1 APPLICATION OF QUEUING THEORY TO SEAPORT CAPACITY OPTIMIZATION

The general purpose-seaport supplies services to ships, cargo and land transport vehicles arriving more or less at random and making different demands on port resources. The short-term demand for port services will therefore vary - one week all resources may be occupied and ships will be waiting in the roads, the next week there may be no ships in the port at all. What is the correct trade-off between the two objectives of a high level of utilization of port facilities and a low likelihood of delays for ships?

A useful tool for tackling port optimization problems is provided by the theory of queuing: ships can be regarded as "customers", while the "service stations" of conventional queuing models can be represented by the berths of a seaport. In fact, alongside telephone services, seaports constitute a major area for the application of queuing theory. The so-called "multi-channel queuing model" gives rise, explicitly or implicitly, to the view that significant economies of scale can be enjoyed in port operations, i.e. that if total throughput increases, the number of berths can be expanded at a considerably lower rate without any negative effect on the expected queuing time per customer.

An important point that will emerge from the following discussion is that queuing theory may not be as readily applicable to seaports as is commonly assumed. One fact that should be remembered is that in the case of transit storage the transfer of goods between sea and land transport is not a single-stage process. In these circumstances the conventional question regarding the optimum number of berths is not altogether relevant: it should be replaced by two interrelated questions regarding the optimum number of cranes and the optimum area for transit storage.

9.1.1 The case of a single-stage, single-channel facility

The mean queuing time for customers (ships) arriving more or less randomly at a single-channel facility (a port section with one berth only) will rise quite steeply already at rather modest levels of capacity utilization. If the customers' (the ships') capacity requirements differ significantly, i.e. if individual service times for loading/ unloading a ship vary substantially, this tendency will be reinforced.

This can be illustrated, using elementary theory on a basis of the following standard assumptions:

1. Customers arrive at random, which means that the distribution of arrivals can be described by the Poisson probability distribution.

2. Similarly, the duration of the service time is a random variable, fitting the "negative-exponential" probability distribution.

Needless to say, these assumptions cannot possibly be generally valid for all individual ports. However, in a fairly large number of cases it seems that the Poisson and negative-exponential distributions do provide fair approximations of real conditions.[1] The third assumption is:

3. There is no upper limit to the length of the queue - customers are indefinitely "patient".

The expected queuing time in statistical equilibrium will then be:

[1] See e.g. "Berth Throughput", Systematic methods for improving general cargo operations. Report by the Secretariat of UNCTAD. TD/B/C.4/109. United Nations, New York, 1973.
De Weille, J. & Ray, A., The Optimum Port Capacity. Journal of Transport Economics and Policy. September 1974.
Gulbrandsen, O., Trafikk- og investeringsanalyse for Bergen havn. TØI, 1973.
Slettemark, R., Optimalisering av Hoveddimensjonerne for havner. TØI, 1974.

$$q = \frac{As^2}{1-As} = s\,\frac{\phi}{1-\phi} \qquad (1)$$

q = expected queuing time per customer (days)

A = expected number of arrivals per day

s = expected service time per customer (days)

ϕ = occupancy rate (= As)

A characteristic of this, as of many other more complicated queuing models, is that, given the occupancy rate, the mean queuing time, q, is proportional to the mean service time, s. As can be seen, q moves towards infinity as ϕ approaches unity. The sharp rise in the mean queuing time as the occupancy rate increases is shown in the middle column of Table 9:1, which gives the ratio of q to s for different values of ϕ.

Table 9:1. Queuing time in the single-stage, single-channel model

Occupancy rate ϕ	The mean queuing time as a proportion of the service time $q/s = \frac{\phi}{1-\phi}$	The expected additional queuing time caused by another arrival as a proportion of the service time $A\,\frac{\partial q}{\partial A}\,/s = \frac{\phi}{(1-\phi)^2}$
0.1	0.11	0.12
0.2	0.25	0.31
0.3	0.43	0.61
0.4	0.67	1.11
0.5	1.00	2.00
0.6	1.50	3.75
0.7	2.33	6.26
0.8	4.00	20.00
0.9	9.00	90.00
1.0	∞	∞

The root cause of the queuing that occurs is, of course, the variability of A and s. The laytime of similar ships varies because a great many more or less random factors significantly affect the actual value of the service time. Such factors include the weather, stoppages due to the breakdown of handling or other equipment. In addition, the type and size of ships and the type of cargo vary a lot.

What difference does it make if this variability can be reduced?

The Pollaczek-Khintchine formula provides a general answer to the question of how the mean queuing time is affected by the distribution of service time.[1] It states that for any arbitrary distribution of s it is possible to express the steady-state mean queuing time, q, as a function of the arrival rate, A, the mean service time, s, and the variance, var(s), of the service time distributions:

$$q = \frac{A\left[s^2 + \mathrm{var}(s)\right]}{2(1-As)} \qquad (2)$$

In the case of a negative-exponential distribution of s, the variance is s^2. It is easily checked that the previous expression (1) for the mean queuing time is obtained by inserting this for var(s) in (2).

In the case of constant service time, the variance of s is zero, and the general formula gives us:

$$q = \frac{s}{2} \cdot \frac{\phi}{1-\phi} \qquad (3)$$

The elimination of service-time variability will, ceteris paribus, reduce the mean queuing time by half.

It is clear from expression (2) above that, given the occupancy rate, the mean queuing time is proportional to the

[1] Saaty, T.L., Elements of Queuing Theory, McGraw-Hill, 1961, pp. 40-43 and 186-189.

sum of the service time and its relative variance [s+var(s)/s]. To reduce the queuing time, it is consequently as important to achieve a reduction in the variability of the service time as it is to achieve a reduction in the mean service time itself. In the case of seaport operations this means that the expected queuing time may be reduced either by increasing the handling speed or by making ship calls more homogeneous, e.g. by specializing in serving a particular type of ship or cargo.

9.1.2 The case of a single-stage, multi-channel facility

Suppose that the seaport consists of n identical berths. Under the same conditions otherwise as in the previous model, the standard multi-channel queuing model can be used for predicting how far queuing time depends on the rate of capacity utilization. Since this model is mathematically more involved, it will be helpful to derive the steady-state mean queuing time in two steps. Let us introduce a symbol p for the probability that an arrival will find all channels occupied.

The mean queuing time can then be written:[1]

$$q = \frac{s}{n(1-\phi)} \cdot p \qquad (4)$$

The occupancy rate ϕ is now = As/n. The left-hand factor s/n(1-ϕ) represents the expected queuing time for those customers who actually meet with a delay, while p is the probability that a delay occurs. This probability is a function of n and ϕ, but is independent of s (given the values of n and ϕ).[2]

[1] See Saaty, op. cit. Chapter 4.

[2] This means that the proportionality between q and s, which applies in the corresponding single-channel model, is retained in this multi-channel model.

The influence on the queuing time of the number of service stations originates from both factors in (4). Given the occupancy rate, the value of the left-hand factor is inversely proportional to n. This is an interesting relationship. If the occupancy rate remains constant when demand increases, i.e. the number of service stations grows in proportion to the demand for service, the *total* queuing time will be equal to a constant multiplied by p. And given the occupancy rate, p decreases continuously as n increases. This is not easy to see from the algebraic expression for p, which can be written as (5) below,[1] but it is a well-established fact.

$$p = \frac{(n\phi)^n}{n!(1-\phi)} \; \frac{1}{\sum\limits_{i=0}^{n-1} \frac{(n\phi)^i}{i!} + \frac{(n\phi)^n}{n!(1-\phi)}} \tag{5}$$

The combined effect of these two relationships indicates that the advantages of multi-channel service facilities are truly remarkable. The *total* queuing time decreases when demand and capacity increase at the same rate.

It should be mentioned that the economies of number can be realized either in the form of lower queuing costs, or lower capacity costs per customer, or a combination of both. It can be shown that the occupancy rate will steadily increase along the "expansion path", while the mean queuing time will decrease. This means that both the capacity cost and the queuing cost per unit of throughput will fall as throughput increases, provided that an optimal factor combination is chosen.

Table 9:2 shows the mean queuing time as a proportion of the service time according to the multi-channel model.

[1] See Saaty, op. cit., p. 166.

Table 9:2.

Queueing time/service time ratios

		Number of berthing points														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
	.050	0.053	0.003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.050
	.100	0.111	0.010	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.100
	.150	0.176	0.023	0.004	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.150
	.200	0.250	0.042	0.010	0.003	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.200
	.250	0.333	0.067	0.020	0.007	0.003	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.250
	.300	0.429	0.099	0.033	0.013	0.006	0.003	0.001	0.001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	.300
	.350	0.538	0.140	0.053	0.023	0.011	0.006	0.003	0.002	0.001	0.001	0.0	0.0	0.0	0.0	0.0	.350
	.400	0.667	0.190	0.078	0.038	0.020	0.011	0.006	0.004	0.002	0.001	0.001	0.001	0.0	0.0	0.0	.400
	.450	0.818	0.254	0.113	0.058	0.033	0.020	0.012	0.008	0.005	0.003	0.002	0.002	0.001	0.001	0.001	.450
	.500	1.0	0.333	0.158	0.087	0.052	0.033	0.022	0.015	0.010	0.007	0.005	0.004	0.003	0.002	0.002	.500
	.550	1.222	0.434	0.217	0.126	0.079	0.053	0.037	0.026	0.019	0.014	0.010	0.008	0.006	0.005	0.004	.550
	.575	1.353	0.494	0.254	0.151	0.097	0.066	0.047	0.034	0.025	0.019	0.014	0.011	0.009	0.007	0.005	.575
	.600	1.500	0.562	0.296	0.179	0.118	0.082	0.059	0.044	0.033	0.025	0.020	0.016	0.012	0.010	0.008	.600
Berth	.625	1.667	0.641	0.344	0.213	0.143	0.101	0.074	0.056	0.043	0.034	0.027	0.021	0.017	0.014	0.012	.625
Occupancy	.650	1.857	0.732	0.401	0.253	0.173	0.124	0.093	0.071	0.055	0.044	0.035	0.029	0.024	0.020	0.016	.650
	.675	2.077	0.837	0.468	0.301	0.209	0.152	0.115	0.090	0.071	0.057	0.047	0.038	0.032	0.027	0.023	.675
	.700	2.333	0.961	0.547	0.357	0.252	0.187	0.143	0.113	0.091	0.074	0.061	0.051	0.043	0.037	0.031	.700
	.725	2.636	1.108	0.642	0.426	0.305	0.299	0.178	0.142	0.115	0.095	0.080	0.067	0.058	0.049	0.043	.725
	.750	3.0	1.286	0.757	0.509	0.369	0.281	0.221	0.178	0.147	0.123	0.104	0.089	0.076	0.066	0.058	.750
	.775	3.444	1.504	0.899	0.614	0.451	0.347	0.276	0.225	0.187	0.158	0.135	0.117	0.102	0.089	0.079	.775
	.800	4.0	1.778	1.079	0.746	0.554	0.431	0.347	0.286	0.240	0.205	0.176	0.154	0.135	0.119	0.106	.800
	.825	4.714	2.131	1.311	0.917	0.689	0.543	0.441	0.367	0.311	0.267	0.232	0.204	0.181	0.161	0.145	.825
	.850	5.667	2.604	1.623	1.149	0.873	0.693	0.569	0.477	0.408	0.353	0.310	0.274	0.245	0.220	0.199	.850
	.875	7.0	3.267	2.062	1.476	1.132	0.908	0.751	0.635	0.547	0.478	0.422	0.376	0.338	0.306	0.278	.875
	.900	9.0	4.263	2.724	1.969	1.525	1.234	1.028	0.877	0.761	0.669	0.594	0.533	0.482	0.439	0.402	900
	.925	12.333	5.926	3.829	2.796	2.185	1.782	1.497	1.285	1.122	0.993	0.888	0.802	0.729	0.668	0.614	.925
	.950	19.0	9.256	6.047	4.457	3.511	2.885	2.441	2.110	1.855	1.651	1.486	1.348	1.233	1.134	1.049	.950
	.975	38.999	19.252	12.708	9.451	7.504	6.211	5.291	4.602	4.068	3.642	3.295	3.006	2.762	2.553	2.373	.975

Source: Calculated by UNCTAD secretariat from queueing theory formula with poisson arrivals and exponential service times with first-come, first-served queue discipline.

A problem that arises in applying the multi-channel model to modern seaports, and one which has often been pointed out in recent literature, is that the definition of a "berth" is no longer self-explanatory. In the old ports consisting of berths between finger piers there is no problem of definition. The question is, however, what constitutes a berth in a port where the quay-front is one straight line, or, in other words, how many "service stations" does such a port contain.

Perhaps an obvious first suggestion could be that a berth is a stretch of quay-line sufficient to accomodate a representative ship, and the total number of berths equals the total number of ships (of a given size) that can be accommodated along the quay. But this is not always necessarily the most appropriate definition. At container terminals and many specialized bulk terminals, the *cranes* are obviously the relevant service stations, because these are sometimes scarce whereas the quay-line is practically never fully utilized. Take the example of the Scandia container terminal in the port of Gothenburg. If its quay-line were fully occupied by berthing ships, the container cranes available would by no means suffice to serve all the ships.

At general-cargo port sections, on the other hand, where the berthing ships are equipped with their own gear, it seems less fitting to count the service stations in terms of shore-based cranes. At such facilities a certain length of quay-line may provide a more adequate definition. Where there are no shore-based cranes, for instance in a berth where truck-loading/unloading via side-ports is practised, this will be the obvious answer.

9.1.3 Towards a multi-stage, multi-channel model

The problem of defining berths is a symptom of a more fundamental problem, namely that a single-stage queuing model is often inadequate as a means of representing a seaport. This inadequacy has been demonstrated very clearly by recent experiences from ports in the "nouveau riche" oil-producing countries. In recent times ships at ports in the Persian Gulf, in particular, but also in a number of ports in West Africa, have been faced with severe queuing problems. A common experience seems to be that transit sheds are crammed, while quay occupancy does not always reach the same excessive level. This reflects at its most extreme an inherent characteristic of both break-bulk and unitized cargo ports: due to the relatively fast productivity increase in the loading/unloading operation proper, the transit storage stage in the throughput process tends to become the bottleneck. This will be very conspicuous in cases where total port investments have lagged behind total throughput demand, perhaps on account of a sudden spurt in general cargo imports. The basic fact is that the transfer of goods between sea and land transport is a multi-stage process whenever the indirect route is chosen, i.e. when transit storage comes in between.

In primitive ports where direct loading and unloading[1] are practised, the single-stage model is quite adequate. The same goes for old ports that consist of berths between finger piers, where the "berth" including transit storage is a self-contained unit; additional berths are created by replication of existing berths.

However, one factor that rendered the finger pier system obsolete, was that the capacity of the loading/unloading operation tended to outstrip the capacity to get the cargo through the transit shed. Thus, cranes were sometimes left standing with nothing to do. The introduction of a lineal quay-line meant that the cranes could be moved from one berth to another (by being mounted on rails), and it thus

[1] The cargo is directly transferred between sea and land transport vehicles without intermediate transit storage.

became easier to adjust the ratio of storage capacity to crane capacity, i.e. the amount of storage space per crane could be increased.

It is therefore no longer quite satisfactory to define a seaport service station in a queuing model context as a "berth" consisting of certain amounts of quay space, crane capacity, and storage space, which remain in constant proportions to one another, since it is not necessary to maintain constant proportions between these factors. A major problem of port design is in fact to determine the most appropriate factor combination in each particular case.

To find the most appropriate modification of the standard single-stage queuing model, it is necessary to take a closer look at the throughput process.

9.1.3.1 Stages in the throughput process

The process of transferring goods between sea and land transportation in a port is often described as a "chain of links". This metaphor is used in two different contexts. One is that of detailed operation analysis where every single action in the handling of cargo is defined as a "link" in the chain. A substantial number of links can be identified, particularly as far as the handling of general cargo is concerned.

Concerning export shipments, the arrival to the unloading platform for cargo by land transportation is one link, which is followed by the discharge of the truck or railcar. The next link consists of moving the cargo into the transit shed, where it may be sorted, weighed and/or measured, and individual consignments marked. After a period in transit storage, the cargo may be prepared for loading by preslinging or palletization, and taken to the quay apron just before the ship arrives. When the ship has arrived and is lying alongside the berth, break-bulk cargo has to be made up in sets to feed the hook of the crane. The crane carries the cargo from the quay apron into the hold of the ship, where it has to be stowed according to a stowage plan. When the ship has unberthed and is steaming out of port, the whole throughput process is completed.

Import cargo follows a similar multi-stage route through the port. Some links may be different or missing altogether. Customs inspection is usually included, and reconditioning of the cargo may be called for.

The point of such a detailed scrutiny of the indirect route of cargo through a berth is, in the first place, to find the weak link or links that reduce the total throughput capacity. This kind of detailed "chain-of-links" approach is normally applied to existing berths, in an attempt to tackle short-term problems in operations and organization.

Another context in which the chain-of-links concept is useful is in connection with the long-run problem of optimizing port capacity, i.e. choosing the optimal number of service stations. For such investment problems a very fine division into links of the throughput chain would become unmanageable, as the number of factors to be taken into account becomes very large when the entire design of the port is variable. However, where the indirect route of cargo predominates, it may seem appropriate - from the point of view of queuing theory - to regard the throughput process as consisting of two main stages with a "waiting-room" in between. The transit storage space is a waiting-room between the two throughput stages of loading/unloading ships and loading/unloading land transport vehicles. The waiting-room acts as a buffer between these two stages, making them independent of one another and thus improving the efficiency of them both.

According to this view, the multi-channel queuing model is applicable separately to the loading/unloading of ships and to the loading/unloading of land transport vehicles. This approach is all right, provided the holding capacity of the "waiting-room" is practically unlimited. However, it is well known that in the transit storage facilities of a great many ports this is certainly not so.

9.1.3.2 The nature of transit storage

The "waiting-room" between two stages in the throughput process will hardly ever have absolutely unlimited capacity. But if the cost of expanding the holding capacity of the waiting-room is small in relation to the cost of production in the preceding and succeeding process stages, the capacity of the waiting-room can be regarded for all practical purposes as unlimited.

In ports, more often than not, this condition does not obtain. The cost of the transit storage capacity depends to a great extent on land values. In older ports situated in or near the core of a city, storage space is a chronic bottleneck. And as soon as the capacity of the waiting-room is no longer unlimited, the performances of the two service stages will be connected. When the waiting-room happens to be completely filled by customers because of some hold-up in production in the second stage, the preceding service stage cannot pass on customers who have been served. The customers have to remain in the first stage, blocking the way for the customers behind them. If nothing is done to relieve the original cause of delay in the second stage, production in the first stage will eventually drop to the level of output in the second stage. It should be mentioned here that one branch of general queuing theory deals with two-stage processes with waiting-room of limited holding capacity between them.

However, this is not the whole story of transit storage in a port. An import consignment that arrives in transit storage cannot move on to the next stage as soon as a service station (= freight delivery station) is vacant, unless the right truck is there to pick it up. This is obviously a decisive condition. The time spent by a consignment in transit storage is not determined primarily by the capacity of the following service stage, but by the time that elapses before it is collected by the importer or, in the case of export cargo, by the interval between its receipt at the port land-side and the arrival of the relevant ship.

Thus, in one respect, the storage facility does not correspond exactly to a waiting-room in the limited sense that is relevant in queuing models. This can be allowed for by defining transit storage as an independent service stage as well as a waiting-room, so that the time spent by the cargo in transit storage can then be regarded as *service* time rather than pure waiting time. If this service stage is fully occupied, it is obvious that the waiting-room function of transit storage is inoperative. The question is, how exactly does this affect the queuing of ships? As far as import cargo is concerned, two slightly different outcomes are conceivable, of which the second seems to be the most relevant from a practical point of view.

1. A ship arriving at the port can be unloaded, provided that a sufficient stretch of the quay-side is unoccupied and that the required crane capacity is available. The cargo has to be placed provisionally on the quay apron, however, and this means that the corresponding berth is effectively blocked for any further ships.

2. A ship arriving at the port cannot be unloaded, simply because the cargo is too large to be put on the quay-side. That is to say, there is no time lag between the event of the transit storage being fully occupied and the queuing of ships.

As regards export cargo taking the indirect route through the port, it is obvious that if the transit storage facilities are full of other cargo, a cargo booked for a particular sailing cannot be lifted until the "service time" of a sufficient quantity of the cargo in transit storage is completed, and room is made for the export cargo concerned.

Assuming the second import alternative to be the most realistic, it can be concluded that a ship will have to queue unless sufficient free space along the quay *and* crane capacity *and* transit storage space are all available when it arrives.

9.1.3.3 The model

As has been mentioned, it is unusual that the quay-line itself is a scarce resource. For the sake of simplicity this possibility will be disregarded here. In cases where the cargo-handling technique involves one crane only per ship, the definition of the service stations of the loading/unloading stage - *stage one* - it self-explanatory. In cases where more than one crane can be used per ship, the question is how many cranes should be said to constitute a service station? This seemingly tricky problem of definition can be resolved by resorting to the useful rule-of-thumb of port operations that "dilution" of resources is generally inefficient. The allocation of cranes is a case in point. Suppose that a particular quay-line is equipped with three quay-cranes, and that it is possible to use all three cranes simultaneously on one ship. Suppose further that the marginal productivity of a crane is constant, regardless of whether one, two or three cranes are working on the same ship. Under this condition it will be optimal, in the sense that the sum of the total queuing time and service time will be minimized, to use three cranes on a ship irrespective of whether or not other ships are waiting to be served. "Dilution" of cranes - one crane for each of three ships, two cranes for one ship, and one crane for another ship, etc. - is inferior as a policy of crane allocation.[1]

It can consequently be assumed that a service station in stage one consists of a *given* number (= k) of cranes; in other words k is independent of the actual number of ships in the port at any particular time.

A service station in *stage two* consists of a part of the total storage area sufficient to hold a representative shipload, plus the area of the quay apron immediately in front of each transit storage sub-area.[2]

This case in which service can be supplied to an arriving customer only if two succeeding service stations are unoccupied is not covered by the existing "serial" models in standard queuing theory. The next step is therefore to develop a model for the two-stage case, thus defined, in order to derive a formula for the expected queuing time of ships.

Lastly, it can be mentioned that land transport vehicles, too, are involved in queuing at seaports. A wholly symmetrical model is appropriate for explaining this queuing; the only adjustment required is to redefine stage one so

[1] A numerical example can clarify this point. Suppose that three identical ships arrive at the same time, and that the service time of a ship will be one day, or one and a half days, or three days according to whether three cranes, two cranes, or one crane are employed.

Crane allocation	Total service time	Total queuing time	Total time
3 0 0	3·1	0 + 1 + 2	6
2 1 0	2·1·5 + 1·3	0 + 0 + 1.5	7.5
1 1 1	3·3	0	9

[2] It can be noted in passing that in the case of terminals without shore-based cranes, the ordinary single-stage queuing model is applicable. However, in this case, too, the quay-line is not normally in the focus of interest; rather, the transit storage sub-areas constitute the relevant service stations for queuing modelling.

that it consists of the freight reception and delivery stations at the land-side of the port. In the following discussion, however, we will be concerned with the queuing costs of ships only.

The basic assumption of the model is that a ship arriving at a port, if it is to be able to discharge its cargo and/or load of the export cargo booked for its next sailing, must find free crane(s), and sufficient free transit storage space must be made available for its cargo.

As before, the mean queuing time per ship call, q, can be calculated as the product of the probability that a delay will occur and the expected queuing time when a delay actually occurs.

We shall use the following notations:

P_1 = probability that all crane capacity is occupied when a ship arrives

P_2 = probability that no storage space is available

q_1 = mean time before sufficient crane capacity for another ship is free on occasions when all cranes are occupied

q_2 = mean time before sufficient space is cleared on occasions when the storage space is fully occupied

The probabilities of four possible events, and their outcomes as regards queuing time, are summarized in Table 9:3.

Table 9:3. Expected queuing time

Event	Probability	Queuing time
Both crane and storage available	$(1-p_1)(1-p_2)$	0
Storage but no crane available	$p_1(1-p_2)$	q_1
Crane but no storage available	$p_2(1-p_1)$	q_2
No crane, no storage	$p_1 \cdot p_2$	$\max(q_1, q_2)$

The expected queuing time per ship is thus:

$$q = (1-p_1)(1-p_2)0 + p_1(1-p_2)q_1 + p_2(1-p_2)q_2 + p_1p_2\max(q_1,q_2) =$$

$$= p_1q_1 + p_2q_2 - p_1p_2\left[q_1 + q_2 - \max(q_1,q_2)\right] \qquad (6)$$

As the value of ships' time in relation to the cost of port resources is normally quite high, the probabilities of delays (p_1 and p_2) *should* be low. A representative value of either of these probabilities can be about 1/10, in which case $p_1p_2 = 1/100$. Therefore, the first two terms of expression (6) are very dominating in most realistic cases. In the first place, the product p_1p_2 is small in relation to either p_1 or p_2, but also the difference $q_1 + q_2 - \max(q_1q_2)$ is smaller than the value of q_1 or q_2, whichever is the lower. The combined result is that the third term will contribute a few per cent only to the total value of q. The basic fact is that in most cases q is *approximately* equal to the sum of the queuing time caused by each of the two service stages, regarded in isolation.

The next step is to establish a relationship between the mean queuing time and the capacity and demand for service, corresponding to that of the model of a one-stage multi-channel facility.

The following notations are now introduced:

n_1 = number of service stations in the first stage (cranes)

n_2 = number of service stations in the second stage (transit storage)

s_1 = mean service time of ships in the first stage, in days

s_2 = mean service time in the second stage, in days (= mean lay-time of cargoes in transit storage)

Assuming that the Poisson probability distribution applies to the arrival rate, A, and that the negative-exponential distribution applies to both service times, s_1 and s_2, it is possible to use the result of the previous single-stage multi-channel model for specifying the relationships between the four components of q, that is p_1, p_2, q_1 and q_2, and the capacity, and rate of capacity utilization of each service stage.

The expression for p_1 and p_2 are rather complex, as was seen in (5) on page 283. It is pointless to reproduce the complete expressions. It is sufficient to point out that p_1 and p_2 are both functions of the applicable occupancy rate and the number of service stations only.

$$p_1 = p(\phi_1, n_1) \tag{7a}$$

where

$$\phi_1 = \frac{As_1}{n_1}$$

$$p_2 = p(\phi_2, n_2) \tag{7b}$$

where

$$\phi_2 = \frac{As_2}{n_2}$$

The expression for q_1 and q_2 can be fully specified (compare (4) on page 282):

$$q_1 = \frac{s_1}{n_1(1-\phi_1)} \tag{7c}$$

$$q_2 = \frac{s_2}{n_2(1-\phi_2)} \tag{7d}$$

An interesting peculiarity of the queuing-time function is that the number of arrivals, A, and the service time, s_1 or s_2, never appears as separate determinants of the *total* queuing time, but always in the form of the products As_1 and As_2. These products represent the total service-days of the ships and the total transit storage-days of the cargoes. By multiplying both sides of (6) by A, we find that the arguments s_1 and s_2 do not appear on their own. This means that one variable can be eliminated when the function for the total queuing time is adopted. Let us now add the three following symbols:

$X = As_1$ = total service time of the ships
$Y = As_2$ = total storage time of the cargoes
$Z = Aq$ = total queuing time

Rather than measuring the output of the port by the number of ship arrivals, A (or, more adequately, the number of ships turned around, a two-fold output measure is adopted by substituting the total service time of the ships, X, for As_1, and the total storage time of the cargo, Y, for As_2.

We can consequently regard the total queuing time, Z, as a function of the service outputs of stage one and stage two as well as the number of service stations in each stage.

$$Z = Z(X, Y, n_1, n_2) \tag{8}$$

9.1.4 Port-capacity expansion paths

The main and ultimate purpose of all queuing models is to provide help in choosing the appropriate level of capacity for a particular service facility and, in particular, in choosing the number of service stations. Given the costs of the service stations and the value of customers' time, a queuing model should answer the question: how much capacity is to be provided for the service of A customers? Since queuing models are mathematically complicated, a common practice is to construct "nomograms" for those relationships that are of the greatest practical interest. By means of the present model it is possible to produce nomograms for port capacity "expansion paths" by assuming different values for the ratio of the capacity unit cost to the value of ship's laytime. However, the calculations are limited here to providing a single example of a capacity expansion path. The most interesting thing in the present context is to get some idea of the extent of the scale economies of port operations, and for this purpose it is quite sufficient to consider a single representative value of the factor-price ratio concerned.

In the present two-stage model two different expansion paths for the number of service stations are applicable, in principle, in each particular case - one for the number of cranes and another for the transit storage space (i.e. the number of standard units of transit storage capacity). Thus, given the relevant factor prices, the two following questions have to be answered:

1. What is the optimum number of cranes of a specified type, when the total crane-hour requirement is kX?
2. What is the optimum transit storage capacity, when the total storage space requirement is Y?

In the present model this is entirely a matter of trading off crane and storage capacity costs against savings in queuing costs. The balancing factors are the number of service stations, n_1 and n_2.

$$\textit{Total capacity costs} = c_1n_1 + c_2n_2$$

$$\textit{Total queuing costs of ships} = vZ(X, Y, n_1, n_2)$$

When the number of service stations is small, there may be significant indivisibilities that have to be taken into account, since the addition of another service station constitutes a substantial relative increase in total capacity. Container terminals provide a good example of this, as typical choices are between one or two, two or three, and three or four cranes. Only the very largest container terminals have more than five container cranes.

Hence, the criteria for determining the optimal number of service stations can conveniently be stated:

Stage one: $c_1 \leq v \left[Z(n_1) - Z(n_1+1)\right]$

Stage two: $c_2 \leq v \left[Z(n_2) - Z(n_2+1)\right]$

where

$Z(n_1)$ = total queuing time with n_1 service stations in stage 1

$Z(n_1+1)$ = total queuing time with n_1+1 service stations in stage 1, etc.

On a basis of these criteria it is possible to trace the "expansion path" for the number of service stations for successively higher levels of demand for port services for any given values of c_1/v and c_2/v. The choice of a representative value of the relative factor prices will inevitably be rather subjective. The value of c_1 depends on the type of cranes that are installed on the quay, and the value of c_2 depends on whether or not a transit shed is needed, and on the applicable value of land. The value of v varies very widely depending on the type and size of ships calling at the port. If relatively big ships are involved, the ratios c_1/v and c_2/v tend to become quite low - 1/10, 1/20, or even lower. In Swedish ports, however, the average size of ships is rather small, because short-sea traffic dominates.

In the following example we have set $c_1/v = c_2/v = 1/5$, which seems appropriate for a container terminal serving short-sea and sea feeder traffic as well as for a modern, well-equipped break-bulk cargo port for ships of moderate size.

In the case of a container terminal the value of v can be represented by the cost per day in port of a container ship carrying 300 TEU, which is about 30 000 Sw.Cr.[1] The purchase price of a 40 ton container-crane is about 20 mill. Sw.Cr. Assuming an economic life of 20 years and a discount rate of 8 per cent, we get a cost per day of approximately 6 000 Sw.Cr. Furthermore, assuming that 5 ha. of paved area is required for an average shipload (export plus import) of containers, and that the land value is 400 Sw.Cr./ha., we get a ratio of c_2/v which is also equal to 1/5.

This factor price ratio can also apply to a break-bulk cargo port, although a wide range of values exists in practice in this case. The value of v is about three times smaller for a conventional general cargo ship than for a container ship of the same tonnage. The price of a conventional quay crane is about 1 mill. Sw.Cr., but each ship may use about five cranes simultaneously. Hence the ratio of c_1/v can be of the same magnitude as in the container case. In the case of c_2/v, the much smaller land requirement per break-bulk berth has the effect of reducing c_2 substantially. An offsetting effect, on the other hand, is the cost of the transit shed which comes into the picture, as compared with the case of a container terminal.

On a basis of the information contained in Table 9:2, and given the factor price ratio, the port-capacity expansion path has been traced. Needless to say, $c_1/v = c_2/v$ is a condition which is unlikely to be exactly applicable anywhere. The reason for making this assumption is simply to save the effort of making the same sort of calculations twice over, and to save graph space. The diagrams of Figures 9:1 and 9:2 can be used for both service stages. However, another simplification is necessary in order to achieve this convenient

[1] "Sjöterminaler för enhetslasttrafik", Kjessler & Mannerstråle AB, 1977.

generality: as pointed out before, the third term of (6) including the product p_1p_2 will be very small in relation to the two preceding terms, provided that these probabilities are small. We have tested the result of disregarding the third term under different realistic conditions and have found that an overestimation of the queuing time of no more than 2-3 per cent is most likely, and that the overestimation is highly unlikely ever to exceed 10 per cent. By disregarding the third term of (6), the following attractively simple and "separably additative" form of the Z-function is obtained:

$$Z \approx p(x,\ n_1)\ \frac{X}{n_1-X} + p(Y,\ n_2)\ \frac{n}{n_2-Y} \qquad (9)$$

where the p-function is given by (5) above.

In Figure 9:1 the development of n_1 as X increases, or of n_2 as Y increases can be found. If the horizontal axis is regarded as measuring total service-days in stage one, the expansion path portrayed would represent the development of n_1, and if service-days (i.e. cargo storage-days) in stage two are considered, the expansion path gives us the development of n_2. However, since s_2 is likely to be longer many times over than s_1, Y is correspondingly greater than X, and the number of service stations in the second stage, n_2, is much greater than the number of service stations in the first stage, n_1. Let us take an example. Suppose that X=400 service-days per annum of the ships. The optimal value of n_1 according to the diagram is equal to 3. Suppose that the mean storage time, s_2, is seven times longer than s_1 - for example, one week as opposed to one day. This means that Y is also seven times greater than X, i.e. 7x400 = 2 800. From the diagram it is clear that in that case the optimal value of n_2 = 12.

It can be inferred from the illustration that the optimal number of service stations increases degressively - and markedly so - with increases in total service-days. However, this feature does not stand out very clearly in Figure 9:1. This important characteristic is demonstrated in two alternative ways.

9.1.5 Capacity utilization and queuing costs along the expansion path

In Figure 9:2 the optimal occupancy rate is given for each number of service stations. Like the previous diagram, this diagram is applicable both to the relationship between ϕ_1 and n_1, and to the relationship between ϕ_2 and n_2. Since n_2 is likely to be much greater than n_1, it is clear that ϕ_2 will also be greater than ϕ_1. Because of the "lumpiness" of the service stations there will be not one but a whole range of optimal values for the occupancy rate of each number of service stations.

The main message of Figure 9:2 is that a substantially higher occupancy rate is optimal - in each of the two stages - at large ports than at small ports.

Another way of making the same point is to calculate the elasticity of the number of service stations with respect to the total service-days. The ratio of the relative increase in the former to the relative increase in the latter is calculated for each number of service stations. The result is presented in Table 9:4 below.

A further important question concerns the way queuing time develops in optimum as total traffic increases. Again, owing to the lumpiness of capacity additions, we will get not a continuous function but a saw-toothed curve. When another service station is added, the queuing time drops abruptly to a lower level and then gradually increases as demand grows until another service station is added, upon which the queuing time drops again, and so on and so on. In Table 9:5 the minimum and maximum values of the mean queuing time per ship in the optimal range are given for each number of

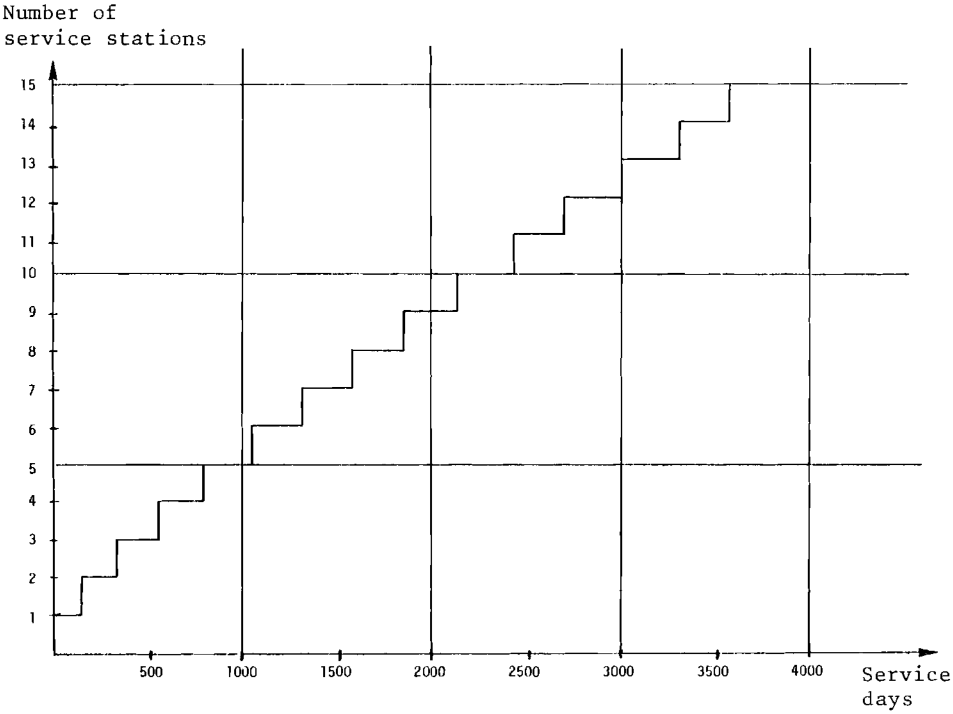


Figure 9:1. Expansion path for the number of service stations

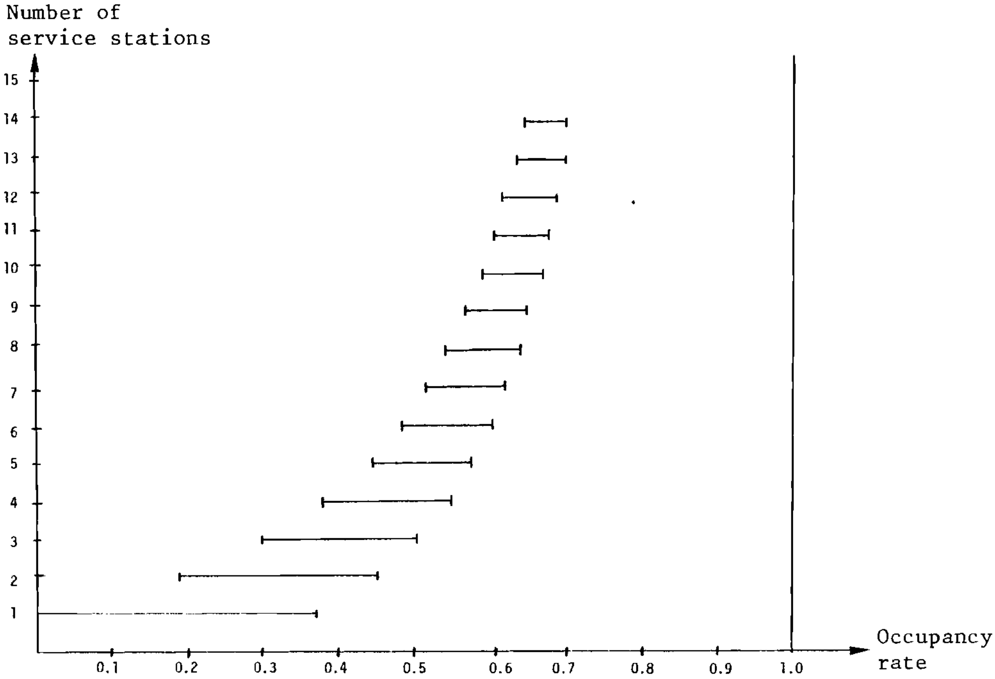


Figure 9:2. Ranges of optimal occupancy rates

Table 9:4.

The elasticity of the number of service stations on the expansion path

Optimum number of service stations	The elasticity with respect to total demand for service
1	0.50
2	0.60
3	0.65
4	0.70
5	0.72
6	0.75
7	0.77
8	0.79
9	0.81
10	0.83
11	0.85
12	0.86
13	0.87
14	0.88

Table 9:5.

The mean queuing time in optimum for different numbers of service stations

n	Range of q/s in optimum
1	0 - 0.59
2	0.04 - 0.25
3	0.03 - 0.16
4	0.03 - 0.12
5	0.03 - 0.10
6	0.03 - 0.08
7	0.02 - 0.07
8	0.02 - 0.06
9	0.02 - 0.06
10	0.02 - 0.05
11	0.02 - 0.05
12	0.02 - 0.05
13	0.02 - 0.04
14	0.01 - 0.04

service stations. As expected, optimal capacity widening results not only in a fall in the capacity costs per unit of traffic but also in a fall in the mean queuing time.

9.2 PORT PRICING

The total price paid for getting cargo through a port is normally divided between the stevedoring charges for the direct labour of loading and unloading the cargo to and from the ship, and port charges for recouping the port capacity costs. In addition, various charges for specific services such as pilotage, towing, customs clearance, cargo reconditioning, etc are often payable by the shipowners or the wareowners.

The division into stevedoring charges and port charges is largely explained by a corresponding organizational division between one or more separate stevedoring companies and the port authority responsible for the capital resources, including the quay-cranes and the transit storage facilities. Here we will exclusively discuss the port charges.

9.2.1 Common principles of port pricing

The philosophy underlying existing port charges seems to include two main ingredients. The proportion of the two ingredients varies from one country to another and, to a lesser extent, from one port to another within a particular country. But these two principles, well known from century-long public-utility pricing-policy discussions, are omnipresent.

1. The cost-of-service principle (CSP)
2. The value-of-service principle (VSP)

The origin (and rationale) of CSP boils down to the fact that port users differ from one another and they make different demands on the facilities. A fully loaded ship of deep draft requires deeper water than other ships. Long ships require wider berths, etc. It seems both "fair" and rational to the proponents of CSP that the port users who "cause" the cost of water deepening, berth widening, etc should also be responsible for those costs. Secondly, it is believed that misallocation of resources will result if this cost responsibility is not exacted. For example, it is claimed that ships will be too big if shipowners are not made responsible for the cost of port investments caused by the necessity of dealing with bigger ships. When a tariff or a recommended revision of a tariff of port charges is claimed to be "cost-based", it is almost invariably an example of applying CSP.

The value-of-service principle appears under a number of different names. "Charging what the traffic will bear", which is a household expression in the transport industries, is basically the same as VSP. There is, however, a slightly different slant in that charging what the traffic will bear indicates outright profit maximization, whereas the connotation of VSP does not necessarily involve more than a differentiation of port charges for a given service according to the individual "value" of the service as revealed by the individual willingness to pay.

The application of either of these pricing principles used to result in very complex tariffs of port charges. In recent times this has given rise to complaints about unnecessary difficulties, and in Sweden a radical streamlining of charging practices is well under way. However, no consistent *principle* of port charging seems to have emerged in place of CSP and VSP. It therefore seems worthwhile as an introduction to the main issues at stake here to present the idea of social optimal pricing applied to seaports in the simplest case of a single-berth port.

9.2.2 The optimal port charge in the single-stage, single-channel case

Queuing models are usually applied to seaports not to help in calculating optimal port charges but to determine optimal capacity. The treatment of the service time of ships as a random variable is symptomatic of this purpose. Thus, the big differences in the actual port-resource requirements of individual ships are disregarded. If we are to use the results of seaport queuing models, just as they are, for calculating optimal port charges, we will have to assume that all ships are the same in all relevant respects, that is to say that the *expected* service time of each ship is identical. It is still possible to assume that the actual service time is a random variable adhering to the negative-exponential distribution depending on various irregular external factors, such as the weather, strikes, etc. However, the port charge should be based on the *expected* queuing cost imposed on other ships, which means that it should be the same for all, irrespective of the actual time that a ship spends at the quay-side. This follows from the assumptions that the expected service time is s for all ships, and that arrivals are strictly random, which can readily be seen from the single-stage, single-channel model we have already discussed.

The optimal port charge per call, PC, is equal to the expected queuing time costs caused by another call, which can be represented by the product of the average value of ships' time, v, the expected number of calls per day, A, and the derivative of the mean queuing time, q, with respect to A.

$$PC = vA \frac{\partial q}{\partial A} = \frac{vAs^2}{(1-As)^2} = vs \frac{\phi}{(1-\phi)^2} \qquad (10)$$

In the right-hand column of Table 9:1 on page 280 the relative development of PC with increases in ϕ is illustrated. As can be seen, the rate of increase in PC is still much greater than of the mean queuing time, q. For example, at 50 per cent capacity utilization the optimal port charge should be equal to the cost of the laytime (queuing time plus service time) per call of an average ship. At this level of capacity utilization the total revenue per day would be equal to the total cost of an average ship-day, v. Previously, it has been argued that a representative value of the ratio of $(c_1+c_2)/v$ is 2/5. It thus follows that, at 50 per cent capacity utilization, the revenue from optimal port charging in the single-channel case would cover the total port-capacity cost twice over at least. At a level of capacity utilization of 30 per cent, the total revenue would fall to $v/4$, which may be insufficient for total capacity-cost recovery. It is apparent that the financial result is highly dependent on the capacity utilization.

9.2.3 Optimal crane and storage charges

In practice the port charges are in turn divided between charges on ships and charges on cargo. The idea is that some port services benefit ships in the first place, and other port services benefit cargo. This accords well with the structure of our two stage queuing model. Two kinds of port charges suggest themselves in this model - a crane-occupancy charge on ships, and a storage-occupancy charge on cargoes. As has been mentioned, the transit storage stage tends to be the bottle-neck in many ports, and the exclusive concern of seaport queuing models with the "sea-side" of ports is somewhat irrelevant.

Since an important point about the two-stage model is that it makes it possible to calculate one charge on the ships for occupying the quay-cranes and another on the cargoes for occupying transit storage space, it is necessary to make a distinction in the queuing time function between ships arriving at service stage one, A_1, and cargoes arriving at service stage two, A_2. If some cargoes are transferred direct between the ships and the land transport vehicles, then $A_1 > A_2$; if all cargoes pass through the port via transit storage, then $A_1 = A_2$.

It has previously been shown that the total queuing time, Z, is a function of the total service-time in stage one, $X = s_1A_1$, the total service-time in stage two, $Y = s_2A_2$, and the number of service stations, n_1 and n_2. The optimal charges are calculated for a given port capacity. As was mentioned in the introductory single-stage and single-channel case, the theory states that the pricing-relevant cost is equal to the expected queuing time caused by a further arrival, either of a ship in stage one or a cargo in stage two. The expected queuing time q can be written:

$$q = \frac{Z(X, Y)}{A_1} \tag{11}$$

The pricing-relevant costs, in stage one, PC_1, and in stage two, PC_2, are obtained as:

$$PC_1 = vA_1 \frac{\partial q}{\partial A_1} = vA_1 \frac{A_1 \frac{\partial Z}{\partial X} \frac{\partial X}{\partial A_1} - Z}{A_1^2} = vs_1 \left(\frac{\partial Z}{\partial X} - \frac{Z}{X}\right) \tag{12}$$

$$PC_2 = vA_1 \frac{\partial q}{\partial A_2} = vA_1 \frac{\frac{\partial Z}{\partial Y} \frac{\partial Y}{\partial A_2}}{A_1} = vs_2 \frac{\partial Z}{\partial Y} \tag{13}$$

One difference between the two occupancy charges is worth noting: the crane occupancy charge is equal to the marginal queuing cost minus the average queuing cost of ships, whereas the storage charge is made up of the derivative of Z with respect to Y without any deduction.

On the basis of Table 9:2 above, the values of Z/X, $\partial Z/\partial X$ and $\partial Z/\partial Y$ have been calculated along the expansion path. On the convenient assumption that $c_1 = c_2$, the expansion paths of n_1 and n_2 coincide, and $\partial Z/\partial X$ is equal to $\partial Z/\partial Y$, provided that X and Y are each measured in service-days. The same cost diagram (Figure 9:3) can be used to illustrate the level of the crane-occupancy charge and the level of the storage-occupancy charge.

Figures 9:3a and 9:3b show the SRMC-curves applicable to the present numerical example. On the horizontal quantity-axis the unit of measurement is a "service-day". The same curve is applicable, regardless of whether ship service-days or cargo storage-days are being considered. It should be noted, however, that the mean service time of ships, s_1, is much shorter than the mean service time of cargoes, s_2. The range of output of Figure 9:3a is therefore relevant primarily to the service-days of ships, while the range of output of Figure 9:3b is relevant to the storage-days of cargoes.

The crane-occupancy charge is obtained as the vertical difference between the marginal queuing cost, $SRMC_i$ (i=1...n, where n is the total number of service stations), and the average queuing cost, AVC_i, for each given number of service stations multiplied by s_1. As can be seen from Figure 9:3a, the optimal crane-occupancy charge can range, because of the marked "lumpiness" of capacity additions, from almost zero to a maximum corresponding roughly to the cost of s_1 ship-days. The more service stations there are, the less wide will be the possible range of the crane-occupancy charge.

The optimal storage-occupancy charge is equal to the whole of $SRMC_i$ multiplied by s_2. On the assumption that the mean service time in the second stage is something like 5-10 times longer than the mean service time in the first stage, it will often be true that the total volume of service-days in the second stage will be found in the range covered by Figure 9:3b, while the total volume of service-days in the first stage will be found in the range covered by Figure 9:3a.

9.2.3.1 The problem of factor indivisibility

It may seem rather odd that there is such a wide span of possible optimal prices, depending simply on what the optimum volume of traffic happens to be. However, the wide range of charges indicated in Figures 9:3a and 9:3b is unlikely to be

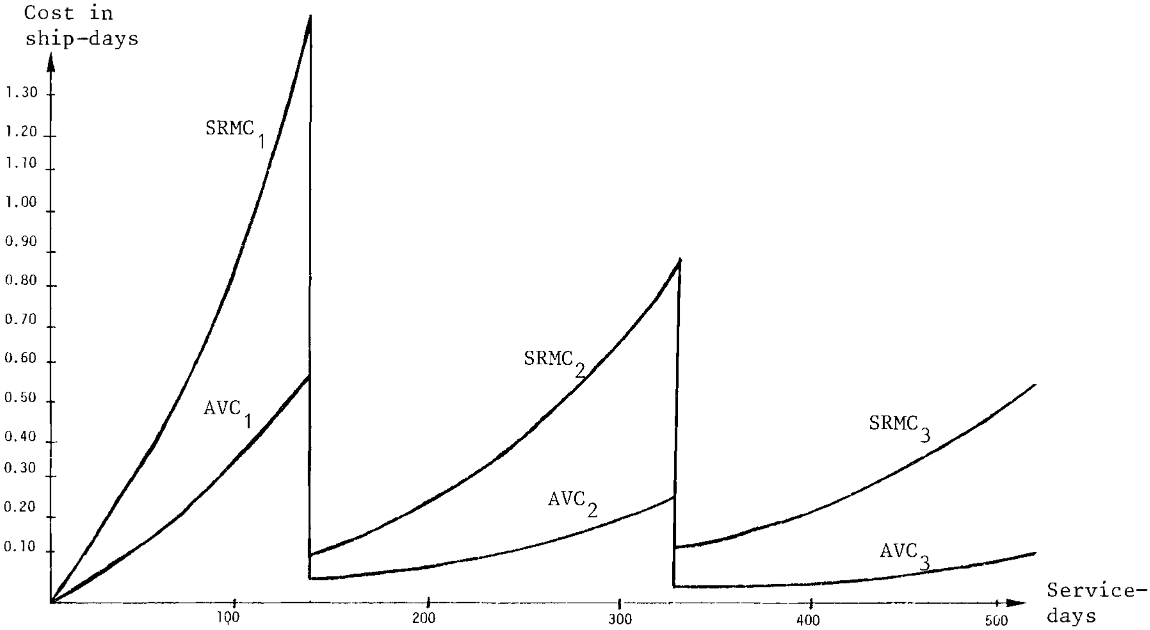


Figure 9:3a. Marginal and average queuing costs

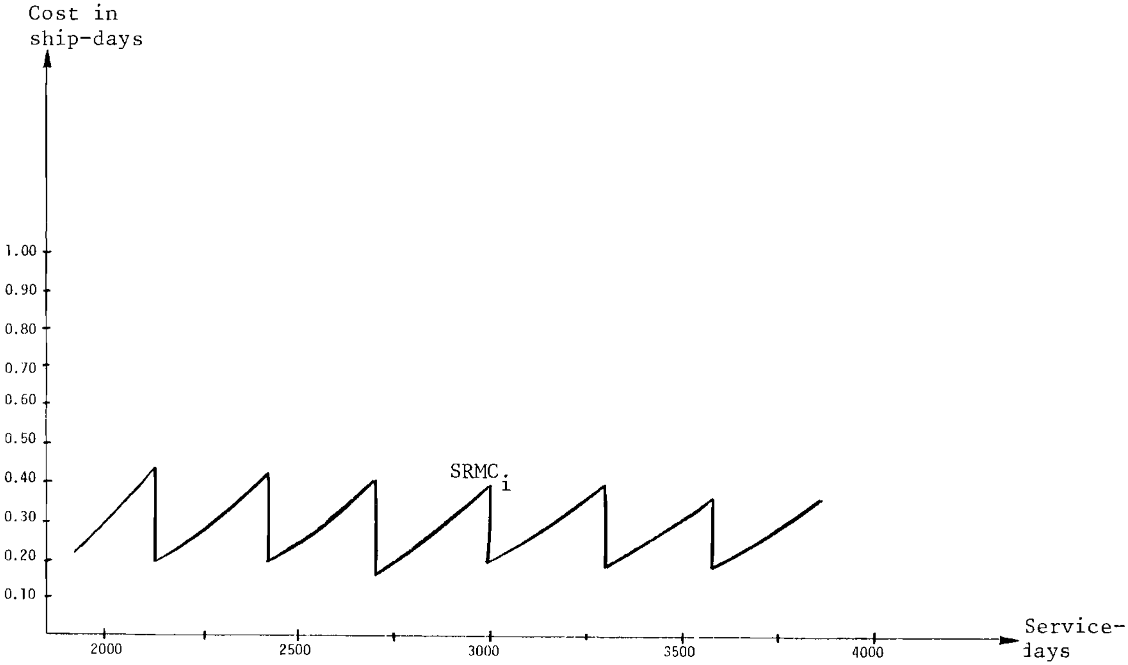


Figure 9:3b. Marginal queuing cost

applicable in reality. A considerably narrower span will be applicable for one or both of the following reasons: (i) In practice it may be possible - and this is not allowed for in the model - to mitigate the marked factor-indivisibility so that each SRMC-segment shows a smaller difference between the highest and lowest values. This will be the result, for instance, if the capacity of each crane can be varied. (ii) If demand is fairly elastic, the two extreme intervals for each SRMC-segment will become inapplicable; only an interval in the middle of each SRMC-segment will be relevant to pricing (apart from the initial range of the SRMC associated with a single service station). In the case of very elastic demand, this last levelling effect can be very significant.

To illustrate this point an explicit demand curve is introduced into the model. It is diagrammatically convenient to let the demand for ship service-days, X, be represented by the marginal net benefit of service-days for shipowners, which is a falling function of X. The choice between the alternative service-day quantities, X_1 and X_3, should be settled by comparing the two shaded areas A and B. If A is larger than B, X_1 service-days does not outweight the additional social cost including the cost of an additional service station, and vice versa.[1] It is obvious that if the relative positions of the demand and cost curves are such that X_1 is relatively close to X_2, X_3 is likely to be optimal, and vice versa. The extreme values of SRMC are unlikely to be relevant to pricing, provided that demand is fairly elastic.

[1] It may seem peculiar that such a conclusion can be drawn from a diagram where the cost curves do not directly include any service-station capital costs. The answer is that they do this *indirectly*. When the level of X_2 (see Figure 9:4) is passed, and a new SRMC-segment starts, the area of the "tooth" immediately behind no longer represents any laytime costs. This area has been transformed, as it were, to crane capital costs (or whatever the service station in question may be). The reason for adding another crane when the level of X_2 is passed, is that the saving in laytime costs is equal to the capital cost of a crane.

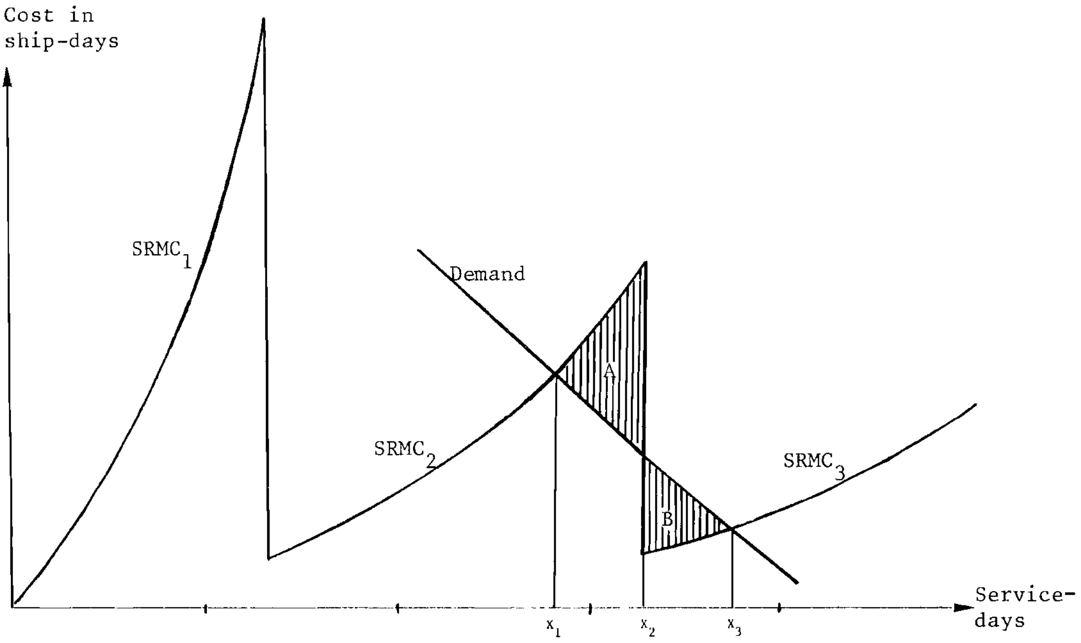


Figure 9:4. What is the optimum of service days?

Under dynamic conditions, with the demand curve shifting continuously over time, the question arises as to whether or not it is desirable and/or practicable to let the occupancy charges change continuously as well. This question has a long history in the general marginal-cost-pricing debate. There seems to be no general or definite answer: in each particular case it is a matter of degree. If the indivisibility is very marked, so that considerable overcapacity is created as soon as a capacity-expanding investment is undertaken, it seems wasteful to discourage the use of the facility by a charge which is not optimal until demand has grown very substantially, which may take several years. Analogously, during the year(s) preceding a planned capacity expansion, the facility may be rather congested. A high charge to relieve the congestion will be beneficial, particularly if customers are also informed that capacity will be expanded and the level of charges reduced at a certain time in the future.

In a case of marked indivisibility, the proponents of average-cost and marginal-cost pricing will be at loggerheads, particularly as regards the development of charges over time. The former will suggest that the charge should be initially high, because the capacity cost per customer is at its greatest then, and that it should be low when capacity utilization is at a maximum, because the capacity cost per customer is then at its lowest. But this is the worst possible way of time-structuring charges from the point of view of optimizing the utilization of capacity. A constant charge corresponding to a time-average of the marginal costs is much to be preferred. This problem is further discussed in an appendix to this chapter.

9.2.3.2 Optimal charges versus capacity costs

As for the financial result, the revenue from the occupancy charges should be compared with the service-station capital cost, $c_1n_1 + c_2n_2$. The relative level of the occupancy charges along the expansion path is indicated by introducing a curve for the average port-capacity cost in the previous diagrams for the pricing-relevant costs. In Figures 9:5a and 9:5b the average capital costs *per service-day*, c_1/X and c_2/Y, are given in terms of ship-days, on our earlier assumption that the cost of a ship-day is five times higher on average than the cost of a service-station day at each stage. In the diagram the producer costs are designated AC^{prod}.

The portion of the total service-station capital costs that would be recovered by levying optimum occupancy charges, is shown by the vertical differences between the pricing-relevant cost PC_1 or PC_2 and AC^{prod}.

In a case where only a few service stations are needed to meet the demand - that is, a situation covered by Figure 9:5a - it is clear that quite a substantial financial deficit

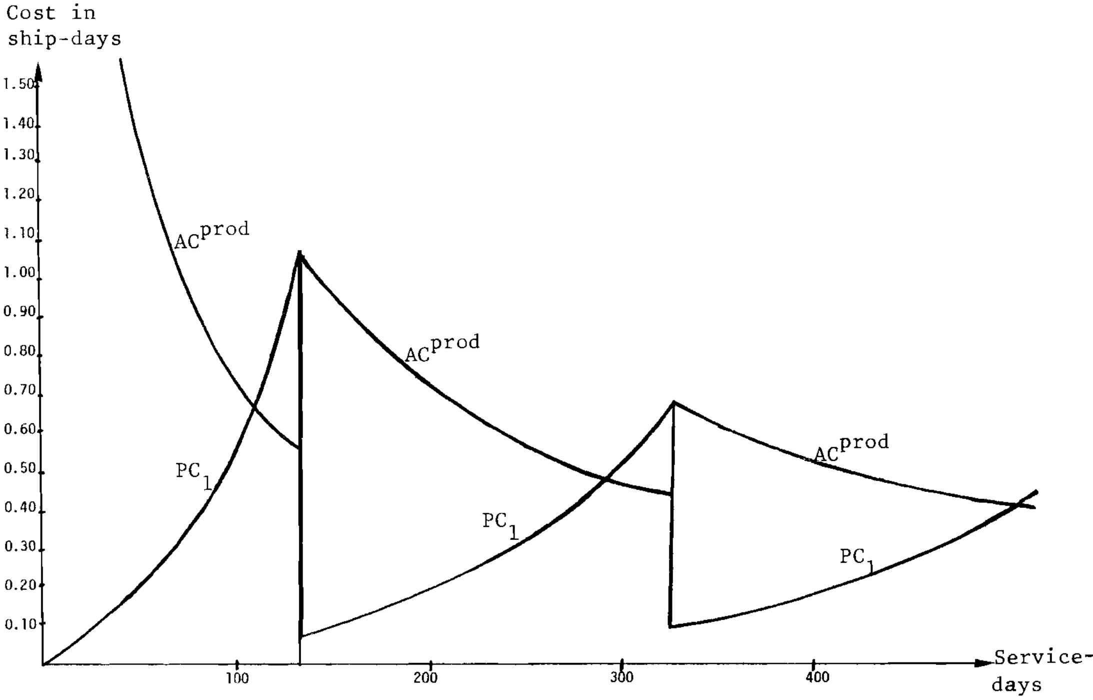


Figure 9:5a. Optimal charges versus port capacity costs: service stage one

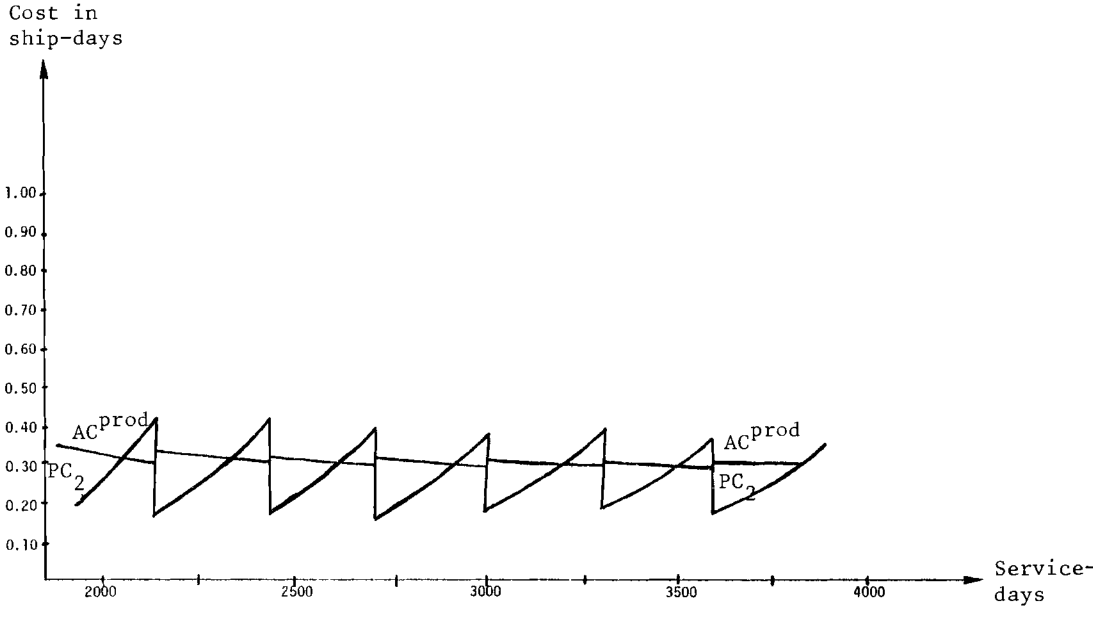


Figure 9:5b. Optimal charges versus port capacity costs: service stage two

is the most likely result. On the other hand, in a situation covered by Figure 9:5b, where 9-15 service stations are required, the financial result from optimal occupancy charges will not be that "bad". It even seems as though a financial surplus might result. However, for the aforementioned reasons the peaks of each PC-segment will rarely be relevant to pricing: small ports, in particular, applying optimal pricing would face a substantial financial deficit.

9.2.4 Price differentiation in ports

The main reason why such a mixture of port charges used to exist (and still exists in some ports) is that the great variety of ships and cargoes makes the equitable differentiation of charges an extremely complicated matter. The resource requirements of different customers are almost infinitely variable. Any very simple system of port charges is therefore bound to cause some complaints of unjust treatment.

It is no exaggeration to say that a large port supplies at least a thousand different services. It is even possible to claim that each individual customer makes virtually "unique" demands on port resources. To produce a port tariff listing the separate charges for each individual customer is out of the question. In order to produce a reasonably slim tariff, this kind of millimetre equity has to be renounced.

9.2.4.1 Input-pricing versus output-pricing

An alternative method of quoting port charges is based on input-pricing (as distinct from ordinary output-pricing). Instead of gearing charges to the final product, charges are based on the inputs used by different customers. Most goods in a market economy are sold, as a matter of course, at prices levied per unit of output. It would be patently ridiculous to present a prospective buyer of a car, for example, with a list of the prices of steel, various kinds of labour, etc. Input-pricing has various advantages so far as services such as those provided by a port. First, the number of inputs is very much less than the number of possible outputs. A tariff of input-prices would quote some five to ten prices for the different inputs used in port operations, and each customer would pay according to the amount of use of each particular input - man-hours, crane-hours, etc. It is clear that an almost perfectly equitable structure of charges (in terms of resource use) would be the result of input-pricing. Secondly, in view of the fact that the buyer of port services takes a very active part in the production of the services, input-pricing provides a certain incentive to economies on scarce resources which output-pricing does not provide. The more active the part played in production by the buyer - that is to say, the greater his freedom of choice of inputs for a particular service - the more appropriate will the input-pricing alternative be. Most port charges which exist in practice can be levied in the form of a price on inputs, or of a price on output. For example, stevedoring charges are usually presented as a tariff of prices per ton of the different goods loaded and unloaded, but in some cases the stevedoring tariff can be supplemented by a price per labour-hour, which becomes applicable under certain circumstances. If the port user is free to choose gang size, and/or the number of gangs that he wants to employ on each particular call, then input-pricing has much to be said for it. The quay, crane and storage *occupancy charges* are important examples of input-prices. There is no practical obstacle to replacing these charges altogether by a charge per ton of throughput - that is to say, by an output-price. In this case, however, the pricing tells port users nothing about the scarcity of these resources in relation to one another or in relation to stevedoring labour and ships' time. Where users are able to vary the combination of inputs required to produce a particular throughput, time-proportional occupancy charges seem to have clear advantages.

One obvious disadvantage of input-pricing should not be forgotten. It can be awkward for users not to know beforehand exactly what the loading or unloading of a particular cargo is going to cost. To make it possible for users to calculate the total final price, information about factor *productivity* has to be provided as well as the factor prices. Moreover, charging the shipowner per hour of stevedoring labour input, for example, has the perverse effect (in the short-run) that the more inefficient the stevedoring, the higher will be the total price per ton, besides which ships' laytime will also be longer. Another less apparent but nevertheless important drawback of input-pricing is that the pricing-relevant cost per unit of input is not constant for different users, but depends a great deal on the total input quantity required, thus destroying the apparent simplicity of input-pricing. This problem will be discussed further in the following sections.

9.2.4.2 Price differentiation by occupancy charges

In the present model a single criterion for the differentiation of charges is relevant - the relative expected queuing time imposed on other ships. However, a strategic assumption in the earlier discussion of optimal port-charging was that the customers - ships and cargoes - are homogeneous. Such variations in service time which nevertheless occur are thus assumed to be due to various external factors rather than to differences in the nature of the ships or the cargoes. This assumption is justified only so long as the purpose is to calculate the average level of optimal port charges with a view to examining the financial result. When different customer categories show systematic differences in their capacity requirements, an undifferentiated charge of PC_1 per call, and an undifferentiated charge of PC_2 per cargo (= shipload) using transit storage, will clearly be unsatisfactory. To apply the same charge to a ship requiring an average of ten days service time as to a ship that can normally be turned around in one day, would be both inequitable and allocatively inefficient. An undifferentiated storage charge on the cargo seems still more inadequate, in view of the fact that the "service time" in stage two is largely determined by the importers' willingness to collect their cargo promptly.

A simple - perhaps too simple - method of achieving a differentiation which is at least a step in the right direction comes readily to mind. By changing the object of the port charge from the "arrivals" of ships and cargoes to the "cargo tons" passing through the port, rough justice will be administered, provided that the actual service time is reasonably closely correlated to the total quantity of cargo being loaded or unloaded. The only necessary modification of the previously calculated pricing-relevant costs, PC_1 and PC_2, is to divide these costs by the average quantity of cargo loaded or unloaded per call.

Inevitably a flat charge per cargo ton would leave much to be desired as regards differentiation, first between shipowners and wareowners, and secondly within each of these two user categories. For example, given the cargo, a modern ship with easily accessible holds and an old ship with narrow hatches may differ widely in the time they take to load or unload. An alternative method of port-charge differentiation, which also comes to mind, is to apply input-pricing, making the charges proportional to the time spent occupying cranes and storage respectively.

In our model the following alternative suggests itself. Dividing PC_1 by s_1 gives the crane-occupancy charge per day, or $v(\partial Z/\partial X - Z/X)$, and dividing PC_2 by s_2 gives the storage-occupancy charge per day, $c\partial Z/\partial Y$. However, it cannot be stressed too much that PC_1 and PC_2, and consequently also the crane and storage-occupancy charges per service day thus derived, are *averages* calculated with an exclusive view to predicting the financial result of optimal port charging. It is *not* generally true that an appropriate differentiation of the charges on ships and on cargoes can be obtained by making the charges proportional to the time of occupancy. This important issue will now be taken up for discussion.

9.2.4.3 Progressivity in the occupancy charges

There are good reasons for assuming that the expected queuing cost imposed on other ships by an additional service-day is far from constant - in fact, it is likely to be increasing with increases in the total service time required per call. An indication of the true degree of progressivity is obtained by considering the Pollaczek-Khintchine formula for the mean queuing time, q, in the single-stage, single-channel case ((2) in this chapter). For any arbitrary distribution of the service time, s, and assuming that arrivals are Poisson-distributed, this formula gives us q as a function of s and the variance of s, var(s).

$$q = \frac{A\left[s^2 + \text{var}(s)\right]}{2(1-As)} \tag{14}$$

Suppose now that the variability of s is explained by different characteristics of the ships (and cargoes). Let us divide the ships into 1 ... j ... m classes. A ship of the j^{th} class takes s_j days to turn around, and the expected number of calls per day of ships of the j^{th} class is A_j. We then have:

$$A = \sum_j A \tag{15}$$

$$s = \frac{\sum_j A_j s_j}{A} \qquad (16)$$

$$\mathrm{var}(s) = \frac{\sum_j A_j (s_j - s)^2}{A} \qquad (17)$$

The product of A and var(s) can be developed as follows:

$$A\ \mathrm{var}(s) = \sum_j A_j s_j^{\ 2} + s^2 \sum_j A_j - 2s \sum_j A_j s_j = \qquad (18)$$

$$= \sum_j A_j s_j^{\ 2} + s^2 A - 2s^2 A = \sum_j A_j s_j^{\ 2} - s^2 A$$

Expression (14) for q can consequently be rewritten:

$$q = \frac{\sum_j A_j s_j^{\ 2}}{2(1 - \sum_j A_j s_j)} \qquad (19)$$

Another call by a ship of the kth class will increase the mean queuing time per call by $\partial q / \partial A_k$ ship-days.

$$\frac{\partial q}{\partial A_k} = \frac{(1 - \sum_j A_j s_j) s_k^{\ 2} + s_k \sum_j A_j s_j^{\ 2}}{2(1 - \sum_j A_j s_j)^2} \qquad (20)$$

Recalling that $\sum_j A_j s_j^{\ 2} = A\ \mathrm{var}(s) + As^2$, and that $\sum_j A_j s_j = As$, (20) can be written:

$$\frac{\partial q}{\partial A_k} = \frac{s_k^{\ 2}(1 - As) + s_k A \left[\mathrm{var}(s) + s^2\right]}{2(1 - As)^2} \qquad (21)$$

Multiplying this by A and the mean value of ships' time, v, gives us the optimal charge per call of a ship of the k^{th} class, PC_k. Observing that As equals the occupancy rate, ϕ, PC_k is found to consist of the two following terms:

$$PC_k = vA \frac{\partial q}{\partial A_k} = s_k^2 \frac{v\phi}{2s(1-\phi)} + s_k \frac{v\phi^2 \left[s^2+\mathrm{var}(s)\right]}{2s^2(1-\phi)} \qquad (22)$$

Given the occupancy rate and the distribution of service time, the left-hand term of PC_k is apparently proportional to the *square* of s_k, while the right-hand term is proportional to s_k. To get an idea of the relative importance of these two components, let us assume that the service time distribution is negative-exponential, and that the mean service time s = 2 days. At an occupancy rate $\phi = 1/3$, the optimal charge on a ship of the j^{th} class is $PC_j = \frac{v}{8}(s_j^2 + 2s_j)$.

The average ship taking two days to turn around should consequently pay 1v. A ship requiring twice that time should pay 3v, and a ship requiring four times more time should pay 10v. On the other hand, a ship requiring just one day to turn around should pay less than half v, or more exactly 0.38v.

The same sort of progressivity should apply to the transit storage charges. To some extent this already exists in a number of ports, namely in those ports where an initial period (up to one week) of transit storage is free. Storage time (for import cargo) exceeding the free period is charged at a flat charge per day. A better pricing principle, however, is to let the storage charge per day continue to rise with the time a cargo is kept in transit storage.

There is another dimension of the relative resource requirements, which can be viewed in much the same way as the different service time requirements. So far it has been assumed that a ship arriving at the port will occupy one, and only one, service station in stage one, and that every cargo is the same size so that the same transit storage space is required. In reality, ships of all sorts of lengths are likely to appear. As a rough rule-of-thumb it can be assumed that the number of cranes that can be usefully employed is proportional to the length of the ships of a particular type. Similarly, the size of cargoes (= shiploads) in fact varies greatly. I have not gone very deeply into this matter, but it seems that it should be possible to apply a structure of charges geared to the number of cranes or the storage space required on the same lines as the charges geared to the service time required. For example, the ratio of the charges on two different import cargoes of which one takes twice as much space and stays twice as long in transit storage, should be not 4:1 but perhaps as high as 10:1 or even higher.

APPENDIX: THE AVERAGE COST OF A MARGINAL SERVICE STATION AS AN APPROXIMATION TO THE MARGINAL COST

A way out of the financial dilemma of MC-pricing has seemed to many people to be to base the pricing on the *long-run* marginal cost. The true meaning of this alternative in a port with markedly indivisible factors of production can be considered with the help of Figure A:1. It can be seen that LRMC coincides with $SRMC_i$ ($i = 1 \ldots n$), but also that each individual $SRMC_i$ has a much wider scope than the segment which constitutes a link of LRMC. The point is that the parts of $SRMC_i$ situated outside the LRMC-segment represent less interesting ranges. Only if the existing number of service stations is inoptimal for the current level of output, can the optimal price differ from LRMC. Suppose that the position of the demand curve is as shown in Figure A:1. The demand curve intersects practically all $SRMC_i$ curves. However, under static conditions the only interesting point of intersection is the encircled one. This point represents the optimal output under the conditions portrayed.

Under dynamic conditions, in which demand increases (decreases) over time, various pricing rules can be optimal under different circumstances. A common characteristic, however, is that the average price level *over time* is likely to come close to the level of LRMC. But the level of LRMC is no "magic number". Only by chance will the average level of optimal prices over time exactly coincide with the level of LRMC.

However, it is not the latter dynamic case that most advocates of LRMC-pricing have in mind. They seem to be under the delusion that a systematically *higher* price level will result from basing the pricing on the long-run rather than the short-run marginal costs. However, this is impossible, since LRMC is nothing but a series of SRMC-segments

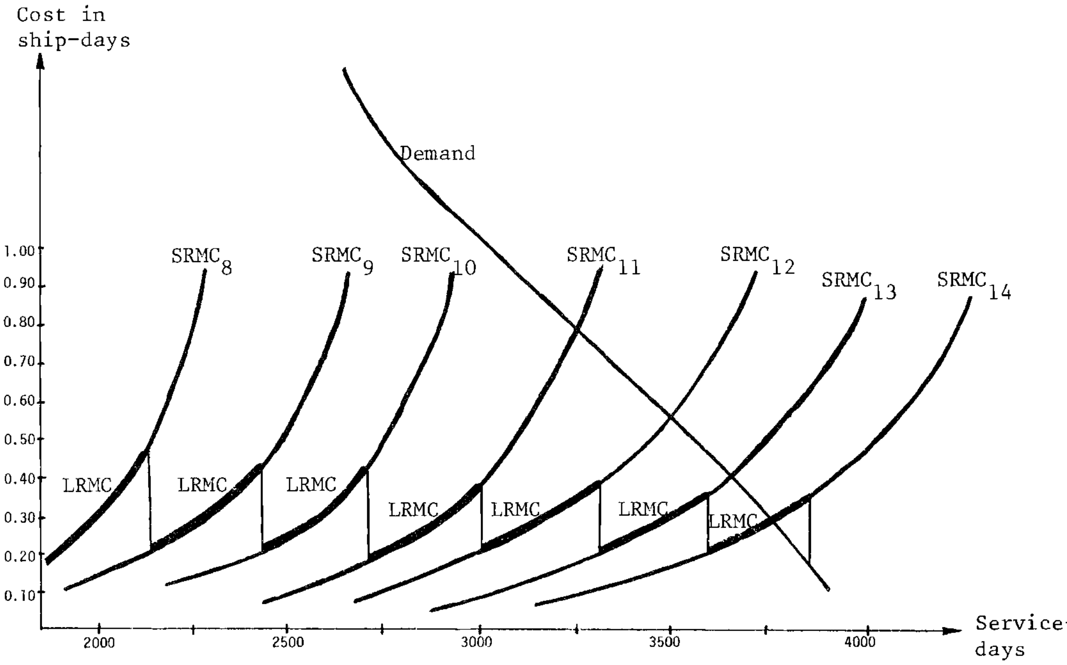


Figure A:1. Short-run and long-run marginal costs

of varying range. (When the production factors can be assumed to be perfectly divisible, LRMC becomes an absolutely smooth, continuous curve, which has only one point in common with each individual SRMC-curve.)

The reason for the persistence of the controversy is that the "long-run proponents" generally calculate LRMC in a different way. Typically, they take the ratio of the capacity cost of an additional unit of capacity to the corresponding increment of output as an approximation of LRMC. One well-known example of this method of calculation is the "average-cost-of-the-marginal-plant" approximation to LRMC, which has been applied for example to tariff-making in the electricity-generation industry in certain countries. Generally speaking, it is a practical method of deriving the marginal cost in *multi-plant* enterprises, and it could be useful in the case of fairly large seaport terminals. In small terminals, where the addition of another service station may mean a doubling of total capacity, it provides only a very rough approximation.

However, this is not the main problem connected with applying the "average-cost-of-the marginal-plant" principle. The main problem rather, is that it is frequently applied in a way that violates a fundamental condition for the validity of the principle - a condition that may seem obvious, but which is only too often unfulfilled: *the quality of the product (service) must remain the same after a capacity addition as it was originally*. It can often be optimal to change (normally to raise) the quality of service as capacity is expanded. In such a case the principle can be correctly applied in two ways:

1. If changes in the quality of service can be reasonably well expressed as changes in user cost, the change (a decrement, in the case of a quality improvement) in the total costs of the original users should be included in the numerator of the ratio of the total incremental cost to the increase in output.

2. If it can be assumed that the original situation was an optimum, it is not necessary to translate quality of service into user costs, even if an improvement (or impairment) in the quality of service actually occurs as a result of an addition to capacity. We can calculate the incremental cost per unit of the additional output in the *hypothetical* case where the quality of service is constant after an addition to capacity. The result will be the same as the result obtained by method (1).[1]

[1] The theoretical base for this idea is provided by Frisch's "indifference principle of the general marginal cost". See Frisch, R., Theory of Production. D. Reidel Publishing Co. Dordrecht-Holland, 1965, p. 166.

The second of these methods has been used for calculating the average cost of the marginal plant in the present model. The purpose of calculating LRMC in an alternative way was to be able to compare the result with the previously derived LRMC-curve and, in particular, to demonstrate that no systematic difference will emerge. We can now describe the calculation procedure in greater detail.

The incremental cost of a service station in the first stage = c_1, and in the second stage = c_2. If the occupancy rate remained constant after the addition of another service station, which would make the incremental port-capacity cost per unit of additional output equal to the average port-capacity cost, it is clear from the queuing-time formulas that the quality of service would rise - the mean queuing time would fall. We therefore let an increase in the occupancy rate accompany the addition of another service station, and this is chosen so that the mean queuing time remains unchanged. Finally, c_1 and c_2 are divided by the chosen increments in X and Y respectively. To summarize:

Approximation of LRMC of an additional ship service-day = $\frac{c_1}{\Delta X}$, where X is chosen such that

$$\frac{Z(X, Y, n_1, n_2)}{X} = \frac{Z(X + \Delta X, Y, n_1+1, n_2)}{X + \Delta X}$$

Approximation of LRMC of an additional cargo storage-day = $\frac{c_2}{\Delta Y}$, where Y is chosen such that

$$\frac{Z(X, Y, n_1, n_2)}{X} = \frac{Z(X, Y + \Delta Y, n_1, n_2+ 1)}{X}$$

As can be seen, there is an obvious difference between $c_1/\Delta X$ and $c_2/\Delta Y$ as regards the conditions for constancy in quality. This difference corresponds to that previously noted, i.e. that the optimal crane occupancy charge should be equal to SRMC minus AVC, while the storage occupancy charge should be equal to the whole of SRMC.

A comparison of the two conditions for constancy in quality indicates that, ceteris paribus, ΔX will be greater than ΔY and, when $c_1 = c_2$, $c_1/\Delta X$ will consequently be smaller than $c_2/\Delta Y$. In Figures A:2a and A:2b the values of these two approximations of LRMC are compared to the previously calculated pricing-relevant costs.

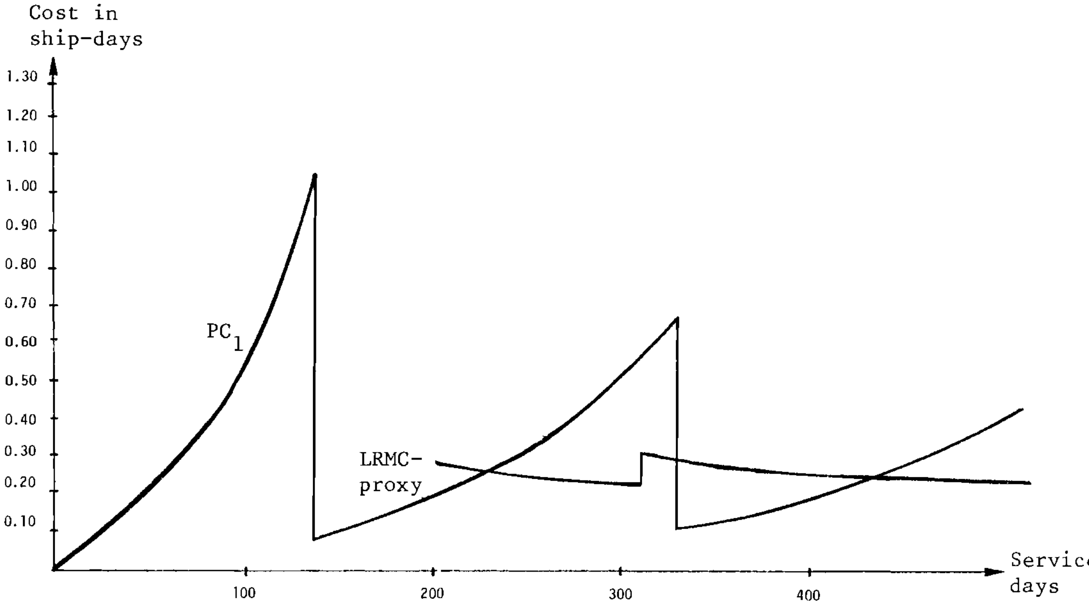


Figure A:2a. Comparison of the pricing-relevant cost per service-day in stage one, and the "average cost of the marginal plant" proxy

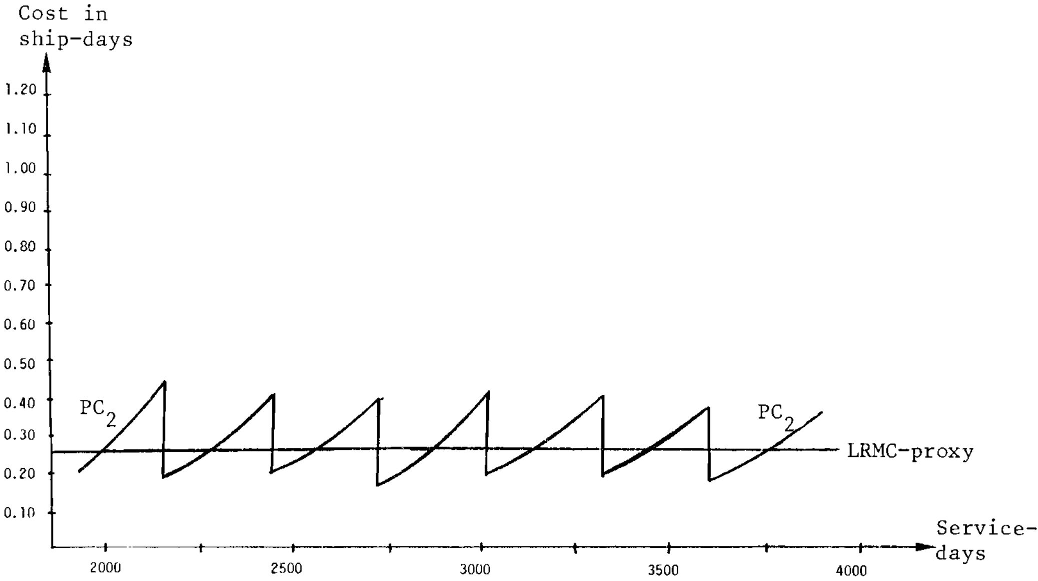


Figure A:2b. Comparison of the pricing-relevant cost per service-day in stage two, and the "average cost of the marginal plant" proxy

10 OPTIMAL ROAD USER CHARGES AND THE COSTS OF THE ROADS

10.1 PROBLEM AND PURPOSE

The question of optimal road user charges is a complicated one, since a number of cost categories are involved which are very difficult to calculate. The cost of accidents poses perhaps the greatest problem. In addition, the wide diversity of vehicles using the roads raises complicated problems of cost allocation. Since the 1950's a number of unusually long-lived public committees of inquiry into the costs and prices of road services have been set up in Sweden as well as in many other countries.

No attempt will be made here to give an exhaustive account of the whole problem area. Instead, the discussion focuses on one question, which seems to me to be the main issue so far as the national road network is concerned: would optimal road user charges generate sufficient revenues to pay for the roads?

For good reasons or bad, the theory of optimal road user charges has come to be divided into two parts which roughly correspond to the administrative division of the total road network.

So far as urban roads are concerned, the theory of optimal pricing - that is the theory of "congestion tolls" -

had already gained acceptance among economists by the mid 1960s.[1]

The main conclusion of this part of the theory in accordance with which only the city of Singapore has acted so far, is that peak motorists in cities pay far too little for using the roads. This conclusion was based not on a comparison of urban road traffic tax revenue and the costs of urban road-building, but on the fact that a huge discrepancy normally exists between the social and private short-run marginal costs of road use. Urban roads are seldom financed by such taxes; instead they are financed by taxes on property-owners and/or on the income of the urban community as a whole - although the municipalities get a share of the road traffic tax revenue indirectly, in the form of state grants to the main urban roads which are, so to speak, integral parts of the national road network.

In urban areas road-building is not the exclusive concern of the motor traffic. Pedestrians, shopkeepers, town-dwellers, etc also represent important groups interested in the location and design of the roads. In many cities investment in roads has almost ground to a halt, and far-reaching traffic regulations have been introduced latterly to safeguard the interests of town-dwellers rather than car commuters working in the town but living outside it. The idea that there should be a one-to-one relationship between urban road user charges and the costs of urban road-building, has never been very prominent.

The interurban road network, which is normally administrated and financed by the central government, is a different

[1] Largely thanks to the pioneering contribution of Walters, A.A., The Theory and Measurement of Private and Social Cost of Highway Congestion. Econometrica, 1961; and the report by the Smeed Committee (of which Walters was a member), Ministry of Transport: Road Pricing, the Economic and Technical Possibilites. HMSO, 1964.

matter. Outside the urban areas the location and design of roads is almost exclusively determined by the demands of the traffic. The roads are simply "plants" for the production of transport by motor cars and trucks. Under these circumstances most people feel that nothing short of full road cost coverage by means of road traffic taxation is equitable.

The question of road finance is politically quite important, and public committees of inquiry into the matter of road traffic taxation are set up with regular intervals. For these committees "the allocation of road track costs" is typically the main issue.[1] Economists approach the problem differently. The first question is, which are the optimal road user charges without regard to a particular budget constraint? If optimal charges would not yield the revenue required, the second question is, which structure of additional "taxes" on the road traffic, or on other parties benefiting from the roads is optimal? The most authoritative work along these lines appeared already in 1968 - Alan Walters' treatise on road user charges.[2]

The cost allocation approach can be meaningful provided that the average level of first-best prices is not too different from the average cost. However, the main message of

[1] A representative example is the British report on Road Track Costs, HMSO, 1968. It is symptomatic for the static state of the art that the recent "Consultation Document" on Transport Policy, HMSO, 1976 also contains a section about "The Allocation of Road Track Costs" with the same basic philosophy as the predecessor. For a commentary on the latter document, see Harrison, A.J., The Track Cost Issue. Journal of Transport Economics and Policy. January 1979.

[2] Walters, A.A., The Economies of Road User Charges, op.cit. Mention should also be made of a report in Swedish by an expert committee set up in 1965, known as "Vägkostnadsutredningen" (VKU) (Vägtrafiken - Kostnader och Avgifter, SOU 1973:32. The high scientific ambition set by VKU made original research necessary in several areas, and the inquiry became unusually lengthy: the report was submitted eight years after the committee was set up.

Walters' study was that this proviso is wholly inapplicable. Instead zero, or near zero, road user charges are optimal both on most links of the interurban road network and on local (rural as well as urban) roads. This conclusion was, and is still rather uncomfortable. It runs counter to the well-established "development cost" school of thought, and above all, it exposes the awkward conflict between equity and efficiency, which public committees of inquiry have typically sought to play down by the cost allocation approach. This may be politically most expedient, but it is unfortunate from a scientific point of view. To my mind, the primary issue is to what extent optimal road user charges would yield the revenue required by the national road administration (NRA) for a road building program aimed at the maximization of the net social benefits of road traffic. The cost allocation studies are begging the main question, and doing so a good solution to the issue of an efficient structure of road user charges on different types of vehicles is unlikely to be found, too.

The most interesting aspect of Walters' work in my view is that he puts forward a theory to explain *why* optimal road user charges would provide the NRA with an extremely poor financial result. But the theory does not seem to have been fully grasped either by its supporters or by its opponents.[1] Was it perhaps too sophisticated an explanation? The main purpose of the present discussion is to try to simplify Walters' theory. An additional advantage of such a simplification is that a certain amount of generalization can be afforded. The "road problem", involving the extreme difficulty of reconciling equity and efficiency, seems to be a general problem, at least so far as the rest of the transport infrastructure is concerned.

[1] An exception is the VKU report which, particularly in an appendix (Bilaga B), includes an interesting discussion along lines similar to Walters.

10.2 SUMMARY OF THE THEORY OF OPTIMAL ROAD USER CHARGES

Walters' conclusion should be qualified in some respects, if it is to be understood correctly. For this purpose we need a brief outline of the elements of the theory of optimal road user charges.

The following classification of the short-run costs of road use provides a useful basis for the exposé.

Table 10:1. Short-run average cost and traffic flow relationships

Cost category	Relationship with traffic flow
1. Costs of the National Road Administration	
Wear and tear resulting from road use	Constant
2. Road user costs	
Time and vehicle operating costs	Increasing
Accident costs	Constant
3. "Third party" costs	
Disturbance costs	Constant or decreasing
Accident spillover effects (grief, medical care, consumption loss)	Constant

The same information can be expressed diagrammatically. In Figure 10:1a the average road user cost is given as a function of the rate of the capacity utilization of a particular road. Since the average "time and vehicle operating cost" is increasing sooner or later, the corresponding marginal cost will eventually exceed the average cost. Close to the capacity limit the gap between the marginal and average road user costs is considerable. In Figure 10:1b the average costs to the NRA and to various "third parties" caused by the traffic on a particular road, are depicted in outline.

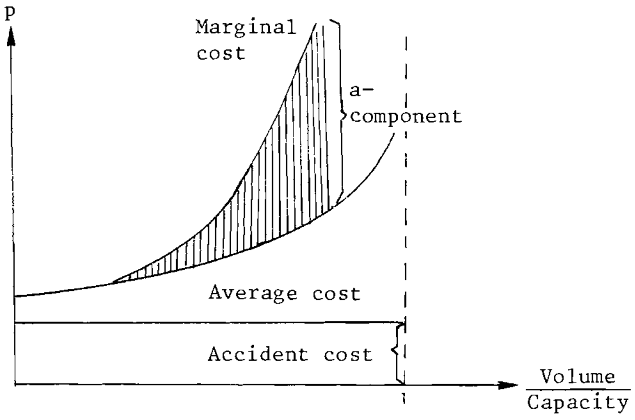


Figure 10:1a. Short-run road user costs

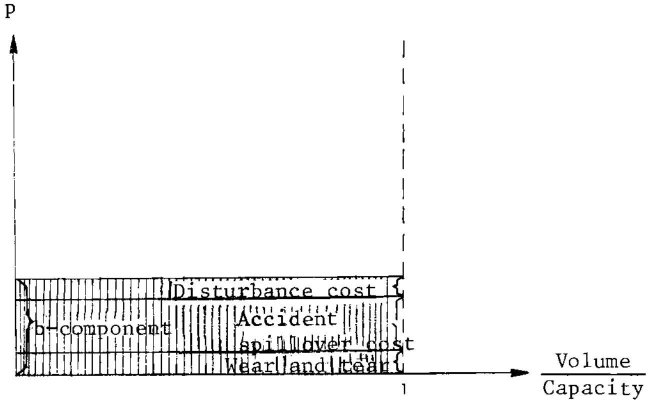


Figure 10:1b. Short-run costs of the National Road Administration and various "third parties"

In discussing optimal road user charges it is helpful to view the optimal price as consisting of two components, namely: the a-component which should be levied on account of the short-run costs borne by the road users themselves (illustrated in Figure 10:1a), and the b-component which should be levied on account of the short-run costs borne by the NRA and various third parties. These two components would have quite different effects on the financial result (for the NRA) of optimal road pricing.

On the fundamental assumption that the optimal rate of capacity utilization is normally well below the maximum capacity, only the a-component can be expected to make any contribution to the road investment costs. The revenue arising from the b-component of the optimal price would just suffice to cover the costs of (i) use-dependent road maintenance, (ii) medical care for the victims of road accidents, and compensation[1] to the relatives of people killed in accidents, as well as compensation[1] to society as a whole for the loss of future consumption caused by casualties on the roads, and (iii) compensation[1] to people suffering from disturbance by traffic.

The a-component of the optimal price is equal to the difference between the marginal (collective) road user cost and the average road user cost. This difference consists of the time and vehicle operating cost increases caused to the original traffic by another unit of traffic. A fitting name for the a-component of the optimal price is the "congestion toll". Algebraically, the relationship between the congestion toll and the road user costs can be expressed as follows: the average road user cost AC^{user} is a function of the traffic volume/capacity-ratio, ϕ, on a given road.

$$AC^{user} = g(\phi) \tag{1}$$

[1] In practice it is, of course, quite possible that some (or all) of these compensations will never be paid out. It is nonetheless inappropriate to regard this revenue as payment for the roads.

The imposed congestion cost constituting the appropriate congestion toll is equal to the product of Q and the increase in AC^{user} caused by an additional unit of traffic.

$$\text{Congestion toll} = Q \frac{\partial g}{\partial \phi} \frac{\partial \phi}{\partial Q} = \phi \frac{\partial g}{\partial \phi} \qquad (2)$$

The total road user costs is equal to $Q \cdot AC^{user}$, and the marginal (collective, as distinct from individual or "private") road user cost, MC^{user}, is obtained as:

$$MC^{user} = \frac{\partial\ Q \cdot g(\phi)}{\partial Q} = \phi \frac{\partial g}{\partial \phi} + g(\phi) \qquad (3)$$

As is seen, the congestion toll is equal to the difference between MC^{user} and AC^{user}.

It can be noted that the accident costs do not contribute to the congestion toll. The available empirical evidence does not support the much favoured hypothesis that the total accident costs increase progressively as the traffic flow increases, because vehicle interaction increases progressively as the traffic flow increases. Instead it seems that, given the type of road, and external conditions such as weather and light, total accident costs are basically proportional to the traffic flow.[1] No addition to the price is justified on account of the accident costs borne by the road users themselves, since the accident costs of the existing traffic on a particular road will not increase as a result of another car using the road. The motorists should pay just for the marginal spillover costs (to third parties) of accidents.

[1] Nilsson, G., Studier av samband mellan olyckor, vägens utformning och trafikens storlek. Statens Väg- och Trafikinstitut, Rapport nr 27, 1973.

10.3 TRAFFIC CONGESTION AND ECONOMIES OF SCALE IN THE PROVISION OF ROAD SERVICES

As regards road finance, the crucial fact is that the revenue arising from the a-component of the optimal price does not correspond (as in the case of the b-component) to a variable cost item - a cost which has to be covered, but which can make a genuine contribution to the road-building costs.

Now the real meaning of Walters' main conclusion can be clearly stated. He suggests that the optimal range of capacity utilization in the lifetime of a typical road (outside the CBD of cities) lies well below the range where the gap between marginal and average road user costs becomes appreciable. Consequently, if optimal road user charges were introduced, their contribution to the road-building costs would be very small. Walters says little about the b-component of the optimal price because, as we have seen, no real contribution to the road capital costs can be expected from this source of revenue, regardless of the level of the costs of use-related road wear and tear, traffic noise and other disturbances, and accident spillover effects. Walters' crucial contention concerning the low relative value of the a-component is based on somewhat impressionistic empirical evidence. However, as motorists we are all participants in the production process in question, and we can see for ourselves that on roads outside the larger towns and cities any congestion worthy of the name is a rare occurrence, except perhaps around public holidays.

It should, however, be pointed out that no systematic information about the existing level of road congestion costs is available at present. The difficulties of estimating a representative figure are obvious, bearing in mind that there are thousands of road links constituting the total network. Each road has an individual history: provided that there are any congestion costs at all on roads outside urban areas, the level of these costs for individual roads can be expected to vary comparatively widely depending on when the road was built, the external conditions (topography, geology, etc.) for the road building, and the subsequent traffic development. Moreover, it is well known that the volume/capacity-ratio fluctuates considerably both between the hours of the day, the days of the week, as well as between seasons.

Under these circumstances another method of predicting the financial result of optimal pricing can be more fruitful. By calculating the degree of economies of scale in the provision of road services, it should be possible to get an idea of the *average* level of optimal congestion tolls.

To my knowledge no estimations of the relationship between traffic volume and the total costs of road traffic have been made anywhere. Admittedly it has often been observed that a four-lane road is less than twice as expensive to build as a two-lane road while its capacity will be more than double, and this has been taken as evidence of economies of scale in the provision of road services. The only question is, how important these are. A problem which cannot be ignored in estimating the economies of scale, is that the quality of service does not remain the same for the road users when the capacity of the roads is increased. It would be very misleading to compare the costs of, let us say, a low-capacity gravel road with a motorway without taking into account the difference in the quality of service. There is only one adequate way of allowing for quality differences: the quality of road services should be translated into user costs - the cost of running a car or a truck, the cost of travel time, and the cost of accidents. Road user costs and producer costs should be added together, and the relevant scale-elasticity should be defined as the inverse of the elasticity of the total costs of road traffic with respect to the traffic volume.

It is sometimes claimed that the production technology of road service is of a markedly discontinuous nature, i.e. that relatively "lumpy" capacity additions are involved if a road is upgraded by one class. It is true that rather a limited number of road classes is in fact given in the norms issued by the Swedish National Road Administration so far as the cross-section of roads is concerned (in Sweden the choice is between no more than 6-8 classes), but the road alignment - the vertical and horizontal profiles - leaves literally endless possibilities for varying the design and consequently the capacity and quality of a road.

The following estimation does not do justice to the wide variations at present existing in road design, but for reasons of data availability it has been limited to the six standard types included in a recent study of road service quality undertaken by the Swedish NRA.[1] The road producer and user costs associated with the six types considered are given in Table 10:2. Three comments on the figures are pertinent: (1) The figures for the average flow per day represent the middle point in the flow interval for which each particular road is the best choice, except in the case of the highest (motorway) class where the flow figure given corresponds to the lower limit plus 10 per cent. (2) The expressing of the road maintenance costs as an annual total cost like the road capital cost, may give the impression that the former cost is a fixed cost in the short run. This is not quite true, although a major part of the road maintenance costs of a given type of road is independent of the traffic. (3) As can be seen, the vehicle running cost per kilometre increases slightly with increases in the quality of road. This may seem counter-intuitive, and is perhaps untrue. The problem is that there is no Swedish data on the influence on running costs of cars and trucks of different road qualities. On the other hand it is clear that, given the road, from about 50 km/h any increase in speed will result in increasing

[1] Statens Vägverk: En studie av vägstandard. P003, 1978-11.

Table 10:2. Road traffic and costs

ROAD CLASS	SPEED AND FLOW			ANNUAL ROAD COSTS PER KM OF ROAD (Sw.Cr.)		ROAD USER COSTS PER VEHICLE-KM (öre)		
Cross-section (paved width in metre)	Speed limit, km/h	Mean speed, km/h	Traffic flow vehicles per day	Road capital costs	Road maintenance costs	Driver's and passengers' time costs (value of time=20.25 Sw.Cr. per vehicle-hour	Costs of vehicles including fuel costs net of tax	Accident costs (cost per accident = 125 000 Sw.Cr.)
5 + 2 x 0.25	70	69	400	70 100	10 200	30	26	6.5
6 + 2 x 0.25	90	78	1 350	78 800	24 600	26	27	7.5
7 + 2 x 0.5	90	83	7 000	102 200	25 600	25	28	6
7 + 2 x 3	90	87	14 500	211 700	52 900	24	28.5	5.5
four-lane rd.	90	90	27 000	394 200	98 600	23	29	5
motorway	110	96	40 000	657 000	73 000	21	30	4.5

running costs (largely due to the rise in fuel consumption per kilometre), and the road-traffic cost model of the Swedish NRA therefore treats the vehicle running cost per kilometre as an item which increases as a road is upgraded.

It is true that the number of observations is too small to warrant any precise conclusions from a regression analysis. The order of magnitude of the economies of scale should, on the other hand, be revealed. Fitting an exponential curve to the six pairs of values for the total costs and traffic flow, gives us the following relationship:

$$y = 858\ x^{.87} \qquad (4)$$

y = annual road producer and user cost in Sw.Cr. per kilometre of road

x = average traffic flow per day

The total cost elasticity is thus .87, which corresponds to a scale elasticity of 1.15. This rather modest figure is somewhat surprising. There are economies of scale in the provision of road services, but they are not extraordinary. Admittedly, the data on which this conclusion is based are not plentiful. However, there is no obvious reason for assuming that this has seriously biased the estimation.

10.4 AN APPARENT INCONSISTENCY

There is thus a problem of reconciliation of two seemingly conflicting facts concerning roads outside urban areas:

- The existing congestion appears to justify but trivial congestion tolls
- The economies of scale in the provision of road services are, relatively speaking, of a moderate magnitude.

This apparent inconsistency has convinced many observers that current road investment policy is inoptimal - is to expansive: if the traffic congestion is negligible outside urban areas, why spend a lot of money on expanding the road capacity? This is, however, not the way Walters argues.

A main purpose of Walters' study is, as a matter of fact, to explain this apparent inconsistency - that is to say, to reconcile the fact that expansion of the interurban road network will continue to be a worthwhile line of investment in the foreseeable future, with the seemingly inconsistent fact that zero or near zero congestion tolls should be charged for the use of the same road network.

The gist of Walters' argument is that roads approximate to being "pure public goods" for certain reasons of road construction technology. This explanation is challenged here. It seems unnecessary to enlarge the family of pure public goods by a somewhat dubious bastard, in view of the fact - to be demonstrated presently - that there is a much simpler, unambiguous explanation at hand. The central idea of the alternative explanation is that the revenue from congestion tolls is a closer equivalent of the total return on the capital invested in a plant, or "*quasi-rent*", than of the total proceeds of the sale of the goods produced at the plant concerned. And the ratio of the quasi-rent to the capital cost is very sensitive to deviations from constant returns to scale, much more sensitive than the ratio of the total revenue to the total cost (assuming that marginal cost pricing is applied).

Before this new idea is developed, it is appropriate to look at Walters' explanation more in detail.

10.5 ARE ROADS PUBLIC GOODS?

In the case of ordinary private goods, a zero or nearly zero optimal price can perhaps be imagined as a temporary expedient, e.g. before demand has grown sufficiently to match the capacity of a new plant. But in the long run such a state of affairs is difficult to imagine. According to well-known theory, it would mean that the economies of scale were enormous. The standard proposition in question is that the ratio of the marginal cost to the average cost is by definition equal to the inverse of the scale-elasticity of output, E. If marginal cost pricing yields a zero price, this must mean that E is almost infinitely large.

As we have seen, this is not true about roads. Walters did not go very deeply into the question of economies of scale in the provision of road services. He chooses another approach by arguing that road services is not like ordinary goods but can be characterized as "pure public goods".[1]

> In practice, it seems that many rural and interurban roads do in fact approximate to being pure public goods; congestion is so infrequent and small that it can be ignored. (Walters, op. cit., p. 17)

Walters' definition of pure public goods is the accepted one in modern litterature. There is non-rivalry in consumption; that is to say, at any level of demand for a public good, an additional consumer's enjoyment of the service in question will neither exclude anyone else from consumption nor cause any disutility to the original consumers. Walters' explanation of why most roads approximate to being pure public goods is briefly stated, that road service quality and road capacity are to a large extent *joint attributes*.[2] Road investments are made chiefly to improve the quality of road service, i.e. to reduce the road users'

[1] It can be argued that this is in effect the same as saying that E is virtually infinite. By defining economies of scale as economies of the *density* of demand, it can be shown that those services which are commonly regarded as public goods - broadcasting, etc. - are characterized by E = ∞.

[2] For an elaboration of Walters' main line of argument, see Jansson, J.O., Medveten överdimensionering av kapaciteten. En för prissättning och finansiering av infrastrukturen relevant produktionsekonomisk egenhet. Socialøkonomen, Oktober 1970.

costs by making the roads shorter, faster and safer. In this process additional road capacity tends to arise as a by-product, unwanted in itself. Note that the continuous improvement in road quality should not include any significant reductions in congestion cost, for the argument to be relevant. The point is that it should not be possible to reduce capacity so much that some congestion effects are felt, without also impairing the quality of service in other respects at the same time. Of course, it is not particularly original to point out that capacity and quality are joint attributes, in the sense that for a given traffic volume more capacity means less congestion costs. This is quite a general characteristic of many types of production, and it certainly does not mean that the products concerned are public goods. The crucial requirement is that *given the volume capacity ratio*, the quality of the service will rise as capacity increases. That is what Walters' argument is all about.

There are some snags to Walters' explanation. First, all roads are manifestly not pure public goods. On the contrary, if optimal congestion tolls were levied, some *urban* roads would generate revenues that could far and away exceed the capital costs. Technologically there is no fundamental difference between rural and urban road construction. The main difference concerns the relative factor prices. Land values and the cost of traffic noise and exhaust fumes are far higher in urban areas than in the countryside. In other words, the nature of roads as public goods is not an inherent technological characteristic, since it is conditional on relative factor prices.

Secondly, it is true that roads are designed in such a way that the quality of the road service will increase, the greater the expected demand. Road transport production becomes less user-time-intensive, the larger the scale of production. However, this is not unique technological characteristic. It is typical of goods manufacturing, too, that large-scale production techniques are less labour-intensive in comparison with the techniques of small-scale production. In goods manufacturing it is generally perfectly possible - but highly uneconomic - to choose techniques in such a way that the labour cost content will in fact increase as the production volume is increasing. Walters never demonstrated that this is *impossible* as regards road traffic, or in other words, it remains to prove that the observed covariation of road capacity and quality reflects truly "fixed proportions", rather than the simple fact that it is economic to choose successively less user-time-intensive techniques of road transport as the traffic volume increases.

Thirdly, the salient feature of pure public goods such as television and radio broadcasting is that, given the range of a transmitter, the capacity to serve watchers and listeners within the range is virtually *unlimited*.[1] This is manifestly not characteristic of roads. The capacity of any road is a perfectly finite entity. Terminologically it seems desirable to reserve the label "pure public goods" to services provided by facilities of literally unlimited capacity.

10.6 AN ALTERNATIVE EXPLANATION OF THE APPARENT INCONSISTENCY

We can find a much simpler way of explaining why optimal road pricing in combination with optimal road investment would result in a huge financial deficit for the NRA, by first making the rather trivial observation that the optimal congestion toll is not an exact equivalent - at least, not formally - of the optimal price of general marginal cost pricing theory. The marginal cost relevant to the pricing of electricity,

[1] Increasing the range of a radio station is equivalent to increasing the length of a road, which is an irrelevant dimension of the capacity in the present connection.

for example, is the short-run marginal cost to the *producer*; in the case of road services, on the other hand, the optimal price is equal to the *difference* between the road *user* short-run marginal cost and the average cost. This is obviously something different from the producers' short-run marginal cost, which constitutes the price in the "ordinary case" of marginal cost pricing. Obviously, this does not mean that there need be any material difference between the ordinary case and the "special case" of road user charges. As a matter of fact, all the economists who have discussed the subject have taken the stand that there is no material difference, and that there is no reason to stress the formal difference either terminologically or otherwise.

Here a different stand is taken. The supposedly immaterial difference between ordinary marginal cost pricing and road pricing will be shown to make a significant difference so far as the financial consequences are concerned.

To put it briefly: compared with the situation in the markets for ordinary products, the revenue from congestion tolls is a closer equivalent of the total return on the capital invested in a given plant, or "quasi-rent", than of the total revenue of the output.

In road transport the road user inputs correspond to the short-run variable factors of production, and the road represents the short-run fixed capital of an industrial plant. Functionally, the road and the vehicles on the road are factors of production related to the same production function. The production process concerned involves the transformation of inputs (road capacity, vehicle-hours, fuel etc.) into the transport of people and goods. Administratively, however, the capital input represented by the roads is provided by a single organization (the NRA), while a multitude of other organizations and individual people control the complementary short-run variable factors. If the functional view of road transport production is taken, it follows that the optimal congestion toll - the difference between the short-run marginal cost and the average variable cost of road traffic - is the exact equivalent of the quasi-rent per unit of output of an industrial plant. The quasi-rent is generally defined as the difference between the total revenue and the total compensation to the short-run variable factors of production.[1] Walters' paradox can be explained quite easily, if we simply look at what determines the ratio of the quasi-rent to the fixed capacity cost, rather than the ratio of the total revenue to the total cost, assuming marginal-cost-pricing.

It will be shown below that the ratio of the quasi-rent to the capacity cost fluctuates much more than the total revenue/total cost ratio, as soon as constant returns to scale cease to apply. In the first of these ratios, values close to zero cannot be described as particularly remarkable; nor can values as high as 2 or even 3 be regarded as exceptional.

Immediately below an intuitive diagrammatic demonstration of this will be provided. As the present problem is a matter of general importance a more formalized analysis of the relationship between the quasi-rent, capacity cost, and scale-elasticity is presented in the following chapter, which concludes Part II.

Consider a standard diagram (Figure 10:2) of the average total cost (ATC), average variable cost (AVC), and marginal cost (MC) of producing "gadgets". In the short-run (SR) the difference between ATC and AVC - the average fixed cost - can be characterized as the "capacity cost". As the capacity limit is approached, all of SRATC, SRAVC, and SRMC

[1] This difference is a "rent" in the short run but not in the long run (hence "quasi"), in the sense that it is a compensation to the short-run fixed capital input which is not strictly necessary for the provision of the services of the capital.

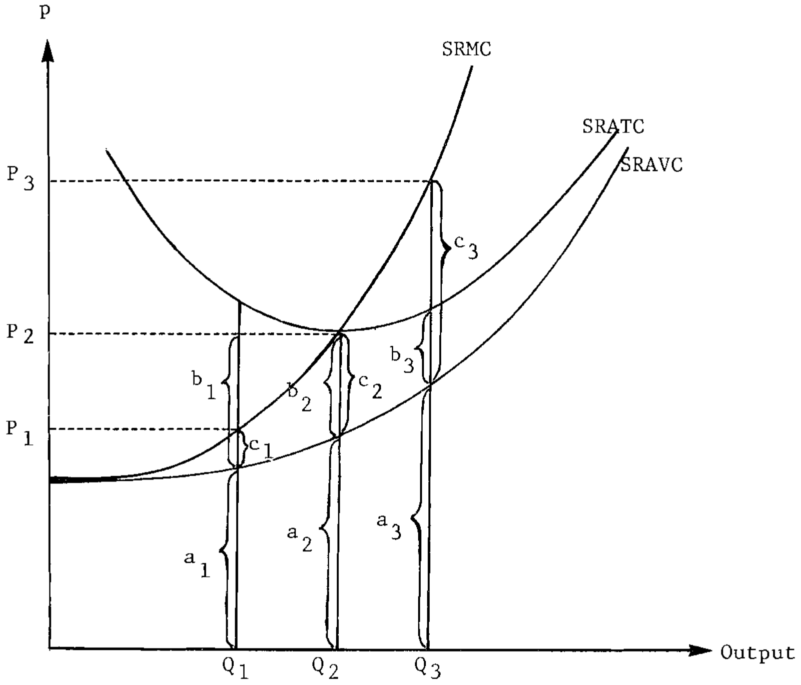


Figure 10:2. The relative size of the quasi-rent (congestion toll) in the case of (1) increasing returns, (2) constant returns, and (3) decreasing returns to scale.

are eventually rising, headed by the latter. If marginal cost pricing is applied, price and output are determined by the point of intersection of the demand curve (not given in the diagram) and SRMC. Let us compare three different points of intersection corresponding to the three price and output pairs, P_1; Q_1, P_2; Q_2, and P_3; Q_3. In these cases the average variable costs are a_1, a_2, and a_3, the average fixed costs are b_1, b_2, and b_3, and the quasi-rents per unit of output are c_1, c_2, and c_3 (see Figure 10:2).

It can be observed to begin with that when the point of intersection between the demand curve and SRMC happens to be where SRMC intersects SRATC, that is in the minimum point of the latter curve, the quasi-rent (c_2) is equal to the average fixed cost (b_2). To the left of this point the quasi-rent is higher than the average fixed cost, and to the right of the minumum point of STRAC, the quasi-rent is lower than the average fixed cost. The ratio of the price to the average total cost obviously behaves in a similar way. One significant difference is, however, that the deviations from unity are much larger in the former than in the latter ratio.

$$\text{Case 1:} \quad \frac{c_1}{b_1} < \frac{P_1}{a_1 + b_1} < 1$$

$$\text{Case 2:} \quad \frac{c_2}{b_2} = \frac{P_2}{a_2 + b_2} = 1$$

$$\text{Case 3:} \quad \frac{c_3}{b_3} > \frac{P_3}{a_3 + b_3} > 1$$

We now assume that each of these costs represent long-run, steady-state optima. For this to be possible it is necessary to assume that the long-run average cost curve takes different shapes in each case. It is clear that case 2 corresponds to that of constant returns to scale. In this case the long-run average cost (and long-run marginal cost) is made up of the minimum points of all SRATC-curves. In the case of economies of scale, or increasing returns, only points of the SRATC-curves to the left of the applicable minimum can coincide with the long-run average cost curve. Case 1 is consequently a case of increasing returns. As is seen in Figure 10:2, the ratio of the quasi-rent to the fixed (capacity) cost, c_1/b_1, can be down even to zero, while the ratio of the price to the average total cost will never go below the ratio of SRAVC to SRATC. Conversely, in the case of diseconomies of scale, or decreasing returns, only the part of SRATC to the right of the minimum point is relevant in the long-run. In case 3, representing an example of decreasing returns, it is seen that the ratio of the quasi-rent to the capacity cost can be very high - in the extreme the quasi-rent can be many times higher than the capacity cost - while the ratio of the price to the average total cost will be of a more moderate size.

Suppose now that the diagram of Figure 10:2 depicts the costs of a road link. The fixed capacity cost represents the road cost - capital and (use-independent) maintenance costs of the road, and the variable cost represents the road user cost. For expository reasons we disregard all possible costs to third parties (as well as the use-dependent maintenance cost). As we explained in the previous summary of the theory of optimal road user charges (see Figure 10:1a on page 334), the optimal congestion toll is constituted by the difference between the marginal (collective) user cost and the average user cost, which corresponds to the difference between SRMC and SRAVC in Figure 10:2. The given examples of quasi-rents per unit of output, c_1, c_2, and c_3, can also stand for three different congestion tolls. The ratios c_1/b_1, c_2/b_2, and c_3/b_3 are illustrative examples of what the financial result can be under different conditions with respect to economies of scale in the provision of road services.

With the diagram of Figure 10:2 before us, the apparently conflicting facts of optimal congestion tolls close to zero, and moderate economies of scale in the provision of road services can now be reconsidered. In this new light the inconsistency emerges as "more apparent than real". In particular, it should be pointed out that the observation that the traffic congestion is so relatively insignificant on the interurban road network that the average level of optimal congestion tolls is only a small fraction of the average cost of the roads, cannot be intepreted as a symptom of excessive road investment. Nor should it lead one to believe that roads are resources of a very special nature. A more pertinent reaction is instead that the particular relationship between the financial result of optimal pricing and the degree of economies of scale - i.e. that the relative deficit will be inversely proportional to the scale-elasticity of output, E - which we are so used to, does not apply to the transport sector. The shape of this relationship concerning transport services is taken up in Chapter 11.

11 SUMMARY AND CONCLUSIONS OF PARTS I AND II

Now it is time to take stock. We now summarize the results of the preceding case studies and, if possible, draw some general conclusions.

The purpose of this study has been two-fold. On the one hand, we have intended to operationalize the general theory of first-best optimal pricing - that is, the theory of marginal cost pricing - to make it applicable to a number of different transport services for which there is no adequate operational price theory. The general conclusions of this research are summarized immediately below under the heading "pricing-relevant costs". On the other hand, we have argued that a complete theory of first-best optimal pricing should also aim at predicting the financial result of charging such prices. In the second part of the present chapter this aspect of the theory is dealt with in two steps. First, a general formula for the financial result of optimal transport pricing is derived and secondly, the previous findings from all the different transport sub-sectors are recapitulated in the light of this formula.

11.1 PRICING-RELEVANT COSTS

In the absence of external effects the general formulation of the pricing-relevant cost, PC, is as follows:

$$PC = MC^{prod} + Q\,\frac{dAC^{user}}{dQ} \tag{1}$$

where Q is the volume of the transport service in question.

A diagramatical illustration of the pricing-relevant cost function is facilitated by splitting up the user cost component into two parts. The product of the transport volume, Q, and the derivative of AC^{user} with respect to Q can be written as the difference between the marginal cost of the users as a collective body,MC^{user}, and the average user cost, AC^{user} (which is normally equal to the private marginal cost of individual users).

$$Q\,\frac{dAC^{user}}{dQ} = MC^{user} - AC^{user} \tag{2}$$

The total pricing-relevant cost, PC, can thus alternatively be written:

$$PC = MC^{prod} + MC^{user} - AC^{user} \tag{3}$$

As far as the transport infrastructure - roads etc. - is concerned, the short-run costs are the pricing-relevant ones, i.e. the costs associated with the use of a given facility. The road pricing discussion has so dominated the modern literature of transport pricing that the case of a rising average cost curve with the corresponding marginal user cost curve lying above to make the difference between MC^{user} and AC^{user} positive, as illustrated in the top chart of Figure 11:1, has, explicitly or implicitly, been taken to be the truly general case. It may give a more balanced view of the theory of optimal transport pricing to consider at the same time the case of a downward sloping average user cost.

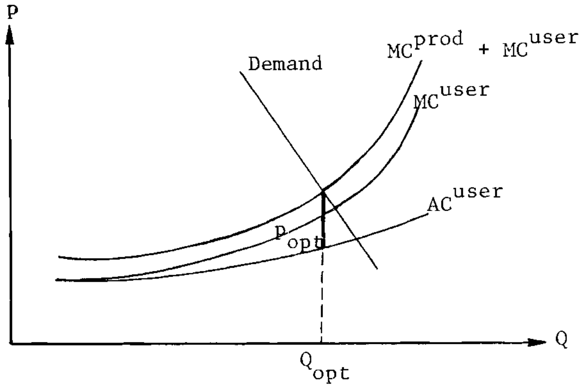


Figure 11:1a. Optimal price and output of services supplied by a piece of transport infrastructure

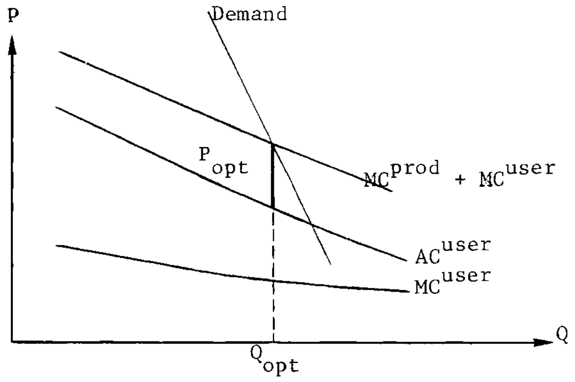


Figure 11:1b. Optimal price and output of scheduled transport services

As soon as the capacity of the facility can be regarded as variable, AC^{user} is downward-sloping in the normal case. It has previously been argued that concerning scheduled transport services, the "medium-run" costs - including above all the full costs of the vehicles with crew - are relevant to pricing. In practice it would seem almost absurd to argue that, for example, railway fares should be based on the short-run costs, implying given schedules and given lengths of trains. The practicioner is probably well aware of the fact that, if the expected level of demand is close to the existing capacity, the pricing-relevant cost including the appropriate user cost component is rather high and, conversely, if the expected level of demand is well below the existing capacity, the pricing-relevant cost is very low. However, before drawing any conclusion about optimal fares, the practicioner wants to know the optimal input of rolling stock. If some additional carriages are added in the former case, and if some carriages are withdrawn in the latter, it is obvious that the short-run "pricing-relevant" costs would change considerably: so, how could one say anything about optimal fares, before settling the question of optimal capacity?

If there is a large gulf between theory and practice - so much the worse for the theory. Our conclusion has been that, as regards all scheduled transport services, positions along any short-run PC-curve *off* the medium-run PC-curve are irrelevant to pricing, because in practice capacity optimization and pricing are carried out simultaneously.

The additional point which must not be overlooked is that, as a rule, there are important economies in the user costs of increasing the number of vehicles in the system (given the rate of capacity utilization).

If AC^{user} is falling with increases in Q in the medium-run, the difference between MC^{user} and AC^{user} is negative, constituting a negative component in PC.

In the bottom chart of Figure 11:1 the falling curve made up of the sum of MC^{prod} and MC^{user} represents the supply curve for scheduled transport services. The point of intersection with the applicable demand curve gives the optimal transport volume. As illustrated in the chart, the pricing-relevant cost is obtained by deducting the difference between AC^{user} and MC^{user} from MC^{prod}. The levels of the three curves of Figure 11:1b can, of course, vary a great deal. In some cases the optimal PC makes up a large part of the total marginal cost, and in other cases the optimal PC constitutes just a minor part thereof. The point conveyed by the diagrams of Figure 11:1 is just that, although the formula for PC - (1) or (3) above - is generally applicable, it makes a big difference to the relative importance of the producer cost and user cost components in the price, whether the short run or medium run is relevant to pricing.

11.2 THE FINANCIAL RESULT OF OPTIMAL TRANSPORT PRICING

One purpose of a theory of optimal pricing is to say what the optimal price should consist of in terms of costs, making the theory fit for empirical development and practical application. A different purpose is to seek to predict the financial result of optimal pricing. As has been argued in Part I, the latter purpose is also well worth striving to fulfil, in formulating a theory of optimal pricing. For the latter purpose a system context is necessary, in which the complete design of the transport system concerned should be assumed to be adaptable. In addition, it should be assumed that the least-cost design of every transport system is aimed at. There are many well-known examples of transport facilities being too big or too small for the existing demand. Nevertheless, there is no more appropriate basic assumption for the purpose than that efficient factor combinations are representative.

The principal point of the present theory of the financial result of optimal transport pricing is that the ratio of the total revenue to the total cost of the transport producer behaves just like the ratio of the quasi-rent to the capacity cost of a plant for storable goods production. As soon as deviations from the neutral state of constant returns to scale exist, the behaviour of the quasi-rent/capacity cost ratio is rather different from the behaviour of the total revenue/total cost ratio. This is shown immediately below by a more formalized model than was presented in the preceding chapter. The link between that model and the question of the financial result of optimal transport pricing is simply that in the short run, the transport user costs are equivalent to the variable costs, and the transport producer costs to the fixed capacity costs of an "ordinary" industrial plant. And as the user cost component of the pricing-relevant cost[1] applicable in the short run - called the congestion cost, crowding cost, or queuing cost, depending on what type of facility is concerned - comes out as the difference between the marginal user cost and the average user cost, the quasi-rent analogy is evident.

It can finally be observed that in the present connection, it is not necessary to maintain the dichotomy emphasized in the previous section by the two diagrams of Figure 11:1. Under the standard efficiency condition the pricing-relevant cost is the same, regardless of the degree of factor fixity. The fact that the medium run rather than the short run seems most relevant to the actual pricing of scheduled transport services is important, partly on account of the empirical problems involved, but it does not restrict the applicability of the following model when it comes to predicting the financial result of optimal pricing.

[1] The producer cost component of the pricing-relevant cost applicable in the short run does not normally make any contribution to the fixed capacity costs, and can be left out of the discussion.

11.2.1 The relationship between the quasi-rent and the capacity cost assuming marginal cost pricing

Consider an "ordinary" public enterprise where the organizational separation of "capital" and "labour" does not exist. That is to say, those factors which go into the provision of the capacity and those factors that are required for the capacity utilization are owned/hired by one and the same organization. Take the simplest case of a two-factor production function:

$$Q = f(K, L) \tag{4}$$

where K stands for "capital" which is a fixed factor in the short run, while L is a short-run variable factor of production - "labour" for short. Assuming that the capacity lasts for ever if properly maintained, the total cost is written:

$$TC = rK + wL \tag{5}$$

where rK is the capital cost and wL is the short-run variable cost. The short-run average variable cost comes to:

$$AVC = \frac{wL}{Q} \tag{6}$$

The objective of the undertaking is assumed to be net social benefit maximization. A necessary condition for this is that the price is set equal to the short-run marginal cost. As long as we are "on the expansion path", the financial result of marginal cost pricing depends only on the degree of the economies of scale in the production. The ratio of the total revenue to the total cost of an enterprise applying marginal cost pricing is equal to the inverse of the scale-elasticity of output.

As well as the relationship between total revenue and total cost, the relationship between the quasi-rent and the capital cost will also be examined. In particular, we will study whether the level of capital cost coverage will be the same as the level of total cost coverage, given a particular scale-elasticity. Intuitively we would regard this as unlikely, bearing in mind that the quasi-rent has the character of a residual: it is the remaining contribution to the covering of the short-run fixed costs, after the total variable costs have been deducted from the revenue.

The short-run marginal cost is equal to the ratio of w to the marginal product of L

$$\mathrm{SRMC} = \frac{w}{\frac{\partial f}{\partial L}} \tag{7}$$

Assuming that the price, p, is set equal to SRMC, the total revenue and the total quasi-rent are written:

$$\text{Total revenue} = pQ = \frac{wQ}{\frac{\partial f}{\partial L}} \tag{8}$$

$$\text{Total quasi-rent} = (p-\mathrm{AVC})Q = \frac{wQ}{\frac{\partial f}{\partial L}} - wL \tag{9}$$

From a financial point of view the most interesting entities are the total revenue *in relation to* the total cost, and the total quasi-rent *in relation to* the total capacity cost, respectively. The following two ratios are thus formed:

$$\frac{pQ}{TC} = \frac{wQ}{\frac{\partial f}{\partial L}\;(rK+wL)} \tag{10}$$

$$\frac{(p-\mathrm{AVC})Q}{rK} = \frac{wQ}{\frac{\partial f}{\partial L}\;rK} - \frac{wL}{rK} \tag{11}$$

These two ratios will be considered on the expansion path, i.e. under the efficiency condition that each volume of output is produced at minimum cost. The efficiency condition can be stated in the following familiar form:

$$\frac{w}{r} = \frac{\frac{\partial f}{\partial L}}{\frac{\partial f}{\partial K}} \tag{12}$$

Using this equality and by substituting $w \frac{\partial f}{\partial K} \Big/ \frac{\partial f}{\partial L}$ for r in (10), the total revenue/total cost ratio can be rewritten:

$$\frac{pQ}{TC} = \frac{wQ}{wK \frac{\partial f}{\partial K} + wL \frac{\partial f}{\partial L}} \tag{13}$$

Abbreviating by wQ, and designating the two partial elasticities of output (with respect to K and L) by E_K and E_L,[1] the total revenue/total cost ratio is obtained as:

$$\frac{pQ}{TC} = \frac{1}{E_K + E_L} \tag{14}$$

The *scale*-elasticity of output E is equal to the sum of the partial elasticities of output with respect to each factor. This well-known expression of the total revenue/total cost ratio is thus finally obtained:

$$\frac{pQ}{TC} = \frac{1}{E} \tag{15}$$

[1] $E_K = \frac{\partial f}{\partial K} \frac{K}{Q}$ $\qquad E_L = \frac{\partial f}{\partial L} \frac{L}{Q}$

A corresponding transformation of the expression for the ratio of the total quasi-rent to the total cost of capital is achieved in the following way. By substituting $1/\frac{\partial f}{\partial K}$ for $w\ /r\ \frac{\partial f}{\partial L}$ in the first term in (11), we are able to write the quasi-rent/capital cost ratio as:

$$\frac{(p-AVC)Q}{rK} = \frac{1}{E_K} - \frac{wL}{rK} \tag{16}$$

The inverse of E_K can be developed as:

$$\frac{1}{E_K} = \frac{1}{E} + \frac{E_L}{EE_K} \tag{17}$$

Now it can be noted that on the expansion path the ratio of the two partial elasticities is equal to the factor cost ratio, wL/rK. The final result is thus:

$$\frac{(p-AVC)}{rK} = \frac{1}{E} + \frac{wL}{rK}\left(\frac{1}{E} - 1\right) \tag{18}$$

When E is unity, the right-hand "additional" term vanishes and the result is, as expected, that the quasi-rent is equal to the cost of capital. However, as soon as either increasing or decreasing returns apply, the additional term becomes operative. The hypothesis suggested at the outset, that the quasi-rent/capital cost ratio is much more sensitive to deviations of E from unity than the total revenue/total cost ratio, is confirmed. A particularly interesting aspect is that the relative size of the short-run variable and the fixed costs is very important. Given a value of the scale-elasticity of output that differs from unity, we find that the ratio of the quasi-rent to the capital cost will fall in the case of $E > 1$ as the variable cost share increases, and that in the case of $E < 1$ it will rise as the variable cost share increases.

11.2.1.1 Cobb-Douglas exemplification

An illuminating example of this relationship is obtained by specifying the production function as a Cobb-Douglas function.

$$Q = A \cdot K^{\alpha} \cdot L^{\beta} \qquad (19)$$

In this case the ratio of the labour cost to the capital cost is equal to β/α and the scale-elasticity is equal to the sum of $\alpha + \beta$. Assuming that marginal cost pricing applies, the result will then be:

$$\frac{\text{Total revenue}}{\text{Total cost}} = \frac{1}{\alpha + \beta} \qquad (20)$$

$$\frac{\text{Quasi-rent}}{\text{Capital cost}} = \frac{1}{\alpha + \beta} + \frac{\beta}{\alpha}\left(\frac{1}{\alpha + \beta} - 1\right) = \frac{1 - \beta}{\alpha} \qquad (21)$$

Take first an example of moderate economies of scale. Suppose that $\alpha + \beta = 1.2$. The ratio of the total revenue to the total cost is then .83, irrespective of the ratio of α to β. The ratio of the quasi-rent to the capital cost, on the other hand, varies greatly depending on the relative size of α and β. When the value of β is five times the value of α, the quasi-rent will be zero, buth when the ratio is reversed, the quasi-rent will be 80 per cent of the capital cost (see Table 11:1).

Take also a case of diseconomies of scale. The result of assuming that E = .8 for the quasi-rent/capital cost ratio at different labour cost shares can be seen in the right-hand column of Table 11:1.

Table 11:1. The relative quasi-rent in the Cobb-Douglas case

Labour cost Capital cost	Quasi-rent / Capital cost	
	E = 1.2	E = 0.8
5	.00	2.50
4	.17	2.25
3	.33	2.00
2	.50	1.75
1	.67	1.50
1/2	.75	1.38
1/3	.78	1.33
1/4	.79	1.31
1/5	.80	1.30
.		
.		
.		
.		
.		
0	.83	1.25

11.2.2 Two causes of the serious conflict between efficiency and equity

The preceding model is, admittedly, very simple. All the different producer inputs are merged into a single one - "capital", and all user inputs are likewise represented by a single factor - "labour". Nevertheless, this simple model is a useful point-of-departure for summarizing the findings of Part II.

Two facts make the classical conflict between allocative efficiency and equity - or financial stringency, if you like - particularly difficult in the transport sector, namely:

1. Omnipresent increasing returns (at least outside urban areas)

2. Large user-cost shares

Let us first briefly recall our position concerning increasing returns with respect to the density of demand for transport.

11.2.2.1 Increasing returns

A peculiar ambiguity about the crucial matter of increasing returns is to be found in the literature of transport policy and economics, as well as in the general debate. On the one hand, the available empirical evidence of bus, railway, and air transport costs has been taken to refute the hypothesis that increasing returns are enjoyed in the production of transport vehicle services. As regards the transport infrastructure, no empirical studies of economies of scale have been made, apart from the investigation of railway track costs by Stewart Joy 15 years ago,[1] which has been interpreted as disproving increasing returns in the case of rail track services. On the other hand, many economists nevertheless seem to take for granted - either by assertion, or by reference to some alternative cause - that there is a serious conflict between MC-pricing and self-financing. The most frequent alternative explanation of this conflict is the prevalence of pronounced "indivisibilities" in the production of transport services. Unfortunately, the exact meaning of this term is not always stated. In an appendix to this chapter it is argued that factor indivisibility cannot be accepted as a basic cause of a negative financial result of first-best pricing, unless - as sometimes happens - the meaning ascribed to the term makes it synonymous with increasing returns.

Anyway, it has been shown here that there is no need for an alternative explanation. However, not until the transport production functions are "completed" by including the user inputs (without which no transports can be produced)

[1] Joy, S., op. cit.

will the issue of increasing returns be seen in the right light. The existence of significant economies of "plant" (vehicle, road link, seaport, etc) *size* in the producer costs in a wide initial size range and the existence of major economies of plant *number* in the user costs will necessarily together give rise to increasing returns, so long as there are no appreciable negative external effects from the concentration of plants for the provision of transport services. Barring this possibility, it follows that the greater the density of demand for transport, the lower the total cost per unit will be. As regards scheduled transport services, the number of economies in the user costs are most conspicuous with regard to the *frequency of service*, which is necessarily proportional to the demand along a particular route, given the size, speed, and load factor of the vehicles.

A symmetrical source of user cost degression is the positive correlation between the *density of service* and the density of demand. It is generally true of scheduled transport services that the average feeder transport distance will be longer, the more coarse-meshed the network of service. A study of maps of networks of railways, airlines, bus services as well as roads gives a vivid impression that the more sparsely populated a region is, the more coarse-meshed the network will be. Similarly, the density of the shipping lines connecting two continents is strongly correlated with the seaborne trade volume per mile of a coastline. Therefore, the cost to the users of scheduled transport services for transporting themselves or their shipments to and from the "stations" can be expected to be related to the density of demand in much the same way as the user costs of infrequency of services are related to the density of demand.

The same principle applies to the services supplied by the transport infrastructure. The more coarse-meshed the main road network, the larger will be the part of a trip of given length that has to be made on feeder roads of lower quality. And the larger the average distance between the seaports along the coastline, the longer on average will be the inland cargo transports. As regards seaports, we have also (in Chapter 9) drawn attention to the very significant berth number economies prevailing in each particular port, which is the main explanation of the economies of port size.

11.2.2.2 Large user cost shares

A new, rather surprising discovery is that the sheer magnitude of the user costs has in itself a decisive effect on the financial result of optimal transport pricing. Admittedly, increasing returns represent the only necessary precondition for the goal conflict between efficiency and equity, but the effect of a large user-cost share can make this conflict really dramatic, as can be seen from the basic formula for the financial result of optimal pricing:

$$\frac{\text{Total revenue}}{\text{Total producer cost}} = \frac{1}{E} - \frac{\text{user cost}}{\text{producer cost}} \left(1 - \frac{1}{E}\right)$$

The scale-elasticity, E, is certainly not constant throughout the transport sector. Our previous findings indicate, however, that the great variations found in the ratio on the left-hand side are still more due to the variations in the user cost/producer cost ratio.

In Chapter 2 the general importance of the user costs in transport was emphasized. It was also pointed out that the user-cost share varies greatly between different services. The following rough ranking list gives us four main groups of transport services, ranked according to a faller user-cost share.

1. Services provided by transport infrastructure
2. Short-distance scheduled passenger transport
3. Long-distance scheduled passenger transport
4. Scheduled freight transport

Let us recall our findings in each of these sub-sectors, and take another look at them in the light of the above basic formula.

11.2.3 The previous results in the light of the formula for the relative quasi-rent

11.2.3.1 Transport infrastructure

The reason why the user-cost share is particularly large in the transport infrastructure sector is, of course, that the user costs include the complete costs of the vehicles with crews using the roads, ports, etc., as well as the costs of time of the passengers or freight being transported by the vehicles.

So far as roads are concerned, it can be seen from Table 10:2 (page 340) that the value of the user cost/ producer cost ratio rises sharply in the initial range, and remains at a fairly constant level in the upper range of traffic volume. It rises from around unity to as much as 10 or more in the upper range. Given this rather high figure, we would expect that, despite the modest level of the scale-elasticity, E, the congestion toll revenue would be very low compared to the road capital cost. If we insert the figure of 1.15 for E as found in the empirical estimation of Chapter 10, we find that the congestion toll approaches zero as soon as the user cost/producer cost ratio exceeds six. Negative values of the congestion toll would obviously be unreasonable. A user cost/producer cost ratio of 10, such as we found for the upper flow range, cannot be combined with an E-value of 1.15, such as we found when the whole flow range was being considered. In the same way as the user cost/producer cost ratio, the value of E is likely to change along the expansion path. The figures in Table 10:2 indicate that when the user cost/producer cost ratio is as high as about 10-12, the value of E is only very slightly greater than unity. However, given the very large proportion of user costs in total costs, the financial result of optimal road charges can nevertheless be a huge deficit.

Similar conditions seem to apply to seaports. This is contradicted to some extent, however, by the analysis in Chapter 9 of seaport capacity optimization and pricing. A particular assumption of the queuing models used in that analysis has proved more crucial - in the present light - than was originally realized. The only user cost that is assumed to be substitutable for the port capacity costs is the queuing cost. The *total* queuing cost is relatively modest at optimum. Hence the financial deficit from optimal port charges turned out to be of such a magnitude as could be expected from a normal allowance for the degree of economies of scale. The queuing models in fact involve a number of important simplifications.

First, given the technique of cargo handling: (i) the service time per ship is assumed to be constant, and (ii) the direct handling cost per ton of cargo is assumed to be constant, i.e. independent of the total throughput and of the rate of capacity utilization.[1]

Secondly, only one technique is allowed for - the problem dealt with is optimal "capacity widening". If a choice of cargo handling techniques were introduced into the model, it is unlikely that either ships' service-time costs or direct handling costs (stevedoring costs) would any longer appear in strictly fixed proportions. That could make a great difference to the present issue.

In seaports three major cost categories are:

1. Port capacity cost (= capital cost of port facilities)
2. Stevedoring costs
3. Ship and land transport vehicle time costs (queuing costs and service time costs)

[1] An investigation of the realism of these standard assumptions of queuing models of seaports is made in Jansson, J.O. & Rydén, I., Swedish seaports - Economics and policy. EFI, 1979. The result was that the constancy hypothesis could not be rejected.

Items (2) and (3) are of a comparable order of magnitude, and both are normally many times greater than item (1). The port user charges (corresponding to the road user charges) are levied on ships and cargoes by the port authority, which is the owner of the port capital. If, as might seem most natural, the port capital and stevedoring labour were owned/hired by one company, the user-cost share would be of a moderate size. However, this structure of organization is fairly rare in seaports; the stevedoring company or companies of a particular port are normally independent of the port authority concerned. In a number of ports the stevedoring companies are owned by the port users, primarily the shipowners. In this case the total stevedoring labour cost can be regarded as a user cost on a par with the laytime costs of the ships and cargoes. Under these circumstances the user-cost share is very substantial indeed. If the port charges were set to attain an optimal utilization of resources, a very poor financial result for the port authority could be expected in view of the fact that, among other things, there are significant economies of "berth number" in seaports where ships arrive more or less at random.

Finally, it should be noted that the dilemma of a large user-cost share is not necessarily a problem of a large financial deficit. A value of E which is less than unity can exist, for instance in the case of urban roads and some airports serving big cities. In such cases relatively large quasi-rents might well be earned, if optimal pricing were applied. For example, if other cities facing severe problems of traffic congestion followed the lead of Singapore and applied optimal road pricing, the financial result in many cases could be that the road capacity costs would be covered many times over.

11.2.3.2 Short-distance public transport

In short-distance scheduled passenger transport the user-cost share is comparatively large, because both the feeder transport cost (normally the cost of walking to and from railway stations or bus stops), and the waiting-time cost are, relatively speaking, major items in the total cost of urban transport systems.

Concerning (conventional) urban bus transport, we have found in Chapter 6 that the financial consequences of charging first-best optimal fares would be very serious indeed. There are two main options to consider. The cost data used in our model are not sufficiently accurate to make a choice. One option is to offer the services free of charge everywhere and at all times except across the critical section in the most busy peak round(s), where a fare should be charged, which would cover the bus size-dependent part (bSN) of the total costs to the bus company, aN + bSN. The bus size-dependent part of the total bus company costs rises from about 10 per cent to 25 per cent, as the bus size increases from 20 to 75 passengers. The required subsidies would consequently be much larger than the taxpayers are burdened with at present. In the other alternative the financial consequences would be somewhat less disastrous, because, besides the peak fare, a basic flat fare would be levied on every rider on account of the imposed costs of boarding and alighting and/or to discourage riding for fun or for the warmth of the bus. There is too close a balance between the collection and impediment costs and the allocative benefits of charging such a basic fare to recommend one or the other of the two alternatives. Anyway, the total revenue would be nearly doubled in the second case, raising the degree of total cost coverage to some 20 to 50 per cent, depending on the level of demand on the route in question.

These orders of magnitude are borne out by the general formula. The average social cost elasticity of bus transport with respect to the passenger flow on a particular route can, first, be calculated on the basis of the figures of Table 5:7 (on page 162). It is found to range from -.367 to -.135, corresponding to a range of value of the scale-elasticity, E, from 1.58 to 1.16. Secondly, according to Tables 5:7 and 5:8, the ratio of the user costs to the producer costs rises from 1.2 to 1.6 as the level of demand increases. Applying these figures in the formula for the relative quasi-rent gives us a range of values of the degree of bus company cost coverage from 19 to 64 per cent.

11.2.3.3 Long-distance public transport

In long-distance scheduled passenger transport the user-cost share is generally smaller than in urban public transport, because, relatively speaking, the feeder transport cost and waiting cost are less important items in the total transport system costs.

Within this sub-sector the user-cost share is likely to vary considerably among the different modes of transport. The bigger (in terms of passenger holding capacity) and slower the vehicles are, the lower the transport producer cost per passenger tends to be. On the other hand, the value of time of passengers going by slower modes can be assumed to be rather less than the value of time of those choosing a faster mode of transport. Therefore, the user-cost share in, for example, air transport and sea transport may not be as different as it may at first appear.

On the other hand, compared with short-distance transport, the economies of scale seem to be very significant in long-distance passenger transport, mainly due to considerable size-economies in the producer costs in a relatively wide size range. Jumbo-jets, ocean liners and trains of the length seen on the main railway routes of the European continent exhibit rather low costs per passenger-mile in comparison to the aircraft, ships and trains employed in short-distance traffic. The costs of boarding/alighting, which increase progressively as the vehicle size increases, would become too large, if the very biggest vehicles were used on shorter distances. This means, in brief, that for long-distance passenger transport the economies of vehicle size in the producer costs is the main source of overall increasing returns, while the economies of vehicle number in the user costs is the main source of increasing returns concerning short-distance, and, in particular, intraurban public transport. Note again, however, that the *coexistence* of these two sources of increasing economies of scale is a prerequisite for overall increasing return.

We have studied neither air transport nor sea transport for passengers and cannot say anything specific about the likely financial outcome of optimal pricing in these sectors, except that, on the basis of the general discussion of "optimal pricing of scheduled transport services" in Chapter 4, it can be expected that appreciable financial deficits would result.

Railway companies are interesting in this connection in that they control the production of both the transport vehicle services and the track services. As regards track services alone, let us suppose that rail transport was organized like road transport, so that the rail track was used by independent rolling-stock owners paying track-user charges to the rail-track owners. In that case it can be envisaged that, just as in the case of roads, optimal "track congestion tolls" would cover only a fraction of the track capital costs of the interurban railway network. With the present form of organization, the total revenue/total cost ratio might not be as low as that, but if someone cared to calculate the "quasi-rent" earned on the track capital, the conclusion would probably be that a high degree of track-cost subsidization is required.

On the other hand, why should optimal pricing of the train services alone produce anything but a poor financial result? The cost per seat-mile of running, say, just one train per day consisting of only a couple of carriages on a "twig-line", is much higher than the cost per seat-mile of the trunk-line services where full-length trains can be run. This has been constantly argued by the railway transport operators themselves, and is, of course, manifest in their pressing for closing down "twig-lines" and smaller branch-lines. It is, on the other hand, not revealed by cross-section studies of the average costs of railway *companies* (in the USA) for the simple reason that most individual companies are running low-volume as well as high-volume services. Nothing else is thus to be expected than that optimal pricing of train services would be involved in the same dilemma as optimal pricing of track services.

Unfortunately, there is a dearth of relevant empirical work on the pricing-relevant costs of railway transport services. In Sweden an attempt was made recently by the Transport Policy Committee of Inquiry, reported in TPUII. In briefest summary, that attempt was based on the assumption that the average social variable cost is a valid proxy for the pricing-relevant cost. However, not all relevant social costs are taken into account by TPUII. The railway transport user costs are not included in the analysis. Since the Committee did not have access to the accounts of the Swedish Railways (SJ), it simply had to assume that the same ratio exists between social marginal cost and producer average cost as is said to exist in the Norwegian Railways, i.e. 2/3. The Committee estimated that as far as the rail track services are concerned, the pricing-relevant cost is only 27 per cent of the average cost. Given that the ratio of the total rail-track cost to the total train service cost is about 1/3, this means that the pricing-relevant cost of train services is assumed to be almost 80 per cent of the producer average cost of train services. This follows from the first assumption that the total marginal cost/average cost ratio is 2/3.

The main problem in judging the reasonableness of this figure (80 per cent), is that it is an average for both passenger and freight transport. A "guesstimate" based on the discussion in Chapter 4, is that this percentage is considerably lower for passenger services, while it may be about 80 per cent or even higher in the case of freight services. In any case, optimal railway transport pricing is an area particularly ripe for theoretical clarification and relevant empirical research. Swedish Railways have suddently become aware of the possibilities offered by a flexible pricing policy, and they also seem to accept that the railway is an enterprise that should be run in the public interest. In this new situation it is very desirable to have a clear notion of what optimal pricing would actually mean. The financial consequences might well be very uncomfortable, but if this were so, it should naturally not be disguised.

11.2.3.4 Scheduled freight transport

In Chapters 7 and 8 liner cargo shipping was taken up as the most apparent contrast - within the group of scheduled transport services - to bus transport, which has been discussed in Chapters 5 and 6.

In the light of the formula for the relative quasi-rent the most important difference seems to be that the user-cost share is many times less in liner shipping. There are economies of trade density of a "normal" order of magnitude. Values of the scale-elasticity from just above unity to 1.15 or possibly 1.20 are suggested by the shipping cost model of Chapter 7. The user-cost share is, however, so relatively small - no larger than about 1/3, and normally much smaller - that the deficit resulting from optimal pricing would be insignificant compared to that of urban bus transport. The main real explanation of this significant difference in the financial result is the *relatively* inexpensive possibility of mustering a full cargo simply by letting it accumulate over time, and/or extending the cargo catchment area. Provided that these means of mitigating the thin trade problem exist for other modes of scheduled freight transport, the conclusion that the classical conflict between efficiency and equity in decreasing-cost activities is not very serious in liner cargo shipping, can be generalized to the whole subsector of scheduled freight transport.

11.3 THE REASON FOR PART III

The main purpose of this study has now been fulfilled. A theory of first-best optimal transport pricing has been formulated in such a way that operationalizations for different transport services are now a straightforward matter.

The likely financial result of charging optimal prices has also been considered for different transport services.

The conclusions in the latter respect have been disturbing. In some important transport sub-sectors first-best optimal pricing would be in sharp conflict with the desideratum of financial self-support, which can also be entertained on grounds of equity and perhaps of X-efficiency.

It thus seems that a very wide selection of alternative transport policies have to be considered by central and local governments alike, ranging from (1) granting the heavy subsidies which are necessary to first-best optimal pricing of the services provided by the transport infrastructure as well as by public transport companies, to (2) stipulating that every single enterprise in the transport sector, whether private or public, should pay its way, despite the fact that substantial losses in social surplus will be made.

The case made by economists for strict marginal cost pricing has not been generally embraced by transport policymakers and decision-makers in the public transport sector, and in practice we are closer to the latter extreme than to the former. This should not make economists abandon the case for first-best optimal pricing, but it should stimulate the further development of the theory into the domain of the second-best. Such a development is well under way. It is not the purpose of this study to cover this line of development of normative price theory. For one particular reason, however, a short third part dealing with second-best pricing theory has been added to the main text: just as the general theory of marginal cost pricing has been applied to the transport sector, there is a tendency that the emerging general theory of second-best pricing is being applied to various transport services without the necessary understanding of the special conditions of the transport sector.

APPENDIX: INDIVISIBILITIES AND ECONOMIES OF SCALE[1]

What exactly does it mean that a production process is characterized by indivisibilities? Factor indivisibility should simply mean, in my view, that the least units of a factor are relatively "big". "Bigness" should not, of course, be interpreted literally as bulkiness, but in terms of *productivity*. When the average (or marginal) product of the least unit of a factor is fairly large in relation to the total output contemplated, then we can speak of factor indivisibility. Under this condition the Long-Run Marginal Cost curve has the saw-toothed shape illustrated in Figure A:1 in Chapter 9 (page 324). The discontinuities are explained there by the technological fact that there is rather a high lower limit to the capacity (in terms of tons handled per day) of a crane that is sufficiently powerful to lift a standard container, and the fact that a fairly large amount of minimum transit storage space is required to house a shipload. The more marked the factor indivisibility, the greater will be the discrepancy between the top and bottom ranges of the constituent SRMC-segments of the curve. Two rather obvious points can be made in this connection: (i) factor indivisibility is by no means a sufficient condition for economies of scale, except in the range where only one factor unit is required, (ii) the optimal pricing of services, where the production factors are characterized by indivisibility, does not necessarily result in a financial deficit.

[1] Some relevant "classical" references are: Chamberlin, E.H., Proportionality, Divisibility, and Economies of Scale. Quarterly Journal of Economics, February 1948. Kaldor, N., The Equilibrium of the Firm. Economic Journal, March 1934. Kaldor, N., The Irrelevance of Equilibrium Economics. Economic Journal, December 1972. Tjalling C. Koopmans, Three Essays on the State of Economic Science. McGraw Hill, New York, 1957.

The second point is really a consequence of the first. Only if there are economies of scale, is the pricing-relevant cost more often than not below the average producer cost.

A much more specific meaning of factor (in)divisibility, which used to be assumed in the earlier literature of production theory, is bound up with a fundamental problem in the definition of factor units. The basic problem boils down to the idea of a "proportionality axiom". This means that if the factor proportions stay the same, output *must* also be proportional to the scale of operations. That is to say, economies of scale are assumed to be non-existent. If increasing returns to scale seem to apply, it must mean - so the argument runs - that factor indivisibility exists, which prevents the factor proportions from remaining constant as output changes. On a closer scrutiny this argument reveals itself as nothing more than a particular definition of the units of factors of production: the factor units are defined such that the marginal product is the same for all units, provided that the complementary factor proportions remain constant. This may seem quite a natural definition for certain factors of production, but is a notoriously difficult requirement to make in the case of capital inputs. Let us take a simple example.

Suppose that a steel container is needed in a particular production process. Suppose also that the required volume of the container is proportional to the output. A natural measure of factor quantity in this case is the weight of the steel. However, this violates the condition that the productivity of each kg should remain constant. It will be found that the marginal product of each additional kg of steel increases steadily with increases in output (and container volume). An alternative factor unit might be container volume; each additional m^3 of container volume will certainly increase the holding capacity of the container equally. The problem now is, of course, that in this case the factor price falls steadily with increases in the factor quantity.

Admittedly some sort of factor indivisibility is involved, causing these definitional problems. In the form of containers, steel is imperfectly divisible. If one ton of steel is used to make 1 container, 2 containers, or ... n containers all with the same total holding capacity, it will soon be found that the walls of the containers have to be made so thin that they will break too easily. It is an open question of semantics, whether this should be regarded as a reflection of "factor indivisibility", or simply as an example of the geometric principle that the surface area of a vessel equals the volume raised to the power of 2/3. In my view it is preferable to reserve the term "factor indivisibility" for the first clear-cut case of a lower limit to capacity. The fact that the costs of capital equipments normally increase less than in proportion to the capacity, is quite a different matter. Even if it can be shown to have something to do with the imperfect divisibility of something or other, it is confusing to use the term "factor indivisibility" to describe this phenomenom as well. At least the difference should be marked by speaking of "factor *productivity* indivisibility" in the latter case.

PART III

SECOND-BEST PRICING UNDER A BUDGET CONSTRAINT

12 THE VALUE-OF-SERVICE PRINCIPLE AND OPTIMAL COMMODITY TAXATION

13 A MODERN VERSION OF THE COST-OF-SERVICE PRINCIPLE

PART III

SECOND-BEST PRICING UNDER A BUDGET CONSTRAINT

The golden pricing rule of welfare economics that price should equal the marginal cost of every good or service cannot or should not be applied when:

a) A budget constraint is imposed on a decreasing cost activity, and

b) prices of substitutes that are outside one's control deviate from the marginal costs.

The latter problem has been a sore point in applied welfare economics for more than two decades, and no easy way out is yet in sight.[1] Depending on what we may think about the position of prices and marginal costs in the economy as a whole, it can even be argued that average cost pricing might be preferable to marginal cost pricing in the transport industry. In that case the former problem would simply disappear.

The stand taken here is that in several cases the discrepancy between the average costs of transport service producers and the pricing-relevant costs is exceptionally large. Moreover, as will be argued here, a substantial part of the total transport industry can be regarded as a

[1] Although this problem had been recognized and discussed previously, it was first high-lighted in the general formulation achieved in Lipsey, R. & Lancaster, K., The General Theory of the Second-Best. Review of Economic Studies, 1957.

largely self-contained sector, which can be optimized as it were in isolation. Under these circumstances we are faced only by one second-best problem - problem (a) above.

In this last part we will consider the modification of the previous results, which are necessary in the presence of a budget constraint. It is clear from the preceding analysis that the modifications will be very substantial indeed: the relevance of the present problem does not require any special pleading.

The discussion is divided into two chapters called, "The value-of-service principle and optimal commodity taxation", and "A modern version of the cost-of-service principle". An old idea in public utility and transport pricing discussions is that a budget constraint is met most expediently by exploiting the often large differences in the maximum rates that different consumer categories can bear. This idea is best known as the value-of-service principle. As the name suggests, the basic philosophy is that the *values* placed on transport services by different users rather than some more or less ambiguous costs should determine the price structure.

The main competing family of pricing principles is, of course, the cost-of-service principle. The distinguishing characteristic of this principle, in comparison with marginal cost pricing, is that according to the former principle the price calculation starts from the total costs which are to be covered, after which a "cost allocation" procedure follows. The basic philosophy of the cost allocation is that those who "cause" - this is the key expression - the various costs should also bear them. This idea is perfectly sound at this general level; the problems arise when causal relationships between costs and traffic categories are to be established.

The present position is that the pricing-relevant costs, as previously defined, should always be the basis of the price calculation; the question is, how the necessary additions to these costs to meet a budget constraint are to be determined. In the two following chapters an answer is sought to this question. In Chapter 12 our main concern is to repudiate the currently prevailing view that the price discimination inherent in the value-of-service principle constitutes the "optimal departure from marginal cost pricing". In Chapter 13 we draw attention to an important characteristic of the demand for a number of transport services, which makes it possible to recommend that a cost-based price structure should be maintained also under a budget constraint.

12 THE VALUE-OF-SERVICE PRINCIPLE AND OPTIMAL COMMODITY TAXATION

The nature of the pricing problem in a decreasing-cost multi-product enterprise subject to a budget constraint is seemingly the same as the old problem of "optimal commodity taxation", which is to raise a given sum of money by imposing commodity taxes that minimizes the allocative distortions. It is therefore common in the literature to call the difference between the price and marginal cost of a product a "tax", irrespective of whether it is due to a fiscal impost or not.

A solution to the problem of commodity taxation was suggested as early as 1927 by Ramsey, who proposed that the optimal tax structure is that which reduces all outputs proportionally from the first-best output levels, i.e. from the levels of output obtained by setting prices equal to the marginal costs.[1] The same rule holds true for a multi-product decreasing-cost enterprise subject to a budget constraint.

12.1 IS PRICE DISCRIMINATION JUSTIFIABLE BY THE THEORY OF OPTIMAL COMMODITY TAXATION?

Ramsey's Rule appears frequently in theoretical discussions of public enterprise pricing, but is rarely applied in actual tariff construction. One reason is that, empirically, it

[1] Ramsey, F., A Contribution to the Theory of Taxation. Economic Journal, March 1927.

is rather demanding. Who knows what the unconstrained (first-best) optimal set of outputs would be in a typical public enterprise currently operating under a budget constraint?

A more popular version of Ramsey's Rule is therefore formulated in terms of deviations of prices and marginal costs. First, it is clear that provided the demand elasticities of the different products are different, prices should generally exceed the marginal costs *dis*proportionally. In the special case of independent demands for individual items in the product line of an enterprise, the corresponding pricing rule is the well-known propostion that the relative excess of price over marginal cost should be proportional to (the absolute value of) the inverse of the own-price-elasticity of demand.[1] Let us call this the "Inverse Elasticity Rule" (IER). This seems to be a less demanding rule, empirically speaking. Price-makers are more likely to have approximate knowledge about the elasticity of demand in the current output range, and it is fairly easy to check whether or not the existing price structure is consistent with IER.

It can easily be shown that profit maximization by a multi-product firm in the special case of independent demands also yields relative price and marginal cost differences, which are inversely proportional to the demand elasticities. The proportionality constant is unity under a profit maximization regime, while it is less than unity if zero profit is stipulated.

A natural question is, however: is not IER inapplicable in practice in view of the fact that the typical multi-product firm produces more or less closely related products?

[1] See e.g. Baumol, W.J. & Bradford, D.F., Optimal Departure from Marginal Cost Pricing. American Economic Review. June 1970.

This rule may be applicable when it comes to taxation of the outputs of different industries, as - by definition of an "industry" - inter-industry cross-elasticities are fairly small in most cases. It is quite a different matter when we consider the range of products supplied by a representative multi-product enterprise. And when cross-elasticities within the product line are significant, IER is no longer of any use for pricing-policy.

The fact is, however, that independence of individual demands is *created* by price discrimination, so that a situation is obtained where IER is applicable. And IER is also widely used in practice.

Of course, IER is nothing but a formalized version of the time-honoured principle of "value-of-service" charging. The history of the value-of-service principle in the transport industry is nearly as old as the industry itself. In the heyday of the railways in the second half of the 19th century this pricing principle already played an important role, and today the value-of-service principle is probably the single most important pricing principle in the transport industry.

For a long time price-makers in the transport industry were mostly on the defensive vis-à-vis the economics profession, which has tended to favour the "cost-of-service principle". However, while early advocates of value-of-service charging had to disprove claims of monopolistic abuses, "discrimination", and lack of cost-awareness, the principle has recently received the blessing of economists exploring "the second-best".

For example, Baumol and Bradford state that[1]

> ... ordinary price discrimination might well set relative prices at least roughly in the manner required for maximal social welfare in the presence of a profit constraint. (page 267)

[1] See next page.

This statement is strongly objected to here. I suspect, however, that my objections are mainly due to different meanings of the term "price discrimination". I think that Baumol and Bradford have in mind what could be called "innocent price discrimination", which is something quite different from the real price discrimination actually practiced in the transport sector. And the reason why this issue is not merely a terminological disagreement is that, more often than not, the price discrimination actually practiced does *not* deserve the blessing of welfare economists. This we seek to demonstrate in the following discussion. However, we must start the discussion with some terminological remarks.

12.2 INNOCENT AND REAL PRICE DISCRIMINATION

A definition of price discrimination often found in older literature is simply that it is to sell identical products at different prices. The problem with this definition is the practically unlimited possibilities of artificial product differentiation. It has proved next to impossible to establish that price discrimination so defined exists in reality as soon as the enterprise in question does not want to admit to price discrimination. Economists have therefore abandoned the strict requirement that exactly identical products have to be involved, and mean that it is sufficient that prices deviate markedly disproportionally from the corresponding marginal costs.

[1] (footnote from foregoing page)
Baumol, W.J. & Bradford, D.F., op. cit. Their idea is certainly not new, and they acknowledge a string of predecessors by introducing the paper in this way: "The need for this paper is a paradox in itself, and indeed it might be subtitled: The Purloined Proportion, or the Mystery of the Mislaid Maxim". Extensive references to earlier literature is also given.

If an enterprise supplies n products with different marginal costs, MC_1, MC_2, ... MC_n, it is commonly said that the enterprise applies price *differentiation*, if the prices P_1, P_2, ... P_n are set equal to the respective marginal costs, or if $P_1/MC_1 = P_2/MC_2 = \dots P_n/MC_n$, or, alternatively, if $P_1 - MC_1 = P_2 - MC_2 = \dots P_n - MC_n$. On the other hand, if the prices are set more or less disproportionally to the marginal costs with a view to exploiting differences in the demand elasticities of the n different products, the enterprise is said to apply price *discrimination*. An often mentioned special case of price discrimination is the charging of a uniform price, despite marginal cost differences between the products supplied. It can be observed that by this wide definition of price discrimination, every multiproduct profit-maximizing monopolist or oligopolist is practicing price discrimination, provided only that all elasticities of the products in the line do not happen to be the same.

Normally, however, the actual "discrimination" involved is of a very innocent nature. Neither the general public, nor the lawyers etc. watching over the price-making in trade and industry, think that the pricing behaviour called "price discrimination" by economists is very discriminatory. I think most people feel that for pricing-policy to be really discriminating, it is necessary that some customers are being tangibly discriminated against, i.e. barred from buying certain products on the same conditions as others. For example, if the same wine is sold in bottles labelled in two different ways - on one label a chateau is depicted, and on the other the plain name of the district of the wine is given - and sold for widely different prices, this may be considered bad commercial morality, or even fraudulent according to the law in certain countries, but it is not price discrimination in the real sense of the word. As long as everybody is free to buy either of the two "different" bottles for the given prices, no one is actually discriminated. Not until a certain category of wine consumers is barred from buying one or the other type of bottle, for example by a stipulation that those of foreign origin cannot be offered the cheap bottle, a clearcut case of price discrimination arises.

12.2.1 Market segmentation is what price discrimination is all about

I propose herewith that the term price discrimination should signify actions on the part of a producer of goods or services, which are truly discriminatory in the pejorative sense of the word, that is to say actions that exclude some customers or customer categories from buying certain products altogether, or for prices that apply to other customers.

It is of secondary importance for the real issue, whether product 1 and product 2, which are charged the different prices, P_1 and P_2, are in fact identical or not, or have marginal costs which are disproportional to the prices. The vital point is instead whether or not all potential customers are free to buy both products on equal terms. Seen from the point of view of the producer/seller seeking to confiscate additional consumers' surplus, the art of price discrimination is not a matter of determining the prices at a given set of products, but to find methods for *market segmentation*. This point sometimes gets lost in economic analysis of price discrimination. For example, in the most common textbook case of price discrimination, i.e. the monopolist charging different prices in geographically separate markets, the reader's attention is drawn to the problem of finding the profit-maximizing pair of prices, *given* that the two markets are truly separate. To rule out the possibility of buying the product in question in the cheaper market and reselling it in the more expensive one, the textbook case often makes the additional point that re-export to (the more expensive) home market is unprofitable due to tariff protection.

In other words, the conditions for successful geographical price discrimination are exogenously given beforehand, which, however, means assuming away the most important and interesting problem of price discrimination: how is the market segmentation to be secured in the first place?

12.3 METHODS OF MARKET SEGMENTATION IN THE TRANSPORT SECTOR

When it comes to transport services, much greater possibilities for price discrimination are available than concerning storable goods. The first thing to bear in mind is that the relevant "products" are services rather than commodities. Both the public utility and transport industries, where price discrimination is more far-reaching than anywhere else, are service industries in the sense that they produce immaterial, generally non-storable "goods". Non-storability is an important prerequisite for successful price discrimination. Non-storable services cannot be resold at all, and the geographical separation of markets is not a necessary condition.

The main method of "charging what the traffic can bear" is simple enough. It is to charge different customers different prices for the same service with a view to confiscating as much of the consumers' surplus as possible. The difficulties arise when it comes to disguising this crude objective, and yet carrying it out.

Very elaborate techniques for the confiscation of the consumers' surplus have evolved in the freight transport sector. The most far-reaching variant is practiced by some railway companies, such as the Swedish Railways, whereby *secret* rate agreements are negotiated with each individual (fairly large) shipper. In scheduled freight transport by sea, air (and sometimes road), where more than one carrier serves each of the more important routes, price cartels (liner shipping conferences, the all-embracing IATA, and trucking rate bureaus) are formed for the explicit purpose of creating "rate stability", and for the implicit purpose of extracting as much consumers' surplus as possible. In order to uphold internal discipline in the cartel, secret agreements between individual cartel members and shippers cannot be practiced. Instead, the common technique (inherited from 19th century railway freight rate-making) is to make the *commodities* the object of the freight rates, rather than the more cost-relevant characteristics of different *packages*. In all its simplicity this is an ingenious way of applying far-reaching price discrimination under the constraint that all or most freight rates have to be published (in the tariff). It is a more sophisticated form of price discrimination than the cruder related method of *ad valorem* charging. The point is that the widely differentiated freight rates in a commodity-based tariff can be claimed to be justified by differences in the cargo handling costs. However, commodities, which are different only in name, and which appear in one and the same type of package, can be charged very different freight rates. In the present container age of sea and air cargo transport, the practice of price discrimination by commodities has led to the slightly ridiculous situation that carriers have to ask shippers politely what is in the containers, in order to carry out the price discrimination. (A significant advantage of container transports is that en-route handling of individual articles is not required during the complete door-to-door transport; it would be self-destructive to require the stripping and re-stuffing of the containers only for the sake of charging what the traffic can bear.)

As far as passenger transport is concerned, price differentiation by customer category for the same service is also fairly common. The most imaginative examples can be found in the airline industry. For example, the time period between the outbound and homebound flights can be used as a criterion for market separation. If you stay away more than x days but less than y days, a much cheaper fare can often be obtained. The lower limit (x) is set sufficiently high to exclude most business trips and the like, and the higher limit (y) excludes those for whom the cost of the flight is of little consequence in comparison to the cost of the trip as a whole. The supposedly quite elastic tourist-travel market has thus been separated from the much less elastic business-travel market and the long-stay travel market used by a range of travellers from visiting professors to emigrants.

Other forms of price discrimination between passengers are ruled out as unlawful, or at least as being against a common sense of justice. To differentiate prices according to race, sex or arbitrary characteristics, such as colour of the hair, would be regarded in most countries as "discriminatory" in the pejorative sense of the word in common usage. On the other hand, age and military status, for example, are often accepted as grounds for charging different prices to different people for the same service. (Rebates for children, pensioners and servicemen are not characterized as price discrimination by anyone but economists.)

To summarize: non-storability of transport services makes far-reaching price discrimination possible, because it is impossible for a favoured customer to buy more of the service than he needs for himself and to resell it to less favoured customers. The problem for the discriminating price-makers is to define customer categories such that (a) the general public and possible regulatory authorities do not take offence, and (b) the definitions of customer categories are "waterproof". A middle-aged man cannot become a pensioner in order to travel more cheaply, and copper cannot be transformed into iron in order to be carried at a lower freight rate.[1] On the other hand, if cargo "in bags",

[1] There are exceptions to this rule as far as freight is concerned. The *degree of processing* of goods in international trade is likely to be influenced by the existing discriminatory structure of liner shipping freight rates. This is discussed in Jansson, J.O. & Shneerson, D., The Effective Protection Implicit in Liner Freight Rates. Review of Economics and Statistics, November 1978.

for example, were charged a considerably higher rate than "boxed" cargo, just because the former cargo category can support a much higher rate for the present, all cargo would come in boxes sooner or later, and the price discrimination would be completely ineffective.

12.4 THE ETHICS OF PRICE DISCRIMINATION

The spirit of real price discrimination is to sell the same product for different prices *to different customers*. The question now is whether innocent and real price discrimination are substantially different in their nature as second-best pricing policies. To make the contrast between the innocent price discrimination practiced by every profit-maximizing multi-product enterprise in the storable goods sector and the real price discrimination in the transport sector as sharp as possible, it is convenient to let the latter case be represented by a single-product enterprise, which, however, charges n different prices, P_1, P_2 ... P_n to n different customer categories for one and the same product.

When it comes to finding a second-best structure of prices under a budget constraint, the problem appears to be identical for these two enterprises. Provided only that the maximand is unweighted total social surplus, it makes no difference whether the outputs, Q_1, Q_2 ... Q_n, represent the total quantities demanded of n different products aggregated over all individual consumers, or the individual demand of n different consumers for one and the same product; on the assumption that cross-elasticities are negligible in the multi-product case, IER comes out as the right pricing policy in both cases. From an ethical point of view, however, it can make a world of difference.

12.4.1 A cautionary tale

Take first the problem of the multi-product enterprise. Suppose that it makes two types of products - coffins and walking-sticks - and has a monopoly position in both lines of its business. For the sake of the example, let us assume that the demand for coffins is almost completely inelastic - it is mainly determined by the death rate - while the demand for walking-sticks is quite elastic. Besides the old, the lame and the crippled, a good number of ordinary walkers think that a walking-stick can be worth having, provided that it is not too expensive.

The application of IER in the multi-product firm results in a substantial addition over the marginal cost in the price of coffins, while the price of walking-sticks is close to the corresponding marginal cost. The "tax" on coffins has nearly the effect of a head tax, which causes minimal allocative distortions. The rich and the poor are hit equally by such a tax, which, of course, can be regarded as undesirable from a welfare distributional point of view, but which is ignored in the model underlying IER and Ramsey's Rule.

The question is, whether the distributional effects, or rather the ethical aspect, can be ignored in the second case of real price discrimination. Suppose that the single-product enterprise in our example is a monopoly producing just walking-sticks. It practices real price discrimination by separating the customers in four classes, which, in order of rising price-elasticity, are (1) lame people, (2) crippled people, (3) old people, and (4) ordinary people. The application of IER leads, of course, to a price structure implying that the lame are charged the highest price, etc.

This somewhat morbid cautionary tale is admittedly rather extreme. But the moral is worth serious consideration all the same.

12.4.2 Discrimination against no-choice transport users

Generally speaking, there are three main factors which can make demand elasticities vary widely among the consumers of a certain product.

1. The urgency of the need
2. The ability to pay
3. The availability of substitutes

The previous example took up point (1): for some people a certain product can be imperative, while others consider it optional. Concerning transport services, I venture the generalization that, in our reasonably egalitarian society, point (3) above is the most important cause of significant elasticity differences among transport service users. Part of the explanation for this is that the natural monopoly position, which may once have been enjoyed by public or regulated private transport enterprises, has been eroded by the growth of private transport, or has in other cases been replaced by a "natural duopoly" (or oligopoly) situation. (Railway and airline and/or bus transport competition on some long-distance routes can be mentioned as an example.)

Under these circumstances users can be divided into those who have a choice and those who have no choice in their source of supply of transport services. The application of IER means that users who have no choice will be exploited to the benefit of users who can choose. Is this fair?

This has been a hotly debated issue in passenger transport for a very long time. Price discrimination against no-choice customers is more common in freight transport where the question of fairness is less marked.

A common pattern applying both to sea, land, and air freight transport prices, is that high prices (above the marginal costs) are charged on commodities whose transport is not exposed to competition, while low prices close to the marginal costs, or below the marginal costs, as we have seen in Chapter 8, are charged on commodities which are exposed to competition from individual modes of transport, and transport for hire.

A typical example can be two sea-ports fairly close to one another, which are rivals in some parts of the traffic but which are each virtually unchallenged in other parts. A grain importer, for instance, may have installed a depot in one of the ports but not in the other. In these circumstances grain may be a "milk-cow" for the former port, being able to bear quite a high port charge. The grain importer has no good means of putting counter-pressure on the discriminating port, because the cost which the importer would incur by installing an alternative depot in the other port is likely to be prohibitively high for the purpose. On the other hand, traffic which can use both ports without very substantial disadvantages are likely to pay comparatively low port charges.

The general public accepts this kind of price discrimination more easily, because it is seemingly in line with ability-to-pay considerations. ("The wealthy grain importer can afford to pay.") This is, of course, largely an illusion. Ultimately it is the consumers of the final goods in which grain is an input who are being discriminated against.

Economists should, however, make it clear that it is unwarranted to attach much relevance to the fact that the loss in aggregated social surplus is minimized by applying the policy of "charging what the traffic can bear". Real price discrimination is not primarily an issue of allocative efficiency.

12.4.3 Price discrimination cannot be judged by the efficiency criterion

The charts of Figure 12:1 illustrate what the issue is all about: suppose that a sum of money equal to A has to be raised by a commodity tax. If the tax is imposed on a product with a relatively elastic demand - D_1 in Figure 12:1a - the loss in social surplus corresponding to the shaded triangle in Figure 12:1a, will be substantially larger than if

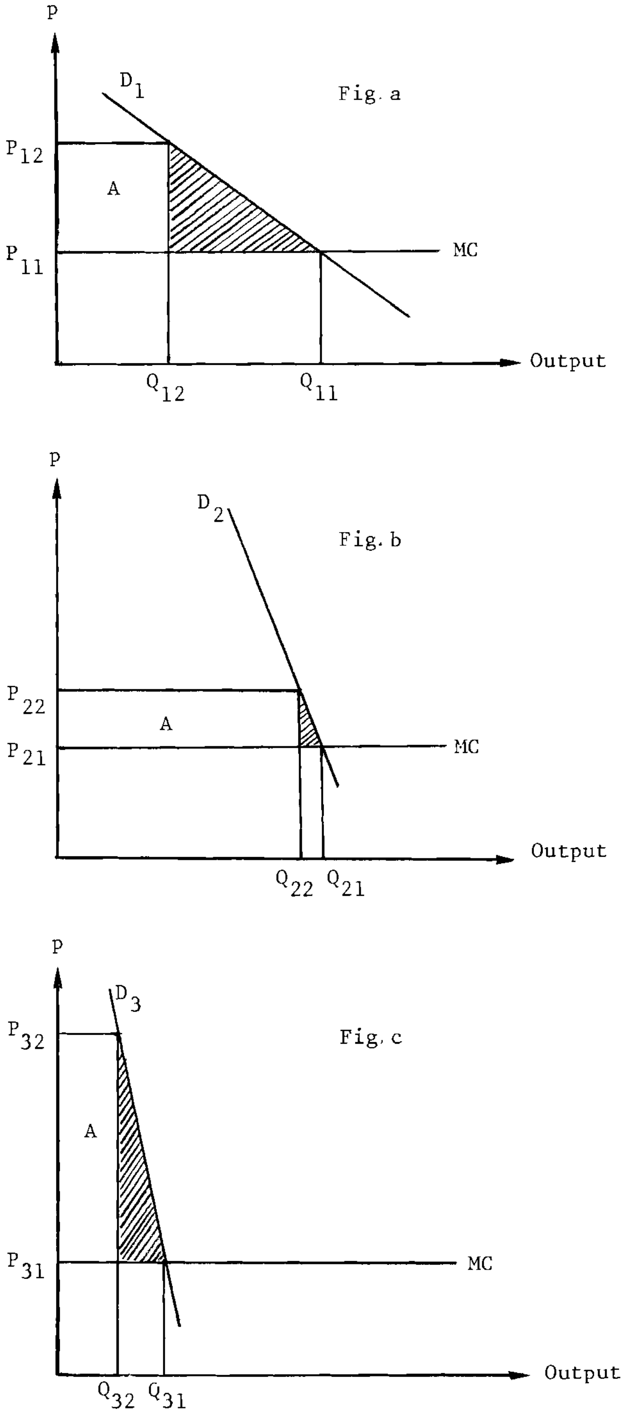


Figure 12:1. The effects of raising a given sum of money by taxing each of three different products, or three different consumer categories

the tax is imposed on a product with inelastic demand as illustrated in Figure 12:1b. This difference is particularly significant if roughly the same consumers are involved more or less equally in each market. Then it can occur that practically everybody gains from shifting the whole tax burden from product 1 to product 2. On the other hand, if D_1 and D_2 represent the demands of two different consumer categories for the same product, it is difficult to argue that taxing the consumers of category 2 is more efficient than taxing the consumers of category 1.[1] It may seem generally more reasonably to impose the tax on category 2 in the illustrative example, as this would mean that the total tax burden is shared by more people or, at least, the tax per unit of output is lower than if the tax is imposed only on category 1. By this reasoning it may seem still more reasonable to spread the tax burden equally over all consumers, i.e. to apply average cost pricing.

The idea that it is somehow "efficient" to tax the consumer category with the most inelastic demand for a certain product looks very suspect when we take a case of a minority consumer group being exploited in such a way as is illustrated in Figure 12:1c. If the tax sum A is raised by charging the small consumer group 3 the markedly discriminatory price P_{32} (while keeping the prices for all other consumers at the marginal cost level), most people would consider this very unfair even if consumer category 3 happens to consist of relatively wealthy people. Such price discrimination

[1] It would be illogical to assume that means are available by which the consumers of category 1 are made to compensate the consumers of category 2 without causing any further distortions, so that everybody is happy compared to a situation where the prices P_{12} and P_{21} are charged. If such means were available there would be no need for a commodity tax in the first place.

smells nevertheless of usury and extortion. And if the consumers of category 3 were ordinary people with the only distinguishing characteristic that they have very urgent needs for travel, but have no choice of mode of transport, an outcry would be raised against such a pricing policy.

The general conclusion is that as opposed to innocent price discrimination, real price discrimination is a question of equity and income distribution with only secondary efficiency implications. It is misleading to treat it as analogous to the problem of optimal commodity taxation.

It is interesting to note, finally, that even in a case where all consumers are free to choose on equal terms among all products supplied, the relevance of the whole idea of IER for social welfare maximization is being reconsidered in recent literature - more exactly in the parallel discussion of optimal commodity taxation (in which real price discrimination is out of the question). The basic assumption of models in the Ramsay tradition of distributional neutrality (often by the single-consumer device) has rightly been called in question:

> Indeed, distributional questions are really of the essence of the commodity taxation problem. After all, if one were merely concerned to achieve a Pareto optimum, one could dispense with commodity taxes altogether: a uniform head tax ensures a distortionless first best optimum. Thus the real relevance of the frequently cited propositions of the efficiency-oriented single-person model remains obscure .... Perhaps the whole approach to the study of optimal commodity taxation as an exercise in social welfare maximization of the usual type is due for reconsideration.
> (Wildasin, 1977)[1]

12.5 PRINCIPLES OF TAXATION OF USERS OF DECREASING-COST TRANSPORT SERVICES

With regard to the original problem of second-best pricing under a budget constraint, the previous conclusion can also

[1] Wildasin, D.E., Distributional Neutrality and Optimal Commodity Taxation. The American Economic Review, December 1977.

be expressed in this way: it is an idle wish that this second-best problem can be solved by taking only efficiency into account. By nature it is a question of taxation, and was pointed out in the preceding quotation, the distributional aspect for questions of taxation is more important than anything else. Let us therefore wind up this chapter by a brief discussion of principles of taxation worthy of consideration in the present case.

The essence of the problem can be put most clearly in the simplest case of a single-product decreasing-cost enterprise: if a budget constraint is imposed on such an enterprise, it is obvious that everyone cannot get away with paying a price equal to the marginal cost. Some or all have to pay a higher price. Who should be burdened with the required tax in the first place? Certainly not necessarily those who have the most inelastic demands. This could be an utterly unacceptable principle of taxation, as we have argued before.

Under certain conditions a head tax levied on all users of a particular transport facility in combination with MC-pricing is both equitable and almost without distortions. The problem with a head tax restricted to the current consumers is, however, that it can be an effective bar to new customers. In the present connection this system is better known as a "two-part tariff".

The two-part tariff is an old device for improving allocative efficiency in a decreasing-cost enterprise subject to a budget constraint. The general idea is that it should contain a variable charge (use-proportional) equal to the marginal cost, and a fixed "subscription fee" for recouping the remaining cost. The two-part tariff is well suited to services where the demand of each individual consumer is more or less recurrent. Only under this condition would it be profitable for most of the users to invest in a subscription. Unfortunately, this condition is seldom fully met in the transport sector. It is quite clear that a substantial subscription fee for the use of different modes of public transport would discourage an appreciable number of potential "less regular users" altogether. To avoid the worst of the ill-effects of compulsory subscription, it is normally considered necessary to sell the services also on a per-unit basis, in which case the price obviously has to be at a much higher level than the variable component in the price of the two-part tariff concerned. (Otherwise hardly anyone would pay the subscription fee.)

This may nevertheless result in the wholesale loss of a particular category of potential irregular users of public transport. A very important category of less regular users of public transport consists of all those who normally have a car at their disposal. It is a real danger that a relatively high ordinary per-unit fare would discourage the majority of this category from ever making use of public transport, as they have a very competitive alternative at hand. It is impossible to generalize in this instance, but it is well known that on both short and long distances, the public transport fare quite often exceeds the private marginal cost of road transport in a private car.

This is rather unfortunate in view of the fact that, especially in off-peak, the marginal costs of public transport services are only a fraction of current fares.

A solution to this problem could be to introduce an *intermodal* subscription fee, but this should be asymmetrical in that it should be compulsory for private car owners and voluntary for public transport users who do not own a car. In essence it would mean that by showing a paid-up licence tax for the year, a car owner would be entitled to go by public transport for a fare equal to the pricing-relevant (marginal) cost. Part of the car-licence tax revenue would be transferred to public transport undertakings.

A remaining problem concerns the irregular public transport users who do not own a car. Admittedly, they are not excluded from using public transport because of the "ordinary fare" alternative to the two-part tariff; but the relatively high level of the fare would rightly be perceived as unjust. This problem can never be wholly eliminated, but it can be mitigated by introducing an intermediary form of subscription, namely different cards entitling the bearer to a *limited* number of trips for a fare equal to the pricing-relevant cost. To ensure that the limit is not exceeded, it would be necessary to register each trip, perhaps by making a hole in the card. The higher the limit, the more expensive the card should be - but of course not so much that the price per trip is not falling.

Distributional neutrality is of course not necessarily desirable under all circumstances. It would in fact be quite an efficient method of real income redistribution to encourage additional consumption of decreasing-cost transport services by certain groups of people which society wants to help. This is certainly price discrimination - but in a positive spirit, as it were. To single out a minority of consumers who can bear substantially higher charges than the majority and squeeze out most of the consumers' surplus can seem very inequitable, while the opposite policy of subsidizing the travel of minorities, like servicemen, families with children, and pensioners, who can be expected to benefit greatly from cheaper fares, is considered to enhance social justice. These groups of people often have great need for transport because of their particular situation, but lack the means for satisfying their wants. If the ordinary level of fares is well above the pricing-relevant costs, selective fare reductions can be an inexpensive line of welfare policy with a high rate of social return in the form of more visits at home by soldiers stationed a long way from home, etc.

12.5.1 Price discrimination as the last resort

Only in one special case, in my opinion, can fully-fledged price discrimination according to IER be worth considering. When "apoly" threatens, that is to say in a case in which a budget constraint cannot be met unless a large proportion of the total consumers' surplus is confiscated, the only solution may be to apply discriminatory prices roughly in accordance with IER. All the transport users involved would probably prefer even the most exorbitant price discrimination to complete discontinuance of the transport service in question.

13 A MODERN VERSION OF THE COST-OF-SERVICE PRINCIPLE

In this chapter we take up the true multi-product pricing problem as opposed to the problem of market segmentation with a view to exploiting differences in demand elasticities among transport users. The conclusion of the previous chapter was that the virtue of price discrimination cannot be judged by the efficiency criterion. It may have to be resorted to when no other means of avoiding threatening apoly is available. Barring this extreme case, we mean that the inherently inequal treatment of the customers implied by price discrimination should not be allowed unless it is unequivocally desirable from an income distributional point of view. There is a further argument which can be raised against current discriminatory pricing practices, namely that the possibility of applying far-reaching price discrimination seems to induce an unawareness of cost.

13.1 TOO LITTLE COST CONSCIOUSNESS IN THE TRANSPORT SECTOR

In contrast to the elaborate classification schemes and consequent markedly discriminatory structures of freight rates and passenger fares which can be observed in land, sea and air transport alike, there is a conspicuous uniformity of rates and fares *in time and space*. Given the customer category, railway charges are typically calculated on the basis of a flat charge per ton-mile or passenger-mile, regardless of when and where in the system the transport takes place.[1]

There is great scope for a socially beneficial transport price differentiation geared to products rather than to customer category. A transport "product" is defined by (i) the object which is transported, (ii) the start and end points of the transport,[2] and (iii) the time at which the transport takes place.

Passengers can with some minor exceptions be regarded as homogeneous objects of transport. Passenger transport undertakings, however, are more or less markedly multi-product enterprises, although in varying degrees depending (1) on the number of routes on which a particular enterprise provides a service, and (2) on the number of stops made on each route. A national railway transport company serves hundreds of spatially separate markets, while some passenger shipping lines may provide a service between one pair of ports only. In both cases the time dimension contributes to product variation. Cyclical fluctuations in demand - seasonal, weekly, or diurnal - are a cause of product differences (manifest as marginal cost differences) which is specific to non-storable goods.

In freight transport an additional reason for product differences is the diversity of sizes, shapes, and other characteristics of the articles presented by the shippers. As has been mentioned, these differences are not generally given due importance in freight rate-making, which focuses

[1] In the years 1978 and 1979 an agreeable innovation in the pricing policy of the Swedish domestic airline company and the Swedish railways has been introduced, which has meant price differentiation over time, or peak-load pricing. It has been a very popular reform, since it has brought big reductions in off-peak prices.

[2] In the case of terminal services, the start and end points - to take the example of a seaport - are a particular berth in a port and the hold of a particular type of ship.

on "commodities" instead, not because the type of commodity necessarily makes a difference to the cost of cargo handling and carriage, but because it is the simplest and most acceptable method of price discrimination.

Why is price discrimination much more popular with price-makers than marginal cost-based price differentiation? This question goes to the heart of the matter. It should be borne in mind that the revenue-boosting possibilities of price discrimination geared to category of customers are tremendous. In liner shipping, for example, the total revenue which can be obtained by a system of commodity-based freight rate discrimination may well be 50-100 per cent greater than the total revenue generated by a flat per-ton freight rate.[1] By inclination, economists have focused on the allocative improvement which seems to be made, thanks to price discrimination, by expanding total output towards the point where average revenue and marginal cost coincide. The price-makers themselves, on the other hand, are naturally concerned with the private profit potentials of price discrimination. Action aimed at confiscating consumers' surplus will generally yield a much higher dollar ratio of additional profit to allocative gain, than action aimed at reducing costs. It appears to me that, in industries where the possibility of price discrimination exists, there is an inherent bias away from a socially desirable concern with costs towards an excessive elaboration of tariffs.

[1] This order of magnitude is suggested by findings in Jansson, J.O. & Shneerson, D., The Transformation of Cartel Profits into Social Losses in Liner Shipping. International Journal of Transport Economics. August 1978. The main purpose of this article is, however, to show that the regulation of freight rates by liner conferences results in cost increases, which deny the conference members any monopoly profits - an example, if you like, of an induced unawareness of costs.

Let us take as an example the diagnosis by the former Chief Economist of British Railways of the disease that afflicted the railways at the beginning of this century, and from which they have never recovered: *ad valorem* rates.

> Between the wars the main-line railways' blind adherence to Acworth's revenue maximization dictum, with an equally blind acceptance of Acworth's notion that costs are relevant only for constructing the profit and loss account, were root causes of the excess capacity, sub-optimal pricing, and other ills inherited by the British Transport Commission on nationalization.
> (Stewart Joy)[1]

13.2 HOW IS THE MULTI-PRODUCT PRICING PROBLEM TO BE APPROACHED?

Transport economists have been worried about distortions of the cost *structure* for a long time, and the value-of-service principle has regularly been brought into challenge. However, Baumol and Bradford's authoritative endorsement of price discrimination has for the moment silenced the advocacy of the cost-of-service principle. The present position is, on the other hand, that advocates of the cost-of-service principle should take courage. There is a case for cost-based prices in important sub-sectors of the transport sector. In the previous chapter we have argued that the real price discrimination applied in the transport sector does not deserve the blessing of welfare economists.

However, I think that what Baumol and Bradford had in mind was not real price discrimination - that is the value-of-service principle - but the innocent type of price discrimination, which I prefer to call simply price differentiation by product. If this interpretation is correct, their problem could be stated simply like this: how should a decreasing-cost multi-product enterprise, facing a budget constraint, differentiate the prices of its products in order

[1] Joy, S., Pricing and Investment in Railway Freight Services. Journal of Transport Economics and Policy, September 1971, p. 236. (Acworth was an authority in railway transport economics in the last century.)

to maximize social welfare? *This* problem is genuinly analogous to Ramsey's problem of optimal commodity taxation, and Ramsey's Rule is much more likely to be an acceptable second-best solution to the problem.[1] The important additional point is, however, that IER is inapplicable in general, because the cross-elasticities are too significant when we have moved from a pseudo-multi-product situation created by market segmentation to a real multi-product situation in which no customer discrimination exists.

The rejection of IER as a rule-of-thumb for second-best transport pricing also in the real multi-product case is worrying, because the need for a simple and yet approximately correct rule-of-thumb is very great indeed. For a century the idea of marginal cost pricing of transport services has been to the fore. But how it should be reconciled with individual budget constraints on decreasing-cost enterprises in the transportation industry, is still a confused issue.

In the transport sector the situation is further complicated by the fact that complementary road, terminal, and vehicle services are often provided by different enterprises. Furthermore, particularly in the freight transport sector, door-to-door transport is notably a "chain of links", which means yet more enterprises complementing one another's services.

Under these circumstances it seems very difficult to find a reasonably simple rule-of-thumb for second-best pricing, *unless* it can be assumed that the aggregate demand for a well-defined group of related products is very inelastic. That this possibility is a reality in the transport sector may never enter one's mind if one views the situation from the standpoint of an individual transport enterprise. However, by moving up one level, and assuming the point of view

[1] Note, however, Waldasin's word of caution quoted in the previous chapter on page 401.

of a regulatory authority or transport policy-maker, this possibility springs easily to one's mind. We argue here that, as an approximation, the total demand in most markets for *freight* transport services can be assumed to be completely inelastic. Commuter traffic is a second important sector of very inelastic demand.

13.3 MARSHALL'S LAW AND THE ELASTICITY OF DEMAND FOR TRANSPORT

The demand for *freight* transport is what is known as a derived demand; freight transport, like other intermediate goods and services, has no intrinsic value, but is required only as an input into the production of various final goods.

According to "Marshall's Law", the elasticity of demand for an input for which there is no substitute is equal to [1]

> the share of the input in the total cost of the final product
> multiplied by
> the elasticity of demand for the final product.

It is the former factor which is the strategic one in the present context. Transport economists and engineers frequently work on the assumption that the freight-rate elasticity of total demand is negligible,[2] justifying this on the grounds that the transport cost element is generally rather low.

[1] For a comprehensive discussion, see Jansson, J.O. & Shneerson, D. Charging What the Traffic Can Bear - Alternative Rule to the Value of Service Principle. International Journal of Transport Economics, Dec. 1979.

[2] We may quote Friedland in this respect: "Transportation is probably characterized by an overall low elasticity and high cross-elasticity between modes. Friedlander, A.F., The Dilemma of Freight Transport Regulation. The Brookings Institution, 1969.

However, the transport-cost portion of the total cost of production for different goods varies very much. As far as manufactured goods are concerned, transport costs generally constitute only 2-4 per cent of the total value of output, while the transport-cost element in relatively "cheap" primary products like iron ore, coal, timber, etc. is much higher - about 20-40 per cent of delivery prices.[1]

On the further assumption that the price-elasticity of the demand for manufactured goods is in the "normal" range, it is rightly taken for granted that the aggregate demand for the transport of different high-value goods is very inelastic.

With regard to low-value goods, opinions are more divided. It should be borne in mind, however, that there is a markedly systematic relationship between the degree of processing and the price-elasticities of demand for different products. The *aggregate* demand for a primary product (as opposed to the demand for a primary product from a particular source of supply) is generally much less elastic than the demand for final products. There is ample empirical evidence to support this claim. As an example, the results of two studies of import-price elasticities are reproduced in Table 13:1.

Given the generally low elasticities of primary product demands, it follows that the freight-rate elasticity of demand for the transport of primary products can be as pronouncedly inelastic as demand for the transport of high-value manufactured goods, although the transport-cost element is relatively high for primary products.

When it comes to *passenger* transports, it can be noted that, as in the case of freight transport, a journey is not normally undertaken for the sake of the transport itself;

[1] See e.g. Mohring, H. & Williamson Jr. H.F., Scale and "Industrial Reorganization" Economies of Transport Improvements. Journal of Transport Economics and Policy, September 1969.

Table 13:1. Elasticities of demand for imports of goods of varying degrees of processing

Import-price elasticities of goods in different categories[a]		Import-price elasticities in different countries[b]		
			Industrial input goods	Manufactured goods
Crude Materials	-.18	United States	-1.18	-3.07
Crude Food	-.21	Canada	-.38	-1.62
Manufactured Food	-1.40	Common Market	-1.77	-2.61
Semi-Manufactured Goods	-1.83	United Kingdom	-.75	-2.71
Finished " "	-4.05	Continental EFTA	-.81	-2.01
		Japan	-.45	-2.53

[a] Source: Houthakker and Magee, "Income and Price Elasticities in World Trade". The Review of Economics and Statistics. May 1969.

[b] Computed from Balassa and Kreinin. The Review of Economics and Statistics. May 1967, p. 130.

it is made to reach a particular destination. Marshall's Law can be used at two levels to get an idea of the elasticity of the aggregate demand for different kinds of passenger trips. At one level different purposes can be distinguished with respect to the degree of necessity of the trip. Trips for purposes that are perceived as "absolutely necessary" give rise to very inelastic aggregate transport demands, especially if the transport cost is very low in relation to the cost incurred or the value gained by fulfilling the purpose of the trip. On this count the important travel category of commuters may be regarded for practical reasons as completely inelastic when all alternative modes of transport are lumped together[1] - and note that trips with alternative origins and destinations are aggregated, too. In other

[1] This assumption is made, for example, in a recent study advocating the application of IER to urban mass transit (unjustifiedly in my view): Train, K., Optimal Transit Prices under Increasing Returns to Scale and a Loss Constraint. Journal of Transport Economics and Policy, May 1977.

words, it is assumed that if the costs of all urban travel were halved or doubled, the total volume of commuter traffic would be affected hardly at all. Similarly, if the cost of car commuting only were halved or doubled, the modal split would be substantially changed, but the total traffic volume would remain largely the same.

On the other hand, long-distance travel for the purpose of visiting relatives and friends, for example, can be expected to be quite elastic, which means that a change in the costs of all modes of long-distance travel would have an appreciable effect on the total number of visits to distant relatives and friends.

At another level Marshall's Law can serve to underpin theoretically the general view that the price elasticity of demand for services supplied by transport infrastructure is invariably rather inelastic. This is because the prices of these services normally represent no more than a small fraction of the total generalized cost of the transport concerned. It is well established that the aggregate demand for road services is very inelastic with respect to the road user charge (petrol tax). The main reason is that the optimal road user charge constitutes only about one tenth of the total generalized cost of private road transport. If the generalized-cost elasticity of the aggregate demand for road transport were as high as (minus) unity, the user-charge elasticity of the demand would be only about -1/10.

To summarize: it thus seems that with respect to demand elasticity the total transport sector can be divided into two sectors, one where individual markets have notably inelastic aggregate demands and one where demands are to varying degrees elastic. The first of these sectors includes all freight transport markets and the markets for home-work travel. In the latter long-distance, non-business travel is the major category.

13.4 THE MEANING OF A "COST-BASED STRUCTURE" OF TRANSPORT PRICES

In the sector characterized by inelastic aggregate demands, it is not necessary to aim just at a second-best price structure in the presence of a budget constraint. A first-best optimum is, in principle, obtainable.

It can be noted as a curiosity to begin with that the inelastic sector includes some notorious loss-making sub-sectors - urban public transport, railway freight transport, seaports. If the level of prices of the services supplied in these sub-sectors were *deliberately* kept below the average cost level, there would be nothing much to complain about. The pricing-relevant costs are, as we have seen before, well below the average producer costs. It is an anomaly, however, in view of the fact that full cost coverage, making financial self-support possible, is desired by most politicians. It is a different matter so far as the sector with more or less elastic demands are concerned. Even if the politician wanted all enterprises to break even also in this sector, a general break-even situation may simply be impossible to obtain, or it can be obtained only at a substantial loss of social surplus. We have discussed this problem in the previous chapter, and some possible second-best solutions have also been indicated.

If we now restrict the analysis to the inelastic sector, the basic idea is simply that if the prices of all related products in a group for which the aggregate demand is very inelastic were raised *in unison*, all current loss-makers would sooner or later make ends meet.

As mentioned, the addition to the pricing-relevant cost in the price of a product can be viewed as a "tax"; thus the problem is to determine a structure of taxes that makes the product choice (the modal split) the same as it would be without taxes. Under some conditions an allocatively neutral structure can be obtained by making all prices proportional to the pricing-relevant costs. Under other conditions the absolute difference between the price and the pricing-relevant cost should be the same for all products. The latter case is relevant when it can be assumed that each individual consumer consumes, per unit of time, a *given number* of product units from the product group concerned. The products within the group are substitutable in varying degrees, but the characteristic relationship is that a consumer will be no better off for being offered x+y units of product A in place of x units of product B, rather than x units only of product A, whatever the size of y. For example, commuters demand no more and no less than five trips per head and per week in each direction, no matter what the prices and qualities of alternative modes of transport are at different hours of the day. Given these prices and qualities, a particular commuter may prefer to go by train, say four days a week, and by car one day a week (when he or she is shopping on the way home). Other combinations of trips by car and train result in a somewhat lower net benefit per week for the commuter in question, and the point is that this also includes all combinations involving *more* than five trips per week in each direction.

In conclusion: in the sector characterized by inelastic group demand it is sufficient, in order to get every transport user to choose the same alternative he would choose if all prices were set equal to the pricing-relevant (marginal) cost, that the *differences* between the pricing-relevant costs of the alternative services in the group are the same after the prices have been raised above the pricing-relevant costs as they were before.

13.5 THE CASE FOR TAXING PRIVATE TRANSPORT SO THAT PUBLIC TRANSPORT BREAKS EVEN IN A FIRST-BEST POSITION

The purpose of making additions to the pricing-relevant costs, of course, is to make financial self-support possible. The break-even requirement may apply to the group as a whole, which can consist of a number of individual enterprises, or to each individual enterprise within the group. In the former case it will only be by chance that all enterprises in the group break even individually. Only if the degree of unexploited economies of scale happens to be the same for all the enterprises, will this be the result.

A common situation, however, is that the relevant group includes (i) an enterprise supplying scheduled transport services at markedly decreasing costs, (ii) a number of enterprises offering transport vehicles for hire - an activity which can be assumed to show constant returns to scale - and (iii) individual modes of transport, i.e. private cars and trucks, whose services are not sold on a market. If it is considered sufficient that the group as a whole breaks even, a pattern of cross-subsidization between the two private modes and the public mode of transport would be the result. Cross-subsidization may be considered an evil in itself, whether it be intersectoral or intrasectoral, and it may be required that even public transport companies break even. As a consequence, the taxes on (ii) and (iii) have to be raised to such a level that AC-pricing of (i) would result in the same modal split as would occur under MC-pricing.[1] The sector characterized by inelastic demand would become a source of general tax revenue. The enterprises in each relevant group enjoying the most economies of scale would break even, and all other enterprises and do-it-yourself transport activities would be net contributors to the public finances.

Largely unintentionally we are steadily approaching this position in Sweden (as well as in a number of other countries) so far as interurban transports are concerned. The "tax" element (in the sense of an addition to pricing-

[1] Compare Braentigan, R.R., Optimal Pricing with Intermodal Competition. American Economic Review, March 1979.

relevant cost) in the prices of private transport outside urban areas is getting larger and larger, as fuel taxes for different reasons unconnected with the issue of second-best transport pricing policy are being raised more or less at the same rate as the oil producers raise their prices. The soaring cost of fuel has already had a notable effect on the relative demand for private and public transports on longer distances.

In urban areas there is still a long way to go. In the peak hours the pricing-relevant cost of private urban transport is generally many times greater than the petrol tax, and the losses made in urban public transport are as large as ever.

REFERENCES

ABRAHAMSSON, B.J., A Model of Liner Price Setting. *Journal of Transport Economics and Policy*, September 1968.

ANDERSEN, A. & Co., A Method of Bus Route Costing. In Costing of Bus Operations. Proceedings of a symposium held at the Transport and Road Research Laboratory. Crowthorne, June 1975.

ANDERSEN, P., Public Utility Pricing in the Case of Oscillating Demand. *Swedish Journal of Economics*, 1974.

BALASSA, B. & KREININ, M.E., Trade Liberalization under the "Kennedy Round": The Static Effects. *Review of Economics and Statistics*, May 1967.

BAUMOL, W.J. & BRADFORD, D.F., Optimal Departure from Marginal Cost Pricing. *American Economic Review*, June 1970.

BAUMOL, W.J. & VINOD, H.D., An Inventory Theoretical Model of Freight Transport Demand. *Management Science*, March 1970.

BECKER, G., The Theory of the Allocation Time. *The Economic Journal*, Vol. 75, 1965.

BENNATHAN, E. & WALTERS, A.A., The Economics of Ocean Freight Rates. Praeger Special Studies in International Economics and Development, 1969.

Berth Throughput. Systematic methods for improving general cargo operations. Report by the Secretariat of UNCTAD. TD/B/C.4/109 United Nations, New York, 1973.

BFR:s arbetsgrupp för trafikforskning. Trafik och bebyggelse. Statens Råd för Byggnadsforskning, January 1977.

BOHM, P., et al., Transportpolitiken och samhällsekonomin. Liber Förlag. Stockholm 1974.

BORTS, G., Increasing Returns in the Railway Industry. *Journal of Political Economy*, August 1954.

BORTS, G., The Estimation of Rail Cost Functions. *Econometrica*, January 1960.

BRAENTIGAN, R.R., Optimal Pricing with Intermodal Competition. *American Economic Review*, March 1979.

BROWN, G. & JOHNSON, B., Public Utility Pricing Under Risk. *The American Economic Review*, 1969.

BRUZELIUS, N., The Value of Travel Time. Theory and Measurement. Nationalekonomiska Institutionen, Stockholm University. Skrift No. 3, 1978.

BURENSTAM LINDER, S., Den rastlösa välfärdsmänniskan. Stockholm, 1969.

CHAMBERLIN, E.H., Proportionality, Divisibility, and Economies of Scale. *Quarterly Journal of Economics*, February 1948.

CUNDILL, M.A. & WATTS, P.F., Bus Boarding and Alighting Times. Report LR521. Transport and Road Research Laboratory. Crowthorne, 1973.

DEMSETZ, H., Why Regulate Utilities? *Journal of Law and Economics*, April 1968.

DESSUS, G., The General Principles of Rate-Fixing in Public Utilities. International Economic Papers, No. 1, 1951. (Also in NELSON, J.R. (ed.), Marginal Cost Pricing in Practice. Prentice-Hall, 1964.)

Department of Transport, Transport Policy. HMSO, 1976.

DEVANNEY III, J.W. et al., Conference Rate-Making and the West Coast of South America. Commodity Transportation and Economic Development Laboratory. MIT, 1972.

DE WEILLE, J. & RAY, A., The Optimum Port Capacity. *Journal of Transport Economics and Policy.* September 1974.

ERNST & ERNST, Selected Commodity Unit Costs for Oceanborne Shipments via Common Carriers (Berth Liner). Undated report (but pertaining to 1964) to the US Department of Commerce.

Federal Maritime Commission, Investigation of Ocean Rate Structures in the Trade between United States North Atlantic Ports and Ports in the United Kingdom and Eire. Docket No. 65-45, 1968.

FERGUSON, A.R., et al., The Economic Value of the United States Merchant Marine. The Transportation Center at the Northwestern University, 1961.

FOSTER, D., The Transport Problem. Revised edition. Croon Helm, London, 1975.

FRIEDLAENDER, A.F., The Dilemma of Freight Transport Regulation. The Bookings Institution, 1969.

FRISCH, R., Theory of Production. D. Reidel Publ. Co. Dordrecht-Holland, 1966.

FUCHS, V., The Growing Importance of the Service Industries. Occasional Paper 96. National Bureau of Economic Research. New York, 1965.

FUCHS, V., The Service Economy. National Bureau of Economic Research. New York, 1968.

GEDDA, S. & KOCH, A., Principer för fraktsättning vid sjötransport av containers. Skipsfartsøkonomisk Institutt, Bergen, 1966.

GOSS, R.O., Some Financial Aspects of Shipping Conferences. *Journal of Transport Economics and Policy*, May 1971.

GULBRANDSEN, O., Trafikk- og investeringsanalyse for Bergen havn. TØI, 1973.

HARRIS, R.G., Economies of Traffic Density. *The Bell Journal of Economics*. Autumn 1977.

HARRISON, A.J., The Track Cost Issue. *Journal of Transport Economics and Policy*, January 1979.

HEGGIE, I., Charging for Port Facilities. *Journal of Transport Economics and Policy*, January 1974.

HIORT, O.C., Kostnadsorienterte bilavgifter. Transportøkonomisk Institutt, Slemdal, 1964.

HOUTHAKKER, H.S. & MAGEE, S.P., Income and Price Elasticities in World Trade. *Review of Economics and Statistics*, May 1969.

JANSSON, J.O., Medveten överdimensionering av kapaciteten. En för prissättningen och finansieringen av infrastrukturen relevant produktionsekonomisk egenhet. *Sosialøkonomen*, Oktober 1970.

JANSSON, J.O., Prissättning av gatuutrymme. EFI, 1971.

JANSSON, J.O., Intra-Tariff Cross-Subsidization in Liner Shipping. *Journal of Transport Economics and Policy*, September 1974.

JANSSON, J.O., MC-Pricing of Scheduled Transport Services. *Journal of Transport Economics and Policy*, September 1979.

JANSSON, J.O., A Simple Bus Line Model for Optimization of Service Frequency and Bus Size. *Journal of Transport Economics and Policy*, January 1980.

JANSSON, J.O. & RYDÉN, I., Swedish Seaports - Economics and Policy. EFI, 1979.

JANSSON, J.O. & SHNEERSON, D., Economies of Scale of General Cargo Ships. *Review of Economics and Statistics*, May 1978.

JANSSON, J.O. & SHNEERSON, D., The Transformation of Cartel Profits into Social Losses in Liner Shipping. *International Journal of Transport Economics*, August 1978.

JANSSON, J.O. & SHNEERSON, D., The Effictive Protection Implicit in Liner Freight Rates. *Review of Economics and Statistics*, November 1978.

JANSSON, J.O. & SHNEERSON, D., Charging What the Traffic Can Bear - Alternative Rule to the Value of Service Principle. *International Journal of Transport Economics*, December 1979.

JOY, S., British Railways' Track Costs. *Journal of Industrial Economics*, Vol. 13, 1964. Reprinted in "Transport", ed. Munby, D., Penguin Modern Economics, 1968.

JOY, S., Pricing and Investment in Railway Freight Services. *Journal of Transport Economics and Policy*, September 1971.

KALDOR, N., The Equilibrium of the Firm. *Economic Journal*, March 1934.

KALDOR, N., The Irrelevance of Equilibrium Economics. *Economic Journal*, December 1972.

KJESSLER & MANNERSTRÅLE AB, Sjöterminaler för enhetslasttrafik, 1977.

KOSHAL, R.K., Economies of Scale in Bus Transport: Some Indian Experience. *Journal of Transport Economics and Policy*, January 1970.

KOSHAL, R.K., Economies of Scale in Bus Transport: Some United States Experience. *Journal of Transport Economics and Policy*, January 1972.

KRITZ, L., Transportpolitiken och lastbilarna. IUI, 1976.

LANCASTER, K.J., A New Approach to Consumer Theory. *Journal of Political Economy*, Vol. 74, 1966.

LAWRENCE, S.A., International Sea Transport: the Years Ahead. Lexington Books, D.C. Heath & Co., 1972.

LEE, N. & STEEDMAN, N., Economies of Scale in Bus Transport. *Journal of Transport Economics and Policy*, January 1970.

LIPSEY, R. & LANCASTER, K., The General Theory of the Second-Best. *Review of Economic Studies*, 1957.

MCDEVITT, P.K., Returns to Scale and Technological Change in Urban Mass Transit. *The Logistics and Transportation Review*, No. 4, Vol. 12, 1976.

MEYER, J., et al., The Economies of Competition in the Transportation Industries. Harvard University Press, 1959.

Ministry of Transport, Road Pricing: the Economic and Technical Possibilities. HMSO, 1964.

Ministry of Transport, Road Track Costs. HMSO, 1968.

MOHRING, H., Optimization and Scale Economies in Urban Bus Transportation. *American Economic Review*, September 1972.

MOHRING, H. & HARWITZ, M., Highway Benefits: an Analytical Framework. The Transportation Center at the Northwestern University, 1962.

MOHRING, H. & WILLIAMSON Jr., H.F., Scale and "Industrial Reorganization" Economies of Transport Improvements. *Journal of Transport Economics and Policy*, September 1969.

MORGAN, R.T. & Partners, Costing of Bus Operations. An interim report of the Bradford Bus Study. July 1974.

NILSSON, G., Studier av samband mellan olyckor, vägens utformning och trafikens storlek. Statens Väg- och Trafikinstitut. Rapport Nr 27, 1973.

PEACOCK, A.T. & ROWLEY, C.K., Welfare Economics and the Public Regulation of Natural Monopoly. *Journal of Public Economics*, 1972.

PIGOU, A.C., Economics of Welfare. London, 1920.

QUARMBY, D.A., Effect of Alternative Fares Systems on Operational Efficiency. British Experience. In symposium on Public Transport Fare Structure: papers and discussion. TRRL Supplementary Report 37UC. Crowthorne, 1974.

RAMSEY, F., A Contribution to the Theory of Taxation. *Economic Journal*, March 1927.

ROTSCHILD, M. & STIGLITZ, J.E., Increasing Risk II: Its economic consequences. *Journal of Economic Theory* 3, 1971.

SAATY, T.L., Elements of Queuing Theory. McGraw-Hill, 1961.

SAMUELSON, P., The Monopolistic Revolution, in KUENNE, R.E. (ed.), Monopolistic Competition Theory: Studies in Impact. Wiley, 1967.

von SCHIRACH-SZMIGIEL, C., Liner Shipping and General Cargo Transport. EFI, 1979.

SIMON, J.L. & VISRABHANATHY, G., Auction Solutions for Airline Overbooking. *Journal of Transport Economics and Policy*, September 1977.

SLETTEMARK, R., Optimalisering av hoveddimensjonerne for havner. TØI, 1974.

Statens Vägverk, En studie av vägstandard. P 003, 1978-11.

STURMEY, S.G., Economics and International Liner Services. *Journal of Transport Economics and Policy*, May 1967.

STROTZ, R., Urban Transportation Parables, in MARGOLIS, J. (ed.), The Public Economy of Urban Communities. Resources for the Future, 1965.

THOMSON, J.M., Modern Transport Economics. Penguin Modern Economic Texts, 1974.

THORBURN, T., Supply and Demand of Water Transport. FFI, 1960.

TJALLING, C.K., Three Essays on the State of Economic Science. McGraw-Hill, New York, 1957.

Trafikpolitik - behov och möjligheter. SOU 1975:66 (TPUI).

Trafikpolitik - kostnadsansvar och avgifter. SOU 1978:31 (TPUII).

TRAIN, K., Optimal Transit Prices Under Increasing Returns to Scale and a Loss Constraint. *Journal of Transport Economics and Policy*, May 1977.

TURVEY, R., Public Utility Pricing Under Risk: comment. *The Economic Review*, June 1970.

TURVEY, R. & MOHRING, H., Optimal Bus Fares. *Journal of Transport Economics and Policy*, September 1975.

United States Congress, Discriminatory Ocean Freight Rates and the Balance of Payments. Hearings before the Joint Economic Committee and the Sub-committee on Federal Procurement and Regulation of the Joint Economic Committee, 88th and 89th Congress, 1963-65.

VICKREY, W., Some Implications of Marginal Cost Pricing for Public Utilities. *American Economic Review*, May 1955.

VICKREY, W., Returns to Scale in Transit: a comment. *The Logistics and Transportation Review*, Vol.13, No. 1, 1977.

VISSCHER, M.L., Welfare-Maximizing Price and Output with Stochastic Demand: comment. *The American Economic Review*, March 1973.

Vägtrafiken - kostnader och avgifter. SOU 1973:32 (VKU).

WALTERS, A.A., The Allocation of Joint Costs. *American Economic Review*, June 1960.

WALTERS, A.A., The Theory and Measurement of Marginal Private and Social Costs of Highway Congestion. *Econometrica*, 1961:4.

WALTERS, A.A., Characteristics of Demand and Supply. European Conference of Ministers of Transport: International symposium on Theory and Practice in Transport Economics. Strasbourg, October 8, 1964.

WALTERS, A.A., The Economics of Road User Charges. World Bank Occasional Papers, No. 5, 1968.

WALTERS, A.A., Marginal Cost Pricing in Ports. *The Logistics and Transportation Review*, Vol. 12, No. 3, 1976.

WILDASIN, D.E., Distributional Neutrality and Optimal Commodity Taxation. *American Economic Review*, December 1977.

WILLIAMSON, O.E., Franchise Bidding for Natural Monopolies - in general and with respect to CATV. *The Bell Journal of Economics*, 1974.

ZVI GRILICHES, Cost Allocation in Railroad Regulation. *The Bell Journal of Economics and Management Science*, Spring 1972.

LIST OF REPORTS PUBLISHED SINCE 1977 BY THE ECONOMIC RESEARCH INSTITUTE AT THE STOCKHOLM SCHOOL OF ECONOMICS

Unless otherwise indicated, these reports are published in English.

1977

ANELL, B., FALK, T., FJAESTAD, B., JULANDER, C-R., KARLSSON, T., STJERNBERG, T., 1977, The study's conceptual framework arrangement, and execution. A report from the research program "The retailing in change". Stockholm. (Mimeographed)[1]

BERGMAN, L., 1977, Energy and economic growth in Sweden - an analysis of historical trends and present choices. Stockholm. (Mimeographed)

BERTMAR, L., MOLIN, G., 1977, Capital growth, capital structure and rates of return - an analysis of a Swedish industrial company.' Stockholm.[2]

DOCHERTY, P., MAGNUSSON, Å., STYMNE, B., CALLBO, K. & HERBER, S., 1977, How to succeed with systems development - an analysis of five cases. Stockholm.[2]

EKLUND, L. & SJÖSTRAND, S-E., 1977, The organization of learning - the Kista case. Stockholm.[1]

FJAESTAD, B. & HOLMLÖV, P-G., 1977, The market for broadband services - interest and planned participation in Televerket's test network for picture-phone, fast tele facsimile and special television. Stockholm. (Mimeographed)[1]

HAMMARKVIST, K-O., 1977, Adoption of new products on the building market. Stockholm.[1]

HAMMARKVIST, K-O., 1977, Buyer behavior in the building industry. Stockholm. (Mimeographed)[1]

HEDBERG, P., JOHANSSON, L. & JUNDIN, S., 1977, Children and marketing communication - a study of small children's wishes as regards products and sources of influence on their wishes. Stockholm. (Mimeographed)[1]

[1] Only published in Swedish.

[2] Published in Swedish with an English summary.

HÖGLUND, M., 1977, Changes in the retail assortment - a study of the trade in rarely bought goods in Sweden. Stockholm. EFI/MTC.[1]

JULANDER, C-R., 1977, Effects among producers of comparative product testing information about washing machines. Stockholm.[1] (Mimeographed)

JULANDER, C-R., LINDQVIST, A. & FJAESTAD, B., 1977, Development of indicators of saving behavior - a survey of the literature and an interview study of 50 households in Stockholm. Report No. 1 from the project "Savings behavior". Stockholm.[1] (Mimeographed)

JULANDER, C-R. & LÖÖF, P-O., 1977, Evaluation of facts about furniture among the consumers. Stockholm.[1] (Mimeographed)

JULANDER, C-R. & LÖÖF, P-O., 1977, Evaluation of facts about furniture among the producers and the retailers. Stockholm.[1] (Mimeographed)

JULANDER, C-R., LÖÖF, P-O. & LINDQVIST, A., 1977, Information for the design of product information - interviews with consumers and analysis of complaints in the furniture market. Stockholm.[1] (Mimeographed)

KARLSSON, H., 1977, The Linden shopping center - execution and design. A report from the research program "The retailing in change". Stockholm.[1] (Mimeographed)

MOSSBERG, T., 1977, Development of key variables. Stockholm.[2]

OLVE, N-G., 1977, Multiobjective budgetary planning - models for interactive planning in decentralized organizations. Stockholm.

STJERNBERG, T., 1977, Organizational change and quality of life - individual and organizational perspectives on democratization of work in an insurance company. Stockholm.

ÖSTMAN, L., 1977, The use of accounting measures in management control - a study of profit- and ROI-measures in multi-divisional companies. Stockholm.[1]

1978

ASPLING, A. & LINDESTAD, L., 1978, In-company training. Reflection of three cases. Stockholm.[1]

EKMAN, E.V., 1978, Some dynamic economic models of the firm. A microeconomic analysis with emphasis on firms that maximize other goals than profit alone. Stockholm.

FJAESTAD, B. & JULANDER, C-R., 1978, Retailers' view of the future. A study of retailers in Norrköping. Report No. 6 from the research program "The retailing in change". Stockholm. (Mimeographed)[1]

FJAESTAD, B., JULANDER, C-R. & ANELL, B., 1978, Consumers in shopping centers in Norrköping. Report No. 4 from the research program "The retailing in change". Stockholm. (Mimeographed)[1]

[1] Only published in Swedish.

[2] Published in Swedish with an English summary.

FORSBLAD, P., SJÖSTRAND, S-E. & STYMNE, B.(eds.) 1978, Man in organizations. A view of co-determination and leadership. Stockholm: EFI/Liber Förlag.[1]

HEMMING, T., 1978, Multiobjective decision making under certainty. Stockholm.

HOLMLÖV, P.G., 1978, News in local press and municipal policy. Stockholm.[2]

HULTCRANTZ G., LINDHOFF, H. & VALDELIN, J., 1978, Chinese economic planning: Characteristics, objectives and methods - an introduction. China's developmental strategy: 7. Stockholm (Mimeographed)

JULANDER, C-R. & EDSTRÖM, E., 1978, Information for the design of product information about flats. An interview study with tenants in Husby/Akalla. Stockholm. (Mimeographed)[1]

JUNDIN, S., LINDQVIST, A., 1978, Telephone as a mail-box. A study of 100 tele facsimile users. Stockholm.[1]

LINDQVIST, A., JULANDER, C-R. & FJAESTAD, B., 1978, Development of indicators of saving behavior - Report No. 2 from the project "Savings behavior". Stockholm. (Mimeographed)[1]

SCHWARZ, B. & SVENSSON, J-E., 1978, Transports and transports research. Stockholm.[2] (Off-print from Transportdelegationen's publication 1978:7.)

STJERNBERG, T., 1978, Retail employees in Linden. A study of employees influence and experiences in connection with the establishment of a shopping center. Report No. 3 from the research program "The retailing in change". Stockholm. (Mimeographed)[1]

TELL, B., 1978, Capital budgeting in practice. Stockholm.[1]

1979

AHLMARK, D. & BRODIN, B., 1979, The book trade in the future. Marketing and distribution. Stockholm.[1]

AHLMARK, D. & BRODIN, B., 1979, State subsidies for the production and distribution of books - an economic analysis. Stockholm.[1]

AHLMARK, D. & LJUNGKVIST, M-O., 1979, The financial analysis and management of publishing houses. Studies of development and behavior during the 1970s. Stockholm.[1]

ANELL, B., 1979, When the store closes down. Report No. 5 from the research program "The retailing in change". Stockholm. (Mimeographed)[1]

ANELL, B., 1979, Consumers and their grocery store. A consumer economic analysis of food buying patterns. Stockholm.[1]

[1] Only published in Swedish.

[2] Published in Swedish with an English summary.

BERTMAR, L., 1979, Wages profitability and equity ratio, Stockholm.[1] (Off-print from SOU 1979:10.)

BORGENHAMMAR, E., 1979, Health care budgeting, goals, structure, attitudes. Stockholm.

ELVESTEDT, U., 1979, Decision analysis - an interactive approach. Stockholm.[2]

ENGLUND, P., 1979, Profits and market adjustment. A study in the dynamics of production, productivity and rates of return. Stockholm.

ETTLIN, F.A., LYBECK, J.A., ERIKSON, I., JOHANSSON, S. & JÄRNHÄLL, B., 1979, The STEP 1 quarterly econometric model of Sweden - the equation system. Stockholm.

FALK, T., 1979, The retail trade in Norrköping. Structure and location 1977. Report No. 7 from the research program "The retailing in change". Stockholm. (Mimeographed)[1]

FALK, T., 1979, Retailing in Norrköping. Structural and locational changes 1951 - 1977. Report No. 9 from the research program "The retailing in change". Stockholm. (Mimeographed)[1]

HEDEBRO, G., 1979, Communication and social change in developing nations - a critical view. Stockholm: EFI/JHS. (Mimeographed)

HOLMLÖV, P.G., FJAESTAD, G. & JULANDER, C-R., 1979, Form and function in marketing. Household appliencies and kitchen carpentry: sales product development and advertising rhetoric 1961 - 1976. Stockholm. (Mimeographed)[1]

JANSSON. J.O. & RYDÉN, I., 1979, Cost benefit analysis for seaports. Stockholm.[1]

JANSSON, J.O. & RYDÉN, I., 1979, Swedish seaports - economics and policy. Stockholm. (Mimeographed)

JULANDER, C-R. & FJAESTAD, B., 1979, Consumer purchasing patterns in Norrköping and Söderköping. Report No. 8 from the research program "The retailing in change". Stockholm. (Mimeographed)

JUNDIN, S., 1979, Children and consumption. Stockholm.[1]

MAGNUSSON, Å., PETERSSOHN, E. & SVENSSON, C., 1979, Non-life insurance and inflation. Stockholm.[1]

Marketing and structural economics, 1979, (ed. Otterbeck, L.,) Stockholm: EFI/IIB/Studentlitteratur.[1]

PERSSON, M., 1979, Inflationary expectations and the natural rate hypothesis. Stockholm.

von SCHIRACH-SZMIGIEL, C., 1979, Liner shipping and general cargo transport. Stockholm.

ÖSTERBERG, H., 1979, Hierarchical analysis of concepts - a technique for solving complex research problems. Stockholm: EFI/Norstedts.

[1] Only published in Swedish

[2] Published in Swedish with an English summary.