USER ECONOMIES OF SCALE

AND OPTIMAL BUS SUBSIDY

Peter Tisato

Economics Department

University of Adelaide

Adelaide, South Australia

December, 1995

A thesis submitted in fulfilment

of the requirements for the

degree of *Doctor of Philosophy*

To my wife, children, parents and family

CONTENTS

Page No.

Chapter 4 : Optimal Pricing, Frequency and Subsidy Formulation

Chapter 5 : Optimal Subsidy and Cross-Subsidy With A Logit Model

Chapter 6 : Service Unreliability and Subsidy

Page No.

Chapter 7 : User Economies of Scale Subsidy in Adelaide

Chapter Appendix :

Appendix B : Parameter Value Derivation

continued next page

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This thesis addresses the *user economies of scale* argument for bus subsidy using the bus system in Adelaide, South Australia, as a case study.

The first area of focus of the study is on the modelling of user cost, a key determinant of optimal subsidy. An improved user cost model is developed, using a logit model to predict user choice between random and planned behaviour, two behavioural modes the user can adopt when catching a bus. Use of a logit model contrasts with the simple deterministic choice models used in several recent subsidy analyses. The term service delay is introduced to describe the user delay caused by the scheduling of bus services, and the concept of stochastic delay is expanded to cover stochastic elements in both user demand and service departure times.

A further aspect of user cost which is investigated is the link between service unreliability (a determinant of user cost) and optimal subsidy, a relationship which has been largely unstudied to date. Unreliability is found to have a significant impact on subsidy results through the influence that changes in unreliability have on the timing and nature of switching by users between random and planned behaviour. Changes in unreliability are found to result in 50% (and greater) changes in subsidy. Service unreliability is therefore shown to be an important determinant of optimal bus subsidy. It is also shown that an increase in subsidy in response to a rise in unreliability may be economically justified, although such a policy recommendation may appear perverse to the community, and therefore difficult to implement. Finally, the trade-offs involved in reducing service unreliability and setting subsidy policy are explored, and a role for road congestion in first best subsidy analysis is established.

The second area of focus is on how introducing a logit choice model for random vs planned user choice affects subsidy analysis. A motivating force for considering this is that recent work has found that use of a bi-modal random vs planned user choice model can result in scope for multiple local optima in the bus optimisation problem, and can result in a break down of the conventional rule that optimal unit subsidy is greater the less patronised a route. The study found that when a logit model is used, the likelihood for multiple local optima largely disappears, and although the conventional negative unit subsidy/patronage relationship may still break down, the likelihood and severity of the breakdown are diminished. It is also shown that, as switching is occurring between random and planned user behaviour, optimal total subsidy can grow at accelerated rates. Further, conventional optimal cross subsidy results are distorted by use of a logit model, making it difficult *a priori* to predict how cross subsidy will vary in changing circumstances.

The third area of focus of the study is the estimation of optimal user economies of scale subsidy for Adelaide buses. Optimal subsidy is estimated at a disaggregated bus corridor level for peak and off-peak periods with recognition of current concession fare policy, the pending introduction of competitive tendering, and the efficiency costs associated with raising public finance (i.e. a second best world). First best optimal subsidies for Adelaide buses are found to be well below current levels, even after the introduction of competitive tendering, with optimal subsidy in a second best world being even lower. The analysis shows that as public finance becomes increasingly costly to raise, optimal peak subsidy can decline sufficiently that it becomes smaller than optimal off-peak subsidy. Ensuring the balance between price and frequency at any given subsidy level is found to be at least as important as achieving optimal levels of patronage and subsidy. Finally, provided that public finance is costly enough to raise, optimal subsidy can approach, and fall to, zero.

DECLARATIONS

This work contains no material which has been accepted for the award of any other degree or diploma in any university of tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

Peter Tisato

December 1995

ACKNOWLEDGEMENTS

My greatest thanks go to my wife Pat, and my children James and Laura, for their love and support during this long, difficult and trying exercise. Without their assistance, this work would not have been completed. To my mother and father and family, thank you for your support and the pride you have shown in my achievements. To my friends Phil and Gayle, my thanks for their gentle reminders and concern.

I am very grateful for the assistance and guidance offered to me by my two supervisors : Dr Christopher Findlay of the University of Adelaide, and Professor Peter Forsyth of the University of New England. Their advice on how best to proceed, comments on chapter drafts and general encouragement were greatly appreciated.

I wish to offer special thanks to Dr Derek Scrafton, Director-General of Transport, South Australia, who has always been prepared to listen, read drafts, discuss issues, and has always offered valued and constructive advice. His continuous support and assistance in this reasearch, and more generally over many years, is greatly appreciated.

I am also indebted to the generous financial support that the South Australian government (via the Transport Policy Unit) has provided me through their joint funding, with the Economics Department at the University of Adelaide, of a Research Fellowship in Transport Policy which I have held in the South Australian Centre for Economic Studies during the course of my research. In this respect, my thanks also go to Professors Cliff Walsh and Jonathan Pincus, Mike Philipson of the SA Transport Policy Unit, and the staff of the South Australian Centre for Economic Studies, particularly Davina Dolman and Cheree Metcalfe, for their assistance in establishing and administering the Research Fellowship, and for generously providing me with the resources for undertaking my research. The Engineering and Water Supply Department provided valuable leave without pay which has enabled me to spend time working in the tertiary education sector and to undertake this reaserch.

The transport agencies in South Australia (the Transport Policy Unit, the Passenger Transport Board, TransAdelaide, the former State Transport Authority, and the South Australian

Department of Transport) each provided data and assistance when requested. I would like to particularly thank Mike Philipson, David Bray, David Willis, Alison Anlezark, Tom Wilson, Adrian Gargett, Kym Della-Torre, Greg Fenn, Roberto DiVisio, Peter Doble, Roger Seaman, Lindsay Oxlad, Nicole Ricketts and Sue Wight for their assistance.

The encouragement and support of numerous colleagues in the Economics Department helped sustain me during my research. My thanks in particular to Ian McLean and John Hatch, and also to Alastair Fischer, John Siegfried, David Round, Tin Nguyen, Penny Neal and Margaret Meyler.

The participants of the 1994 Australian PhD Conference in Economics provided valuable comments on a paper I delivered based on my research. The suggestions and advice of Dr Colin Gannon proved particularly valuable. I offer each of them my thanks.

My thanks also to a number of other colleagues who provided advice and assistance of one form or another during the research, in particular John Dodgson, Paul Kerin, and Mark Turnquist.

Chapter 1
INTRODUCTION

This thesis addresses a particular argument for public transport subsidy, known as the *User Economies of Scale* (*UES*) subsidy argument, using the bus system in metropolitan Adelaide, South Australia, as a case study.

The motivation for selecting this topic has several bases. First, the performance of the Australian economy, and its state economies, has been such in recent years that the country has entered, and is currently undergoing, an important microeconomic reform phase. One of the prime areas of focus in this reform is a reassessment of financial performance of the public sector. The need for such a focus has been particularly necessary in several states, including South Australia, where massive negative shocks have occurred in the public finance system. As a result, all areas of public finance, including public transport subsidy, have come under close scrutiny (Commission of Audit, 1994). In this general context, public transport subsidy is therefore a worthwhile and important topic for research and investigation.

The case study is limited to urban bus transport only. The main reason for this is that in Adelaide, like most other major Australian cities, bus transport is the dominant form of public transport (Industry Commission (IC), 1994).[1] In addition, bus transport has been the focus of recent reform proposals in Adelaide where the government has made the planned introduction of competitive tendering the key plank of its urban passenger transport policy reform program (Liberal Party of South Australia, 1993; Passenger Transport Board, 1994). The government's intention is to use the anticipated cost reductions from the introduction of competitive tendering to reduce subsidy, and thus take an important first step towards improving the financial situation of public transport in

[1] In Adelaide, travel by bus accounted for 83% of all public transport boardings in 1992/93 (STA, 1993a).

Adelaide. This raises the question, however, of how subsidy should be subsequently managed once the gains from competitive tendering have been reaped. In particular, what is the appropriate level of Adelaide bus subsidy which can be justified on economic efficiency grounds and which policy makers could target in the longer term ?

Although a range of arguments have been put forward in favour of public transport subsidy in the past, including a number of economic efficiency arguments, only a single one will be considered here, *UES*. In brief, the logic of the argument is as follows. Bus transport is usually characterised by constant returns to scale in producer costs, with marginal and average producer costs therefore being equal. If pricing was based purely on the nature of producer costs, efficient pricing at marginal cost would therefore deliver a financial breakeven outcome, and thus no need for subsidy. As Mohring (1972) first pointed out, however, the determination of an efficient outcome requires consideration of *all* costs, both producer costs and user costs. The existence of economies of scale in some user costs then generates a case for subsidy.

The relevant user costs are time costs which are related to the frequency of service. The relevant time costs are waiting time spent at bus stops, and any other user delays caused by services departing at times which do not suit users. An increase in frequency, and thus the scale of operation, will cause these user costs to fall for all users. The marginal user cost of additional frequency is therefore below average user cost, that is, there are scale economies in user costs (or "user economies of scale", *UES*)[2]. As a result, as in all cases of economies of scale, efficient pricing at marginal social cost will result in a financial deficit and a need for subsidy.[3]

A number of reasons exist for focusing on the *UES* subsidy argument. First, many of the other arguments proposed in support of subsidy do not have a sound basis (Amos and Starrs, 1984). Second, the *UES* argument has been a central element in the major economic efficiency based studies of subsidy overseas, yet, surprisingly, it has received little attention in Australia. Third, the

[2] To the author's knowledge, the term "user economies of scale" was first coined in the Travers Morgan (TM) study of urban public transport subsidy in New Zealand (TM, 1988).

[3] This assumes that uniform pricing continues in public transport. If two-part tariffs were adopted, allocative efficiency could be achieved without subsidy (Brown, J.B. and Sibley, D.S, 1986; Allen, 1987).

subsidy argument which has received most focus in Australia over the years, that subsidy is a second-best policy for managing road congestion when roads are unpriced, is being increasingly recognised as being of limited importance, particularly in relatively low road congestion cities like Adelaide. The *UES* of scale argument therefore appears to be an important subsidy argument which has received limited attention in Australia, suggesting that focus on this topic is therefore warranted.

Within this general framework, this study has two main thrusts. The first, which is the concern of chapters 3 to 6, involves extending in a number of ways the literature on, and methodology for investigating, *UES* subsidy. This is warranted in its own right since it will be shown that there are a number of unresolved general issues in the literature on this topic. In addition, these extensions provide an improved framework for the study's second and central thrust, the estimation of *UES* subsidy and other optimal policy settings for the Adelaide bus system, reported in chapter 7.

The next chapter, chapter 2, places the study in a broad context, and outlines its scope. It considers the economic situation in Adelaide which makes the investigation of subsidy an important issue, it reviews the arguments for public transport subsidy that have been presented in the past, it provides an overview of the literature on user economies of scale (with a more detailed review occurring on a gradual basis in subsequent chapters), it reviews the previous subsidy work that has occurred in Adelaide, and establishes the need for the subsequent work reported in the study. Chapter 2 concludes with a chapter plan which provides a brief summary of the detailed content of the thesis.

The appendices at the end of the thesis provide a (foldout) glossary of abbreviations and notation used in the report, and summarise the derivation and collection of data and parameter values.

Chapter 2
BACKGROUND, CONTEXT AND SCOPE OF STUDY

2.1 Public Transport Subsidy and the Policy Environment

Public transport operating subsidies have for many years been a feature of the public transport systems of numerous cities around the world, with the size of these subsidies growing steadily over time.

Public transport subsidies in Australia consist of two components. The first is community service obligations (CSO's) which fund fare concessions offered to specific groups of travellers (e.g. pensioners and students) as part of social justice and equity policy. The second component is the subsidy required to cover the deficit arising from the general level of (non-concession) fares being below unit costs. In Adelaide, in 1992-93, concession subsidy accounted for 15% of total public transport subsidy, and deficit funding for the other 85%.[1]

Subsidies have been a particularly pronounced feature of Australian urban public transport systems for many years. For 1991-92, the Industry Commission (IC, 1994) reports that the combined urban public transport operating deficits of the capitals cities in Australia totalled $2.6 billion,[2] which amounts to about $450 per household, or $1360 per passenger. When compared to operating costs, farebox revenues yield an average cost recovery ratio of around 30%[3], which is typically below cost recovery levels found overseas (IC, 1994).

[1] The corresponding figures for buses were 20% and 80%.

[2] Excludes depreciation, local government subsidies, a return on assets *and* concessional fare CSO's. Inclusion of depreciation and local government subsidies raises the figure to around $3 billion.

[3] Sydney has the highest level of cost recovery (44%) and Perth the lowest (17%), with revenues in Adelaide covering about 19% of operating costs.

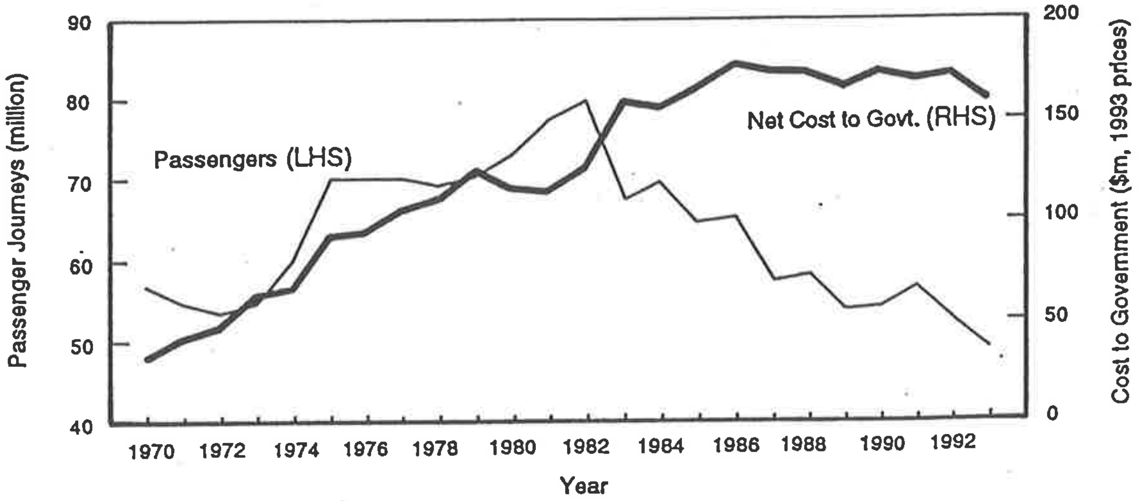
Bray (1995) reports the history of public transport subsidy in Adelaide, illustrated here in Figure 2.1. Total subsidy, or the net cost to government of public transport, grew from (in 1993 dollars) $32 M in 1970 to $177 M in 1986, and then fell slowly in subsequent years to $159 M in 1993. From 1970 to 1982, public transport usage also grew (although much more slowly than subsidy), but importantly, usage has fallen dramatically since 1982, causing subsidy per journey to rise more rapidly than in the previous decade.

Bray (1995) points to a number of factors which have influenced these results : the rises in oil prices in 1973 and 1979, making car travel relatively less affordable in the 1970's; reduced productivity, increases in vehicle-kms and increase in costs (partly due to the existence of subsidy support) of public transport service delivery; and, rising incomes and falling real fuel costs during the 1980's making car travel more attractive. Added to this, the community and government perception has been that public transport is an essential good which must be kept affordable for all users. Governments have also been strongly lobbied by strong special interest groups (e.g. pensioners) to keep fares low. As a result, fare rises have been limited, and have failed to keep pace with cost increases.

Perceptions of the growth in public transport subsidy has varied over this period. In the 1970's, the growth in subsidy was an outcome which could be accommodated given the relatively prosperous state of the Australian and state economies. As the 1980's unfolded, however, general economic conditions declined, public finance began to become relatively less abundant, and the need to control the growth in public transport subsidies became apparent (Scrafton, 1985; Fielding, 1988).

In the 1990's, a major recession and generally poor national economic performance, and a major public finance shock in South Australia in the form of massive financial losses by the government owned State Bank, has made public funds increasingly scarce, and has dramatically altered the policy environment for the whole public sector. Within this context, a blueprint for major reform of the public sector in South Australia (Commission of Audit, 1994) has been largely accepted by government and is being implemented. The Commission of Audit, and the equally important Industry Commission review of urban transport in Australia (IC, 1994), have pointed to the size of public transport subsidy as a problem which requires attention. Making the matter

Table 2.1 : Public Transport Patronage and Subsidy in Adelaide, 1970-93



Source : Bray (1995).

worse, evidence suggests that the prime goal of subsidisation over recent decades, namely the achievement of improved equity, is not being met, with equity in fact being made *worse* as a result of the way subsidies are delivered (e.g. Travers Morgan (TM), 1984; Duldig and Gaudry, 1993).

Overall, public transport subsidies have increasingly been viewed as an area of urban transport and public sector policy which requires reform.

2.2 Current Reform - Improving Production Efficiency Through Competitive Tendering

One of the reasons why subsidies have been higher than they need be is the existence of production inefficiencies in the delivery of public transport services, which raises the cost of service provision. The *identical* service could be provided in alternative ways which draw on fewer economic resources, thus releasing resources for other productive functions in the economy.

The existence of production inefficiencies and higher costs is usually explained as being due to the public monopoly ownership and operation of public transport services, the mode of operation which has dominated public transport provision in most industrialised countries for decades, including Adelaide.[4] Two reasons are usually given for why public monopoly costs are high. First, if the monopoly operator operates with the support of subsidy from government, there can be a "leakage" from subsidy to costs (e.g. Turk and Sullivan, 1987; Pucher et al, 1983). The operating environment facing the operator is such that the incentive to minimise costs is lower than it would be in the absence of a subsidy, and as a result services are provided at above minimum cost. A second, and more general argument has been that the public sector is less efficient than the private sector, although as Domberger (1993) argues, this need not always be the case. An important point here is that ownership is not necessarily an important determinant of production efficiency, but that

[4] In Adelaide, a public monopoly in public transport has existed since 1972. Prior to then train and tram services were government provided, but about half the bus services were delivered by the private sector. When this research commenced, the public monopoly operated under the name of the State Transport Authority (STA), with a full ranges of policy, planning and operational functions. In 1994, the government, in it reform agenda, separated these functions, with the new Passenger Transport Board (PTB) becoming responsible for policy, regulation and funding matters, whilst operational matters became the sole function of a new operating agency, TransAdelaide.

the incentive structure faced by the operator, and the level of actual or potential competition it faces, is important.

Following the important Fielding report in 1988 (Fielding, 1988) a number of reforms have been introduced aimed at improving production efficiency. In the late 1980's and early 1990's, organisational changes and the adoption of a commercialisation approach saw the introduction of service units and business units in the STA with the aim of eliminating non-core activities (Commission of Audit, 1994). More importantly, the current government, as part of its public sector reform agenda driven by the Commission of Audit, has committed itself to the introduction of competitive tendering in public transport service delivery as a way of introducing competitive pressures and improving production efficiency.[5] The initial focus is on competitively tendering the bus system, with similar reforms proposed for the other public transport modes once reform in the bus sector is complete (Liberal Party of South Australia, 1993).

Under the new system, the Adelaide bus system has been divided into a number of service areas (Passenger Transport Board, 1994), with tenders to be invited from any suitable party from private or public sectors to bid for the right to deliver bus services to a given area for a contract period of between 3 to 5 years[6]. The government will specify minimum service standards, will control and coordinate timetables so that service integration and coordination occurs across the entire bus system, and will set fare policy.

Experience elsewhere has demonstrated the potential cost savings that are attainable from the introduction of competitive tendering. Bus services in New Zealand, for example, have benefited from such reforms (Wallis, 1991). A more general review by Stanford (1992), and specific reviews of public transport in Australia by Hensher and Daniels (1994) and TM (1994), suggests that cost reductions in the order of 20% to 30% are achievable.

Evidence that the process is likely to deliver cost reductions appeared recently when the bus drivers of the incumbent, TransAdelaide, made an offer to management of a voluntary wage

[5] Reducing production inefficiencies is a central plank in process of microeconomic reform occurring in Australia (Forsyth, 1992).

[6] The current bus fleet has been retained in public ownership, and will be leased to successful bidders if required.

reduction to ensure that its tender bids are competitive with expected private sector bids. The threat of competition has therefore resulted in a reassessment of the incumbent's operations to reduce the chances of it losing the market which it currently services. Thus even if the incumbent retains the right to service some or all of the network, competitive tendering will have brought about an improvement in the production efficiency of that service delivery.

2.3 Other Reform Issues

The focus of current reform has rightly been on improving production efficiency given the substantial magnitude of the potential savings which could be made. Beyond this, there are a range of other issues and reform options in the area of urban public transport, related to subsidy, which will require addressing in the future. Of these, only the question of optimal subsidy will be addressed in this study. Although the remainder will not be formally assessed or analysed here, it is important to note them as a means of establishing the full context within which the issue of optimal subsidy can be considered.

Better Targeting and Delivery of Assistance

An issue of great concern is whether current public transport subsidies are effective in improving equity, arguably the prime goal of subsidy policy in Adelaide to date. In essence, is subsidy being effectively targeted and delivered to those who most need it ? The available evidence (e.g. TM, 1984; Duldig and Gaudry, 1993) suggests that subsidy is in fact currently poorly targeted and delivered, with the current system actually *worsening* equity. These studies show that the biggest per trip subsidies go to peak hour commuters, yet these users have relatively higher levels of income and wealth than other public transport users. The result is that scarce subsidy dollars, rather than being distributed to users who deserve financial assistance, are instead going to users who are better off and have greater capacity to pay for the cost of service provision.

If transport provision is to be a tool for improving equity in society, then a strong argument can be made for improving the way in which the achievement of this goal is approached. Assistance should be better targeted towards those users who are genuinely in need, namely the transport disadvantaged, for example those in financial need. Further, assistance should be provided in the most appropriate manner, which may or may not involve public transport. Alternative approaches (focusing on user subsidies rather than the current producer subsidies) which may have more merit include direct income supplementation, travel vouchers, taxi trip subsidies, or assistance with the payment of car purchase costs and/or registration fees. The question of what is the most appropriate form of assistance is not yet fully resolved (IC, 1994), and will not be considered in this study.[7]

The Role of Public Transport

The future role of public transport is also currently somewhat unresolved. Traditionally, public transport has had two key roles to perform. First, it has provided an essential source of mobility for those who cannot afford private travel, particularly the poor and users eligible for concession fares. In Adelaide in 1993, about 60% and 70% respectively of travellers in the peak and off-peak were concession travellers (STA, 1993a). The transport poor are likely to continue to be a major share of the public transport market.

The second key role of public transport has been to provide mass transit of large numbers of people with relatively common travel destinations, especially travel to the CBD. However, the nature of urban development over a number of decades has been such that the proportion of travel destinations outside the CBD has gradually increased. A much greater proportion of travel needs are now for highly differentiated cross suburban movement, something to which traditional public transport is poorly suited (Bray, 1995). Increasingly, there is a need for new forms of public transport which are better suited to dispersed travel patterns. Taxis and other forms of transport between the car and conventional public transport, often called community transport or paratransit, may and could play an increased role, provided they are not excessively regulated.[8]

Interestingly, urban planners (Australian Urban and Regional Development Review, 1994, 1995), public transport lobby groups, the environmental lobby, and potentially the wider community (largely in response to perceived environmental problems from car use), are supporters of an

[7] Making the policies usually preferred by economists, such as direct income supplementation, work successfully is heavily dependent on State and Federal governments in Australia cooperating, particularly in relation to access to information such as taxable income.

[8] The taxi experience in developed countries, where entry restriction with fare and quality regulation have been the norm, illustrate the dangers of excessive regulation.

increased role for public transport.[9] However, if environmental problems exist with car use, these are probably best addressed by direct measures aimed at improving cars and the way they are used (IC, 1994). In addition, increasing the role of public transport will be difficult. First, the existing inertia would suggest a continued relative decline of public transport usage. Second, with public transport trips accounting for just 6% of all trips in Adelaide, of which two-thirds are captive public transport users, even if the number of choice users were to double this would only raise the proportion of public transport trips to 8% (Bray, 1995), an increase of only one third. Urban planners often argue that land use patterns should be modified to increase the chances of public transport playing a bigger role. However, the changes required would be large, and it is not clear that land markets, fuelled by the desire of consumers for space, can be controlled to such an extent.

It is difficult to predict how the role of public transport will evolve. Community support for public transport is likely to grow if recent trends are any indication. Yet, unless fares are allowed to rise to bring about increased cost recovery, increasing the role for conventional public transport will also lead to a rise in deficits and subsidy. In the longer term, innovations in service delivery may offer transport solutions at lower cost to the taxpayer.

More Competition

The Industry Commission (IC, 1994) has recently argued that competitive tendering should be seen as a first step in the process of reforming service delivery. It argues in favour of eventual introduction of "open access" competition, also called "on road" competition, where any suitable operator can run a service at any frequency, at any price, with any vehicle. Whilst competitive tendering generates competition for the market, i.e. for the right to service a route/area, open access involves competition in the market, with operators competing for a market share.

Both forms of competition will impose the right incentives on operators, and potential operators, to force them to deliver services closer to minimum cost than uncontested monopoly

[9] The 1995 Federal Government budget also provided a substantial increase in funding for public transport projects through the Better Cities Program.

operation.[10] There are, however, both claimed benefits and disbenefits of progressing to open access competition. A benefit claimed is the potential for market forces to better reflect preferences of consumers in the services delivered, and thus increase allocative efficiency and, through innovation, dynamic efficiency. The basis of this argument is that markets are more effective in extracting information which reveals consumer preferences than other mechanisms such as governments interpreting what these may be. This philosophy played a key role in the case mounted in favour of bus deregulation in the UK (White Paper, 1984; Beesley and Glaister, 1985a, 1985b).

On the other hand, Evans (1990) claims that urban public transport may be a natural monopoly with respect to user costs, basing this observation on the experience in the UK post-deregulation that competition did not appear to be sustainable. The natural monopoly argument is based on average user costs being lower when a single operator is providing services rather than a number of operators. On a route basis, one coordinated timetable is more convenient to users than several independent timetables. At a network level, a single operator may provide better coordination and integration of connections between services and modes, and of ticketing systems, again increasing convenience for the user. In short, there may be economies in user cost from service/information coordination and integration offered by having a single operator.[11,12] The White Paper (1984) and the Industry Commission (IC, 1994) argue, however, that it would be in the interests of operators to provide service coordination, and that a cooperative of operators may be better placed to deliver this than the government. Others, however, question whether the market would deliver effective coordination, or coordination at all (Nash, 1988; Evans, 1990; Hensher, 1993).

In this study, open access is not assessed. The competitive tendering model planned for introduction in Adelaide, which will allow the benefits of service coordination and integration to be reaped along with production efficiency improvement, is considered instead.

[10] The potential for cost reduction under competitive tendering has already been discussed. In the case of open access, Evans (1990) found reductions in operating costs per bus-km to be the most significant effect of deregulation in the UK.

[11] As Evans observes, note the similarities with Mohring's user economies of scale.

[12] Hensher (1993) extends the concept of natural monopoly and economies of integration with a workable concept of contestability, benchmark contestability.

Pricing Reform

The recent reviews of urban transport (IC, 1994; Commission of Audit, 1994) also argue that there is a need for pricing reform, proposing that the structure of public transport prices be modified in a number of ways : fares should be distance based to reflect the relationship between operator cost and distance[13]; fare differentials should exist between peak and off-peak[14]; and the overall level of fares should be increased to improve the financial performance of the industry.

It is this last issue, the level of fares, which is the prime focus of this study. Is full cost recovery warranted as a long term target, or can an ongoing level of subsidy be justified on economic efficiency grounds ?

2.4 Arguments for Subsidy

Public transport subsidy has previously been justified on a number of grounds (Kerin, 1987).

2.4.1 Non-Efficiency Arguments

The promotion of equity has been the main justification given by governments for subsidy. As discussed above, however, equity is in fact worsened through current subsidisation, with more direct delivery of financial assistance to the needy being necessary. Subsidy has also been justified *inter alia* on grounds that urban transport is a merit good, that subsidy generates macroeconomic benefits through (reduced inflation and multiplier effects), that it encourages energy conservation, that it improves urban form, and as a tool for attempting to arrest declines in public transport usage.

Kerin (1987) and others (e.g. Bly et al, 1980), however, dismiss these arguments as being poorly founded. The merit good argument, that subsidy is good for groups like pensioners and the

[13] Kerin (1990; 1992) shows, however, that when user costs are also considered, a flat fare structure may in fact be optimal, so the current situation in Adelaide where prices do not vary with distance travelled (other than a lower fare for trips < 3.2 km) may be adequate. Kerin does note, however, that the literature also reports proposals for inverse and positive optimal fare-distance relationships (Mohring, 1972; Cervero and Wachs, 1982).

[14] Peak/off-peak differential are absent in a number of Australian cities. In Adelaide, a differential has been used for a considerable period of time, although the scale of the differential has varied over time.

unemployed since it encourages them to 'get out' and mix with people etc., is not only paternalistic, but assumes that government knows better than the individual what is in that individual's own best interest, a debatable assumption. Macroeconomic goals are best attained with the most appropriate instruments. Public transport subsidy is unlikely to be such an instrument. The fact that taxes need to be raised elsewhere in the economy to finance the subsidy means that it is not at all clear whether subsidy has any positive beneficial impact on inflation or employment.

Energy conservation could be improved if subsidy actually led to significant transfer of car users to public transport. In practice, however, reduced public transport fares tend to encourage very few people to switch out of car use, suggesting that conservation policies targeted directly at car use may be more appropriate. Urban form may be improved by subsidies, but more direct land market policies can achieve the same outcomes more effectively. Finally, declining public transport usage has been a common long term trend in much of the developed world in recent decades due to a whole host of reasons, including increased incomes, consumer preference for car travel and urban sprawl of population and activities. It is probably unrealistic to expect public transport subsidies to counter the effects of these strong forces.

2.4.2 Efficiency Arguments

A number of allocative efficiency based arguments have often been advanced in favour of urban public transport subsidies.

Option Value

One argument is that public transport delivers an external benefit to non-users in that it provides non-users the option of using the service if they need it in future, and thus generates an option value (Kain, 1981). Since the operator cannot capture these external benefits, public transport may be underprovided without subsidy. This line of argument seems to rely, however, on the service not being provided at all unless subsidy is available. If, however, a reasonable level of service can in fact be provided without subsidy, then the option value would still exist, especially since it is unlikely that all non-users would want to suddenly commence using the service at the same time. To the extent that an option value actually does exist, its size can be measured using contingent valuation techniques, although lack of data prevents its estimation in Adelaide (Della-Torre, 1994).

Producer Economies of Scale

Economies of scale in *producer* costs (i.e. the costs of service delivery) have been advanced as an argument in support of subsidy. Indeed if economies of scale do exist, and thus marginal costs are below average costs, then subsidy may be justified in a first best setting.[15] However, empirical evidence tends on the whole to mainly imply constant returns to scale (see summary of literature by Small, 1992), although some evidence of increasing returns, or economies of scale, exists for urban rail transport (Travers Morgan (TM), 1988; Nash, 1982) due to significant indivisibilities and fixed costs. In the case of urban bus transport, the general consensus is that constant returns to scale exist (Kain, 1981; Windle, 1988; Evans, 1990; Hensher, 1993), although both diseconomies of scale (Obeng, 1985) and slight economies of scale (TM, 1978, in a study of Adelaide buses) have also been observed. On the whole, the case for bus subsidy due to producer economies of scale appears to be weak.

Second-Best Pricing

Another argument for subsidy is that, where road use is not priced, subsidy can be used as a second-best instrument for managing road congestion (e.g. Sherman, 1971, 1972; Jackson, 1975; Glaister and Lewis, 1978)[16]. When urban roads are unpriced, users face the average user cost. However, urban road use is characterised by congestion and environmental negative externalities, with marginal cost exceeding average cost, and thus the level of road use is above optimal, resulting in road use deadweight efficiency losses. By introducing public transport subsidies, marginal users can be encouraged to switch away from car travel to public transport. Since the road use externality increases at the margin, the mode switching would result in a reduction in road sector efficiency

[15] Where public fund raising has no distortionary effects. Once this assumption is relaxed, the strength of this and other efficiency arguments for subsidy presented here is diminished by the extent to which the raising of public funds to finance the deficit may generate inefficiency costs elsewhere in the economy. This issue is addressed later.

[16] The argument was developed on a stand alone basis by Sherman, and Glaister and Lewis. The model has also formed part of a more comprehensive framework for subsidy analysis containing a number of subsidy arguments (e.g. Dodgson, 1985; Glaister, 1987; TM, 1988; Bly and Oldfield, 1987; Kerin, 1990).

loss. Although the subsidy will at the same time introduce some allocative efficiency losses in the public transport sector due to prices being distorted away from marginal cost, these losses will initially be small, thus allowing a net overall reduction in efficiency losses in the combined roads/public transport sectors. A second best optimum occurs when the rise and fall of the two efficiency losses are equated at the margin.

The road congestion based argument has proved to be particularly popular in subsidy analysis in New Zealand and Australia, including Adelaide, largely due to the work of the consulting firm Travers Morgan Pty Ltd (TM). Using the work of Glaister and Lewis (1978) as a starting point, Travers Morgan have extended the Glaister and Lewis model and applied it extensively : in Adelaide (Amos and Starrs, 1984), New Zealand (TM, 1988), Brisbane (TM, 1991a) and Perth (TM, 1991b). Starrs (1984) also used the framework to analyse optimal subsidy. In the case of Adelaide, Amos and Starrs (1984) reported that, in 1981/82, only 20% of the then actual level of subsidy could be justified on second-best road congestion management grounds.

Although considerable attention has been given to this subsidy argument, it has received considerable criticism, especially in recent times. The objection raised has been that although the argument is sound theoretically, it will only lead to significant subsidies in cases where road congestion is relatively high (e.g. Glaister, 1981; Kerin, 1987). The reason is as follows. First, the cross-price elasticity of car travel with respect to the price of public transport is generally regarded as being very low. Dodgson (1985) indicates that there is general consensus on this point, but that there is limited precise evidence on its actual size. Lewis (1978), Hensher (1986), Hensher and Bullock (1979), and others (see review in IC (1994) Appendix B) each find the cross-price elasticity to be below 0.1, and often well below this figure. As a result, subsidising public transport encourages only a small proportion of car users to switch to public transport. Second, given the low cross-elasticity, congestion levels must be relatively high to generate any substantial reduction in road use deadweight loss.

As a result, whilst, in theory, reductions in road congestion are possible through subsidisation, in practice these reductions will in most circumstances be only modest in size. It will only be in the most congested of cities that road congestion will be severe enough for subsidy to have any substantial impact on reducing road congestion efficiency losses. This view would suggest that, at least in the case of Adelaide where road congestion is currently considered to be relatively modest, this argument for subsidy is likely to be quite weak, with only a small level of subsidy being justified.

In addition, advocating subsidy as an instrument for managing road congestion has been criticised on other grounds (IC, 1994). Not only does it take the focus away from introducing first-best policies (such as proper road pricing), but there may be superior second-best road congestion management policies (e.g. possibly parking surcharges and restrictions, although this approach is also likely to have problems (SA Government, 1993)). Kerin (1992) also points to other possible flaws in the argument, for example, the common neglect of the role of other modes (walking, cycling and car pooling) in managing road congestion, and the possible distortions of urban spatial structure that may result from subsidies.

Overall, one could question the validity of arguing for subsidy as a second-best instrument to manage road congestion, especially in a low congestion city like Adelaide.

User Economies of Scale, UES

A further efficiency argument in favour of subsidies is based on the existence of user economies of scale, *UES*, briefly introduced in chapter 1 as the focus of this study. Economies of scale exist when average cost (*AC*) declines with increases in patronage, and thus marginal cost (*MC*) is smaller than *AC*. Efficient *MC* pricing will then result in prices being below *AC*, thus generating a financial deficit and the need for subsidy (qualified by footnote 1 in chapter 1). As discussed above, public transport, and particularly bus transport, tends to be characterised by constant returns to scale in producer costs. Notwithstanding this, economies of scale exist in frequency related user costs, i.e. *user economies of scale, UES*. These scale economies arise because in order to cater for higher levels of demand, service frequency must increase, thus reducing the time costs incurred by all previously existing users of the service[17]. As a result, marginal user cost will be below average user cost, marginal social cost will in turn be below average total cost, and efficient *MC* pricing will lead to a deficit and the need for subsidy.

[17] These time costs can take the form of either waiting time, e.g. at a bus stop, or the inconvenience cost of delays to users caused by service departure times not matching the desired departure times of users.

2.5 The Significance of User Economies of Scale

Mohring : A Case for Significant Subsidy

The relative importance of the user economies of scale argument, and the size of the subsidies which are justified on these grounds, has been debated for a couple of decades since the seminal work of Mohring (1972) in which the notion of increasing returns in user cost, and the resulting first best case for optimal subsidy was initially introduced. Mohring suggested that previous analysis of urban bus transport had made a serious oversight in failing to understand the true role of frequency related user costs in optimal outcomes. Pointing to the existence of economies of scale in user costs, Mohring estimated, through a series of simulation runs using bus system data from Minneapolis in the United States, that first best optimal subsidies of up to 60% of bus costs could be justified.

Since Mohring's original paper, a number of contributions have extended the Mohring user economies of scale framework, finding support for significant optimal subsidy.[18] Turvey and Mohring (1975) explored in a general way the influence of passenger congestion effects, such as the impact of boarding and alighting. Their key conclusion was that optimal fares should rise the fuller buses become, because of the greater impact of boarding and alighting and the greater chance of users being delayed because buses arrive full, thus dampening the scale of optimal subsidy.

Jansson, J.O. (1979) illustrated how the *UES* characteristics previously observed in urban buses applied to some degree or form in all scheduled passenger and freight transport services, and proceeded to present a generalised analysis of *UES* for the general transport category, scheduled transport services. Jansson's main conclusion was that :

[18] Reference should be made at this point to a number of studies in which user economies of scale have played a key role, sometimes in combination with the road congestion reduction argument for subsidy, but in which the optimal level of subsidy was *not* addressed (e.g. Glaister, 1982; Glaister, 1984; Dodgson, 1985; Glaister, 1987; Bly and Oldfield, 1987). The focus in these studies has alternatively been on determining whether there are net benefits (or disbenefits) from expanding subsidy from its existing level, and to address the question of whether dollars of subsidy are being efficiently allocated, at the margin, between fare reduction and frequency enhancement, and between competing cities. Whilst this has been an important area of analysis and research, it is not addressed directly in this study given its lack of focus on optimal subsidy.

> *" ... the coexistence of vehicle size economies in producer costs, and vehicle number [frequency] economies in user costs, can, regardless of the mode of transport, 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."*[19,20]
>
> Jansson (1979), p. 288

Larsen (1983) and Else (1985) were able to express optimal user economies of scale subsidy in terms of price and service elasticities, and noting typical values for these, suggest that, in the optimum, fares would cover between a third and a half of the costs of service provision. Nash (1988) finds from a number of simulations that

> *"... massive economies of scale ... are apparent ... even when produced under conditions of constant returns to scale [in producer costs] ..."*
>
> (Nash, pp. 104 & 109)

with fares covering between 25% and 50% of operating costs in a first-best setting.

There are two notable exceptions of commentators who claim that Mohring's user economies of scale phenomenon leads to optimal subsidies which are much more modest in size than the literature tends to suggest. Both challenges argue that conventional analyses of the bus optimisation problem are limited in either scope, or through the assumptions which underlie them, and that it is these constraints which are responsible for producing large optimal subsidy outcomes.

[19] The concurrent existence of any producer economies of scale increases the magnitude of justified subsidy.

[20] Jansson's analysis is general, looking at a general scheduled transport service. Whilst the user cost economies are a function of waiting time for passenger transport, in freight transport they are a function of storage time. Jansson observes that passenger waiting times are inversely proportional to service frequency (assuming random user arrival at loading points), whilst freight storage time are inversely proportional to the square root of frequency (Baumol and Vinod, 1970). Thus, the impact (in time units) of infrequency is greater in passenger transport than freight, although planned user behaviour in passenger transport would tend to diminish this difference. Jansson then notes (p. 288) that once values of time savings are accounted for, given that time savings are valued many times higher in passenger transport than freight transport, infrequency is therefore much more costly in passenger transport than freight. This leads Jansson to conclude that whilst scheduled passenger transport is a pronounced decreasing-cost industry resulting in a relatively large financial deficit in the optimum, with respect to freight transport, on average, the decreasing-cost characteristic (and thus by implication the justified *UES* subsidy) will be less pronounced, with variation between low and high valued goods.

A parallel literature has also developed in the case of scheduled airline services, where user costs, user economies of scale, and the corresponding argument for subsidy forms an integral part of optimal analysis (Douglas and Miller, 1974: DeNeufville and Mira, 1974; DeVany, 1975; Forsyth and Hocking, 1978; Panzar, 1979; Forsyth, 1983; Findlay, 1983).

The Walters Critique

The first of these challenges came from Walters (1982). Walters, pointing to the key role played by minibuses in developing countries, argued that Mohring's results (and by implication those of much of the subsequent literature discussed above) of large first-best bus subsidies relied on two assumptions : monopoly provision of bus services, and a fixed conventional (large) bus size. Relaxing these assumptions, Walters argued that more frequent services with smaller buses would result. With frequency now substantially greater (for any given level of usage), user costs would be smaller, and thus the gap between average and marginal user costs (the driving force behind *UES* subsidy) and the associated justification for subsidy would be "small and probably trivially small" (p.72).

Although the Walters critique was shown to be biased by an analytical error which led to the odd result that optimal bus size was inversely proportional to patronage (Gwilliam *et al*, 1985), the critique did have an impact. Mohring (1983) acknowledged the restrictive nature of his implicit monopoly and bus size assumptions, and undertook further simulations with the assumptions relaxed. Mohring concluded that, in a first best setting, user economies of scale would justify considerably smaller subsidies for minibuses than for standard buses, whilst in a second-best world of unpriced roads, Mohring found that more substantial optimal subsidies were justified for both minibuses and conventional buses.

These results are consistent with the earlier work by Jansson (1980) which investigated, for a simple bus line model, the trade-off between bus capacity costs and user costs to achieve a total cost (producer plus user) minimisation outcome for any given patronage level. Jansson found that the resulting bus service would look markedly different from typical bus services : service frequency would be greater and bus size would be smaller, particularly on low demand routes such as off-peak services (a point made evident by Waters' (1982a) diagrammatic exposition).

Kerin (1992) also acknowledges that *if* optimal bus size is smaller than conventional bus size, then user economies of scale and the associated subsidy would also be smaller. Kerin questions, however, whether optimal bus size is in fact smaller than conventional bus size, pointing to the lack of consensus in the literature. Oldfield and Bly (1988) find for typical urban conditions in the United Kingdom that optimal bus size for a monopoly bus service lies between 55 and 65 seats, and that even if reductions in unit operating costs could be achieved, bus size would be unlikely to drop below 40 seats. Nash (1988) presents a unique analysis in that in addition to optimising service frequency and bus size, the number (and spacing) of routes in the system is also optimised. Nash finds that, even up to high patronage levels, optimal bus size is below conventional bus size, and that allowing the number of routes to be optimised lowers the optimal bus size below that found in a single route model such as Jansson (1980). Kerin (1990) also investigates optimal bus size using bus route data for Adelaide. Kerin finds that for regular bus services that stop along the full length of the route, optimal bus size is not far off the average bus size in Adelaide of around 50 seats, and in some circumstances even bigger, a result which is consistent with that of Oldfield and Bly for the United Kingdom. Kerin also found that when the system consists of a mixture of all-stopping and express buses, smaller buses do have a greater role to play in serving low density outer areas and on short-haul routes.[21]

The conclusion that can be drawn is that there is no firm consensus that conventional buses are excessively large, nor what the optimal size of urban buses is, although in the right circumstances, smaller buses may be desirable. Given this, it is also unclear to what extent the Walters critique, that buses should be smaller thus making user economies of scale subsidy small, is a valid and lasting objection.

The Kerin Critique

A second, and more substantial, challenge to the significance of *UES* subsidy has come from Kerin (1990, 1992). Kerin identifies a number of problems with the UES subsidy argument, drawing the conclusion that optimal subsidy is far less substantial than the literature would tend to suggest. First, Kerin points to the fact that most analyses ignore (with the exception of Mohring (1979, 1983)) the fact that, although increasing service frequency may reduce average waiting times, the resulting increase in the number of buses on the road also adds to road congestion,

[21] As predicted by Glaister (1986) and others, minibuses have played a significant role in recent years in the United Kingdom following urban bus deregulation in 1985. As Glaister points out, however, bus size in a competitive market would be below that obtained from a social optimisation which accounts for all costs, including those costs outside the operators concerns such as road congestion costs and user time costs.

especially in the peak, thus increasing in-vehicle time of both bus users and road users. These scale *dis*economies may offset any economies in waiting time, thus eliminating the argument for subsidy (as is the case in some of the simulations in Mohring 1979, 1983).

Kerin's second objection is that most analyses model waiting time, a critical analytical input in *UES* subsidy analysis, on the assumption that users arrive at loading points (e.g. a bus stop) in a random fashion, failing to consult a timetable. However, empirical evidence (e.g. Seddon and Day, 1974) suggests that, other than at quite high service frequencies, this assumption is unlikely to hold, with waiting time, and the associated *UES* subsidy thus being overestimated. This has been confirmed in other recent work (Tisato, 1992; Jansson, K., 1993) where the choice of assumption about whether users arrive at a loading point in a random fashion, or in a planned[22] fashion coordinated with scheduled departure times, has been found to make a substantial differences to *UES* subsidy outcomes.

A further major objection levelled by Kerin is that most studies ignore the fact that, in reality, raising public funds to finance public transport subsidy is not costless. Besides administration costs, public fund raising causes efficiency losses elsewhere in the economy as taxation distorts the choices of economic agents away from preferred outcomes (Browning, 1976; Stuart, 1984; Dodgson and Topham, 1987; Freebairn, 1995). Findlay and Jones (1982) estimate that for every $1 of public finance raised in Australia, the resulting marginal efficiency loss is in the range $0.23 to $0.65, with Freebairn deriving potential values as high as $0.73. Acknowledging these efficiency costs in bus optimisation results in lower optimal subsidy levels.

Kerin also argues that the subsidy literature tends to ignore external costs generated by public transport, such as road damage and pollution, although the difficulties in quantifying these effects makes it difficult to estimate the impact of their inclusion. Further, as discussed in section 2.2, subsidy may foster production inefficiency and a subsequent leakage into costs.

Kerin (who also provides a stinging critique of the second-best road congestion reduction argument for subsidy) concludes that it is difficult to draw firm generalised conclusions about optimal bus fares, and by implication bus subsidies, pointing to the strong influence on outcomes of

[22] Using the terminology of Tisato (1991).

the various assumptions that have been made in analyses in the literature. Notwithstanding this, Kerin (1992) draws the tentative conclusion that, at least during peak periods (the focus of his analysis), there is

> *"little case for subsidies on first-best* [*UES*] *grounds"*

whilst there is

> *"probably a case for limited subsidies on second-best* [*road congestion reduction*] *grounds (although these would be much smaller than mathematical models have typically estimated) ...".*

2.6 Scope For Research in This Study

Kerin's critique is an important one, bringing together a considerable literature, and highlighting the importance of clearly stating the assumptions that underlie subsidy analysis. Whilst many of his objections appear to be valid, his tentative conclusion, that there is only a limited case for subsidies, needs some qualification.

First, Kerin's conclusion partly relies on the link between subsidy and production inefficiency. The point that a monopoly bus service supplier operating with the support of subsidy may result in subsidy leakage into costs does have support. However, in the case of Adelaide, the case study in the current investigation, the recent government reform in bus service delivery should overcome Kerin's productive efficiency concerns. The planned introduction of competitive tendering in the delivery of bus services in Adelaide is expected to provide a renewed and commercial focus for bus service providers. In this framework, the threat of losing the right to service a route or an area on contract termination should provide the incentive for firms to deliver bus services efficiently, even if, as is likely in the foreseeable future, the bus system continues to run at a deficit supported by subsidy.

Second, with Kerin's analysis focused on the peak period, an important question is whether Kerin's conclusion is robust in the off-peak, in which much of the existing subsidy occurs. With respect to the road congestion argument for subsidy, if, as Kerin finds, the argument for peak subsidy is modest at best, then justification in the off-peak will obviously be even less substantial given the lower off-peak road congestion levels. Kerin's conclusion with respect to the second best road congestion reduction argument therefore seems robust. However, with respect to the user economies of scale argument for subsidy, it is less clear that Kerin's conclusion is robust. This is the case since an important, and possibly understated, result in the literature (Jansson, J.O., 1980; Waters, 1982a; Gwilliam *et al*, 1985; Nash, 1988; Jansson, K., 1993), is that the user economies of scale effect is stronger (on a unit subsidy basis) the thinner, or less patronised, the route, i.e. there is a negative relationship between unit subsidy and route patronage level. Thus user economies of scale is likely to play an important role on feeder routes, routes to low density areas, and off-peak routes (Waters, 1982a; Jansson, K.,1993). It is therefore unclear whether Kerin's conclusion for high demand peak routes carries over to lower demand off-peak routes. There is some justification therefore for further research to test Kerin's conclusion in the off-peak.

There is also justification for further focus on user cost modelling, a key input into subsidy analysis (Tisato, 1992). Although Kerin (1990; 1992), Tisato (1990; 1991) and Jansson, K., (1993) incorporate user cost models in their subsidy analyses which are an improvement on the conventional simple random user behaviour model, objections can be raised about the models used in each case. The model used by Kerin, the empirical relationship estimated by Seddon and Day (1974) for Manchester, seems somewhat arbitrary for use as a general model. The models of Tisato and Jansson are superior in this respect, in that they predict user cost as an outcome of a cost minimising bi-modal choice process (between random and planned behaviour), the nature of which can vary between different situations. Whilst these recent bi-modal choice models have increased the realism of user cost modelling, they characterise the choice in very simplistic deterministic manner.

An implication of this is that the deterministic choice models predict some very sudden, or knife-edge, behavioural changes within the population of users. This in turn leads to some interesting outcomes (Jansson, 1993; Tisato, 1990) : multiple local optima can occur in the bus optimisation problem (one optimum each for random and planned user behaviour); and there can be a sudden increase in optimal unit subsidy when mode switching occurs, which in turn may distort the conventional negative relationship between optimal unit subsidy and patronage level. These outcomes create difficulties. First, the existence of multiple solutions adds an extra level of complication for the analyst of the bus optimisation problem, requiring a distinction between local optima and the global optimum. Second, the conventional unit subsidy/patronage relationship has provided a simple and important rule of thumb for describing how subsidy per trip varies between bus routes of different demand density, and has been a useful mechanism for explaining a key policy outcome of *UES*. Thus its potential demise is of concern.

Given these concerns, there is some justification for investigating the extent to which Jansson's results are a product of using a simple deterministic framework to predict random vs planned choice, and whether the results still occur even if a more appropriate choice model is used. Developments in the literature in recent decades (Beesley and Kemp, 1987; Small, 1992) suggest that a probabilistic choice model may be a more fruitful way of modelling choice between random and planned user behaviour. One such model, the logit model, will be used in this study.

There is also justification for estimating *UES* subsidy for Adelaide public transport. The user economies of scale argument for subsidy has received quite limited attention in Australia. A few studies have considered *UES*. The most important of these was by Dodgson (1985; 1986) who analysed subsidy in the five major Australian cities, including Adelaide. Whilst an important piece of work, Dodgson did not address the question of optimal subsidy levels. His focus was instead, like that of Glaister (1987) and others in the United Kingdom (see footnote 18), on the optimal use of a given (existing) subsidy.[23] In addition, the study also jointly considered the road congestion reduction argument for subsidy, and was undertaken at a highly aggregated level, considering daily average conditions across the whole network, and thus did not consider the important relationship between subsidy and demand level. A more disaggregated analysis, which focused exclusively on user economies of scale was undertaken by Chalmers (1990). Chalmers determined optimal subsidies for a couple of representative bus and train services in Adelaide. Unfortunately, a serious problem with this work was the use of the simple assumption of random user arrivals at loading points. The other work related to Adelaide was that by Kerin discussed above, in which Adelaide data was used for model calibration, although the analysis was undertaken at a fairly general level.

There is a genuine lack, therefore, of a study of Adelaide (and other Australian cities) which adequately estimates optimal *UES* subsidy. Such a study will be undertaken here (for buses only),

[23] Hensher (1989a) used the same framework to assess the effectiveness of the use of subsidy in Sydney.

with a focus on improved user cost modelling, analysis at a disaggregated level to allow the relationship between subsidy and patronage level to be observed (and the robustness of Kerin's peak subsidy conclusion in the off-peak to be tested), and which acknowledges that public fund raising can be costly.

2.7 Summary

Overall, a number of conclusions can be drawn from the review in this chapter. First, public transport subsidy remains an important issue in Australian cities, including Adelaide, partly due to public finance from which subsidies are financed becoming relatively scarcer, placing pressure on the reduction of subsidies. Second, in the short term, the introduction of competitive tendering is expected to reduce production inefficiency in bus service provision, reducing the cost of bus services and allowing subsidy to be reduced. Third, in the longer term, the question remains of how subsidy policy should evolve. Without doubt, attention is required to ensure, whatever the level of subsidy, that it is delivered to those who most need the financial support it provides (contrary to the current situation). In addition, an equally important question is the level of subsidy which can be justified on economic efficiency grounds. In this respect, the user economies of scale argument for subsidy is an important argument, yet has received little attention in Australia. A review of the literature has revealed a continuing debate about the scale of optimal user economies of scale subsidies.

Analysis of user economies of scale in Adelaide therefore appears justified, and forms the focus of this study. In doing so, the analysis should focus on improving on some of the areas where the literature to date has weaknesses or is still unresolved. The above discussion suggests a number of prime areas for analysis : the relationship between unit subsidy and route patronage, the modelling of user cost, and estimation of the level of subsidy justified for Adelaide. The following chapter outline concludes this overview by setting out in more detail the nature of the subsequent chapters.

2.8 Chapter Outline

The following chapter outline provides an indication of the focus of each chapter of this study. A brief overall summary is provided in Table 2.1.

Chapter 3: *User Cost Modelling*. This chapter aims to develop an improved working model of user cost, for use in subsequent analytical chapters, which overcomes a number of deficiencies in current models. The improved model consists of a refined set of user cost definitions, an expanded set of user cost components, the recognition of multiple modes of user behaviour (random and planned), and a theoretical logit model for predicting user choice between these behavioural modes.

Chapter 4 : Optimal Pricing, Frequency and Subsidy Formulation. The aim of this chapter is to set out and solve the bus optimisation problem from which the *UES* subsidy argument arises. This provides a sound theoretical foundation for the estimation of optimal outcomes in subsequent chapters. The chapter also better integrates and relates previously used optimisation frameworks, taking a taxonomic approach with respect to load factors and bus sizes, presenting improved diagrammatic presentations of *UES* using envelope curves to relate short run and long run analysis, and clarifying previous presentations.

Chapter 5 : Optimal Subsidy and Cross-Subsidy with a Logit Model. In this chapter, the impact on optimal unit subsidy, total subsidy and cross-subsidy[24] of using the logit choice model developed in chapter 3 is investigated. In particular, the chapter tests whether the important results of Jansson's (1993) recent work (multiple local optima, and sudden increases in optimal unit subsidy) persist when a logit choice model is used.

Chapter 6 : Service Unreliability and Subsidy. The aim of this chapter is to explore the link between *UES* subsidy and service unreliability, a key determinant of user cost. The literature on *UES* subsidy has virtually ignored service unreliability as a determinant of user cost, and thus its link ́to subsidy. This link is analysed for random, planned and logit user behavioural models, followed by an exploration of the impact of changes in unreliability on optimal subsidy.

[24] The inverse relationship between optimal unit subsidy and route demand level also leads to a related argument in favour of optimal cross-subsidy between low and high demand routes (e.g. Gwilliam *et al*, 1985).

Chapter 7 : User Economies of Scale Subsidy in Adelaide. The final chapter develops estimates of optimal *UES* subsidy for Adelaide buses, and considers the implications and problems of moving to such outcomes. The analysis uses the improved logit user cost model developed in chapter 3, and builds on the subsidy formulations considered in chapter 4. The analysis is undertaken at a disaggregated bus corridor level for peak and off-peak periods, with recognition of current concession fare policy, the pending introduction of competitive tendering, and the efficiency costs of public fund raising. The chapter also assesses the off-peak robustness of Kerin's (1992) claim that *UES* subsidy is small, draws comparisons with previous *UES* subsidy work in Adelaide, and provides an assessment of the gap between current and optimal subsidy outcomes in both first best and second best (costly public fund raising) settings.

Table 2.1 : Summary Chapter Outline

Chapter	*Main Issues/Focus*
1	general introduction
2	background, context and scope of study
3	develop an improved user cost model
4	optimal *UES* pricing, frequency and subsidy formulation for a range of load factor/bus size cases
5	impact of using a logit user behavioural mode choice model on optimal unit subsidy, total subsidy and cross subsidy
6	the relationship between service unreliability, user cost and optimal subsidy
7	estimation of optimal *UES* subsidy for Adelaide buses using a disaggregated analysis focusing on bus corridors in peak and off-peak periods
8	overall summary and conclusions from the study

Chapter 3
USER COST MODELLING

3.1 Introduction

Chapter 2 established the fact that frequency related user costs play a central role in the economies of scale (*UES*) concept and the related argument for subsidy. The significance of this role has recently been demonstrated in the works of Tisato (1992) and Jansson, K. (1993), in which the conventional assumption, that users arrive at a public transport loading point in a random fashion, was relaxed. Tisato found that, as a result, optimal subsidy fell by as much as 60%.[1] Jansson found that relaxing the random user assumption generated a number of important new results, including a potential reversal of conventional wisdom about the relationship between unit subsidy and demand level.

The aim of this chapter is to develop a manageable working model of user cost for use in subsidy analysis in this study. The recent model developments of Tisato and Jansson are taken as the starting points for this task, with the intention to further extend and improve their model. The chapter :

- develops an expanded and more consistent set of user cost definitions;
- reviews conventional models, and more importantly, the recent developments by Tisato and Jansson;
- establishes a working model for each of several components of user cost for two modes of user behaviour, random and planned;

[1] This does not imply the conventional simple random behaviour user cost models has not played a useful role. The model has facilitated important subsidy analyses which have produced many important general results.

- brings these user cost components together into a unified framework in which the mode of user behaviour (random vs planned) is predicted as the outcome of a probabilistic logit choice model, and the associated user cost estimated.

3.2 The Concepts of User Cost and Rescheduling Cost

In its broadest sense the concept of user cost includes any monetary or non-monetary cost/disutility incurred by users of transport. However, the term most generally tends to be used to refer to non-monetary costs only, and this is the interpretation adopted here. The most prominent components of user cost tend to be time-related user costs. These arise from travel technologies imposing time constraints on a user's time allocation problem, the solution of which is the realm of the economic theory of time allocation (Becker, 1965; DeSerpa, 1971; Bruzelius, 1979; MVA, 1987).

That theory indicates that, if a technological time constraint is binding, the user has to undertake a non-preferred, or "intermediate" (DeSerpa, 1971), activity, e.g. travelling in a train, waiting at a bus stop, being delayed at home waiting for a bus to depart, etc). At the same time, the user must give up a preferred non-binding activity, also called "pure leisure" (DeSerpa, 1971). More precisely, time is "transferred" from pure leisure to intermediate activities (Truong and Hensher, 1985).

Technological transport time constraints therefore lead to activity rescheduling, with a preferred activity being lost and a non-preferred activity being gained, thus generating an inferior pattern of activities. The resulting loss of utility, or gain in disutility, can be called an (activity) rescheduling cost, or user cost. Throughout this study, the term "rescheduling cost" will be used as a *generic* concept for referring to the general category of time related transport user costs. More specific terms for individual types or components of user cost are discussed in this chapter.

In summary, a transport user cost is a(n) "(activity) rescheduling cost" and is defined as follows : the cost of rescheduling activities away from a superior activity pattern (which would be achievable in the absence of transport time constraints) to an inferior one resulting from binding transport time constraints.

3.3 Frequency-Related Delay and User Cost Definitions

A range of types of user cost has been previously used and modelled in the context of urban public transport. In this study, with the focus on *UES*, it is those user cost components which are *frequency-related* which play a leading role. This chapter focuses on modelling these frequency-related user costs. Other user costs are treated exogenously (- these consist of in-vehicle travel time and walk time)[2], and are discussed in appendix B.

Frequency related user cost, and its components, have not always been defined and used in a consistent manner in the transport literature. This can be a source of confusion, and so, before proceeding to model selection and development, it is worth spending some time reviewing the past use of user cost components and definitions, and establishing a clear set of definitions for use here.

3.3.1 Review

In the literature, frequency-related user cost has been defined as consisting of one or more of the following components : waiting time, and schedule delay[3] (which in turn has two components: frequency delay and stochastic delay, both of which will be defined shortly).[4]

Waiting time, which has been very commonly used in models of user cost, has tended to refer to time spent at a loading point waiting for a service to arrive. This seems an unambiguous definition and will continue to apply here.

[2] Travel time and walk time are of course also important variables in the context of comprehensive optimisation of the transport system. The latter, for example, is relevant to the question of optimal route spacing. Although the focus here will be on optimal frequency for a *given* route density, a more comprehensive optimisation would require jointly optimising both frequency and route spacing (Nash, 1988).

[3] For the moment, the term "schedule delay" can be interpreted by its usual meaning in the public transport literature, namely the delay arising from scheduling of services at non-zero time intervals. Shortly, however, it will be argued that this term could more usefully be exclusively reserved for an alternative meaning.

[4] Note that waiting time and service delay are times concepts measured in *physical* units (minutes). User cost, on the other hand, is a *monetary* concept reflecting the valuation/costing of the physical time units. Physical terms like waiting time and delay will, however, be loosely used at times to refer to the associated user costs which they generate.

The concept of *schedule delay*, on the other hand, was first introduced by Douglas and Miller (1974) in the context of air travel, and it, and/or its two components, have been widely used in the air transport literature (e.g. Forsyth and Hocking, 1978; Panzar, 1979; and Findlay, 1983; Smith and Street, 1992). In broad terms, schedule delay refers to the fact that, due to the scheduling of services at non-zero time intervals, a user will in general *not* be able to board public transport at his/her preferred time, denoted t_p. As a result s/he suffers a time "delay", requiring the user's activity pattern to be modified, thus generating a rescheduling cost.

Schedule delay has two components. The first, *frequency delay*, is the delay generated by the fact that *scheduled* service departure times may not coincide with, t_p, the preferred time at which the user would like a service to depart. The user will catch the least inconvenient scheduled service, with the resulting frequency delay usually measured as the time difference between t_p and this least inconvenient scheduled departure time.[5]

The second component of schedule delay, *stochastic delay*, is the additional delay caused by the fact that, due to user demand being stochastic in nature, the least inconvenient scheduled service may arrive full, resulting in the user having to wait for the next service to arrive.[6]

The majority of the urban public transport subsidy literature has mainly tended to model frequency-related user cost as consisting of waiting time (Mohring, 1972; Dodgson, 1985; Glaister, 1987; Travers Morgan, 1988). In addition, the idea of stochastic delay has also been used (although without this term being explicitly used) when waiting time has been modelled as a function of load factor (e.g. Glaister, 1987; Bly and Oldfield, 1987). More recently, the frequency delay concept has played an increasing role (Evans, 1987; Tisato, 1991; Jansson, K., 1993).

The use of these user cost terms and components has not, however, always been consistent. Waters (1982a) refers to the notion of frequency delay as used in the air transport literature and

[5] It is worth noting that Douglas and Miller originally defined it as the time gap to the *nearest* scheduled departure. Forsyth and Hocking (1978) criticise this, arguing that the nearest departure may not be the most convenient for the user, and would therefore not be chosen.

[6] In the case of urban public transport, the user has no choice but to catch the next service that arrives. In the case of air travel, however, with the facility to purchase a ticket in advance of flight time, the user could catch either an earlier or later flight.

denotes the urban transport equivalent as waiting time, and refers to stochastic delay as a queuing cost. Evans (1987) uses the concept of frequency delay but refers to it as rescheduling cost. Whilst this is not incorrect, it does conflict with the use of rescheduling cost as a generic term as discussed earlier. Tisato (1991), distinguishes between waiting time and frequency delay, with the first referring to time spent at a loading point, and the latter referring to time spent in rescheduled activities away from the loading point arising from non-preferred scheduling of services. Jansson (1993) uses the term frequency delay to cover all time spent in non-preferred activities due to the scheduling of services, irrespective of where these activities take place.

3.3.2 Definitions Adopted for this Study

A refined set of definitions, presented below, are adopted in this study. Hopefully, these will avoid some of the inconsistencies discussed above, and provide a clear set of definitions for this study, and future work in this area.

(a) *Activity rescheduling and schedule delay*

Problems can arise when using the term schedule delay as defined by Douglas and Miller. In their use of the term, the word schedule refers to the scheduling of transport services. This conflicts, however, with the use of the term "schedule" as used in rescheduling cost which I suggested earlier (section 3.2) should focus on the scheduling of activities since this underlies the whole concept of user time costs. In addition, the Douglas and Miller approach conflicts with use of the term schedule delay elsewhere. In particular, the term schedule delay has also been used to refer to activity scheduling changes arising out of activity time constraints. For example, Small (1982) considers the schedule delay of arriving to work late when fixed work hours exist. A similar concept applies when considering the introduction of flexible or staggered working hours (e.g. Henderson, 1981).

I conclude that using the term schedule delay to refer to the delay resulting from *activity* rescheduling in general may be more fruitful, at least in the urban context, than using it to refer to delay caused by service scheduling. This is the approach adopted here. Schedule delay will therefore be used as a *generic* term referring to the time spent in non-optimal activities due to binding transport and activity constraints. The existence of such delays generate the *generic* rescheduling costs discussed in section 3.2.

(b) *Service delay and waiting time*

The service scheduling delay concept which Douglas and Miller describe in their use of the term schedule delay continues, however, to be a critical one. I propose the new term *service delay* to describe this Douglas and Miller delay concept, thus service delay comprises the two components frequency delay and stochastic delay.

Service delay, through its two components, fully describes the *source* of frequency-related delay (and thus user cost). As a result, the term frequency-related user cost and service delay cost are *synonymous*. Waiting time (as defined in section 3.3.1), on the other hand, is a term which focuses on the *location* of delay time. As will become clear below, stochastic delay is always spent waiting at a loading point (i.e. waiting time), and is frequency delay in some circumstances (i.e. when users arrive at a loading point in a random fashion). Waiting time is therefore *not* a separate component of user cost, but rather it is a *subset* of service delay. The distinction between waiting time and delay time spent at other locations becomes crucial, however, when assigning unit values to frequency delay and stochastic delay (see section 3.6).

(c) *Frequency delay and stochastic delay*

The definition of frequency delay given in section 3.3.1 will continue to apply.

With respect to stochastic delay, a sound argument can be made for adopting a broader definition. In practice, scheduled services are nearly always unreliable (to varying degrees). Service unreliability creates extra delays for passengers over and above situations where services run on time. This service unreliability delay is caused by the stochastic nature of service departure times, and so it would be convenient and relevant to refer to this delay as a type of stochastic delay. Unfortunately, however, this term is already currently used for the case of delay caused by the stochastic nature of user demand.

To overcome this definitional problem, the definition of stochastic delay is extended here to cover both types of stochastic-related delay. Delays caused by the stochasticity of user demand will be referred to as "stochastic demand delay" (*SDD*), and delays caused by stochasticity of service reliability as "stochastic supply delay" (*SSD*). Both of these delays, *SDD* and *SSD*, will result in time spent at the bus stop waiting for a service to arrive, although for quite different reasons.[7] The explicit recognition of service unreliability is something rarely done in existing models[8], and is one of the features of this study.

In summary, frequency-related delay is captured by the concept of service delay (*D*), which has three components : frequency delay (*FD*), stochastic demand delay (*SDD*), and stochastic supply delay (*SSD*) :

$$D = FD + SDD + SSD \qquad (3.1)$$

Frequency-related user cost, denoted *UC*, is then derived by costing each of these delay components. Thus :

$$UC = UC\,(D\,(FD,\ SDD,\ SSD),\ \text{unit costings of } FD,\ SDD \text{ and } SSD) \qquad (3.2)$$

Unit costings are discussed in sections 3.5 and 3.6.

3.4 Bi-Modal User Behaviour and Choice : A Recent Development

A strong assumption that has underpinned the user cost models used in many subsidy analyses to date, particularly early studies (Mohring, 1972; Turvey and Mohring, 1975; Vickrey, 1980; Chalmers,`1990), has been the notion that users arrive at a loading point in a *random* manner, i.e the relationship between scheduled service departure time and user arrival times is a random one. Empirical evidence suggests (e.g. Holroyd and Scraggs, 1966; Seddon and Day, 1974; Jolliffe and Hutchinson, 1975; Turnquist, 1978; Bowman and Turnquist, 1981), however, that such an assumption breaks down in many circumstances. The evidence tends to support random behaviour

[7] It is interesting to note that, in the case of air travel, service unreliability may result in delay time spent away from, rather than at, the loading point.

[8] Tisato (1990; 1991) being the exception, where user cost was a function of service unreliability, although a formal distinction between *SSD* and *SDD* was not drawn, in fact *SDD* was not modelled. The need to model service unreliability was also previously noted by Douglas and Miller (1974, p.7, footnote 7).

at low headways[9] only. At higher headways, outcomes are more consistent with the hypothesis that users coordinate their arrival time with scheduled departure times, behaviour which will be referred to here as *planned* behaviour.

Several recent subsidy studies (e.g. Dodgson, 1985; Travers Morgan, 1988; Kerin, 1990) have relaxed the random users assumption, using instead user cost (- waiting time -) models fitted to observed data. One such relationship, which has been widely used, is that developed by Seddon and Day (1974) which fits a non-linear relationship to waiting time data from Manchester, England, for a range of headways. The user costs predicted by the model are consistent with random behaviour for headways below 10 to 12 minutes, but for greater headways it predicts user costs which are increasingly lower than those predicted by assuming random arrivals.

Two recent contributions by Tisato (1990; 1991) and Jansson (1993), also develop models in which random user behaviour only occurs in some circumstances. The advantage of these recent models compared to the Seddon and Day model is that they begin to explain user behavioural mode observations (random vs planned user behaviour) as outcomes of an optimised *choice* process by users. The Tisato and Jansson models are very similar (and are thus referred to as the Tisato/Jansson model from hereon), with behavioural mode choice outcomes being governed by the principle of user cost minimisation.

The key aspects of the Jansson/Tisato user cost minimisation modelling approach are as follows. Users have a choice of behaving in either a *random* or *planned* manner with respect to the time they arrive at a loading point. For *random* behaviour, as outlined above, the user is unaware of scheduled service departure times, and therefore user arrival at the loading point is not coordinated with scheduled departure times. On the other hand, under *planned* behaviour, users acquire information about departure times from a timetable (at a cost denoted the information cost, I, the costs associated with obtaining, carrying and studying a timetable)[10], and are thus able to coordinate their arrival time at the loading point with scheduled service departure times.

[9] Headway (H) is the time (minutes) between scheduled services, with $H = 60/F$ where F is service frequency per hour.

[10] See section B.1.4 of appendix B for further discussion of I.

For both behavioural modes, the actual size of user cost experienced by the user will vary from one situation to the next. It is always the *expected* user costs,[11] however, which the user compares when choosing between modes. The user therefore faces a discrete binary choice under uncertainty. Users are assumed to be utility maximisers, or in this case user cost minimisers, and are assumed to be risk-neutral[12]. They therefore choose the behavioural mode with the lower user cost, thus minimising user cost.

The choice situation facing the user in the Tisato/Jansson model is illustrated in Figure 3.1. The figure plots the expected user cost experienced by the user under random and planned behaviour, denoted UC_r and UC_p respectively.[13] In the model, delays tend to zero as headway (H) tends to zero. Thus $UC_r = 0$ when $H = 0$, but under planned behaviour, users always incur the information cost, I, irrespective of H, thus $UC_p = I$ at $H = 0$. As H grows above zero, a positive relationship exists between UC and H under both behavioural modes (i.e. the slope of the UC vs H schedule is positive for both modes), however as H increases, UC_r grows more rapidly than UC_p (i.e. the UC_p schedule is flatter than the UC_r schedule). If the headway at which UC schedules intersect, and where $UC_r = UC_p$, is then denoted as the critical headway, H_c, then the user cost minimising user will :

- be indifferent between modes when $H = H_c$ since $UC_r = UC_p$
- will choose random behaviour when $H < H_c$ since $UC_r < UC_p$
- will choose planned behaviour when $H > H_c$ since $UC_p < UC_r$

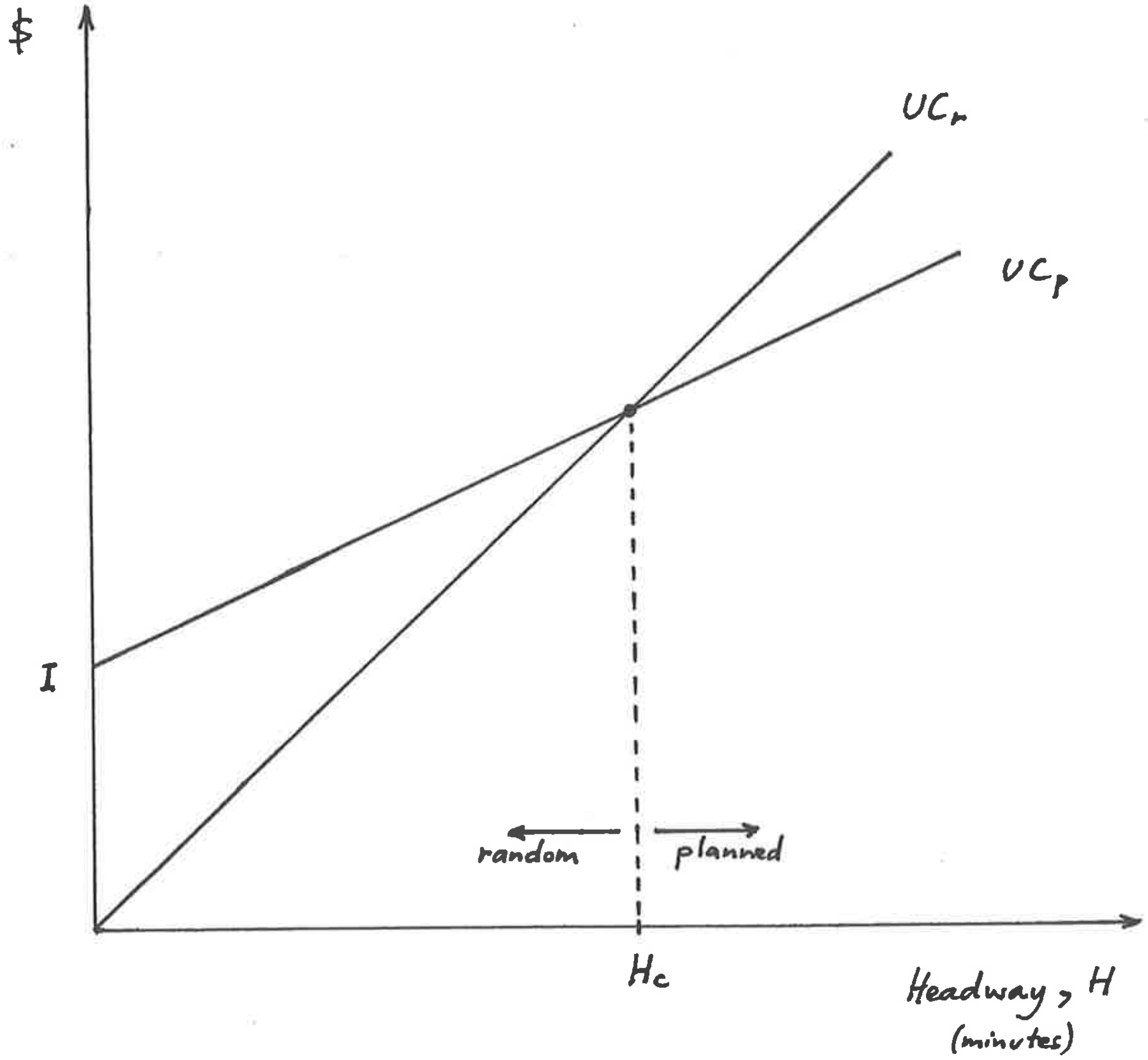
Although the Tisato/Jansson bi-modal user cost minimisation discrete choice model has added a useful new dimension to frequency-related user cost models, both versions of the model arguably have some limitations. These are :

[11] For convenience, the term"expected" will not be used repeatedly to describe user cost or its components, but it implicitly applies throughout unless otherwise stated.

[12] This assumption will continue to apply throughout this study.

[13] For simplicity, UC_r and UC_p are shown at this stage as linear schedules. Jansson (1993) adopts this approach, whereas Tisato (1991) uses schedules of non-linear form. The general nature of the choice situation is the same, however, irrespective of functional form.

Figure 3.1 : The Jansson/Tisato Model of User Choice Between Random and Planned Behaviour



(1) Both models assume that, in the case of random behaviour, users arrive at a loading point exactly at the user's preferred time, t_p, without substantiating this result.

(2) Both models fail to account for *SDD* as a frequency related user cost component. The Jansson model also fails to model *SSD*.

(3) Although the Tisato model takes into account *SSD*, it does so in a fairly restrictive way for the case of planned behaviour. Tisato modelled planned *SSD* after a comprehensive model developed by Bowman and Turnquist (1981). Unfortunately, the complexity of the latter makes it cumbersome to use, with Tisato using an approximation based on a simple functional form. The usefulness of the simpler model may be limited, however, since it was fitted to a small number of Bowman and Turnquist model results over quite a small *H* range only.[14]

(4) Both Tisato and Jansson assume a purely deterministic model to predict the discrete choice between random and planned behaviour by users. This is acceptable for an individual user, and would apply if users were perfectly homogeneous. However, at an aggregate level across all users, the model is less useful. It predicts the same choice outcome for seemingly identical users in identical situations, yet in practice differing choices are often observed (Ben-Akiva and Lerman, 1985), explained by the fact that the analyst is never able to identify all aspects of heterogeneity between users. A probabilistic choice model (e.g. a logit choice model) is likely to better predict aggregate choice outcomes.

In this study, the Tisato/Jansson discrete choice model framework will form the basis for a working model of user cost. However, it is considered beneficial to extend and develop the model further to overcome the limitations outlined above. The aim for the remainder of this chapter is to undertake this further development, and thus generate an improved working model of user cost for use in subsidy analysis. This is done as follows. Sections 3.5 and 3.6 outline a working model of the components of user cost for random and planned user behaviour respectively, with section 3.7 providing a summary of model equations. Section 3.8 then extends the choice framework from a

[14] This was the case since Bowman and Turnquist only reported their results over a small *H* range (see extended discussion in section 3.6.3)

deterministic one to a more realistic and useful one of probabilistic choice, using a theoretical logit model to predict choice outcomes between random and planned behaviour.

3.5 User Cost Under Random Behaviour

3.5.1 Delay, Unit Costing and User Cost

Service delay under random behaviour, D_r, consists of three components :

$$D_r = FD_r + SSD_r + SDD_r \qquad (3.3)$$

It is well known that if services are running to schedule (and thus $SSD_r = 0$), and a user can board the next service that arrives (i.e. vehicles are not full), and thus $SDD_r = 0$, a user behaving randomly can expect to, on average, suffer a wait at the loading point equal to half the headway (Seddon and Day, 1974; Bowman and Turnquist, 1981). This expected wait, is clearly due solely to service scheduling at a positive headway. This non-stochastic component of delay is what will be referred to as the frequency delay, FD_r, associated with random behaviour. When SDD_r and/or SSD_r become positive, D_r will rise accordingly above FD_r . Functional forms for the components of D_r are given shortly.

To convert D_r and its components into user cost units, it is necessary to factor delay by the user's average/unit rescheduling cost. With all three components of D_r consisting of waiting at the loading point, the unit rescheduling cost of each component (and of D_r itself) will therefore be based on the value of waiting time savings, v_w. The size of this unit rescheduling cost may vary, however, with the direction in which activity rescheduling takes place. There are two directions in which activity rescheduling can occur : backwards and forwards. Backward rescheduling occurs whenever the user arrives at, or departs from, a loading point *before* t_p (recalling that t_p is the preferred time at which the user would like a bus to depart). Forward rescheduling occurs whenever the user arrives at, or departs from, the loading point *after* t_p. Drawing on this distinction, the unit rescheduling cost in these two directions can be denoted as v_{wB} and v_{wF}.

For the working model for use in this study, assume that rescheduling costs are linear in time, thus v_{wB} and v_{wF} are constant, and that on average, across the population of users, $v_{wB} = v_{wF} =$

v_w, i.e. rescheduling is equally costly at the margin in both directions. Random user cost, UC_r, is the sum of FD_r cost (FDC_r), SSD_r cost ($SSDC_r$) and SDD_r cost ($SDDC_r$)[15] :

$$UC_r = FDC_r + SSDC_r + SDDC_r \tag{3.4}$$

where

$$FDC_r = FD_r \,.\, v_w \tag{3.5}$$

$$SSDC_r = SSD_r \,.\, v_w \tag{3.6}$$

$$SDDC_r = SDD_r \,.\, v_w \tag{3.7}$$

thus

$$UC_r = D_r.\, v_w \tag{3.8}$$

3.5.2 FD_r and SSD_r

Detailed expressions are now required for the components of random user cost. Consider FD_r and SSD_r first.

The literature (e.g. Holroyd and Scraggs, 1966; Jolliffe and Hutchinson, 1975; Bowman and Turnquist, 1981) provide the following expression for the expected wait time, $E(W_r)$, at loading points under random behaviour :

$$E(W_r) = \frac{H}{2}\left(1 + 2\left(\frac{\sigma}{H}\right)^2\right) \tag{3.9}$$

where σ is the standard deviation of bus departure times, an indicator of the level of service unreliability[16]

$E(W_r)$ increases with σ. The reason for this is as follows. With $\sigma > 0$, actual headways will be non-uniform, being either greater than, equal to, or smaller than, the uniform scheduled headway. That

[15] As in the Tisato/Jansson model, user cost (and its components) for both planned and random behaviour are *expected* costs.

[16] Bowman and Turnquist (1981) actually express $E(W_r)$ as a function of σ_H, the standard deviation of headway, where $2\sigma^2 = \sigma_H^2$ (Turnquist, 1982), and thus $E(W_r) = \frac{H}{2}\left(1 + \left(\frac{\sigma_H}{H}\right)^2\right)$. The parameters σ and σ_H differ because σ is the standard deviation of the *time* of bus departure, whilst σ_H is the standard deviation of the *gap* between consecutive buses.

is, there will be a *distribution* of actual headways centred on the scheduled headway.[17] Average (or expected) wait is therefore greater than $\frac{H}{2}$ because :

> *"... more passengers arrive during the long* [*headways*] *where the average wait is greater than half the average headway, and fewer passengers arrive in the short* [*headways*] *where the wait is correspondingly shorter,* [*thus average wait time*] *increases* [*above H/2*] *as headways become* [*non-*]*uniform ..."*
>
> (Bowman and Turnquist, 1981, p.465).

By splitting the above expression for $E(W_r)$ in two components, one independent of, and the other a function of, service unreliability, $E(W_r)$ can be seen to be the sum of FD_r and SSD_r, where :

$$FD_r = \frac{H}{2} \qquad (3.10)$$

and

$$SSD_r = \frac{\sigma^2}{H} \qquad (3.11)$$

Based on the above quote, SSD_r is the size of the upward biasing of $E(W_r)$ above $H/2$ resulting from service failing to depart at the scheduled time. Notice that SSD_r falls as H increases. The reason for this is as follows. For a given σ value, the bigger H is, the smaller the *relative* difference between short and long headways, thus the smaller is the relative difference in the number of users arriving in shorter vs long headways, and thus the smaller the upward biasing of $E(W_r)$ above $H/2$.[18]

3.5.3 <u>SDD_r</u>

Three *SDD* models were found in the subsidy literature surveyed : Glaister (1982; 1987), Bly and Oldfield (1987) and Chalmers (1990). The common element in each is the key role played by load factor, *LF*, the ratio of the number of passengers on buses to the carrying capacity of those buses : when $LF = 0$, $SDD = 0$; as *LF* becomes positive and increases, so too *SDD* becomes

[17] I am grateful to Colin Gannon for making this point to me.

[18] Two interesting polar cases should also be noted. When buses are perfectly reliable and always depart at the scheduled time (i.e. $\sigma = 0$), $SSD_r = 0$ and thus $E(W_r) = FD_r = \frac{H}{2}$. Alternatively, if buses are perfectly unreliable and thus depart randomly, Holroyd and Scraggs (1966) state that $\sigma_H{}^2 = H^2$, thus $\sigma^2 = \frac{H^2}{2}$ and in turn $SSD_r = H/2 = FD_r$, thus $E(W_r) = H$. Therefore, SSD_r will always lie in the range 0 to $H/2$, and $E(W_r)$ in the range $H/2$ to H.

positive and grows progressively, rising increasingly rapidly at higher *LF* values and tending to infinity as *LF* approaches a critical *LF* value.

Glaister (1982) reports a London based model developed by London Transport. In the Glaister model, *SDD*[19] is a function of a complex growth factor $1 + z/(1+z)$ where z is in turn a function of, amongst other things, *LF*. Chalmers (1990) found that an alternative, relatively simpler, growth factor of the form $1/(b_1 - b_2 LF)$, where b_1 and b_2 are constants, could be calibrated to fit the Glaister results rather well.[20] Values of $b_1 = 1.25$ and $b_2 = 1.65$ were found to provide the best fit. *SDD* is then :

$$SDD = LFmult \frac{H}{2} \tag{3.12}$$

where *LFmult* is the load factor multiple

$$= \left(\frac{1}{b_1 - b_2 LF} - 1 \right) \tag{3.13}$$

This simplified form of the Glaister model is the one used in considerations here.

Chalmers (1990) was dissatisfied with the rate at which *SDD* grew in the Glaister model, and its critical *LF* value of 0.76. Based on advice from the public transport operator in Adelaide, Chalmers proposed alternative parameter values for Adelaide, namely $b_1 = b_2 = 1.0$, forcing the critical *LF* value to 1.0, although no supporting evidence of any form was provided to support this amendment. The resulting model is referred to here as the Chalmers model.

Bly and Oldfield adopt a similar functional form in their analytical modelling, in which London was also the case study. In their model, *LFmult* takes the form :

$$LFmult = \left(\frac{1}{1 - LF^x} - 1 \right) \tag{3.14}$$

with $x = 5$

[19] The Glaister model, and the other two models surveyed, actually model overall waiting time, the sum of FD_r and *SDD*. The *SDD* models reported here were derived by merely substracting off $FD_r = H/2$. Note that the Glaister model also includes passenger "discomfort" costs, which increase as *LF* grows.

[20] The simpler functional form does, however, generate $SDD < 0$ for $LF <$ approximately 0.1. In these cases, the value of *SDD* was forced to zero.

It is important to note that Bly and Oldfield define *LF* differently to Glaister, resulting in quite different *LF* values. Bly and Oldfield effectively define *LF* as the *LF* at the maximum load point. In contrast, Glaister defines *LF* as the average *LF*. To illustrate the difference, think of the simple bus route example considered by Bly and Oldfield, with users uniformly distributed along the route, and with all users having as their destination the terminal at one end. When buses just become full at the terminal, the average *LF* in the Glaister model will be 0.5, but the Bly and Oldfield model will predict a *LF* of 1.0. The *LF* values therefore differ by a factor of 2.

In this study, the Glaister *average LF* definition is adopted. To enable the models to be compared, (3.14) needs to be modified to yield consistency between models. Based on the simple bus route example just discussed, to yield consistency, (3.14) should be replaced by :

$$LFmult = \left(\frac{1}{1-(2LF)^x} - 1\right) \qquad (3.15)$$

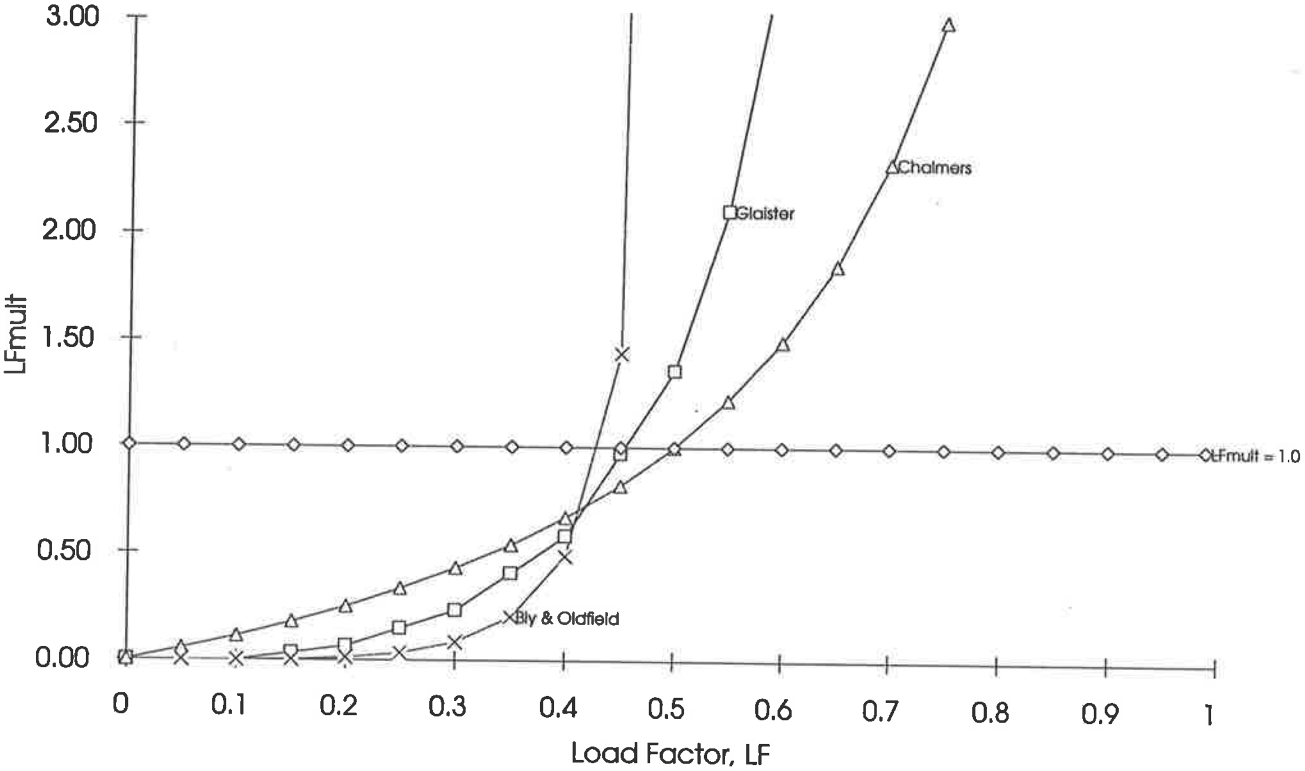
$$\text{with } x = 5$$

where *LF* is now the *average LF*.

Figure 3.2 illustrates the behaviour of *LFmult* (which is indicative of the behaviour of *SDD*) for the three models. Figure 3.2 also plots a comparative baseline at *LFmult* = 1.0, which coincides with $SDD = H/2 = FD_r$, to gauge the relative size of *SDD*. The three *SDD* curves each illustrate the same general behaviour outlined at the start of this sub-section. They differ to some extent in shape, however, due to the critical *LF* value differing between the models. Note that *SDD* tends to infinity at *LF* = 0.5 in the Bly and Oldfield model by definition, since Bly and Oldfield chose to set the critical *LF* to coincide with capacity being reached at the maximum load point.

Several considerations influenced the selection of one of these models for use in subsequent chapters in this study. Considerable weight was given to fact that the Glaister model was based on actual modelling work. Although the Chalmers model tried to modify the Glaister model parameters to better suit the Adelaide situation, there is a genuine lack of documentation in support of this change. Further, Chalmers found in his optimisation work that switching from the Glaister *SDD* model to his alternative model affected optimal results only marginally. The fact that the Bly and Oldfield model predicts *SDD* tending to infinity as buses just become full at the maximum load point was considered to be an overestimate. One would expect *SDD* to begin to rise rapidly in this

Figure 3.2 : Models of Stochastic Demand Delay (*SDD*)



situation, but not to tend to infinity. Overall, in the absence of a model for Australian conditions, the Glaister model was adopted for use in this study[21].

Before proceeding, a more precise definition of *LF* is required. Glaister (1982) does not provide a detailed definition, but closer inspection of the related computer program of the model (Glaister, 1984) indicates that *LF* is interpreted as the *average* load factor, measured as the ratio of passenger-kms (*PK*) to seat-kms (*SK*)[22] An unresolved issue, however, was whether *LF* should be determined for the peak direction of user flow, or the average over both directions. The approach adopted was to determine *LF* as the weighted average across both directions of flow, weighted by the size of the flows.

Adopting the following notation for a given bus route for a given time period, say per hour :

q is the number of users that board buses on the route (users/hour)

i denotes either the linehaul (*LH*) or backhaul (*BH*) directions

q_{LH} is *LH* boardings

q_{BH} is *BH* boardings

L_t is the average trip length (kms)

F is service frequency (buses/hour)

N is the bus size, i.e. the maximum allowable load per bus of seated and standing passengers (users/bus)

and L_r is the route length (kms)

then :

$$LF = \frac{LF_{LH}q_{LH} + LF_{BH}q_{BH}}{q} \tag{3.16}$$

where

$$LF_i = \frac{PK_i}{SK_i} \tag{3.17}$$

$$PK_i = q_i L_t \tag{3.18}$$

and

$$SK_i = FNL_r \tag{3.19}$$

[21] Hensher (1989a) also adopted the Glaister model in his analysis of Sydney.

[22] My thanks to John Dodgson for assisting me in clarifying this point.

If d is the proportion of total boardings that occurs in the LH direction, i.e.

$$q_{LH} = dq \tag{3.20}$$

then (3.16) reduces to :

$$LF = \frac{qL_t}{FNL_r}\left(1-2d+2d^2\right) \tag{3.21}$$

As d varies, the directional flow multiple in brackets changes in accordance with Table 3.1. The polar cases are d = 0.5 (balanced directional flows) and d = 1.0, flow in one direction only. As directional flow becomes increasingly peaky, so too the multiple, and thus average LF grows in size, at an increasing rate. The model is therefore able to predict pronounced differences in LF between peak and off-peak. It is useful for later analysis to also express (3.21) in the following form :

$$LF = \frac{qA}{FN} \tag{3.22}$$

where

$$A = \frac{L_t}{L_r}\left(1-2d+2d^2\right) \tag{3.23}$$

Table 3.1 : Load Factor Directional Flow Multiple

d	$1\text{-}2d\text{+}2d^2$
0.5	0.50
0.6	0.52
0.7	0.58
0.8	0.68
0.9	0.82
1.0	1.00

3.5.4 User Arrival Time, t_a

To complete the discussion of user cost under random behaviour, consider the following question : at what time, t_a, will the user choose to arrive at the loading point under random behaviour, and does this influence user cost ?

To assess this question, the most important thing to note from the outset is that the three components of service delay (D_r) outlined above are *independent* of t_a. That is, irrespective of when the user arrives at the loading point, D_r will always be the same, and will always be incurred. Next, provided the user arrives in the time interval starting from D_r minutes prior to t_p (recalling that t_p is the time the user would ideally like to see a bus depart) and ending at t_p, i.e., t_p-$D_r \le t_a \le t_p$, then the only delay incurred will be D_r. If, however, the user arrives before or after this time range, i.e. t_p-$D_r \ge t_a \ge t_p$, then the user will face an additional delay. For example, if the user arrives at the loading point $D_r + \xi$ minutes prior to t_p, they will on average depart on a bus ξ minutes prior to t_p (after spending the D_r minute service delay at the bus stop), thus generating an additional frequency delay of ξ minutes over and above the service delay. If, on the other hand, the user arrives at the loading point ξ minutes after t_p, they will also experience an additional frequency delay of ξ minutes over and above the D_r minutes of service delay. In both cases, the additional frequency delay, ξ, is unrelated to the provision of the service. It is purely due to the user making a sub-optimal decision. The conclusion that can be drawn is that a cost minimising user will always arrive *within* the time interval t_p-$D_r \leq t_a \leq t_p$, thus avoiding the cost ξ, and thus limiting delay to D_r.

Where, however, will t_a lie in this interval ? This will depend on the magnitude of v_{wB} and v_{wF} (which are defined in section 3.5.1). The cost minimising user will always choose t_a (and thus allocate the units of delay in forward and backward rescheduling directions) in a manner which satisfies the general "equal marginal rule", i.e. at the margin, rescheduling costs are equated in both rescheduling directions. In the case considered here, with v_{wB} and v_{wF} constant (and thus average and marginal rescheduling costs are equal), they will reschedule completely in the direction in which the unit rescheduling cost is lowest. Consider two cases.

(i) $v_{wB} = v_{wF} = v_w$

This is the general case on which the working model was developed earlier in this section. In this case, with unit costs equated at the margin, users will be indifferent between directions of rescheduling, and indifferent between t_a values in the range t_p-$D_r \leq t_a \leq t_p$. On average across the population of users, the expected value of t_a is likely to be the midpoint in the range, i.e. $t_a = t_p - \frac{D_r}{2}$.

(ii) $v_{wB} \neq v_{wF}$

This is a more general case. If v_{wB} <(>) v_{wF}, with all units of backward (forward) rescheduling being less costly at the margin than forward (backward) rescheduling, the user will reschedule totally backwards (forwards), with $t_a = t_p$-D_r ($t_a = t_p$).

These results suggest that the assumption made by both Tisato (1990; 1991) and Jansson (1993) in their models, that random users will arrive at t_p, i.e. $t_a = t_p$ (see deficiency (1) listed in section 3.4), will *not* hold as a general rule. Their assumption only holds for the special case where

$v_{wB} > v_{wF}$. Notwithstanding this, however, provided that t_a lies in the range $t_p\text{-}D_r \leq t_a \leq t_p$ (with any t_a outside this range being sub-optimal anyway), user delay will always equal the service delay, D_r. Therefore, although the t_a assumption used by Tisato and Jansson can be inappropriate, it does *not* bias the size of the delay and user cost under random behaviour.

3.6 User Cost Under Planned Behaviour

3.6.1 Delay, Unit Costing and User Cost

Service delay under planned behaviour, D_p, has three components :

$$D_p = FD_p + SSD_p + SDD_p \tag{3.24}$$

Recall that with random behaviour all three user cost components had a unit cost of v_w. This is also true for SSD_p and SDD_p which both involve waiting at a loading point. In contrast, however, planned users will opt to spend frequency delay time away from a loading point, preferring instead to be at locations where more useful activities can be undertaken (e.g. at home). As a result, unit planned frequency delay cost, denoted f, will be much lower than v_w[23].

As for the case of random user cost (see section 3.5), considerations here are limited to the case where rescheduling costs are *linear* in both rescheduling directions, and thus unit rescheduling costs are constant. In addition, it is assumed from the outset that, on average, unit forward and backward rescheduling costs are of equal magnitude, i.e. $f_B = f_F = f$, and $v_{wB} = v_{wF} = v_w$.

Planned user cost, denoted UC_p[24], is the sum of FD_p cost (FDC_p), SSD_p cost ($SSDC_p$) and SDD_p cost ($SDDC_p$) plus the information cost, I, incurred in obtaining and using timetable information :

$$UC_p = FDC_p + SSDC_p + SDDC_p + I \tag{3.25}$$

where

$$FDC_p = FD_p \,.\, f \tag{3.26}$$

$$SSDC_p = SSD_p \,.\, v_w \tag{3.27}$$

[23] This outcome is widely noted in the literature. Jansson (1993) refers to empirical work in Sweden which suggests that v_w is several times larger than f. See appendix *B* for further discussion of f and v_w values.

[24] Note again footnote 15.

$$SDDC_p = SDD_p \, . \, v_w \tag{3.28}$$

Under planned behaviour, the user will select between the services arriving immediately before and immediately after her/his preferred departure time, t_p. The user chooses the service which minimises expected planned user cost. SSD_p and SDD_p will be the same irrespective of which of these two buses the user catches. Therefore the task of minimising planned user cost reduces to minimising FDC_p.

3.6.2 $\underline{FD_p}$ [25]

The following notation will be used throughout :

L refers to the service immediately before t_p

R refers to the service immediately after t_p

t_L is the scheduled departure time of service L

t_R is the scheduled departure time of service R

Consider the FDC_p associated with each service, L and R.

(a) $\underline{FDC_p \text{ of catching service } L \text{ at } t_L}$

FD_{pL} is the time difference between t_p and t_L, i.e.

$$FD_{pL} = t_p - t_L \tag{3.29}$$

Normalising the time scale by setting $t_L = 0$ for convenience, then

$$FD_{pL} = t_p$$

and from (3.26)

$$FDC_{pL} = t_p f \tag{3.30}$$

(b) $\underline{FDC_p \text{ of catching service } R \text{ at } t_R}$

FD_{pR} is the time difference between t_p and t_R, i.e.

$$FD_{pR} = t_R - t_p \tag{3.31}$$

Now, $t_R = t_L + H$, $= H$ since $t_L = 0$, thus

[25] The model of FD_p used here is based on the earlier definition, and has been used by Evans (1987) and Tisato (1990; 1991).

$$FD_{pR} = H - t_p$$

$$\text{and} \qquad FDC_{pR} = (H - t_p) f \qquad (3.32)$$

(c) Choosing Between Service *L* and Service *R*

As just noted, the user chooses the service with the lower FDC_p, i.e. the smaller of FDC_{pL} and FDC_{pR}. Figure 3.3, which plots the cost schedules FDC_{pL} and FDC_{pR} (3.30 and 3.32) as functions of t_p, illustrates this choice situation for the range of possible t_p values. Defining t_p* as the critical t_p value at which $FDC_{pL} = FDC_{pR}$,[26] :

- if $t_p < t_p$*, $FDC_{pL} < FDC_{pR}$, so the user will choose service *L*, and $FDC_p = FDC_{pL}$
- if $t_p > t_p$*, $FDC_{pR} < FDC_{pL}$, so the user will choose service *R*, and $FDC_p = FDC_{pR}$
- if $t_p = t_p$*, $FDC_{pL} = FDC_{pR} = FDC_p$, so the user will be indifferent between *L* and *R*

The shape *ABC* in Figure 3.3 therefore defines FDC_p facing the user for the full range of t_p values in the headway interval t_L to t_R.

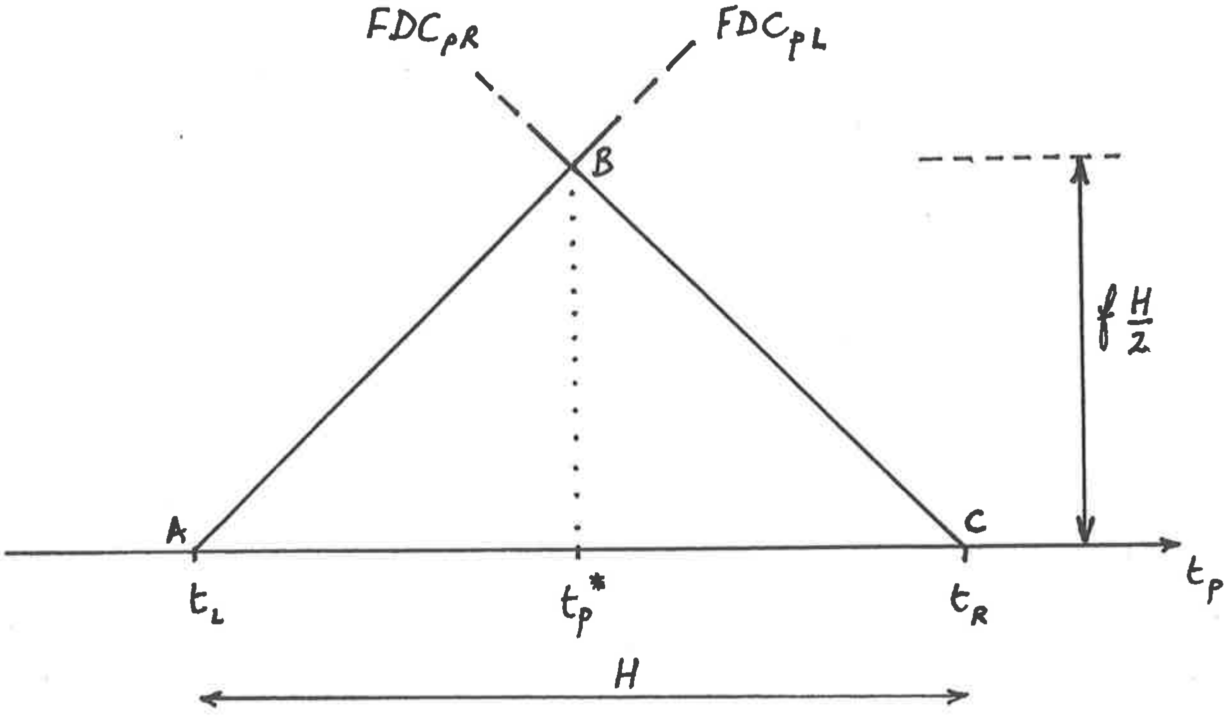
(d) The Representative User

As with random behaviour, we require expressions of planned user cost for the representative user for use in subsequent chapters. As mentioned above, FDC_p will vary from one user to the next as the t_p value varies between users. To determine FDC_p for the representative user, an assumption is required about the distribution of t_p values. Following previous analyses (Evans (1987) and Tisato (1991) for urban public transport, and Panzar (1979) for air transport), it is assumed that t_p is uniformly distributed over the headway time interval t_L to t_R. From Figure 3.3, the maximum FDC_p an individual can incur is $fH/2$, and the minimum is zero, thus with t_p uniformly distributed, the average, or expected, value will be :

$$FDC_p = f\,H/4 \qquad (3.33)$$

[26] In this case, where $f_B = f_F = f$, equating FDC_{pL} and FDC_{pR} yields $t_p^* = H/2$. In other cases, if f_B >(<) f_F, the FDC_{pL} schedule is steeper (flatter) than the FDC_{pR} schedule, and so t_p^* <(>) $H/2$.

Figure 3.3 : Frequency Delay Cost Schedules, Planned Behaviour



3.6.3 Background to SSD_p Modelling

Users plan to catch a certain scheduled bus. If $\sigma = 0$ (where σ is the standard deviation of bus departure times, an indicator of service unreliability), then users would arrive at the scheduled bus departure time, resulting in no waiting at the bus stop. In reality however, actual bus departure times are stochastic, and thus $\sigma > 0$. As a result, arriving at the scheduled departure time may result in the user missing the bus. To reduce the chances of missing a bus, and the long associated delay for the next bus, users are therefore prepared to arrive in advance of the scheduled departure time even though this may result in the user spending some time waiting at the bus stop, and knowing they may still miss the bus on some occasions.

The wait from arriving in advance of scheduled departure time, and the wait when a bus is missed, are both examples of stochastic supply delay under planned behaviour, SSD_p. Bowman and Turnquist (1981) [27] model the expected time at which users arrive at a bus stop, and the resulting expected SSD_p across the population of users, when services are unreliable. An overview of the Bowman and Turnquist model is presented in this section, with a simplified version, for use in this study, developed in the following section.

The conceptual basis of the Bowman and Turnquist model is as follows (Bowman and Turnquist, 1981). First, the utility $U(t_a)$ to the user of arriving at the bus stop at any given time, t_a, is computed. A probabilistic choice model of the form[28] :

$$f(t_a) = \frac{e^{U(t_a)}}{\int_o^H e^{U(x)}dx} \qquad (3.34)$$

is then used to predict the probability, $f(t_a)$, of the user deciding to arrive at t_a. It is then assumed that the distribution of arrival times across the population will follow the distribution of computed probabilities, $f(t_a)$. Finally, denoting $E(W(t_a))$ as the expected wait incurred by arriving at time t_a, SSD_p is then the weighted wait across the population of users, i.e.

$$SSD_p = \int_0^H f(t_a).E(W(t_a)).dt_a \qquad (3.35)$$

[27] With the compendium notes provided in Turnquist (1982).

[28] This is a continuous form approximation of the discrete choice multinomial logit model.

Bowman and Turnquist express $U(t_a)$ as a function of the expected wait :

$$U(t_a) = a\, E(W(t_a))^b \tag{3.36}$$

with a and b being parameters estimated empirically.

A sub-model for $E(W(t_a))$ is therefore a key requirement in the Bowman and Turnquist model. $E(W(t_a))$ is directly dependent on the nature of bus departure time stochasticity. Bowman and Turnquist assume that the departure times of consecutive buses are independent random variables, with bus i always departing before bus $i+1$.[29] There are, therefore, two possible outcomes facing the user arriving at t_a : either the user misses the intended bus since it has *already departed* (*AD*) prior to t_a, or the user is able to catch the intended bus since it has *not yet departed* (*NYD*) at t_a. $E(W(t_a))$ is then the sum of the expected waits of these two possible outcomes :

$$E(W(t_a)) = E(W_{AD}(t_a)) + E(W_{NYD}(t_a)) \tag{3.37}$$

With actual bus departure time being stochastic, a given bus may depart at any of a range of times around the scheduled departure time t_1. Denote the first and last times of this time range as t_s and t_e. The probability of the bus departing at any given time, t, within this range is described by the density function $p(t)$, with the cumulative probability density function denoted, $P(t)$, given by :

$$P(t) = \int_{t_s}^{t_e} p(x)dx \tag{3.38}$$

Then,

$$E(W_{AD}(t_a)) = P(t_a)\, W(t_a \mid AD) \tag{3.39}$$

where $W(t_a \mid AD)$ is the expected wait <u>given</u> the intended bus has *AD*

$$= E(t_2) - t_a \tag{3.40}$$

t_2 is the scheduled arrival time of the next bus

and $$E(t_2) = E(t_1) + H \tag{3.41}$$

and

$$E(W_{NYD}(t_a)) = (1-P(t_a))\, W(t_a \mid NYD) \tag{3.42}$$

where $W(t_a \mid NYD)$ is the expected wait <u>given</u> the intended bus has *NYD*

$$= \int_{t_a}^{t_e} (x - t_a).\, p(x)\, dx \tag{3.43}$$

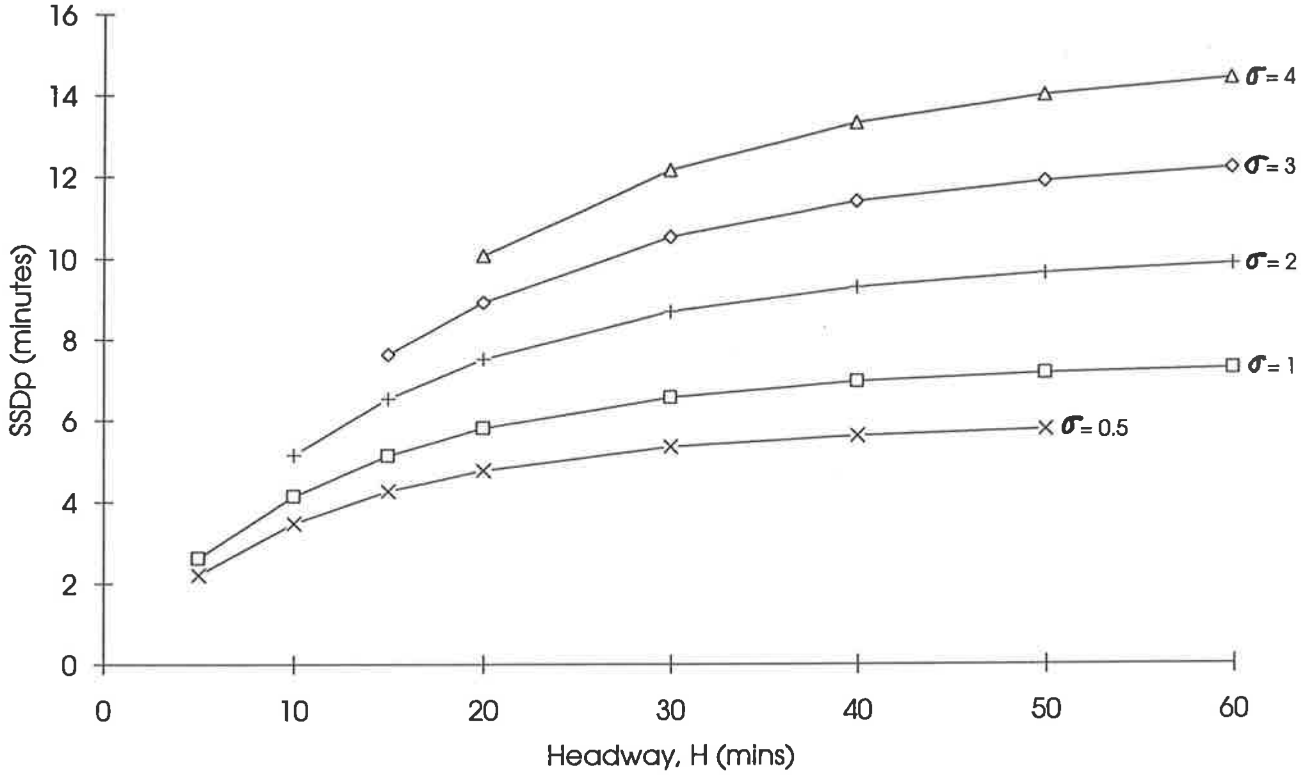
[29] This is a simplifying assumption which does not always hold in practice. For example, bus i may be running late and could arrive after bus $i+1$. Relaxing the assumption increases the complexity of the model considerably (Turnquist, 1982).

Combining (3.38) to (3.43) and substituting into (3.37) yields a complex expression for $E(W(t_a))$. To quantify $E(W(t_a))$, Bowman and Turnquist approximated the bus departure time probability distribution, $p(t)$, with a simple symmetrical triangular distribution.[30] Finally, using empirical data from seven different locations in Chicago and Evanston, Illinois, Bowman and Turnquist calibrated the model parameters (a = -1.0, and b = 0.55), with the resulting model producing close correlation with observed expected waiting times.

Figure 3.4 presents the family of SSD_p curves which result from Bowman and Turnquist's model, each curve representing a different σ value. Several points can be noted. First, even when buses are highly reliable (i.e. small σ), there is still a non-negligible base level of SDD_p. MVA Consultancy (1987) explain this as being due to users allowing a safety margin to cater for errors in clock time, etc. In other words, users are prepared to arrive at least a few minutes early to avoid missing a bus due to the user's watch/clock and that of the bus driver not being synchronised. Second, at any given headway, SSD_p increases with unreliability (σ). The more unreliable a service, a sensible user will increase the length of time that he/she arrives in advance of the scheduled departure time in order to avoid the lengthy delay penalty (i.e. waiting for the next bus) associated with missing the intended bus. Third, for a given level of unreliability (σ), SSD_p increases as H increases. This is because, as H increases, so too does the size of the delay penalty from missing the intended bus. Fourth, whilst SSD_p increases with H, it does so at a decreasing rate (i.e. the SSD_p vs H schedule flattens off as H increases). This is because although an increase in H results in an equivalent increase in the delay penalty from missing a bus, users react by arriving earlier, thus reducing the proportion of users incurring a delay penalty. Thus, expected wait increases at a slower rate than the increase in H. Fifth, the bigger is σ, the bigger the benefit (i.e. a fall in SSD_p)

[30] Studies of bus departure times tend to suggest that in practice bus departure time distributions are skewed rightwards (i.e. towards late departures) rather than being symmetrical. Strathman and Hopper (1993) and Adebisi (1986) identify a number of studies which report a range of rightward skewed distributions, namely, log-normal, gamma, and exponential. On the other hand, symmetric distributions have been used elsewhere (e.g. Lesley (1975) uses a normal distribution), and in fact a symmetric distribution applies in Adelaide (where departure times are almost normally distributed (see section B.3.1 in appendix B). Given the complexity of this modelling problem, Bowman and Turnquist's use of a simplified distribution shape seems reasonable, and is continued here.

Figure 3.4 : Bowman and Turnquist Model of Planned Stochastic Supply Delay (*SSD_p*)



of a fall in H. This is because the bigger σ is, the bigger will be the expected penalty of missing a bus, and so the bigger the reduction in that expected penalty from a fall in H.

3.6.4 Developing A Simpler Model of SSD_p

A drawback of the Bowman and Turnquist (*BT*) model is that its complexity makes it cumbersome to apply (identified as deficiency 3 in section 3.4), which suggests that a simpler model which approximates it would be useful. Although, as indicated in section 3.4, Tisato (1990; 1991) attempted to produce such a simpler model, the resulting model was calibrated on *BT* results over only a limited headway range (H = 0 to 20 mins)[31]. The aim of this sub-section is to develop a new simple model using *BT* results over a much wider H range.[32]

There are several advantages of a less complex model for use in policy studies such as this one. First, it is well known (e.g. Jansson, 1979) that optimal user economies of scale pricing and subsidy is strongly influenced by the rate at which user cost falls when H is reduced[33], i.e. the slope of the UC vs H schedule. To quantify the slope, UC needs to be differentiable. The Bowman and Turnquist (*BT*) model for SSD_p (outlined in section 3.6.3), which forms part of UC_p, may be differentiable, but its complex nature suggests that differentiation will be a complex task and a highly complex derivative will result. A simpler SSD_p expression will make differentiation, and its output, more manageable.

A second advantage of a simpler model is that it allows easier estimation of SSD_p in more general applications outside the subsidy considerations of this study (e.g. the cost benefit analysis of shelters at bus stops).

These sentiments on using a simpler SDD_p model are shared by both Turnquist (1982) and Tisato (1990; 1991), with both generating alternative simpler models that approximate the complex *BT* model. Turnquist used a piecewise linear model, consisting of two linear segments over two H

[31] As that was the limited range for which Bowman and Turnquist (1981) reported their results.

[32] An expanded set of *BT* model results was recently forwarded to me by Mark Turnquist along with a companion paper Turnquist (1982).

[33] The whole issue of optimal pricing and subsidy formulations is addressed fully in chapter 4.

ranges, to approximate the non-linear *BT* model. Tisato, on the other hand used the following simple non-linear single equation functional form :

$$SSDp = AH^{\phi}\sigma^{\gamma} \qquad (3.44)$$

where A, ϕ and γ are estimation constants

and $0 < \phi < 1$ and $0 < \gamma < 1$ in order to generate the appropriate non-linear response of the *BT* model illustrated in Figure 3.4.

Both simplifying approaches were able to approximate the *BT* model results reasonably well. However, both have deficiencies which may pose significant difficulties in this study. Turnquist's piecewise linear model, since it is not continuously differentiable, will generate significant discontinuities in the slope of SSD_p (and thus the slope of UC_p), whilst Tisato's model is relevant over only a limited H range (as discussed above). A new simple model of SDD_p therefore needs to be investigated, including giving consideration to alternative functional forms.

Six non-linear functional form models were tested in total in an attempt to find a suitable approximation of the *BT* model. The models considered are listed in Table 3.2.

Table 3.2 : Candidate Functional Forms for Planned Stochastic Supply Delay (SSD_p) Model

Model	*Functional Form*
A	The *BT* model - the comparison base
B	$SSD_p = A.H^{\phi}.\sigma^{\gamma}$ (the Tisato (1991) model)
C	$SSD_p = A(\ln H)^{\phi}\sigma^{\gamma}$
D	$SSD_p = A(1-\frac{1}{H})^{\phi}\sigma^{\gamma}$
E	$SSD_p = A+BH+C\sigma+DH^2+E\sigma^2+FH\sigma$
F	$SSD_p = A.H^{\phi}+B.\sigma^{\gamma}$
G	$SSD_p = A(1-\frac{1}{\ln H})^{\phi}\sigma^{\gamma}$

These functional forms were selected because they were known to generate (with appropriate parameter values and coefficients) non-linear schedules of the form shown in Figure 3.4, with their main common characteristics being that as H increases, the SDD_p schedule becomes progressively flatter, and the whole schedule is lifted when σ increases. Model *A* (the actual *BT* model) is the base for comparison. Models *C* and *G* are slight variations on models *B* and *D* respectively, with in each case *lnH* replacing *H* as an argument in an attempt to further increase the degree of non-linearity.

Each model was estimated using the *BT* model results as input data. Ordinary Least Squares Estimation (after transformation where necessary) was used for all models, except for model *F* which was estimated using Non-Linear Estimation. The *H* value range 0-60 is the type of range over which an SSD_p model is likely to be applied, however, a *H* range of 0-90 was used in model estimation regressions in order to improve the fit at the high end of the application *H* range.

A summary of the correlation coefficient, R^2, for the models is given in Table 3.3. Based on R^2 alone it would be difficult to choose a preferred model given that all models produce high R^2 results (although model *E* does appear to give the best fit overall). Given this, model preference was determined by comparing plots of SSD_p results generated by the various models with *BT* model results. In assessing preference, it was desirable to match as closely as possible two things : SSD_p itself; and secondly, the slope of SSD_p, since this is what plays a critical role in subsidy analysis (see discussion at start of this sub-section).

Table 3.3 : Correlation Coefficient Summary

Model	R^2
B	0.969
C	0.966
D	0.975
E	0.992
F	0.961
G	0.976

Results for models *C* and *G* were only marginally different from those of the models on which they were based, *B* and *D*. As a result, models *B* and *D* would always be preferred to *C* and *G* due their simpler functional form. Models *C* and *G* were therefore eliminated. For the remaining models, Table 3.4 provides a subjective summary of the goodness of fit of the models. The lower (higher) the rating number given to a model, the better (worse) the correlation with the *BT* model and its slope. Based on this assessment, model *B*, the model previously used in Tisato (1990; 1991), provides the best fit at the lower *H* values. In the medium *H* value range, models *B* and *E* both perform quite well, with model *B* performing slightly better. Finally, the best fitting model at medium to high *H* values is model *E*.[34]

Table 3.4 : Summary Rating Of Models for Goodness of Fit

		Models			
σ	*H Range*	*B*	*D*	*E*	*F*
1	Lower	1.5	2.5	4	3
	Middle	1.5	1	2	2
	Higher	2.5	1	1.5	3
2	Lower	1.5	2.5	4	3.5
	Middle	1	2	1	2.5
	Higher	2.5	1	1.5	2.5
3	Lower	1.5	2.5	4	4
	Middle	1.5	2.5	1.5	2.5
	Higher	2	2	1	1.5
4	Lower	*	*	*	*
	Middle	2	3	2	3
	Higher	1.5	2.5	1.5	1.5
Total	Lower	4.5	7.5	12	10.5
	Middle	6	8.5	6.5	10
	Higher	8.5	6.5	5.5	8.5

Note: The higher the rating number, the worse the correlation with the BT model and its slope.

Model *B* was chosen as the overall best model for use in the remainder of this study since it is the best model at lower and middle *H* values, still performs reasonably well at higher *H* values, and has the simplest functional form. It does, however, tend to overestimate the (important) slope of SSD_p at higher *H* values when σ is small, a point noted in later analysis.

Although the functional form of the final model (*B*) is the same as in Tisato (1990; 1991), the parameter values of model *B* here have been re-estimated based on a more comprehensive set of *BT* model results over a wider *H* range, and is therefore an improved model. The final estimated model to be used in this study is :

$$SSD_p = 1.881H^{0.357}\sigma^{0.389} \qquad (3.45)$$

[34] Models *E* and *F* perform quite poorly at lower *H* values due to the fact that they cut the vertical axis at non-trivial values whereas one would expect $SSD_p = 0$ at $H = 0$ for all σ values. Note also that model *D* cuts the horizontal axis at $H = 1$ rather than the origin.

3.6.5 $\underline{SDD}_p$

It is assumed that *SDD* will be identical for both behavioural modes, i.e. $SDD_r = SDD_p = SDD$. The expressions for *SDD* given in section 3.5.3 therefore also apply here for planned behaviour.

3.7 Random and Planned User Cost Summary

It is convenient for future reference during the study to summarise in one location the final models of (expected) user cost, and its components, for each of the two behavioural modes from the above two sections. It is also useful to draw a distinction between those user cost components which are influenced by *LF*-determined passenger congestion effects, the sum of which are denoted as u, and those which are not, denoted as v. This distinction will prove particularly useful in the optimisation work in chapter 4. The full set of model expressions are listed below. Note that (as mentioned in section 3.6.5), the same *SDDC* model applies for both behavioural modes.

Random Behaviour

$$UC_r = u_r + v_r \tag{3.46}$$

$$u_r = FDC_r + SSDC_r \tag{3.46a}$$

$$v_r = SDDC_r \tag{3.46b}$$

$$FDC_r = \frac{H}{2} v_w \tag{3.47}$$

$$SSDC_r = \frac{\sigma^2}{H} v_w \tag{3.48}$$

$$SDDC_r = LFmult\,\frac{H}{2} v_w \tag{3.49}$$

$$\text{where } LFmult = \left(\frac{1}{b_1 - b_2 LF} - 1\right) \tag{3.50}$$

where $b_1 = 1.25$, and $b_2 = 1.65$

Planned Behaviour

$$UC_p = u_p + v_p \tag{3.51}$$

$$u_p = FDC_p + SSDC_p + I \tag{3.51a}$$

$$v_p = SDDC_p \tag{3.51b}$$

$$FDC_p = \frac{H}{4} f \tag{3.52}$$

$$SSDC_p \;=\; 1.88H^{0.357}\sigma^{0.389}v_w \tag{3.53}$$

$$SDDC_p = SDDC_r \tag{3.54}$$

Note that the expressions for $SSDC_r$ and $SSDC_p$ can also be expressed in the following common format :

$$SSDC_i = A_i H^{\phi_i}\sigma^{\gamma_i}v_w \tag{3.55}$$

where $i =$ r or p depending on whether behaviour is random or planned

and A_i, ϕ_i and γ_i are constants which take the values given in Table 3.5.

Table 3.5 : Stochastic Supply Delay Cost (*SSDC*) Model Parameter Values

i	A_i	ϕ_i	γ_i
r (random)	1	-1	2
p (planned)	1.881	0.357	0.389

3.8 Behavioural Mode Choice : A Probabilistic Choice Approach

3.8.1 Deterministic and Probabilistic Choice Frameworks

The cost minimisation user cost model recently developed by Tisato (1991) and Jansson (1993), discussed in section 3.4, is a purely *deterministic* model of user choice between random and planned behaviour. In such a model, seemingly identical users in identical situations always make identical choices between random and planned behaviour (in accordance with the cost minimisation decision rules outlined in section 3.4). In reality, however, individuals who are seemingly identical in the eyes of the analyst are often observed to behave differently in identical situations, in conflict with the prediction of the deterministic model (Ben-Akiva and Lerman, 1985). This conflict is due to the fact that the analyst is never able to identify all aspects of heterogeneity between users, and thus is unable to identify all the variables which affect choice outcomes. A purely deterministic choice model would suffice if all users are perfectly homogeneous. However, for the more realistic case of heterogeneous users, a deterministic model therefore has limitations (identified as deficiency 4 in section 3.4), and an alternative model is required.

Random utility theory, a probabilistic choice theory, offers an alternative, superior, choice framework. The choice objective function (usually utility) is now treated as a random variable to reflect the analyst's uncertainty about the true objective function. The choice between discrete alternatives is described in terms of choice probabilities. The probability, $P_n(i)$, that individual n will choose alternative i rather than alternative j is (Ben-Akiva and Lerman, 1985) :

$$P_n(i) = \Pr\left[U_{in} \geq U_{jn},\ all\ j \in C_n\right] \tag{3.56}$$

where U_{in} is the utility individual n gains from alternative i

and C_n is the choice set.

In the choice situation considered in this chapter, the $_i$ and $_j$ subscripts can be replaced by $_r$ and $_p$ denoting random and planned behaviour as the two choice alternatives. In addition, given the focus on user cost minimisation, its is more appropriate to express the choice criterion in (3.56) directly in terms of user costs rather than utility. Further, it is also useful to express the choice criterion in terms of only those user cost components which actually influence choice, the sum of which shall be defined as *choice* user cost, *CUC*. A more appropriate expression for $P_n(i)$ instead of (3.56) is thus :

$$P_n(i) = \Pr\left[CUC_{rn} \leq CUC_{pn}\right] \tag{3.57}$$

where CUC_{rn} and CUC_{pn} are the choice user costs incurred by individual n under random and planned behaviour.

Let CUC_{in} (where $i = r$ or p) be a random variable :

$$CUC_{in} = DUC_{in} + \varepsilon_{in} \tag{3.58}$$

where DUC_{in} is the *deterministic* component of CUC_{in}

and ε_{in} is the *stochastic* disturbance component of CUC_{in}

The user cost equations in section 3.7 provide deterministic models of user cost. Inspection of the user cost components in section 3.7 reveals, however, that, with $v_r = v_p$, only component u, the user costs unrelated to *LF* effects, influences behavioural mode choice. Thus

$$DUC_{rn} = u_{rn}\ (= FDC_r + SSDC_r) \tag{3.59}$$

and

$$DUC_{pn} = u_{pn}\ (= FDC_p + SSDC_p + I) \tag{3.60}$$

Substituting (3.59) and (3.60) into (3.58), and in turn (3.58) into (3.57) yields :

$$\begin{aligned} P_n(r) &= \Pr\ [\ u_{rn} + \varepsilon_{rn} \leq u_{pn} + \varepsilon_{pn}) \\ &= \Pr\ [\ du_{pr} \geq \varepsilon_n\ ] \end{aligned} \tag{3.61}$$

$$\text{where } du_{pr} = u_{pn} - u_{rn} \qquad (3.62)$$

$$\text{and} \quad \varepsilon_n = \varepsilon_{rn} - \varepsilon_{pn} \qquad (3.63)$$

Once a functional form, $f(\varepsilon_n)$, has been chosen for the distribution of ε_n, $P_n(r)$ can be determined as the cumulative density function of $f(\varepsilon_n)$:

$$P_n(r) = \int_{-\infty}^{du_{pr}} f(\varepsilon_n) d\varepsilon_n \qquad (3.64)$$

3.8.2 A *Logit* Choice Model of Random/Planned Mode Choice

Use of the logistic distribution as the functional form for distribution $f(\varepsilon_n)$ yields what is known as a *logit* choice model. The logit model has been widely used in the transport field (and other fields) due to its appealing properties and analytical convenience (Ben-Akiva and Lerman, 1985; Beesley and Kemp, 1987), and will be used here to predict the choice between random and planned behavioural modes.

In a logit model, the general choice probability expression (3.64) reduces to :

$$P_n(r) = \frac{1}{1+e^{-\mu du_{pr}}} \qquad (3.65)$$

Noting that $du_{pr} = -du_{rp}$, then :

$$P_n(r) = \frac{1}{1+e^{\mu du_{rp}}} \qquad (3.66)$$

$$\text{where } du_{rp} = u_r - u_p \qquad (3.67)$$

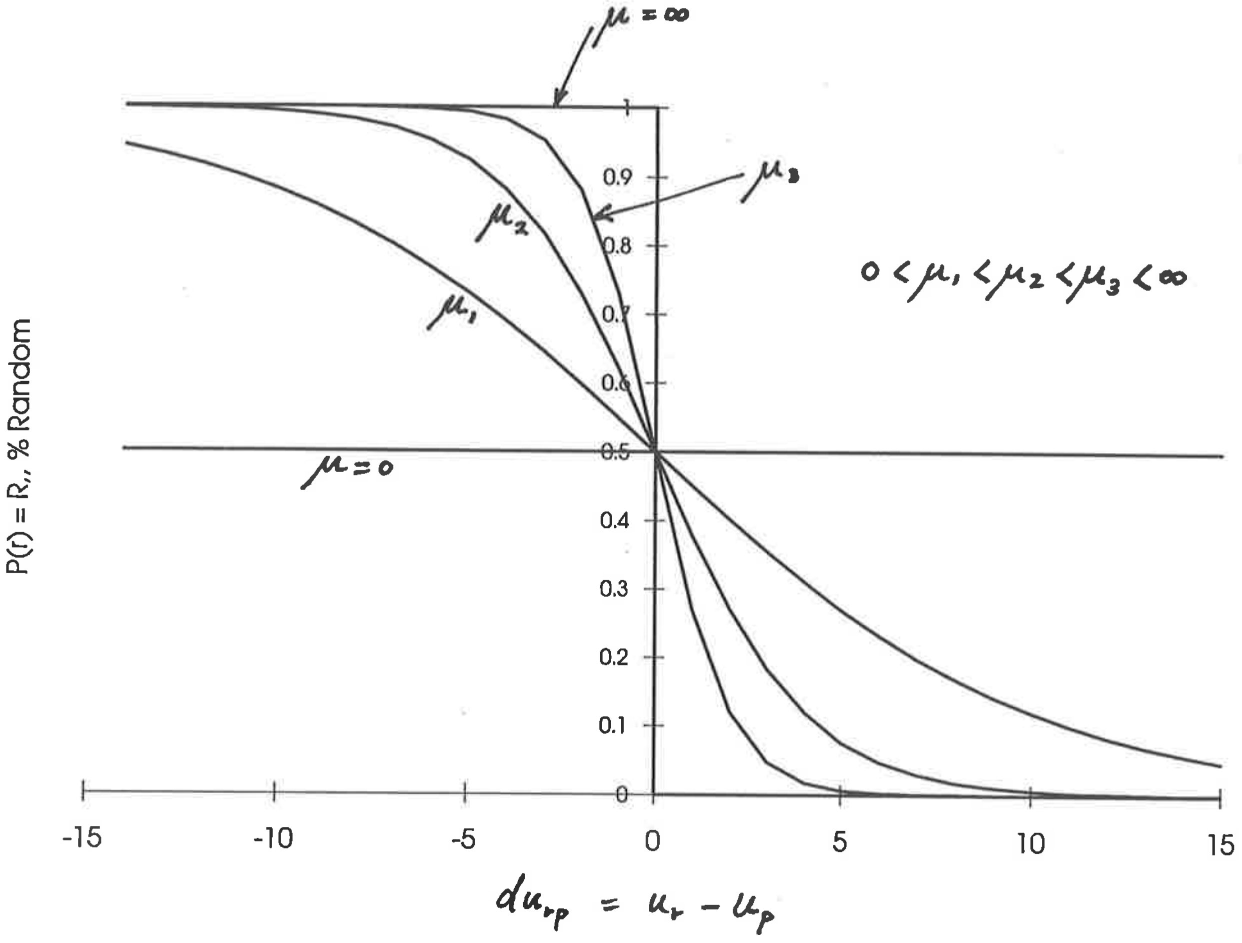
In addition,

$$P_n(p) = 1 - P_n(r) \qquad (3.68)$$

The probability of random behaviour being chosen, $P_n(r)$, predicted by the logit model is summarised in Figure 3.5 which plots $P_n(r)$ for several different values of the parameter μ, called the scale parameter since it acts as a scalar of du_{rp} in expression (3.66). By varying μ (and thus μdu_{rp}), one is able to vary the relative influence in choice outcomes of the deterministic user cost difference between modes (du_{rp}), and thus the relative role played in the random/planned choice process by deterministic and stochastic influences.

The $P_n(r)$ curves in Figure 3.5 are most easily understood by focusing first on the two polar cases, $\mu = 0$, and $\mu = \infty$. When $\mu = 0$, $P_n(r) = P_n(p) = 0.5$ in all circumstances, that is, a user is

Figure 3.5 : Logit Model Probability of Random Behaviour



equally likely to choose random behaviour as s/he is to choose planned. The deterministic component of choice user cost is therefore playing no role in choice, with choice outcomes being fully determined by unobserved stochastic differences between users. In contrast, when $\mu = \infty$, if $u_r < u_p$, $P_n(r) = 1.0$, i.e. only random behaviour prevails, and if $u_r > u_p$, $P_n(r) = 0$, i.e. only planned behaviour occurs. In this case, there are no stochastic (unobserved) variations between users, with choice outcomes therefore fully predicted by deterministic user cost. Thus, when $\mu = \infty$, the logit model coincides exactly with the simple purely deterministic Tisato/Jansson choice model.

The remaining three S-shaped logit choice probability curves in Figure 3.5 correspond to three intermediate μ value cases between these polar extremes.[35] In these cases, choice outcomes can be explained by a *combination* of deterministic and stochastic elements. The bigger is μ, the greater the tendency towards the deterministic polar case.

3.8.3 Making the Logit Model Operational

Three things are required to make the logit choice model operational : a method of aggregation must be selected; deterministic user cost models must be specified for the two behavioural modes; and a working value must be selected for the scale parameter μ.

(a) Aggregation Method

Ben-Akiva and Lerman (1985) discuss five methods of aggregation. In this study, the simplest of these approaches is adopted, namely, the "average individual" approach, where the characteristics of the average individual are assumed to apply across the whole user population[36], with the parameter values of the average user applied to the deterministic user cost model.

[35] Note that the size of μ is partly dependent on the units in which u_r, u_p and thus du_{rp} are expressed. For example, the same choice situation can be modelled with $\mu = \mu_1$ if du_{rp} is expressed in \$ units, or with $\mu = \mu_1 /100$ if du_{rp} is in cents units.

[36] The n subscript in (3.66) and (3.68) can thus be dropped from hereon. Note that as user parameters such as value of time savings vary between users, users will have their own unique critical headway. A probability (e.g. logit) model is therefore an appropriate way of predicting aggregate outcomes in given service situations.

Whilst a more detailed aggregation approach of modelling subsets of users separately was not undertaken here, before moving on it is worth noting briefly some of the potential differences between users and circumstances which are likely to lead to differing choice outcomes :

- First, users are more likely to act in a random manner in peak periods since headways are smaller in those times to accommodate the higher patronage levels. This is reflected in the results of the Adelaide case study in chapter 7.
- Second, the bifurcating nature of bus routes as they eminate outward from a city centre results in the headways faced by users increasing as one moves outward from the city centre, and conversely decreasing as one moves towards the city centre. Thus the closer that the user lives to the city centre, the more likely they are to arrive randomly at bus stops. This would be especially true for inner city residents given the high level of route duplications in inner suburbs. Another example of users being more likely to behave randomly due to route duplications is the North East busway in Adelaide, where a host of bus routes, which commence in outer suburbs, are funnelled onto an exclusive right-of-way busway which passes through inner and middle suburbs. As a result, users boarding along the busway face very high frequencies (low headways) (see further discussion in sections B.2.1 and B.2.3 of appendix *B*).
- Third, the regularity with which users travel is also likely to be an important factor. The more often one travels at the same time of day, the more likely that person would be of acting in a planned manner. An example is commuters who regularly start work at the same time. However, the trend towards flexible working hours, particularly for much of the employment found in the *CBD*, is likely work against this factor.
- Fourth, travel on transfer legs of a trip are likely to lead, in some circumstances, to delays as if users had acted in a random fashion. This is the case where it is not possible to coordinate the various services used in a trip (this issue is discussed further in section 7A.1 of the appendix to chapter 7).
- Finally, the variation in service unreliability across the bus network will also generate differing user responses. The greater the level of unreliability (especially when headway is relatively low to moderate), the less incentive there is for users to act in a planned manner. The Adelaide

experience is that service unreliability tends, on average, to be greater in the PM peak (Wills, 1995) which suggests that random behaviour may be more likely in the PM peak than the AM peak.

The choice probabilities now also reflect the proportion of the population of users acting in each behavioural mode, i.e. denoting R as the proportion of users acting in a random manner, and P as the proportion of users acting in a planned fashion, where $P = 1 - R$, then :

$$R = P(r) \tag{3.69}$$

and

$$P = P(p) = 1 - P(r) \tag{3.70}$$

The logit probability curves of Figure 3.5 therefore also indicate the relative split between random and planned behaviour over the range of possible deterministic user cost differences (du_{rp}). In this context, a further interpretation can be attached to the scale parameter μ : μ reflects the *rate* at which switching occurs between random and planned behaviour as du_{rp} varies. In a purely deterministic model ($\mu = \infty$), switching is perfectly rapid, with all switching occurring in a knife-edge manner at $du_{rp} = 0$. In a purely stochastic model ($\mu = 0$), no switching occurs since there is always an equal split between random and planned behaviour. In the intermediate cases ($0 < \mu < \infty$), the smaller the μ value, the more gradual is the switching between modes.

(b) Deterministic User Cost Specification

The second requirement for an operational logit model, working models of u_r (= FDC_r + $SSDC_r$) and u_p (= FDC_p + $SSDC_p$ + I), has already been satisfied via the specifications of user cost components presented in sections 3.5 and 3.6, and summarised in section 3.7. Figures 3.6, 3.7 and 3.8 summarise the deterministic choice user cost schedules for an illustrative set of parameter values ($v_w = 14, f = 3, I = 5$, and $\sigma = 2$, see Table B.4 in appendix B). Figures 3.6 and 3.7 plot the components of u for each behavioural mode, whilst Figure 3.8 brings together the u_r and u_p schedules, the key schedules which influence choice.

With u_r and u_p specified, du_{rp}, the vertical gap between these two schedules (the critical variable in the logit model) is now known for any given headway (H) value. The logit model output, the proportion of random users R (= $P(r)$), can therefore now be reported in a more useful form, as a function of H (a policy variable). Figure 3.9 illustrates the relationship between $P(r)$ and

Figure 3.6 : Components of Deterministic Choice User Cost Under Random Behaviour

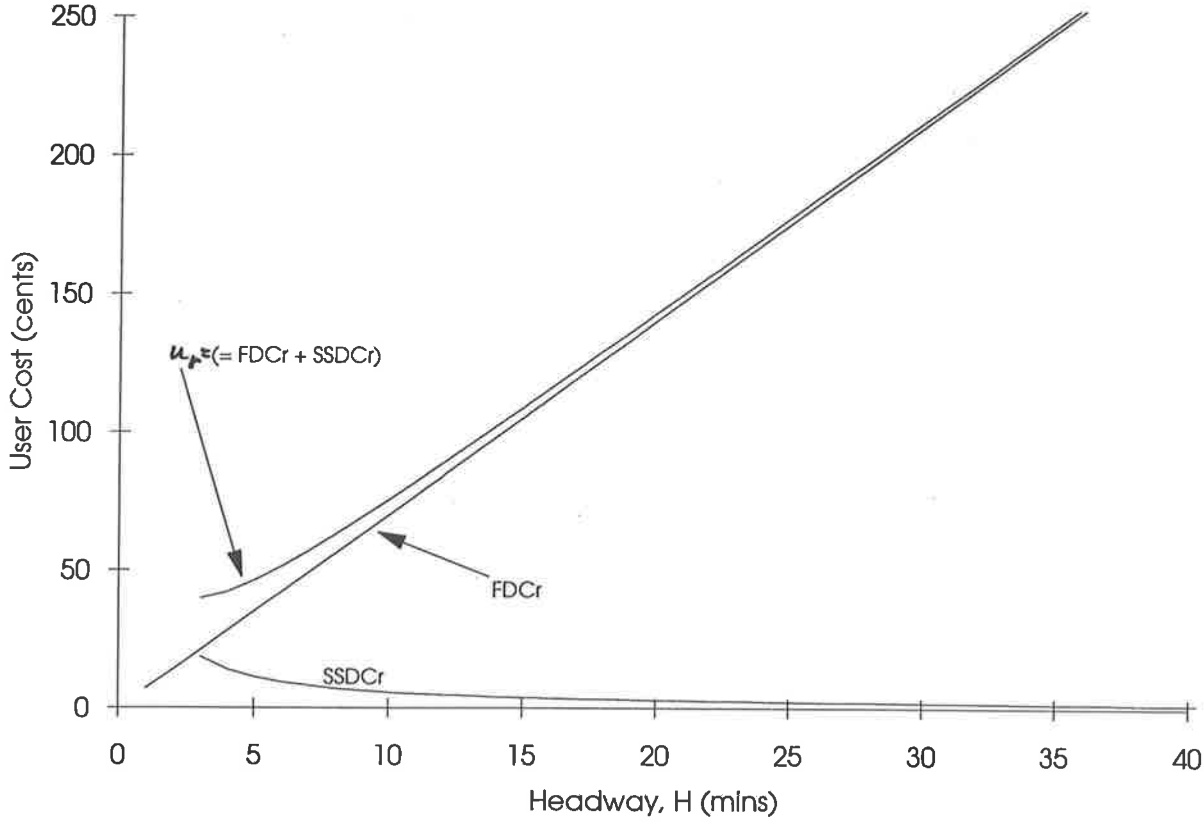


Figure 3.7 : Components of Deterministic Choice User Cost Under Planned Behaviour

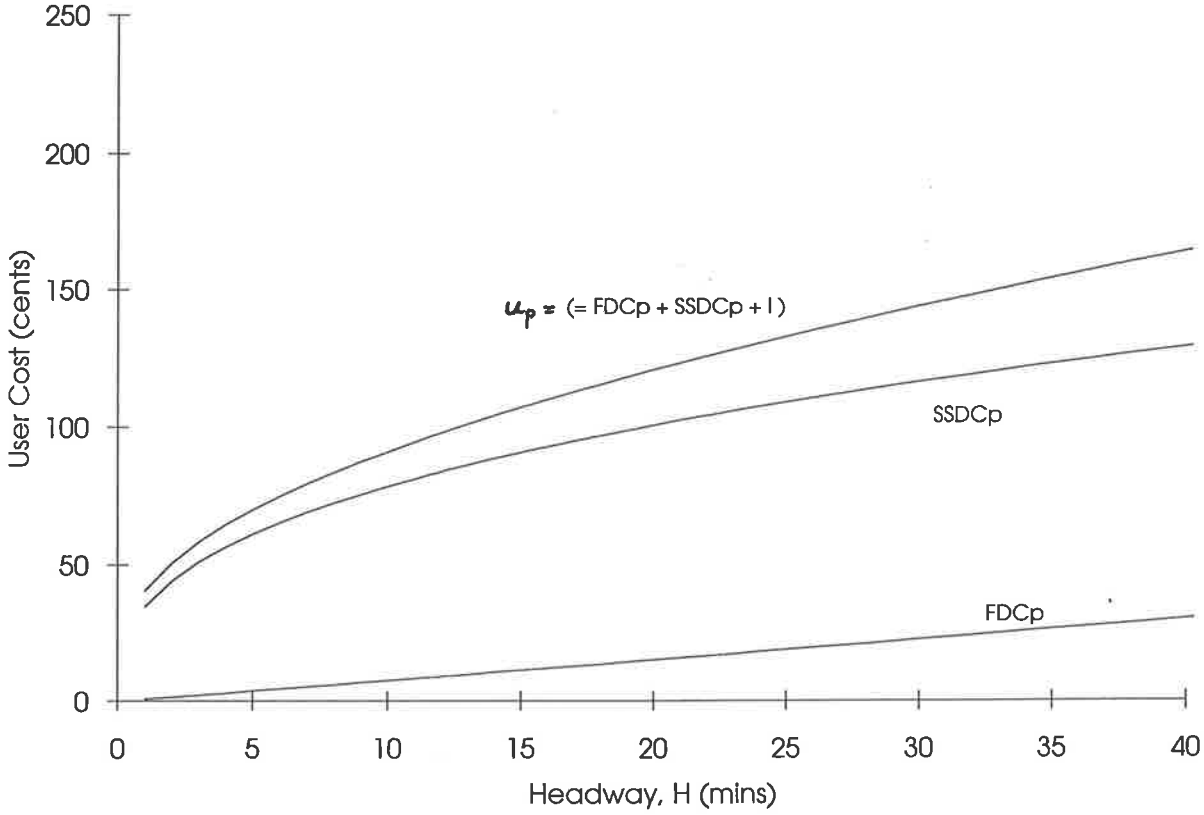


Figure 3.8 : Random and Planned Deterministic User Cost Schedules

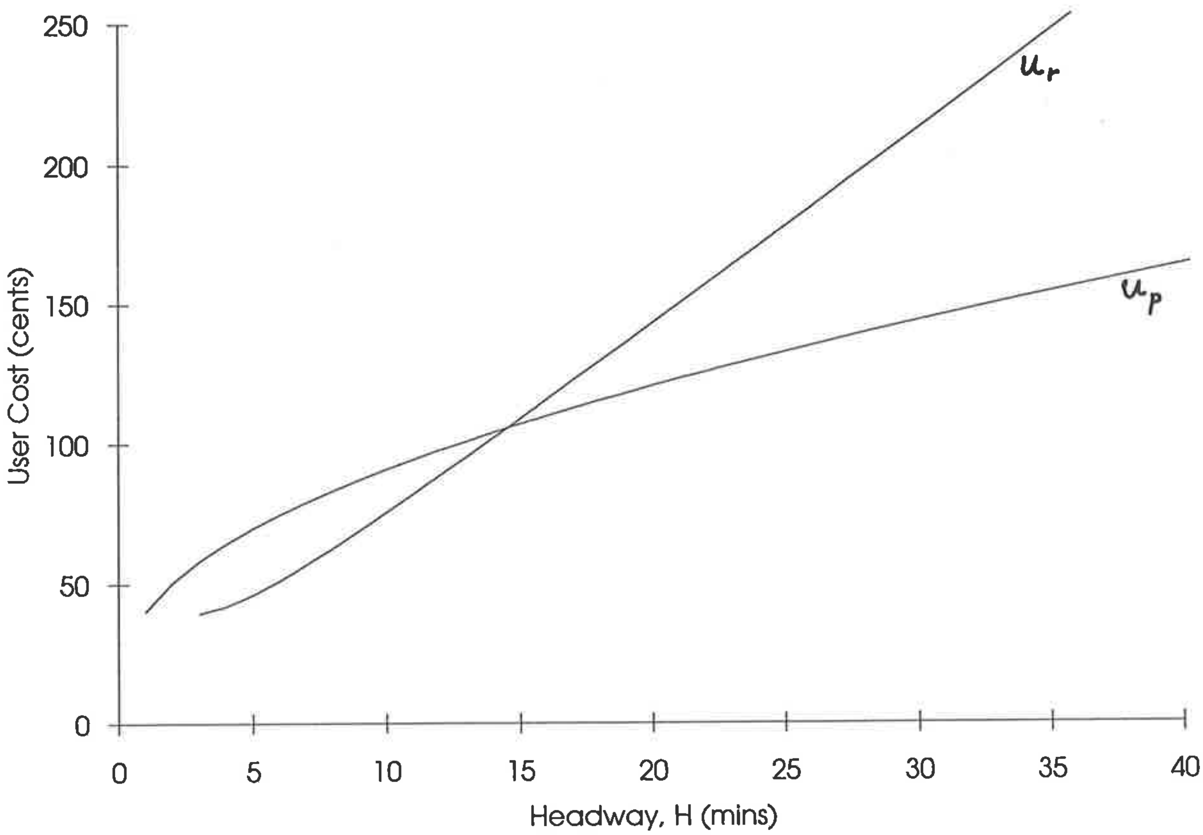
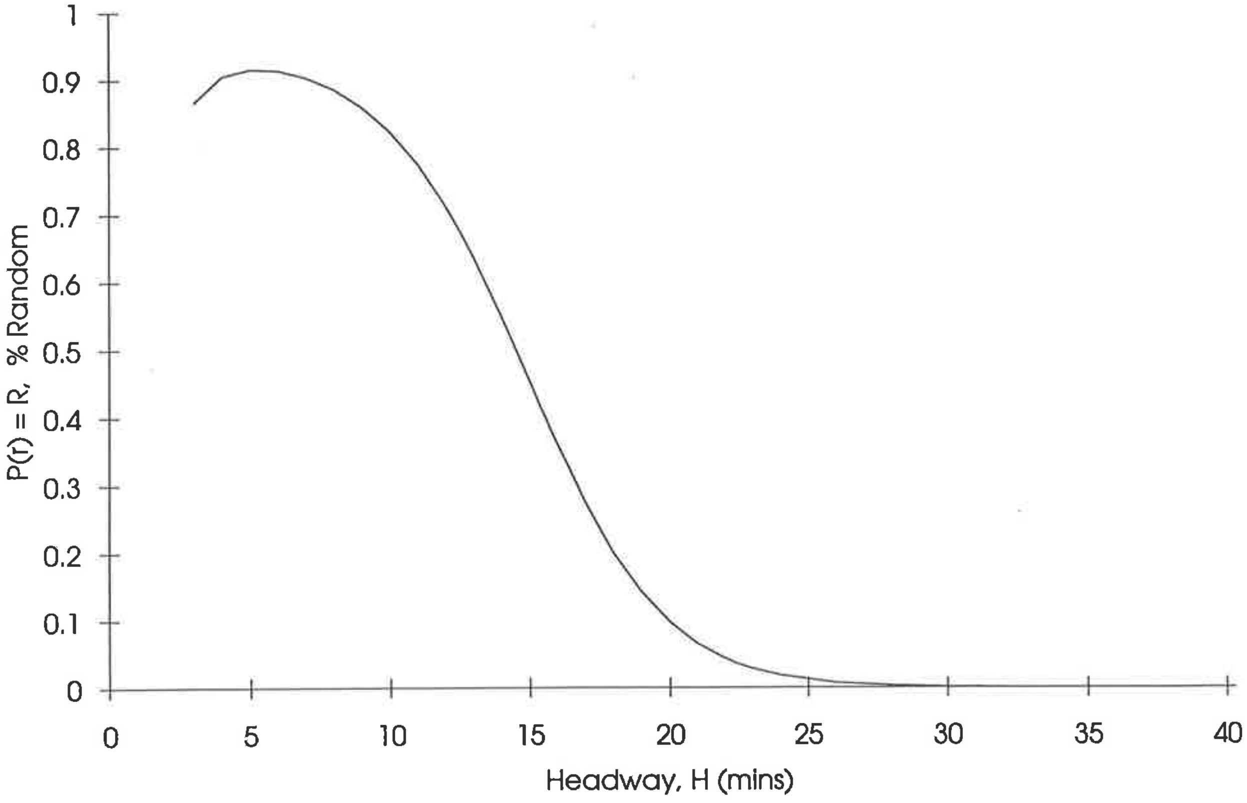


Figure 3.9 : Logit Model Relationship Between Probability of Random Behaviour and Headway (for $\mu = 0.1$)



H for an illustrative μ value of $\mu = 0.1$. In the discussion that follows, Figures 3.8 and 3.9 are considered together to explain outcomes as H varies.

At $H = H_c$ in Figure 3.8, $du_{rp} = 0$, thus from (3.66) $P(r) = 0.5$ and so random and planned behaviour are equally likely. As H grows above H_c, du_{rp} becomes positive and grows in size, resulting in $P(r)$ becoming increasingly smaller than 0.5, i.e. random behaviour becomes increasingly less likely. Conversely, as H declines below H_c, the situation is somewhat different. At first, du_{rp} becomes negative and grows in size, resulting in $P(r)$ growing increasingly above 0.5, i.e. random behaviour becomes increasingly likely. However, if H becomes small enough, the gap du_{rp} stops growing in size with $P(r)$ reaching a peak, with further reductions in H leading to du_{rp} becoming smaller and $P(r)$ declining. This peaking of $P(r)$ is due to two factors observed in Figure 3.8 at low H values : u_p falls at an *increasing* rate as H declines, and u_r falls at a *declining* rate as H declines.

Although the peaking of $P(r)$ at low H values is plausible, the assumptions on which the model is based suggest that this outcome may not be robust. The model assumes that each scheduled service runs independently of all other scheduled services. In reality, however, particularly when H is low and σ is high, there will be interaction between services, e.g. bunching and overtaking of services, which would require more sophisticated modelling of delays. Consequently, less confidence should be placed on results generated for very low H values, including the peaking of $P(r)$.

(c) Choosing the Scale Parameter, μ

The final requirement to make the logit model operational is the selection of a value for the scale parameter μ. The ideal approach is to calibrate μ using a set of data on choice outcomes. Unfortunately, such data was not available for this study. In addition, the value of μ is likely to vary between different settings. The approach taken was therefore an exploratory one which identified μ values that generate certain reference switching patterns. Four reference switching patterns were considered, with the rate of switching varying between patterns. Reference pattern 1 achieves $P(r)$

= 0.1 at $H = H_c + 2.5$,[37] that is, once H reaches $H_c + 2.5$, only 10% of users are behaving in a random manner, with 90% adopting planned behaviour. Reference patterns 2, 3 and 4 achieve $P(r)$ = 0.1 at $H = H_c + 5$, $H_c + 10$ and $H_c + 15$ respectively. That is, as H increases, switching from random to planned behaviour is progressively more gradual.

To test the sensitivity of μ to the parameter values chosen, μ was determined under three separate sets of parameter values, selected to yield a range of H_c values. The parameter values which influence H_c are v_w, f, σ and I. The values of I and v_w were kept constant throughout, but the other two parameters were varied. The three parameter value sets used are summarised in Table B.6 in appendix B.

The resulting μ values are presented in Table 3.6. Although there is some variation in μ values generated by the different parameter value sets, on the whole, the variation is relatively modest. Changing parameter value sets therefore does not have a huge influence on μ. The variation in μ is greater, however, as one moves from one switching pattern to the next, i.e. as the rate of switching changes. Selection of a μ value would therefore depend on the type of switching pattern which prevailed in a particular situation. For the analysis in this study, results are generated for the range of μ values given in Table 3.6 (specifically for the median parameter value set 2) so that the sensitivity of results to variation in the rate of switching is established.

Table 3.6 : Sensitivity of Scale Parameter μ

	PV[1] Set 1	*PV Set 2*	*PV Set 3*
H_c	20.4	14.5	11.8
μ for switching pattern 1[2]	0.23	0.22	0.2
μ for switching pattern 2[2]	0.11	0.11	0.1
μ for switching pattern 3[2]	0.055	0.05	0.045
μ for switching pattern 4[2]	0.035	0.03	0.028

Note : *1. PV stands for parameter value.*
2. Switching patterns 1, 2, 3 and 4 result in P(r) = 0.1 at $H = H_c + 2.5$, $H_c + 5$, $H_c + 10$ *and* $H_c + 15$ *respectively.*

[37] Recall that $P(r) = 0.5$ at $H = H_c$ where $du_{rp} = 0$.

3.8.4 The Logit Model User Cost Schedule

The expected "choice user cost" across the population of users, *E(CUC)*, or simply *CUC*, will simply be the sum, across both modes, of deterministic choice user cost for each mode (expressions (3.59) and (3.60)) weighted by the proportion of users choosing each mode (expressions (3.69) and (3.70)) :

$$
\begin{aligned}
CUC \text{ (i.e. } E(CUC)) &= R \,.\, u_r + P \,.\, u_p \\
&= R \,.\, u_r + (1 - R) \,.\, u_p \qquad (3.71)
\end{aligned}
$$

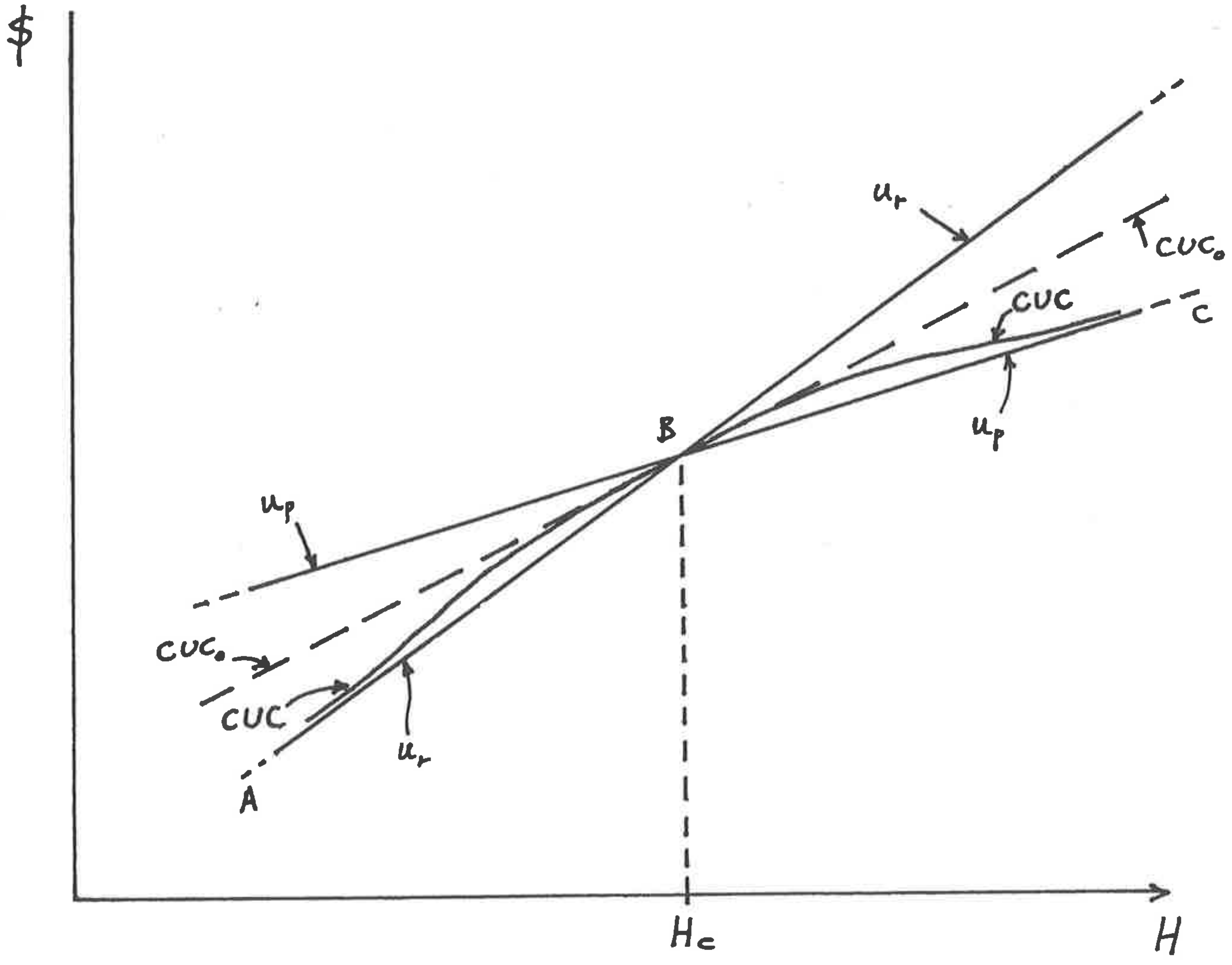
In accordance with (3.71), the *CUC* vs *H* schedule consists of a weighted combination of the u_r and u_p vs *H* schedules. Figure 3.10 presents a schematic illustration of the *CUC* schedule (for a given μ value) in relation to the u_r and u_p schedules. For simplicity and clarity, the u_r and u_p curves are drawn as straight lines, and only a portion of the curves are drawn.

At H_c, where the u_r and u_p curves cross, $R = 0.5$, and so the slope of *CUC* is the average of the slopes of u_r and u_p at that point. As *H* declines below H_c, with random behaviour becoming increasingly likely, *R* increases and *CUC* tends increasingly towards u_r and away from u_p. The reverse occurs as *H* grows above H_c, with the likelihood, and thus proportion, of planned behaviour increasing, and *CUC* tending progressively towards u_p. The rate at which *CUC* tends towards the single mode deterministic curves (u_r and u_p) is dependent on the scale parameter μ. The bigger μ is, the more rapid the convergence of *CUC* onto the single mode curves, and vice versa.

One extreme is when $\mu = 0$, for which $P(r) = 0.5$ throughout, with the *CUC* curve (curve CUC_0) lying exactly half way between u_r and u_p at every *H* value, and its slope equal to the average of the slopes of u_r and u_p. The other extreme outcome is when $\mu = \infty$, where *CUC* (curve *ABC*) coincides with u_r for $H < H_c$, and with u_p for $H > H_c$. This last outcome is precisely the purely deterministic model used by Tisato and Jansson. In such a model, all users suddenly switch from random to planned behaviour for a small increase in *H* from just below to just above H_c. The benefit of using the logit choice model is that it predicts a more gradual transition between user behavioural modes as *H* changes.

Finally, the overall total expected user cost across the population of users, *E*(*UC*) or simply *UC*, which is the key input into the optimisation work of subsequent chapters, can be defined. It is

Figure 3.10 : Logit Model Expected Choice User Cost Schedule



simply the sum of *CUC* (expression (3.71)) and the cost component which had no influence on mode choice, *SDDC* (expression (3.49)), i.e.

$$UC = CUC + SDDC \tag{3.72}$$

3.9 Chapter Summary and Conclusions

User costs play a central role in the analysis and understanding of optimal user economies of scale pricing and subsidy. The importance of user cost has been demonstrated by recent developments in the user cost modelling field which have resulted in a significant impact on user cost estimation and optimal subsidy. These developments have consisted of the conventional simple assumption, that public transport users behave in a random manner when accessing services, being relaxed. Instead, users have been modelled as choosing between two user behavioural modes, random vs planned behaviour (where the user uses timetable information in the latter, but not the former), with choices being made according to a user cost minimisation principle.

Given the importance of user cost models in the analysis of optimal user economies of scale analysis, the aim of this chapter has been to develop a working model of user cost (for use in subsequent chapters in this study) which is cognizant of, and attempts to further improve upon, these recent developments. In developing a working model, the chapter has made several contributions, including addressing a number of deficiencies (listed at the end of section 3.4) in existing models.

First, an expanded and more consistent set of user cost definitions was developed. A case was made for the term schedule delay being used only for describing the concept of activity scheduling, rather than being also used, as it is currently in some of the literature, to refer to the delay caused by the scheduling of transport services. For the latter, a new term, *service delay*, was proposed. In addition, the concept of stochastic delay was expanded into two components : stochastic *demand* delay, which results from stochastic user demand, and is the conventional stochastic delay found in the literature; and a new concept, stochastic *supply* delay, which results from stochastic service delivery (i.e. services not departing on time).

Second, it was argued that the assumption made in existing models, that users acting in a random manner always arrive at a loading point at time t_p, the time at which a user would most prefer a service to depart, does not hold in general. It was demonstrated in the chapter that a user aiming to minimise user cost could arrive over a range of times depending on the relative size of unit user cost for forward and backward activity rescheduling. More importantly, with users experiencing, on average, the same service delay at each arrival time in this range, user cost was therefore shown to be independent of variation in arrival time. As a result, although existing models of random user cost err in their prediction of arrival time, this error does not bias their prediction of user cost.

A third development was in the area of stochastic supply delay modelling under planned user behaviour. Although a rigorous formal model of this delay currently already exists, its complexity suggests there is a role for a simpler fitted model for use in policy analysis. A previous attempt to provide such a simplified model was shown to have limited application, and a superior simple model was developed. The new model has a number of benefits : its simple and continuous functional form facilitates straightforward mathematical differentiation, an important consideration in optimal subsidy determination; and it has been estimated over a comprehensive range of possible headway values, thus ensuring application to a wide range of service circumstances.

The final contribution of the chapter was to extend the random vs planned user behaviour choice from its existing purely deterministic context, to one of probabilistic choice based on random utility theory, a framework which has been widely used in the transport field (and other fields) to model discrete choices. A theoretical logit binary choice model was adopted as the working model for predicting random vs planned choice and outcomes across the population of users. The benefit of this development is that it predicts a more gradual shift between random and planned behaviour as service frequency varies, rather than the knife-edge switching at one specific frequency which occurs in the simpler deterministic model.

Chapter 4
OPTIMAL PRICING, FREQUENCY AND SUBSIDY FORMULATION

4.1 Introduction

This chapter establishes, and derives optimal solutions to, the bus optimisation problem. The motivation for the focus of the chapter comes from several sources. First, there is a need for a clear enunciation of general optimal results for use in later chapters. Second, the literature has reported a number of different approaches to the public transport optimisation problem, with a range of assumptions and constraints used. As Kerin (1990) points out, the results of bus subsidy analysis have been quite sensitive to the differing analytical models and assumptions used. Third, only limited attention has previously been given to diagrammatic presentation and illustration of the user economies of scale (*UES*) concept, and the resulting subsidy justification.

With these points in mind, this chapter has several aims :

- to set up and solve the formal first-best[1] bus optimisation problem from which *UES* subsidy results, including the derivation of optimality conditions which can then be used in later chapters of this study;

[1] Throughout this chapter, a *first-best* world is assumed, where price equals marginal cost in all other sectors of the economy (including for close complements like car travel), and non-distorting lump-sum transfers are possible. In addition, it is assumed there are no financial constraints limiting the amount of available subsidy which can be directed to supporting bus operations. As discussed in chapter 2 (sections 2.4 and 2.5), the analysis is quite different in a second-best world. If roads are unpriced, a road congestion management argument for subsidy arises (which is not addressed in this study). If public fund raising is distortionary, subsidies are harder to justify. Distortionary public finance is introduced into the analysis in the Adelaide case study in chapter 7.

- to present and inter-relate user economies of scale analysis for a number of constraint cases, namely cases where load factor (*LF*) and/or bus size (*N*) are fixed or variable, and consider how subsidy and other optimal results vary between constraint cases; and
- to illustrate graphically *UES* and associated subsidy for the various cases considered, including relating these to existing presentations and thus attempt to better integrate and relate previous analyses.

4.2 Definitions and Assumptions

4.2.1 Defining the Task

It is assumed that the aim of policy development is to set policy variables which will best serve the public interest, which is interpreted here to mean maximising the generic concept of social welfare. It is also assumed that the concept of social welfare is adequately represented by the commonly used applied welfare economics measure economic surplus, *ES*, or social net benefit.

There are many dimensions to the social welfare maximising public transport optimisation problem. The best possible (ideal) approach would be to optimise over all dimensions that can arguably be varied. As in all optimisations, the greater the scope to vary parameters (and thus the fewer the constraints on the optimisation) the greater will be net benefit, and the greater the generality of the results. There are two reasons, however, why it may not be feasible, or desirable, to optimise over a very wide set of variables.

First, in practice, a whole range of political, technical or financial constraints may prevent some items from being varied, at least in the short term. For example, it may be difficult to alter bus size in the short to medium run : there may not be a ready market for the sale of the current fleet; it may appear to the community to be wasteful if the current fleet is relatively new; etc.

Second, the greater the number of dimensions over which we optimise, the greater will be the complexity of the analysis. As is always the case, a trade-off therefore exists between generality and complexity. A sensible approach to dealing with this trade-off, which is adopted in this study, is to ensure that the key variables relevant to the consideration at hand are being modelled, but otherwise minimise the complexity of the analysis.

With user costs playing such an important part in this study, the user cost models developed in chapter 3 contain considerable detail. In order to allow the study to focus most strongly on the impact of user costs on subsidy, a number of simplifying assumptions are adopted in the optimisation formulations. These are :

- fixed route spacing
- fixed vehicle quality with respect to comfort
- fixed operating strategy by the operator
- uniform demand pattern with no peaks (to be relaxed later in chapter 7)
- a given road network on which buses travel

In all chapters of this study, analysis is undertaken at the representative bus route level. In this and the next two chapters, a route with illustrative parameter values is considered. In chapter 7, however, where subsidy is assessed for the Adelaide bus network, a disaggregated analysis is undertaken for the representative route in each of thirteen areas of the metropolitan bus system, with subsequent aggregation to yield optimal results for Adelaide.

What then are the variables over which the optimisation will take place ? This will depend on the particular analysis at hand. Several optimisations will be discussed below, the differences between these being the differing constraints to which the optimisations are subject. The variables to be optimised will, therefore, vary between optimisations. There is a consistent set, however, of potential optimisation variables, with the two key ones being service frequency (F) and price (P).[2] Given that bus size (N) has been at the centre of much debate in the subsidy literature (see "the Walters critique" discussion in section 2.5 of chapter 2), bus size is included as a third optimisation variable. In the discussion that follows in this chapter, one, or more (depending on the constraints that exist), of these three variables are optimised at any one time.

[2] The level of use of the service, or patronage, q, is sometimes optimised rather than P (e.g. Gwilliam *et al*, 1985). These two approaches yield identical results since (for a given F) P implies q, and vice versa.

4.2.2 Costs and Demand

Before considering optimisation formulations, it is useful to first set out some cost and demand definitions, and address issues of functional form, which are central to all the analyses to be undertaken.

The following definitions and notation will apply throughout :

C_p is the total cost incurred by the producer of the service, i.e. the operator

C_u is the total (time) user cost incurred by all users

g is the generalised cost of travel (the sum of money and non-monetary costs, in money units)

From chapter 3, L_t is average trip length (kms), L_r is route length (kms), and q is route patronage (boardings/hour).

An important distinction to make from the outset is between intermediate and final services (Small, 1992). The producer produces an intermediate service (vehicles-kilometres of service) by combining labour resources (e.g. the driver) with capital resources (the bus). To produce a final service, passenger-trips (or passenger-kilometres), which can be "consumed" by the user, the user's own time resource must be added to the intermediate service. As a result, the consumer/user is involved in both production and consumption.

(a) Producer Cost, C_p

The *short run* producer cost function,[3] C_p is a direct function of vehicle-kms of service, VK, (i.e. $\frac{\partial C_p}{\partial VK} > 0$). As discussed in section 2.4.2 of chapter 2, for urban buses, there are constant returns to C_p with respect to bus veh-kms (VK), thus $C_p/VK = \partial C_p/\partial VK$ is constant,[4] and :

$$C_p = \overline{\frac{C_p}{VK}}\, VK \tag{4.1}$$

It follows that, with $VK = 2L_rF$, then :

$$C_p = \frac{C_p}{VK} 2L_rF \tag{4.2}$$

[3] C_p and C_u are both *short run* cost functions since they define costs for any given set of frequency (F) and bus size (N) values.

[4] Values for use in this study are derived in section B.3.3 of appendix B.

Although C_p/VK does not vary with respect to VK, it is well known that C_p/VK is influenced by vehicle size (e.g. Nash, 1982; Kerin, 1990). Appendix B (section B.3.3) established a relationship between C_p/VK and N (expression (B.10)) which, when substituted into (4.2), yields[5] :

$$C_p = (c_1 + c_2 N).2L_r.F \tag{4.3}$$

where c_1 and c_2 are constants

(b) User Cost, C_u

User time costs are experienced by all users. For example, if catching a bus results in users having to wait for the arrival of a bus, this wait applies to all users who want to use the service. The convention is, therefore, to define the user cost experienced by each user as the average user cost, AC_u. Thus[6] :

$$C_u = AC_u q \tag{4.4}$$

As noted in chapter 2, user economies of scale are driven by *frequency-related* user costs (e.g. waiting time). It is useful, therefore, to formally distinguish in the notation used here between these and other user cost components (e.g. in-vehicle time, walk time, etc). This is done by using the subscripts F and O to refer to frequency-related user costs, and other non frequency-related user costs, respectively. Thus :

$$AC_u = AC_F + AC_o \tag{4.5}$$

AC_F has already been discussed at great length in chapter 3, although it was referred to there by the user cost notation UC[7]. From hereon, the term AC_F will be used instead of UC, but the equivalence of the two terms should be noted throughout.

[5] In some analyses, C_p is also influenced directly by q (e.g. Else, 1985; Gwilliam *et al*, 1985; Evans, 1987). Where this approach is used, it is generally assumed that there are also constant returns to C_p with respect to q. This direct influence of q on C_p will be much smaller than that of F, and therefore, for simplicity, it is *not* modelled here. A similar approach has been used elsewhere (Mohring, 1976; Findlay, 1983; Nash, 1988; Small, 1992). It turns out that, with constant returns to scale, this assumption has no impact on optimal subsidy results (Tisato, 1990, 1992).

[6] Note that from the perspective of overall social cost, i.e. producer cost plus user cost, variable cost (VC) consists exclusively of user cost, C_u, whilst fixed cost (FC) consists exclusively of producer cost, C_p (given by (4.3)).

[7] Expression (3.72) in section 3.8 of chapter 3 defines the UC (= AC_F) model.

As others have noted (Mohring, 1972; Turvey and Mohring, 1975), AC_F varies with each of the three potential policy variables (q, F and N). From chapter 3[8] :

$$AC_F = AC_F[FDC(F), SSDC(F), SDDC(LF(q, F, N), F)] \qquad (4.6)$$

i.e.

$$AC_F = AC_F\,(F, LF(q, F, N)) \qquad (4.7)$$

The direct impact of F (through all three components of AC_F) is the "Mohring" effect, or user economies of scale, where an increase in F leads to lower user cost for all users, i.e. $\frac{\partial AC_F}{\partial F} < 0$. In addition, $SDDC$, which is a passenger congestion cost (i.e. the risk of missing the first bus that arrives because it is full), increases with load factor, LF. From (3.12), (3.13) and (3.22), $\frac{\partial AC_F}{\partial q} > 0$ and $\frac{\partial AC_F}{\partial N} < 0$.

In the optimisation analyses that follow in this chapter, LF will on occasions be kept constant. It is convenient, therefore, to express AC_F as consisting of two components, denoted u and v,[9] where u is that part of AC_F which is not influenced by LF-determined passenger congestion effects, varying purely with F, whilst v is the component of AC_F which is influenced by LF-determined passenger congestion effects,[10] i.e. from (4.6) :

$$u = FDC + SSDC \qquad (4.8)$$

and

$$v = SDDC \qquad (4.9)$$

Thus :

$$AC_F = u(F) + v(LF(q, F, N), F) \qquad (4.10)$$

[8] The user cost expressions derived in chapter 3 are expressed as functions of headway, H. Recalling that $H = 60/F$, they can equally be expressed as a function of F. This is the convention followed in this chapter.

[9] This split of AC_F into two components u and v has already been introduced in chapter 3 (see sections 3.7 and 3.8) where only u influenced user choice between random and planned behaviour. Thus, the exact expression for u will vary depending on whether the behaviour of users is explained by random, planned or logit models of user behaviour.

[10] It should be noted that boarding and alighting has not been included in this study as a cause of delays to users on buses. Although boarding and alighting has played a role in other subsidy studies (e.g. Mohring, 1972; Bly and Oldfield, 1987; Chalmers, 1990; Jansson, 1993), Bly and Oldfield (1987) find that other user cost components, such as stochastic demand delay, have a much more significant impact on subsidy analysis. As a result, in order to place some limits on the complexity of the analysis, boarding and alighting effects have been ignored.

(c) Demand, q

On the demand side, let patronage, q, be given by :

$$q = q(g) \tag{4.11}$$

$$\text{where } g = P + AC_u = P + AC_F + AC_o \tag{4.12}$$

$$= P + u(F) + v(LF(q, F, N), F) + AC_o \tag{4.13}$$

Two functional forms for q have been used in the subsidy literature : the constant elasticity demand function (Bly and Oldfield, 1987); and, the exponential demand function, where elasticity varies rather than being constant (Glaister, 1982, 1987; Evans, 1987; Hensher, 1989a; Tisato, 1990, 1992). The exponential functional form will be used in this study. Its form is as follows :

$$q = \alpha \exp(-\beta g) \tag{4.14}$$

where α = the "potential" demand level, i.e. q when $g = 0$

and β = a constant

4.3 A History of Optimisation Formulations

Many versions of the public transport optimisation problem being considered here have been reported in the literature. Two important characteristics of these optimisations is the degree to which LF and N are fixed or allowed to vary, and second the analytical optimisation approach adopted.

(a) The Treatment of Load Factor (LF) and Bus Size (N)

One form of analysis is to adopt a *target load factor*. This approach consists of the optimisation being subject to a specified target load factor (LF_T) being met.

If N is also fixed (Jansson (1979), Waters (1982a); Kerin (1990)), for any given N and LF_T, from (3.22) F is directly proportional to q. If q increases, F increases in a linear fashion. As a result, F cannot be optimised when maximising ES since it is already bound by a fixed relationship with q, and thus the sole policy variable to be optimised is then P (or alternatively q (see footnote 2)).

If, on the other hand, N is allowed to vary (Kerin (1990)), from (3.22) F is now proportional to q/N (rather than just q). Note that for a given q and LF_T, F and N can vary, but not independently of each other. That is, F and N can only vary in a manner which ensures LF_T. There is, therefore, greater scope in optimisation when N is variable, but one can only optimise either F or N, not both independently, since one implies the other.

The target load factor approach, whilst clearly not optimising globally, has a couple of advantages. First, it is a relatively simple optimisation approach, making it easier to explain than more comprehensive optimisations, yet it is still able to capture the important Mohring effect. Second, aiming for a target load factor is an easily understood and attractive simple operational strategy used in the public transport industry (including in Adelaide (Kerin, 1990)) as one dimension of operating policy.

The alternative approach is to allow a *variable load factor*. If LF is allowed to vary, it becomes an integral part of the optimisation process. Further, F is no longer tied linearly to q or q/N, so it can now be genuinely and independently optimised. The variable LF approach is more complex and difficult to explain, but with fewer constraints it enables a greater degree of optimisation to be achieved.

(b) Analytical Approaches

At least three analytical approaches have been used to solve the public transport optimisation problem.

(i) *Two stage optimisation*

In this approach, the optimisation consists of two stages : cost minimisation; followed by maximisation of economic surplus (ES) subject to cost minimisation. Cost minimisation involves optimising either F alone, or F and N simultaneously, to yield minimum total cost (producer costs plus user costs) for any given q, thus yielding *long run* costs. The resulting optimal F and N values are denoted F^* and N^*[11]. Given these cost minimising responses, the optimal P (or alternatively g) which maximises economic surplus is determined for any given q (and thus F^* and N^*). This has

[11] Use of * will denote an "optimal" value throughout this study.

been the most widely adopted optimisation framework[12] (e.g. Mohring, 1972; 1976; Findlay, 1983; Nash, 1988; Small, 1992).

(ii) *One-stage optimisation*

This approach involves a more direct maximisation of *ES* without having to first minimise costs (e.g. Gwilliam *et al*, 1985; Else, 1985; Evans, 1987; Tisato, 1990, 1992; Jansson, 1993). This consists of determining first order conditions with respect to the variables to be optimised. However, by definition, economic surplus cannot be maximised without choosing optimal variable values which at the same time also yield minimum costs, so this approach yields identical results to approach (i) above.

(iii) *"Value for money" approach*

The third approach is the one pioneered by Glaister (1982; 1987) for use in policy making in the UK on the question of how funds allocated nationally for public transport subsidy should be distributed between cities, and has been applied in Australia by Dodgson (1985) and Hensher (1989a).[13] Rather than optimising, this approach consists of determining the net benefit of using additional subsidy to fund variation in policy variables from their current setting, i.e. reducing *P*, or increasing *F*. The value for money approach facilitates two things. First, it allows the determination of the direction in which each policy variable should be varied in order to increase *ES*. Second, it indicates whether balance exists between current policy variable settings with respect to their marginal contributions to economic surplus. Of course, if one were to alter the policy variables until the net benefit of further change in each variable is zero, then the results would match with those of approaches (i) and (ii) above.

In this study, use is made of approaches (i) and (ii). The choice of approach is dependent on the particular application or consideration at hand in later chapters.

[12] Which is not surprising since it is closest to the traditional neoclassical form of microeconomic analysis of market operation.

[13] See related discussion in footnote 18 of chapter 2.

4.4 First-Best Optimisation : The Problem and Its Solution

4.4.1 The Problem and a General Solution

Some versions of the optimisation problem (e.g. Mohring, 1972; Forsyth and Hocking, 1978; Small, 1992) do not formally model user congestion costs (i.e. component *v* in (4.10)). As a result, the optimisation thus requires an inequality bus capacity constraint to ensure that buses do not carry more users than the bus capacity, and is solved using the Kuhn-Tucker technique. When user cost component *v* is included in the optimisation problem, however, the need for a capacity constraint disappears, since the passenger congestion user costs resulting from the number of users approaching bus capacity is now captured through *v*, and forms an integral part of the optimisation. The problem can then be solved by simple non-constrained optimisation.

The economic surplus (*ES*) maximisation problem is then :

$$\max_{\forall j} \quad ES$$

where *j* is a general designator of the policy variables to be optimised

$$ES = CS + GS \tag{4.15}$$

CS is consumer surplus

GS is government surplus[14] = -*S*

S is public transport subsidy.

Restating this,

$$ES = CS - S \tag{4.16}$$

A general first order condition (FOC) can then be expressed as follows :

$$\frac{\partial ES}{\partial j} = \frac{\partial CS}{\partial j} - \frac{\partial S}{\partial j} = 0 \tag{4.17}$$

This means that, *at the optimum, each policy variable being optimised must be set to simultaneously ensure that, at the margin, the impact on consumers (users) of altering j is exactly offset by the impact on taxpayers (the raising of subsidy).* With :

$$S = C_p - Pq \tag{4.18}$$

[14] The usual formulation contains producer surplus in place of government surplus. Although the analysis is not dependent on this choice, there are several reasons why the latter better describes the situation. First, with competitive tendering in service delivery about to be introduced in Adelaide, the government will be able to accrue any rents that may be implicit in the final policy settings. Secondly, the optimal outcome as predicted by the literature is to have a financial deficit, thus requiring a need for subsidy (to be shown shortly).

then
$$\frac{\partial ES}{\partial j}=\frac{\partial CS}{\partial j}-\frac{\partial C_p}{\partial j}+\frac{\partial (Pq)}{\partial j}=0 \qquad (4.19)$$

from which a first order condition can be evaluated in turn for each policy variable.

In the following sub-sections, a number of optimisation formulations are derived which cover the range of possible *LF* and *N* fixity cases,[15] progressing from the simplest analytical cases to the most complex. There are two reasons for presenting this taxonomic framework. First, it allows the most general case to be gradually built up from relatively simpler optimisations, making it easier to understand the mechanics of optimal outcomes. Second, each different optimisation has a role to play somewhere in optimal public transport analysis, depending on the application or consideration at hand. Four *LF/N* cases are considered in total, as summarised in Table 4.1.

Table 4.1 : Load Factor (*LF*)/Bus Size (*N*) Fixity Cases

Case	*LF*	*N*
1	Fixed	Fixed
2	Fixed	Variable
3	Variable	Fixed
4	Variable	Variable

4.4.2 Case 1 : Target *LF*, Fixed *N*

In this analysis, a target load factor, LF_T, is maintained throughout. In addition, bus size *N* is fixed at any given point in time. As a result, from (3.22) :[16]

$$F = qA/\left(\overline{LF_T}.\overline{N}\right) \qquad (4.20)$$

where $A = A(L_t, L_r, d)$ is given by (3.23) and is constant for any given L_t, L_r and *d*. Thus, frequency (*F*) is directly proportional to patronage (*q*). If *q* increases, *F* increases in a linear fashion. As a result, *F* cannot be independently optimised to maximise *ES* since it is already bound by a fixed relationship with *q*. The only policy variable which can, therefore, be optimised is *P*[17].

[15] Another dimension of fixity is to keep *F* constant. This is not considered here, but is referred to in the discussion in relation to Mohring's original contribution.

[16] Throughout this study, a horizontal bar above a term indicates it is being held constant.

[17] It is immediately relevant to ask whether this analysis relates to the short run or long run. The broadest interpretation of the long run in microeconomics is one where everything is variable, with the short run being when one or more variables cannot be varied. In reality, it is rarely the case that everything is variable. A practical interpretation is, therefore, nearly always required. The most common interpretation is that capital is fixed in the short run, with the quantity of capital usually being the dimension of fixity. The dimensions of

footnote continued on next page

With $F \propto q$, and with LF constant, user cost components u and v become $u(q)$ and $v\left(\overline{LF},q\right)$, and from (4.13) generalised cost, g, becomes :

$$g = P + u(q) + v(\overline{LF},q) + \overline{AC_o} \qquad (4.21)$$

Further, substituting for F from (4.20), (4.3) reduces to :

$$C_p = \left(c_1 + c_2 N\right)\frac{2L_r A}{LF_T N}q$$

$$= \left(\frac{c_1}{N} + c_2\right)\frac{2L_r A}{LF_T}q \qquad (4.22)$$

which is linear in q for any given L_t, L_r, d, LF_T and N.

The first order condition for P, $\frac{\partial ES}{\partial P} = 0$, (which is derived in section 4A.1 of the chapter appendix) is :

$$P = \frac{\partial C_p}{\partial q} + q\frac{\partial AC_F}{\partial q} \qquad (4.23)$$

(to be interpreted shortly).

To confirm that this result ensures an optimal outcome, consider the result expressed in generalised cost terms. Substituting (4.23) for P in (4.12) :

$$g = \frac{\partial C_p}{\partial q} + q\frac{\partial AC_F}{\partial q} + AC_F + \overline{AC_o} \qquad (4.24)$$

The first term is the marginal producer cost, MC_p. Next, note that marginal frequency related user cost, denoted MC_F, is :

$$MC_F = \frac{\partial\left(AC_F q\right)}{\partial q} = AC_F + q\frac{\partial AC_F}{\partial q} \qquad (4.25)$$

Thus, the second and third terms of (4.24) combined are MC_F. The fourth component of (4.24) is also the marginal user cost for other (non frequency-related) elements of user cost, MC_o (since AC_o is constant). Then, with marginal social cost, MSC, being the sum of marginal producer cost (MC_p) and marginal user cost (MC_u), i.e.

$$MSC = MC_p + MC_u$$

$$= MC_p + MC_F + MC_o \qquad (4.26)$$

fixity in this chapter are load factor and/or bus size. Cases 1, 2 and 3, where either load factor and/or bus size are given, have distinct short run characteristics, whilst case 4 can be thought of as a long run case.

the FOC (4.23) therefore ensures that $g = MSC$. Then, with equilibrium requiring that $g = MB$ (the height of the generalised cost inverse demand curve), the FOC is thus consistent with $g = MB = MSC$ and thus, by definition, is an optimal outcome. Denoting the corresponding optimal level of usage as q^*, (4.23) and (4.24) can be expressed as :

$$P^* = \frac{\partial C_p}{\partial q}(q^*) + q^* \frac{\partial AC_F}{\partial q}(q^*) \qquad (4.27)$$

and

$$g^* = \frac{\partial C_p}{\partial q}(q^*) + q^* \frac{\partial AC_F}{\partial q}(q^*) + AC_F(q^*) + \overline{AC_o} \qquad (4.28)$$

Now consider the financial outcomes which result in the optimum. Unit government surplus (i.e. unit profit), $GS/q = P - AC_p$, where AC_p is average producer cost (i.e. C_p/q). Substituting for P from (4.23), with $\partial C_p/\partial q$ expressed as MC_p, thus yields :

$$\frac{GS}{q} = MC_p + q\frac{\partial AC_F}{\partial q} - AC_p \qquad (4.29)$$

Then, noting (4.25), and noting the following general expression for MC_p :

$$MC_p = \frac{\partial(AC_p q)}{\partial q} = AC_p + q\frac{\partial AC_p}{\partial q} \qquad (4.30)$$

(4.29) becomes :

$$\frac{GS}{q} = q\left(\frac{\partial AC_p}{\partial q} + \frac{\partial AC_F}{\partial q}\right) \text{ or } (MC_p - AC_p) + (MC_F - AC_F) \qquad (4.31)$$

Expression (4.31) is a general expression for unit surplus. The difference between MC_p and AC_p allows for the possibility of producer economies of scale, whilst the difference between MC_F and AC_F represents user economies of scale.

Interpreting (4.31) for case 1, from (4.22) $MC_p = AC_p$, thus (4.31) reduces to :

$$\frac{GS^*}{q} = q^* \frac{\partial AC_F}{\partial q}(q^*) = MC_F(q^*) - AC_F(q^*) \qquad (4.32)$$

Now, $\frac{\partial AC_F}{\partial q} = \frac{\partial AC_F}{\partial F}\frac{\partial F}{\partial q}$, $\frac{\partial AC_F}{\partial F} < 0$, $\frac{\partial F}{\partial q} > 0$, $q^* > 0$, thus $\frac{GS^*}{q} < 0$, i.e. *a subsidy is required to attain the optimal outcome*. Denoting s^* as the optimal unit subsidy, then for case 1:

$$s^* = -\, q^* \frac{\partial AC_F}{\partial q}(q^*) = MC_F(q^*) - AC_F(q^*) \qquad (4.33)$$

Finally, total optimal subsidy, S^* is given by :

$$\begin{aligned} S^* &= s^* q^* \\ &= -\left(q^*\right)^2 \frac{\partial AC_F}{\partial q}(q^*) \end{aligned} \tag{4.34}$$

The source of subsidy is the positive impact that each additional user has on the frequency-related user cost of all other users (AC_F), i.e. $\frac{\partial AC_F}{\partial q} < 0$, i.e. the "Mohring" effect. The literature describes this in two ways. Mohring (1972) originally referred to the fact that AC_F, and thus AC_u, declines with increases in q as economies of scale in user cost, i.e. "*user economies of scale*".[18] Some of the literature (e.g. Walters, 1982) alternatively describes the Mohring effect as a *positive externality* between the marginal user and inframarginal users. The size of this positive externality at the margin is $-q\frac{\partial AC_F}{\partial q}$ (i.e. the difference between MC_F and AC_F). Thus the FOC (4.23) indicates that in the optimum, price should be set below marginal producer cost by an amount equal to the marginal user cost positive externality.

A useful diagrammatic presentation, which greatly assists in the discussion and explanation of user economies of scale, was developed by Waters (1982a) for the target *LF*/fixed *N* case (case 1). A similar diagram is presented here as Figure 4.1. The figure shows average total cost, *ATC*, where :

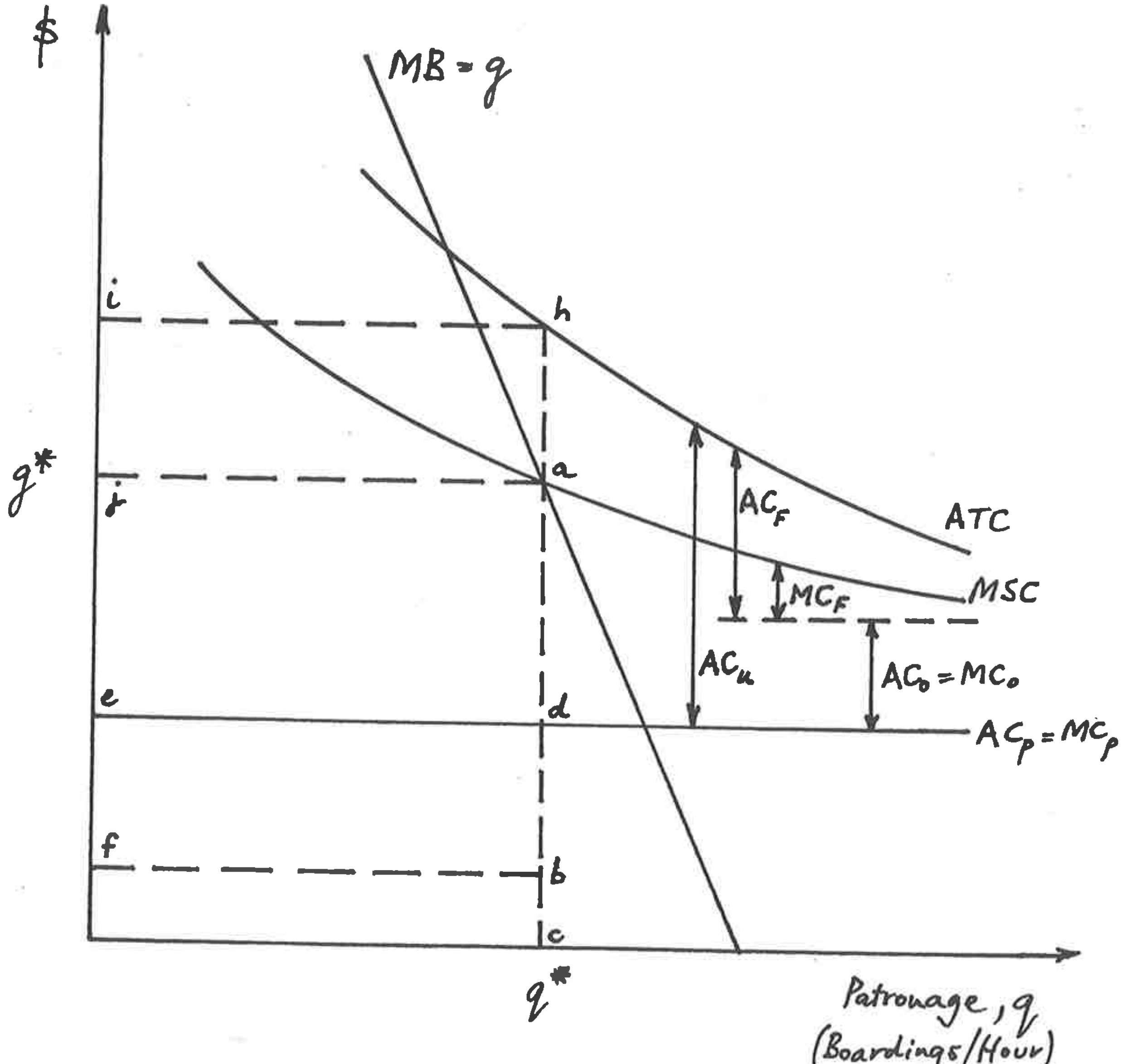
$$ATC = AC_p + AC_F + AC_o \tag{4.35}$$

declining as q increases, and thus economies of scale, with *MSC* declining[19] and lying below *ATC*. With $AC_p = MC_p$ and constant, and $MC_o = \overline{AC_o}$, the economies of scale are due purely to AC_F declining with q, i.e. user economies of scale. The optimal outcome is at point a (i.e. q^*, g^*), where the *MB* (= g) and *MSC* schedules intersect. Noting from (4.12) that $P = g - AC_u$, Waters notes that

[18] Note that the decline in AC_F is due to declines in both components of AC_F, u and v, thus u and v are *both* sources of user economies of scale.

[19] In a subsequent correction, Waters (1982b) points out *MSC* will be constant (rather than falling) in the special case where users arrive at a bus stop in a random manner (since MC_F is then = 0, and thus all components of *MSC* are then constant). The material presented in chapter 3 suggests, however, that this statement needs some qualification. To have the special case of $MC_F = 0$, requires two further assumptions to hold beyond requiring random behaviour. First, it requires v not to be modelled, or if it is then *LF* must be constant. Second, it requires $\sigma = 0$, i.e. that services run perfectly according to schedule. If, either of these assumptions are broken, and/or user behaviour is planned rather than random (see chapter 3), then $MC_F \neq 0$.

Figure 4.1 : User Economies of Scale, Waters Presentation (*LF/N* Case 1)



the optimal price/quantity outcome will be at point *b*, a distance AC_u below point *a*, i.e. optimal price (P^*) is the distance *bc*. Optimal price is therefore clearly less than average producer cost (AC_p = distance *cd*), with optimal unit subsidy being the distance *bd*, and optimal total subsidy area *bdef*.

An alternative way of illustrating subsidy in this diagram, and the approach adopted throughout this study, is to note that the financial position at any q can be inferred by the sign of the gap between MB and ATC. This is easily shown, since with $MB = g = P + AC_u$, and $ATC = AC_p + AC_u$, thus $MB - ATC = P - AC_p$, i.e. financial surplus/deficit. It then follows that :

- at q^*, where MB and ATC intersect, $MB = ATC$, and thus $P = AC_p$, a breakeven outcome;
- at lower q values, where $MB > ATC$ and thus $P > AC_p$, a surplus exists; and
- at higher q values, where $ATC > MB$ and thus $AC_p > P$, a deficit (and thus subsidy) results.

Therefore, the gap between the MB curve and the ATC curve measures unit financial surplus/deficit outcomes. Consequently, at the optimal solution at point *a* in Figure 4.1, MB (= MSC) < ATC, and so a subsidy outcome exists, with *optimal unit subsidy equal to ah,*[20] *the vertical gap between ATC and MSC*, and optimal total subsidy equal to the area *ahij*.

4.4.3 Case 2 : Target *LF*, Variable *N*

As useful as Figure 4.1, and case 1 on which it is based, is for demonstrating user economies of scale and the associated argument for subsidy, the argument can be generalised in a number of directions. One important direction is the relaxation of the fixed bus size constraint.[21]

In this case,

$$F = qA/\left(\overline{LF_T}.N\right) \qquad (4.36)$$

Thus, $F \propto q/N$ (rather than $\propto q$ as in case 1 above). Consequently, there is a fixed relationship between F and N, making it impossible, and unnecessary, to independently optimise both variables (since one implies the other for a given q). Here N will be optimised, making this the second optimisation variable, along with P. With $F \propto q/N$, u and v become $u(q/N)$ and $v(\overline{LF}, q/N)$, and so (4.13) becomes :

[20] Which is exactly equal to *bd*, the unit subsidy as shown by Waters.

[21] Chapter 2 (section 2.5) discussed how bus size has played an important part in the user economies of scale literature.

$$g = P + u(q/N) + v(\overline{LF}, q/N) + \overline{AC_o} \tag{4.37}$$

Expression (4.22) for C_p still applies, except now both q and N are variable. Thus :

$$C_p = C_p(q, N) \tag{4.38}$$

The derivation of the FOC with respect to P, i.e. $\frac{\partial ES}{\partial P} = 0$, is identical to that in case 1 above, yielding (4.23). The FOC with respect to N, $\frac{\partial ES}{\partial N} = 0$, (which is derived in section 4A.2 of the chapter appendix) is :

$$-q\frac{\partial AC_F}{\partial N}\bigg|_{\bar{q}} = \frac{\partial C_p}{\partial N}\bigg|_{\bar{q}} \tag{4.39}$$

This can be interpreted as follows. For a given q, if N falls, F will increase. As a result, user cost (AC_F) falls, generating benefits for all q users. However, the additional buses increase C_p. Expression (4.39) indicates that an optimal outcome results only when, at the margin, these benefits (the LHS of (4.39)) and costs (the RHS of (4.39)) are equated.

Solving expressions (4.23) and (4.39) simultaneously yields the optimal policy variables P^* and N^* (which in turn implies F^*), which in turn allows q^* and s^* and S^* to be determined. In determining s^*, the general surplus expression (4.31) applies. As will be discussed in section 4.5 below, for case 2, AC_p declines with increases in q (thus $AC_p \neq MC_p$), thus subsidy is due to both user and producer economies of scale.

4.4.4 Case 3 : Variable *LF*, Fixed *N*

The target load factor assumption is now relaxed to allow *LF* to vary and be optimised along with other variables. With *LF* variable, *F* is no longer constrained to varying linearly with *q* (as in case 1 in section 4.4.2) or with *q*/*N* (as in case 2 in section 4.4.3), so it can now be independently optimised. To commence with, consider case 3, where *N* is given. The policy variables to be optimised are, therefore, *P* and *F* (Mohring, 1976; Else, 1985; Gwilliam *et al*, 1985; Evans, 1987; Tisato, 1990, 1992; Jansson, 1993).[22]

[22] The Mohring (1976) analysis provides a simple illustration of this case (but with component v of AC_F not modelled).

The two FOCs, $\frac{\partial ES}{\partial P}=0$ and $\frac{\partial ES}{\partial F}=0$ (which are evaluated in section 4A.3 of the chapter appendix), yield :

$$P=q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} \qquad (4.40)$$

and

$$-q\left(\frac{\partial u}{\partial F}+\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial F}\bigg|_{\bar{q}}+\frac{\partial v}{\partial F}\bigg|_{\overline{LF},\bar{q}}\right)=\frac{\partial C_p}{\partial F} \qquad (4.41)$$

The RHS of (4.40) is the *marginal passenger congestion negative externality* associated with an additional user, that is the increase in the user cost component v across all q users. (4.40) then suggests that, in the optimum, price should be set equal to this marginal negative externality, or in other words there is a need for a passenger congestion tax. This is a general result which arises in optimisation work of congested facilities, for example, the case for a congestion tax in the case of vehicles travelling on roads (Walters, 1961; 1968). Expression (4.41) can be interpreted as follows. The RHS is the marginal cost of frequency enhancement. The LHS consists of the direct benefits of frequency enhancement : the improvement in user cost component u for all q users; and the direct reduction in v for all q users.[23] Thus, at the margin, the marginal direct benefit and marginal cost of frequency enhancement must be equated.

Next consider the financial implications of these optimal outcomes. The derivation of optimal unit subsidy, s^*, is the key consideration. This has been derived and expressed in a number of different, although equivalent (as will be shown below), ways in the literature. Once again, unit subsidy is given by :

$$s=AC_p\text{ - }P \qquad (4.42)$$

Substituting for P from (4.40) yields (Mohring, 1972; Tisato, 1990) :

$$s=AC_p\text{ - }q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} \qquad (4.43)$$

Further, noting from (4.3) that, for any given F, $MC_p = 0$, noting that $MC_o = AC_o$, and noting (4.25), then from (4.26) and (4.35) the difference between ATC and MSC is :

[23] The remaining indirect component of $\partial v/\partial F$ on the LHS of (4.41), i.e. $-q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial F}$, cancelled out with an equal and opposite component representing the marginal gain in revenue to the operator from the increase in F (see line prior to expression (4A.17) in section 4A.3 of the chapter appendix).

$$ATC - MSC = AC_p - q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} \tag{4.44}$$

Expressions (4.43) and (4.44) are identical, thus :

$$s = ATC - MSC \tag{4.45}$$

i.e. optimal unit subsidy is measured by the gap between the "long run" average total cost and marginal cost curves (Mohring, 1972; Findlay, 1983).

Next, noting from section 4.2.2(a) that $\frac{C_p}{VK} = \frac{\partial C_p}{\partial VK}$, and $\frac{\partial VK}{\partial F} = 2L_r$ then (4.3) can be written as $C_p = \frac{\partial C_p}{\partial VK}\frac{\partial VK}{\partial F}F = \frac{\partial C_p}{\partial F}F$. Substituting into (4.43) :

$$s = \frac{\left(\partial C_p/\partial F\right)F}{q} - q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}$$

Then, substituting for $\partial C_p/\partial F$ from (4.41) yields (Findlay, 1983) :

$$s = -\left(\frac{\partial u}{\partial F} + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial F}\bigg|_{\bar{q}} + \frac{\partial v}{\partial F}\bigg|_{\overline{LF},\bar{q}}\right)F - q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} \tag{4.46}$$

Rearranging :

$$s = \frac{\partial u}{\partial F}F - \left(\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial F}\bigg|_{\bar{q}}F + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}q\right) - \frac{\partial v}{\partial F}\bigg|_{\overline{LF},\bar{q}}F \tag{4.47}$$

Then, with v being homogeneous of degree zero (Else, 1985; Gwilliam *et al*, 1985), the term in brackets in (4.47) = 0, and (4.47) reduces to (Else, 1985; Gwilliam *et al*, 1985) :

$$s = -\left(\frac{\partial u}{\partial F} + \frac{\partial v}{\partial F}\bigg|_{\overline{LF},\bar{q}}\right)F \tag{4.48}$$

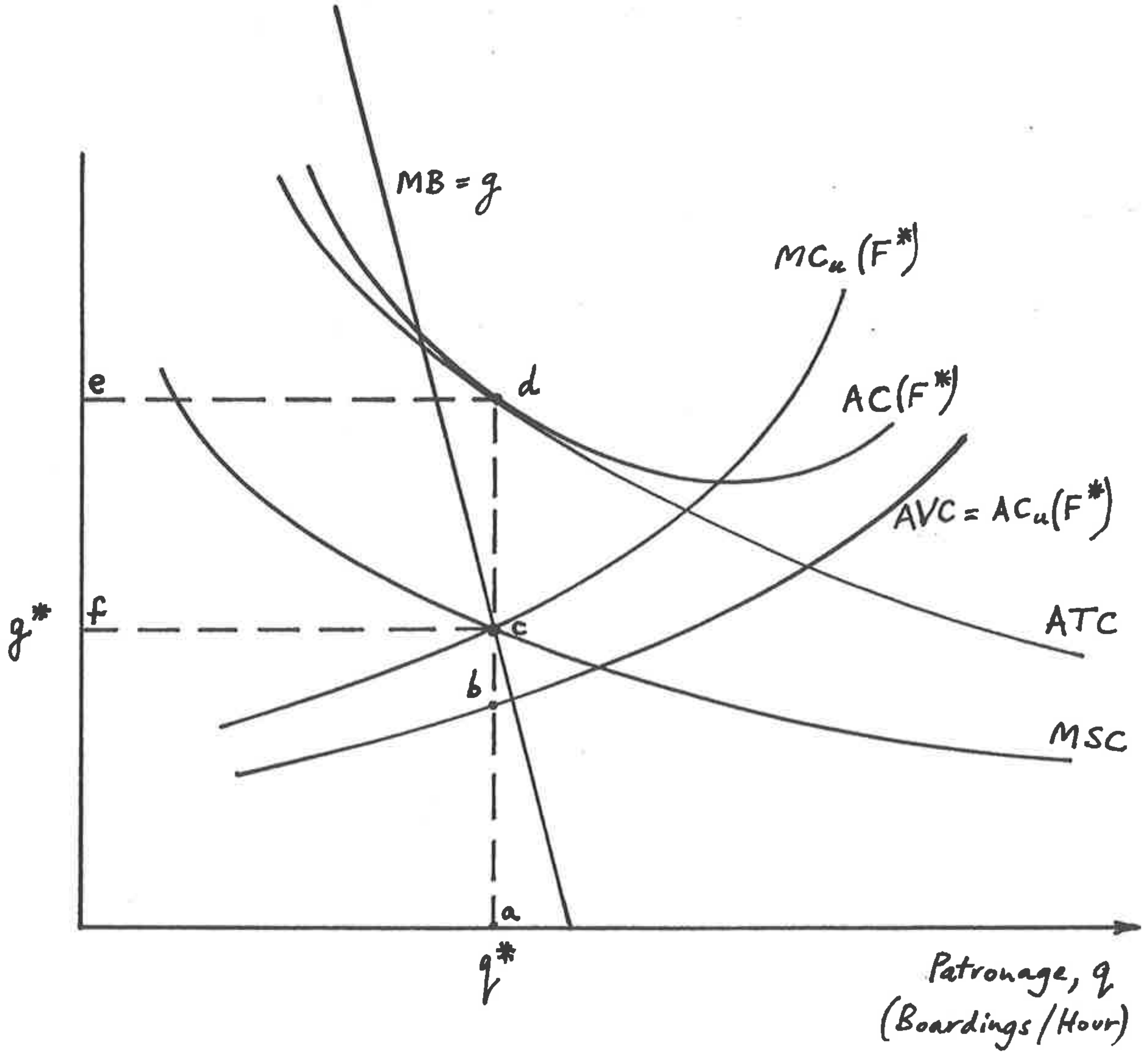
i.e.

$$s^* = -\left(\frac{\partial u}{\partial F}(F^*) + \frac{\partial v}{\partial F}\bigg|_{\overline{LF},\bar{q}}(F^*)\right)F^* \tag{4.49}$$

Here, $\frac{\partial u}{\partial F} < 0$, $\frac{\partial v}{\partial F} < 0$ and $F^* > 0$, thus $s^* > 0$.

The case 3 outcomes have been illustrated diagrammatically by Mohring (1972) and Findlay (1983) as illustrated here in Figure 4.2. In the figure, the *ATC* and *MSC* curves plot average total cost and marginal cost based on frequency (*F*) having been optimised at any given q level. With bus size (*N*) given, *F* is equivalent to plant size in conventional micro cost analysis, so *ATC* and *MSC* equate to long run average and marginal cost curves. For the generalised cost demand curve shown (*MB*), q^* coincides with the intersection of *MB* and *MSC*. F^* is chosen to ensure a tangency

Figure 4.2 : User Economies of Scale, Mohring Presentation (*LF/N* Case 3)



between the short run and long run average cost curves, $AC(F^*)$ and ATC, at q^*. Recalling that all variable costs are user costs, and all fixed costs are producer costs, the average variable cost curve (AVC) measures average user cost, whilst the gap between the $AC(F^*)$ and $AVC(F^*)$ curves measures average fixed cost, AC_p. Thus at q^*, with $AC_u(F^*)$ equal to distance *ab*, a price equal to distance *bc* is required to ensure q^* is achieved. Note that this price is exactly equal to the marginal negative passenger congestion externality, the distance between curves $MC_u(F^*)$ and $AC_u(F^*)$. The optimal unit subsidy is the distance *cd*, i.e. AC_p less the marginal negative externality.[24] Finally, optimal total subsidy is area *cdef*.

4.4.5 Case 4 : *LF* and *N* Both Variable

The final case to be considered, case 4, is where both *LF* and *N* can vary. This adds a third FOC, $\frac{\partial C_p}{\partial N} = 0$, to the two already given in section 4.4.4. The third FOC (evaluated in section 4A.4 of the chapter appendix), yields :

$$-q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial N}\bigg|_{\bar{q}} = \frac{\partial C_p}{\partial N} \tag{4.50}$$

All three FOCs (4.40), (4.41) and (4.50) must now hold for optimal solutions. The subsidy equations of case 3 also apply here, but yield different results because the optimal policy variables will take different optimal values.

4.5 Comparisons of Load Factor/Bus Size Cases

Having established the optimal solutions to the range of *LF/N* fixity cases discussed above, an important question is to what extent do subsidy and other optimal results vary from one case to another, and can the various cases be related diagrammatically. The purpose of this section is to commence to address this question through a quantitative comparative analysis of the various cases.

For the illustrative purpose at hand here, the analysis is presented using a simple model of frequency related user cost (AC_F), the key influence on user economies of scale. The model

[24] The unit subsidy exists due to the positive externality in user costs, but the size of unit subsidy is now also partly offset by the negative passenger congestion externality.

adopted is the random user behaviour model (discussed in section 3.5 of chapter 3).[25] It is also assumed that $\sigma = 0$ for simplicity, thus $SSDC_r = 0$, leaving $u = FDC_r$ and $v = SDDC$. The resulting user cost model is therefore the conventional simple random model used in much of the subsidy literature. With $H = 60/F$, from (3.47) and (3.49), u and v become :

$$u = \frac{30v_w}{F} \tag{4.51}$$

$$v = LFmult\frac{30v_w}{F} = LFmult.u \tag{4.52}$$

Thus
$$AC_F = u + v = \frac{30v_w}{F}(LFmult + 1) \tag{4.53}$$

These expressions can be used to quantify cases 3 and 4. In case 1 and 2, where LF is given, $F = qA/(\overline{LF_T}.N)$, thus (4.53) becomes :

$$AC_F = \frac{30v_w\overline{LF_T}N(\overline{LFmult} + 1)}{qA} \tag{4.54}$$

Equation (4.54) can be used to quantify case 1. For case 2, solving the FOC (4.39) using (4.54) and (4.22) yields :

$$N^* = \sqrt{\frac{L_r c_1 q A^2}{15v_w LF_T^2(LFmult + 1)}} \tag{4.55}$$

and thus from (4.36)
$$F^* = \sqrt{\frac{15v_w(LFmult + 1)q}{L_r c_1}} \tag{4.56}$$

Thus in case2, F and N are now both $\propto \sqrt{q}$, in contrast to case 1 where $F \propto q$. A similar square root type relationship has been previously reported for the case of variable LF and fixed N (Mohring, 1976), i.e. case 3 here.

The quantitative comparative analysis of the various cases was undertaken using the parameter values derived in appendix B. To keep the analysis general, a stylised route was considered, with passenger flow in one direction only (thus the directional split parameter, $d = 1.0$), but with trip and route length characteristics of Adelaide buses ($L_t = 8$ kms, and $L_r = 16$ kms, see appendix B). Results for cases 1 and 3 were evaluated for the range of bus sizes outlined in section B.3.2 of appendix B (mini, midi, rigid, and artic). For cases 2 and 4, on the other hand, N is optimised. The outcomes of the comparative analysis are summarised via a number of Figures (4.3

[25] The sensitivity of the chapter's results to alternate use of a planned or logit user behaviour model is discussed toward the end of the chapter.

to 4.11) discussed below. The format that follows consists of a discussion of each figure, with the key points drawn from the analysis summarised as a series of "results" statements (indented and in italics) at relevant points in the text.

Throughout the discussion, the fixing of *LF* and/or *N* will be referred to as cases of constrained optimisation. The comparative base for the analysis is the results derived from the unconstrained case 4, where both *LF* and *N* are variable. With no constraints, the case 4 optimisation problem will yield the highest possible net benefits. If economic efficiency is the aim, then case 4 optimal outcomes are the outcomes one would wish to strive for.[26] The results of the other cases (1, 2 and 3) then represent outcomes when one or more constraints enter the optimisation problem. Using the long run vs short run distinction of microeconomics cost analysis, case 4 represents true long run analysis, with the other cases being different versions of short run analysis. The discussion first compares case 4 with the single constraint cases (case 2 (*LF* fixed) and case 3 (*N* fixed)), and then extends the comparison to the double constraint case (case 1, where both *LF* and *N* are given).

Consider first Figure 4.3, which presents a family of average total cost, *ATC* (or simply *AC*), curves for cases 3 and 4.[27] All curves but one are for case 3. Each *AC3* curve plots *AC* for *one* bus size. On the other hand, the single *AC4* curve plots *AC* with *N* optimised, with each point along *AC4* coinciding with a different optimal bus size. The important thing to note from Figure 4.3, is that the *AC4* curve is the *envelope* of the *AC3* curves. A similar envelope pattern can be observed in Figure 4.4, where cases 2 (fixed *LF*) and 4 are compared. There, the *AC4* curve forms an envelope to the various *AC2* curves. Thus the *AC* curve for the no constraint case (4), forms an envelope to all the *AC* curves of the two single constraint cases (2 and 3).

This enveloping behaviour is also apparent when one compares case 1 (a two constraint case) with case 2 (a single constraint case). This is evident in Figure 4.5, which plots one of the

[26] Of course, in some situations, it is not always possible to strive for case 4. For example, bus size may be difficult to alter in the short run for a host of reasons (see discussion in section 4.2.1). In this respect, the results of all the cases considered have some role to play in subsidy analysis.

[27] The notation adopted throughout is to follow the abbreviation of the variable being considered by a number (1 to 4) representing the relevant *LF/N* case which applies. Thus, for example, *AC1* is average total cost for case 1.

Figure 4.3 : Average Total Cost Curves, *LF/N* Cases 3 and 4

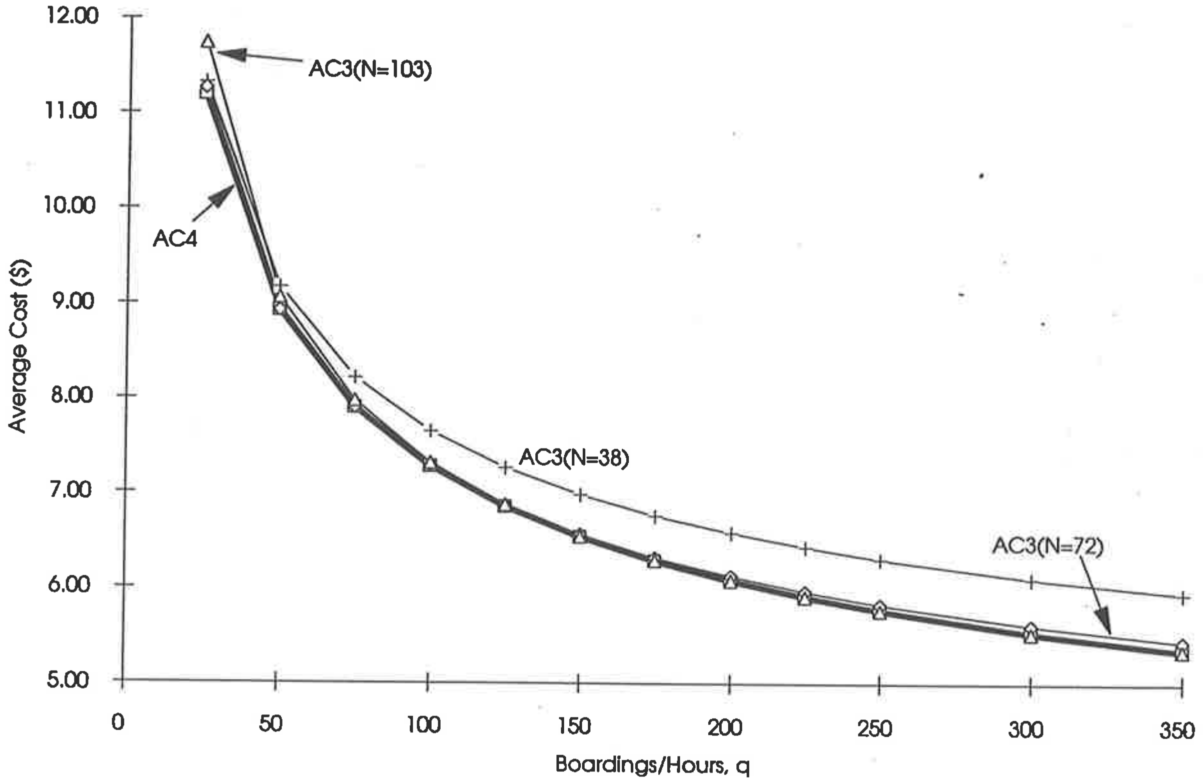


Figure 4.4 : Average Total Cost Curves, *LF/N* Cases 2 and 4

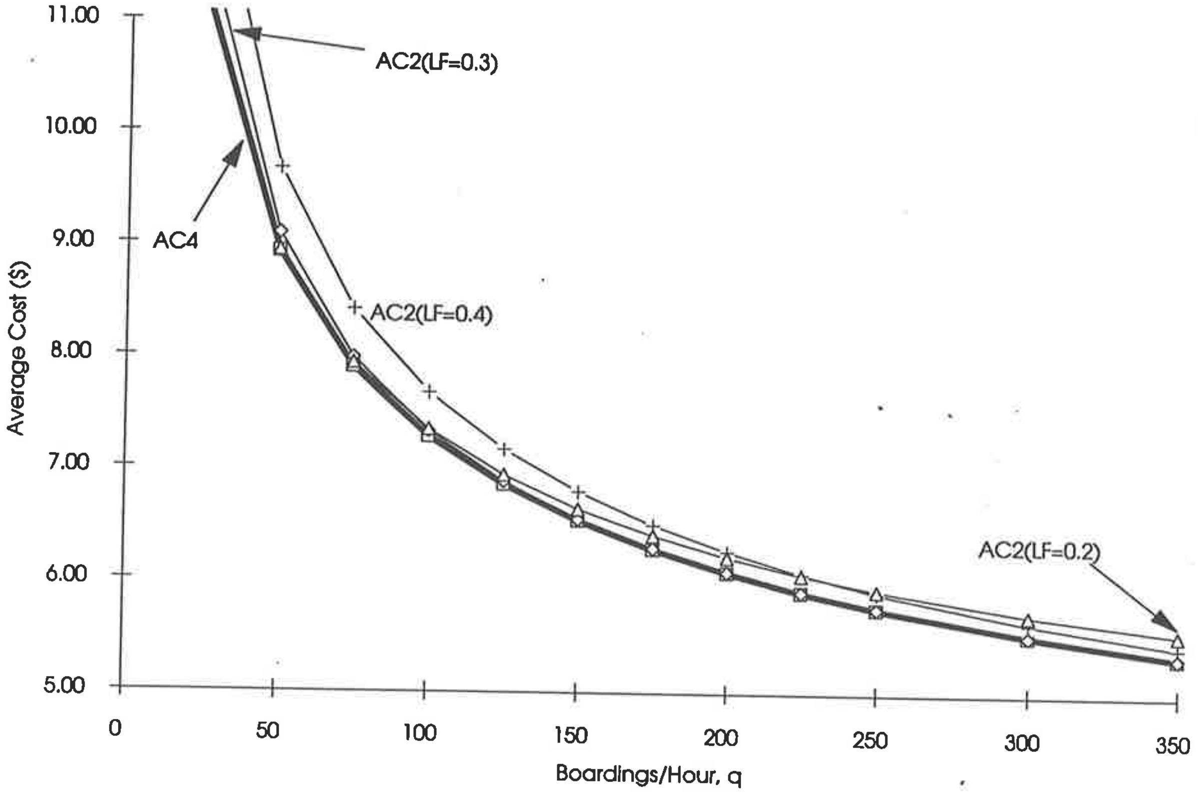
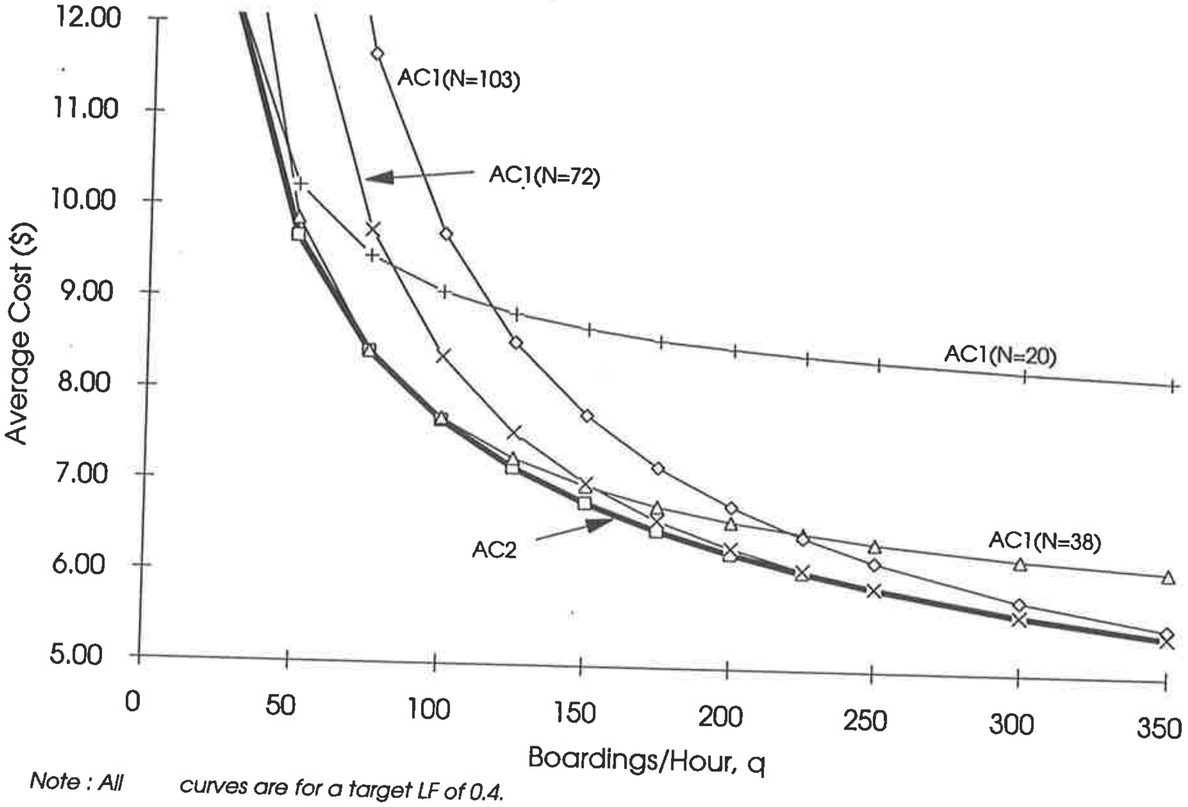


Figure 4.5 : Average Total Cost Curves, *LF/N* Cases 1 and 2



Note : All curves are for a target LF of 0.4.

AC2 curves from Figure 4.4 (i.e. for *LF* = 0.4) acting as an envelope to a family of *AC1* curves for various bus sizes at that given *LF*.

The general result that can be observed from these three figures is :

> ***RESULT 1:*** *Whenever a constraint is removed in the bus optimisation problem, the resulting average cost (AC) curve forms an* <u>*envelope*</u> *to the family of AC curves generated for various values of the previously constrained variable. Thus the AC curve resulting from the optimisation problem with i constraints forms an envelope to the various AC which result from the optimisation problem with i+1 constraints.*

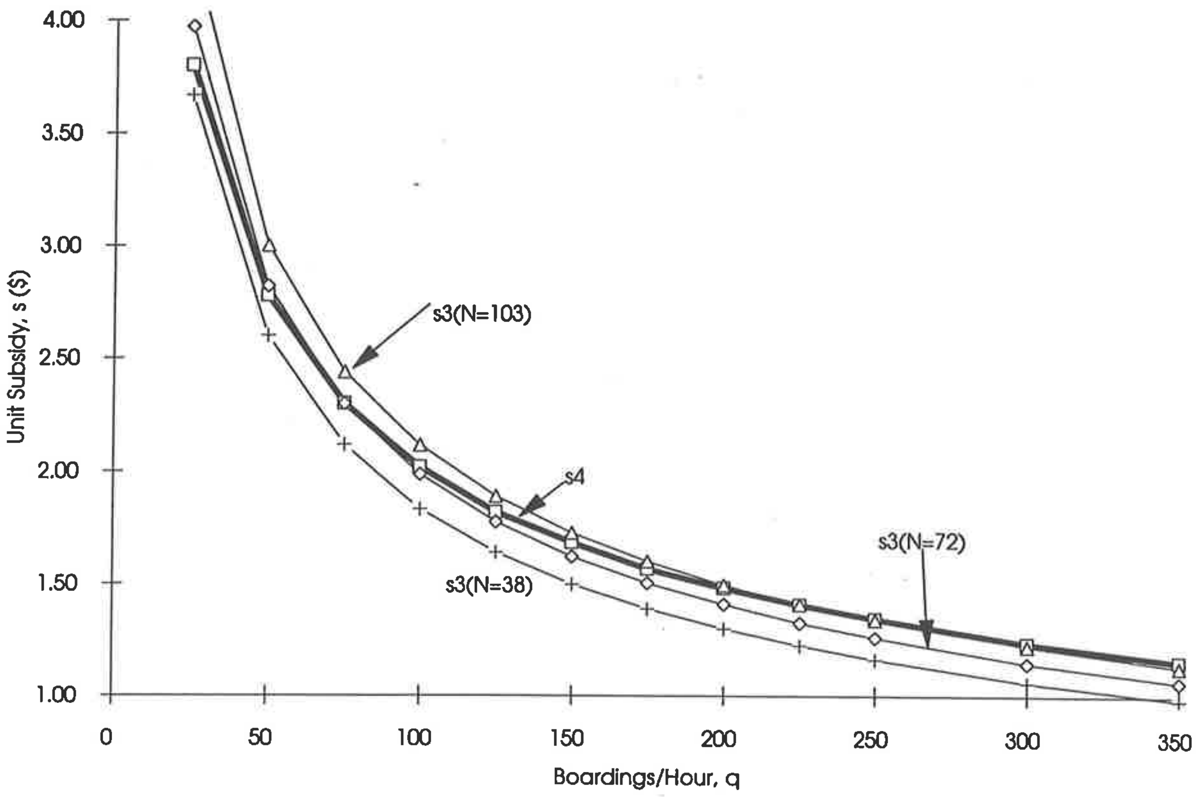
Note the consistency in the relationship between the *AC* curves of various cases above, with the relationship between short run and long run *AC* curves in standard microeconomic cost analysis where the envelope theorem is a key feature.

Result 1 is important because, through Figures 4.3 to 4.5, it is now possible to draw a link between the key diagrammatic presentations of the user economies of scale subsidy argument presented to date, namely Figures 4.1 and 4.2. In Mohring's figure (4.2 here), his long run average cost curve *ATC* coincides with one of the *AC3* curves in Figure 4.3, the curve for the bus size considered by Mohring. On the other hand, the *ATC* curve in Waters' figure (4.1 here) coincides with one of the *AC1* curves in Figure 4.5, the curve for the bus size assumed by Waters.

Next, consider the behaviour of optimal unit subsidy. Figure 4.6 reports unit subsidy (*s*) results for cases 3 and 4. As was the situation in the average cost comparisons, the Figure plots a family of unit subsidy curves,[28] with case 4 contributing only a single *s* curve, *s4*. There are two things to note from Figure 4.6. First, note the decline in unit subsidy as patronage (*q*) increases, the conventional negative relationship between unit subsidy and demand reported in the literature and discussed in section 2.6 of chapter 2. Second, the *s4* curve cuts through the family of *s3* curves, cutting each *s3* curve once. Close inspection reveals that each *s3* curve cuts the *s4* curve at a *q* value which coincides with the tangency in Figure 4.3 between the corresponding *AC3* curve and

[28] The notation of the relevant case for each curve is the same as that used in the *AC* curve discussion (see footnote 27).

Figure 4.6 : Optimal Unit Subsidy, *LF/N* Cases 3 and 4



the *AC4* curve.[29] Each *s3* curve therefore has only one point on it which is genuinely optimal, namely, the point at which the given bus size (*N*) would have resulted as an optimal outcome if *N* had been allowed to also be optimised. The *s4* curve is thus the locus of these single optimal points. A further important feature in the figure is that the case 4 unit subsidy curve is flatter than all the case 3 unit subsidy curves (see result 2 shortly).

A similar pattern of results is also apparent in Figures 4.7 and 4.8 which plot unit subsidy for comparisons of cases 2 and 4, and cases 1 and 2 respectively. In each figure, the unit subsidy curve for the case with fewer constraints cuts the family of more constrained curves at points coinciding with average cost tangency points in the associated Figures 4.4 and 4.5. In addition, the less constrained is the case, the flatter is the unit subsidy schedule. The end result is that, although all unit subsidy curves in Figures 4.6, 4.7 and 4.8 display the conventionally reported negative slope, the *s4* curve is flatter than the *s2* and *s3* curves, and the *s2* curves are flatter than the *s1* curves. Thus :

> ***RESULT 2 :*** *The less constrained the bus optimisation problem (i.e. the more "long run" it is), the less pronounced will be the rate of decline in unit subsidy as patronage level increases.*

The reason for this is that, with unit subsidy being a function of the rate of change in average cost (*AC*), the less constrained the case, the less pronounced was the variation in the slope of the *AC* schedules in Figures 4.3 to 4.5, and thus the less pronounced the variation in unit subsidy. Result 2 is important because, with the negative relationship between optimal unit subsidy and patronage playing an important role in prescribing how optimal unit subsidy should vary between routes of different demand density (see discussion in section 2.6 of chapter 2), result 2 indicates that the strength of this relationship varies depending on the degree of constraint which applies in the optimisation problem.

Another couple of important results can be gleaned from Figures 4.3 to 4.8. First, note that when one constraint is applied (either a given *LF* (case 2) or given *N* (case 3)), the divergence in average cost and unit subsidy results from those obtained in the unconstrained case 4 are relatively

[29] Note again the close similarity with conventional microeconomic short vs long run cost curve analysis. The relationship here between *s1* and *s2* curves is identical to the relationship between short and long run marginal costs in micro cost analysis.

Figure 4.7 : Optimal Unit Subsidy, *LF/N* Cases 2 and 4

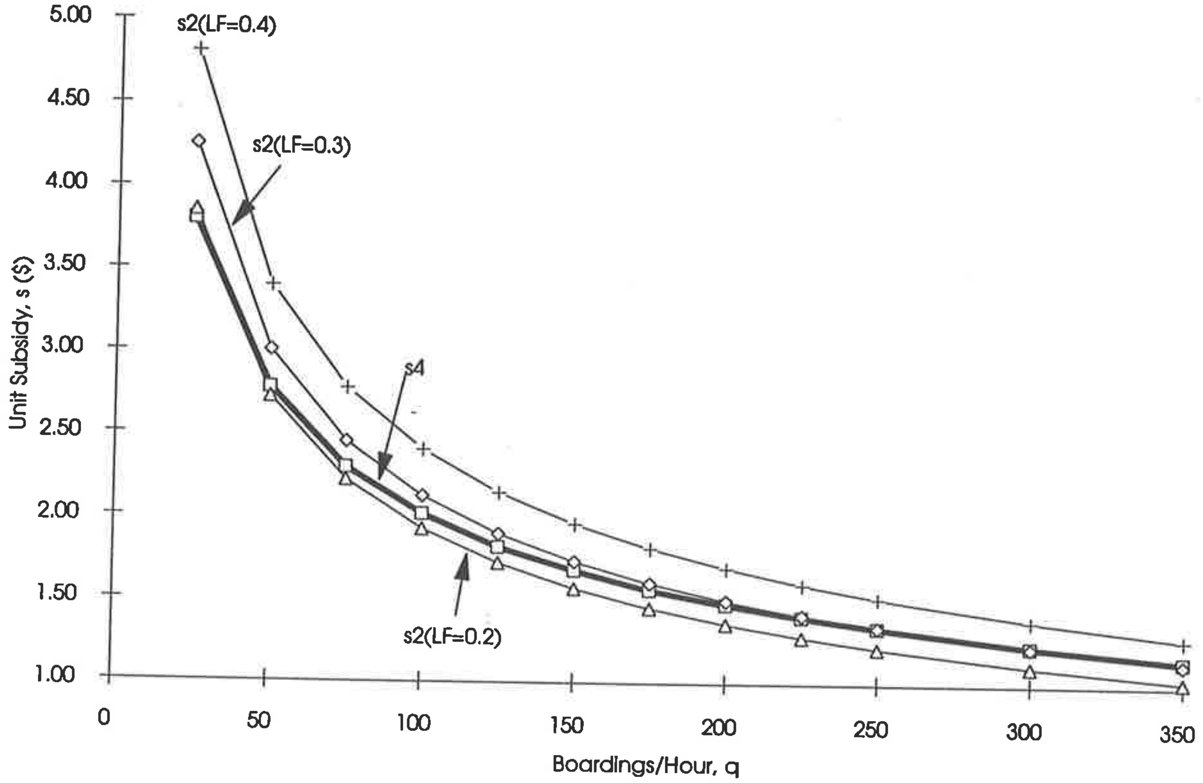
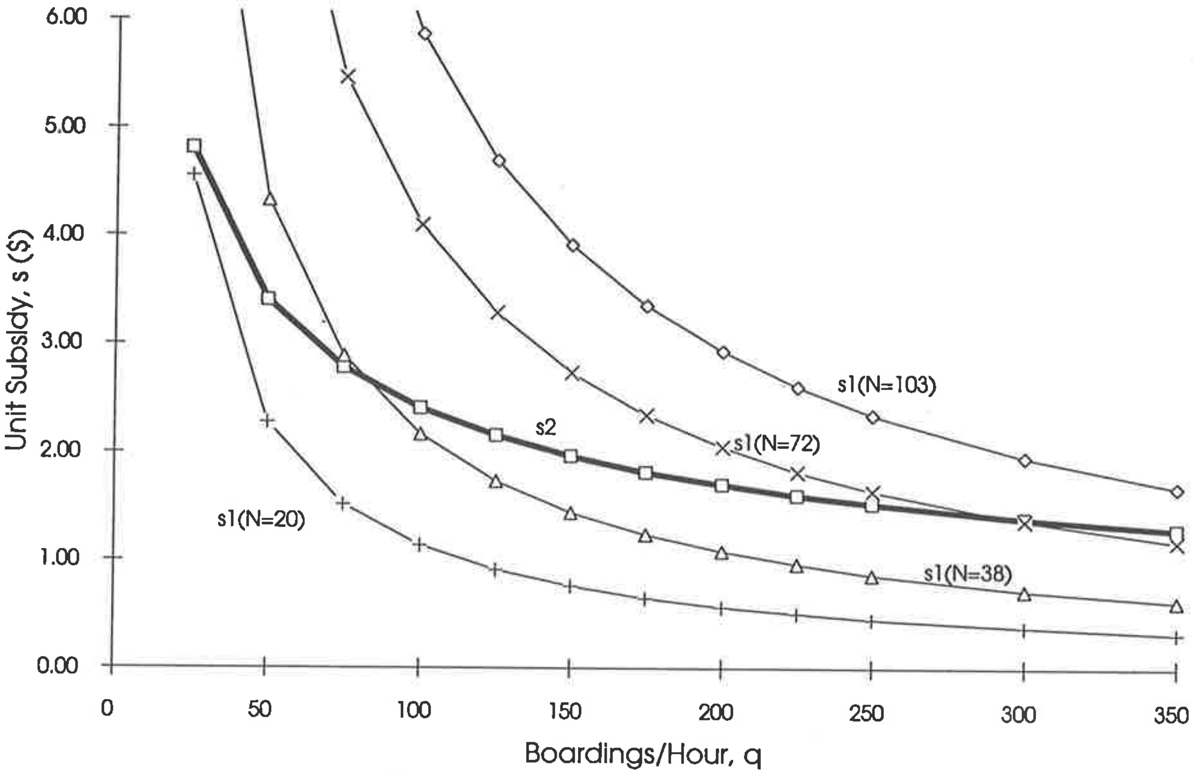


Figure 4.8 : Optimal Unit Subsidy, *LF/N* Cases 1 and 2



Note : All curves are for a target LF of 0.4.

modest. The close bunching of curves in Figures 4.3, 4.4, 4.6 and 4.7 indicate this. In contrast, adding a second constraint, i.e. moving to case 1 (where both *LF* and *N* are given), distorts average cost and unit subsidy results away from case 4 results much more substantially. This is evident from Figures 4.5 and 4.8. Thus :

> ***RESULT 3 :*** *Introducing a single constraint in the bus optimisation problem (either a given load factor or bus size) distorts average cost and unit subsidy results away from unconstrained results by a relatively modest amount. Single constraint optimisations are therefore a reasonable approximation of the unconstrained problem. Introduction of a second constraint, however, has a much more substantial impact, making an optimisation with both load factor and bus size fixed a much poorer approximation of the unconstrained problem.*

A final point to note from the above Figures results from Figures 4.6 and 4.8. Recall (from section 2.5 of chapter 2) Walters' (1982) critique of Mohring's (1972) original analysis, that unit subsidy would be much smaller when small buses are used. Figures 4.6 and 4.8 are both consistent with this idea since lowering the bus size leads to a monotonic fall in the unit subsidy schedule. However, the size of this fall is quite modest if *LF* is allowed vary (Figure 4.6).

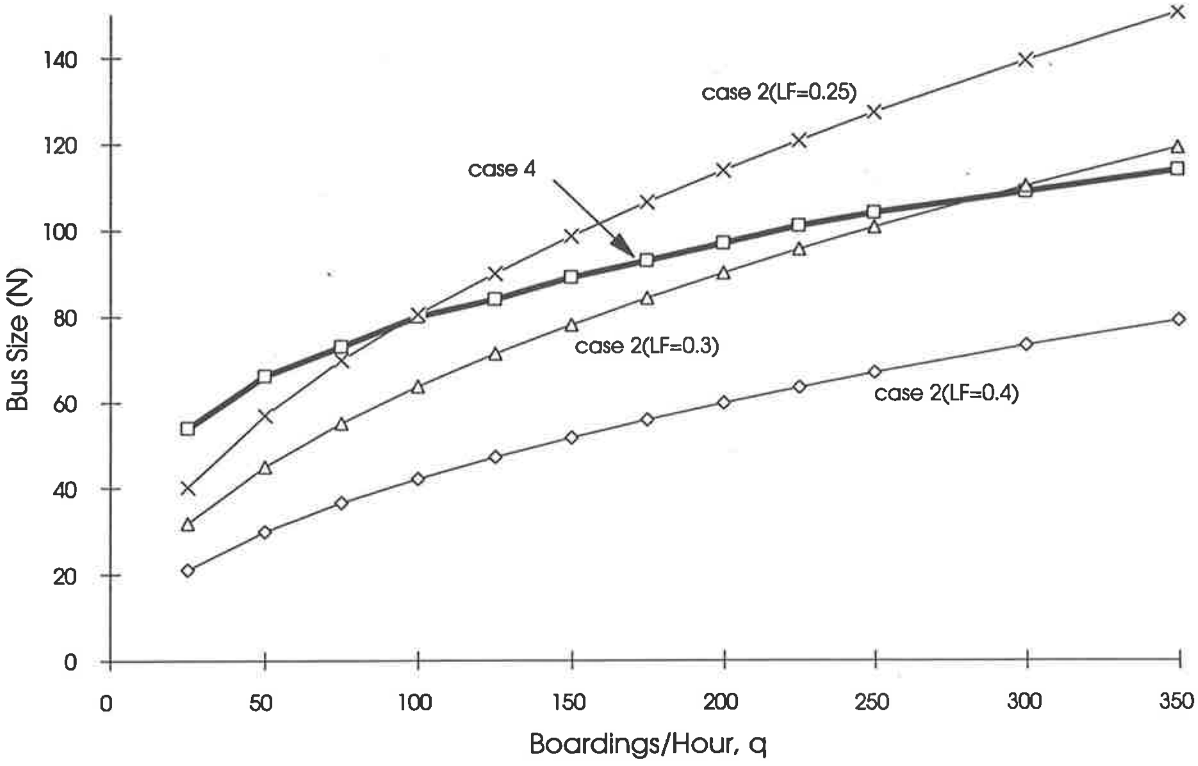
Next, consider the bus sizes generated in the cases where bus size is optimised, i.e. cases 2 and 4. Figure 4.9 reports four optimal bus size (N^*) schedules. Three schedules are for case 2 (where *LF* is given), for *LF* values of 0.4, 0.3 and 0.25. A fourth schedule reports results for case 4, where *LF* and *N* are simultaneously optimised. There are several features to explain.

First, for case 2, for any given load factor, as patronage increases, so too does N^*. This outcome can be explained with reference to the trade-off between user costs and producer costs which is inherent in the optimality condition (4.39). Dividing both sides of (4.39) by q, (4.39) becomes :

$$-\left.\frac{\partial AC_F}{\partial N}\right|_{\bar{q}} = \left.\frac{\partial AC_p}{\partial N}\right|_{\bar{q}} \qquad (4.57)$$

That is, at the margin, the reduction in frequency related user cost brought about by a reduction in bus size must be equated to the corresponding increase in average producer cost (which (4.22) indicates will occur). Commencing at an optimal position where (4.57) holds, an increase in patronage can be accommodated by an increase in frequency (F). However, the rise in F causes

Figure 4.9 : Optimal Bus Size, *LF/N* Cases 2 and 4



user cost to fall, thus reducing the marginal returns of bus size reduction, the LHS of (4.57). With (4.57) then out of balance, bus size must then be increased, raising the LHS of (4.57) and reducing the RHS, until an optimum is reached once more. Thus, for a given *LF*, when patronage increases, it is optimal to accommodate the increase with an increase in *both* frequency and bus size.

A second feature of the case 2 curves in Figure 4.9 is that, for a given *q*, the higher is the target *LF*, the lower is N^*. The reason for this is as follows. Once again start in an optimal situation. If target *LF* suddenly increases, the same passengers can be accommodated with fewer buses. The lower frequency, however, raises the marginal benefit of bus size reduction (the LHS of (4.57)), making (4.57) out of balance. Reducing bus size causes the LHS of (4.57) to fall, and the RHS to rise until a balance is restored.

Finally, and most importantly, optimal bus size is less responsive to changes in patronage level in case 4 than in case 2, particularly compared to low *LF* case 2 situations. Thus :

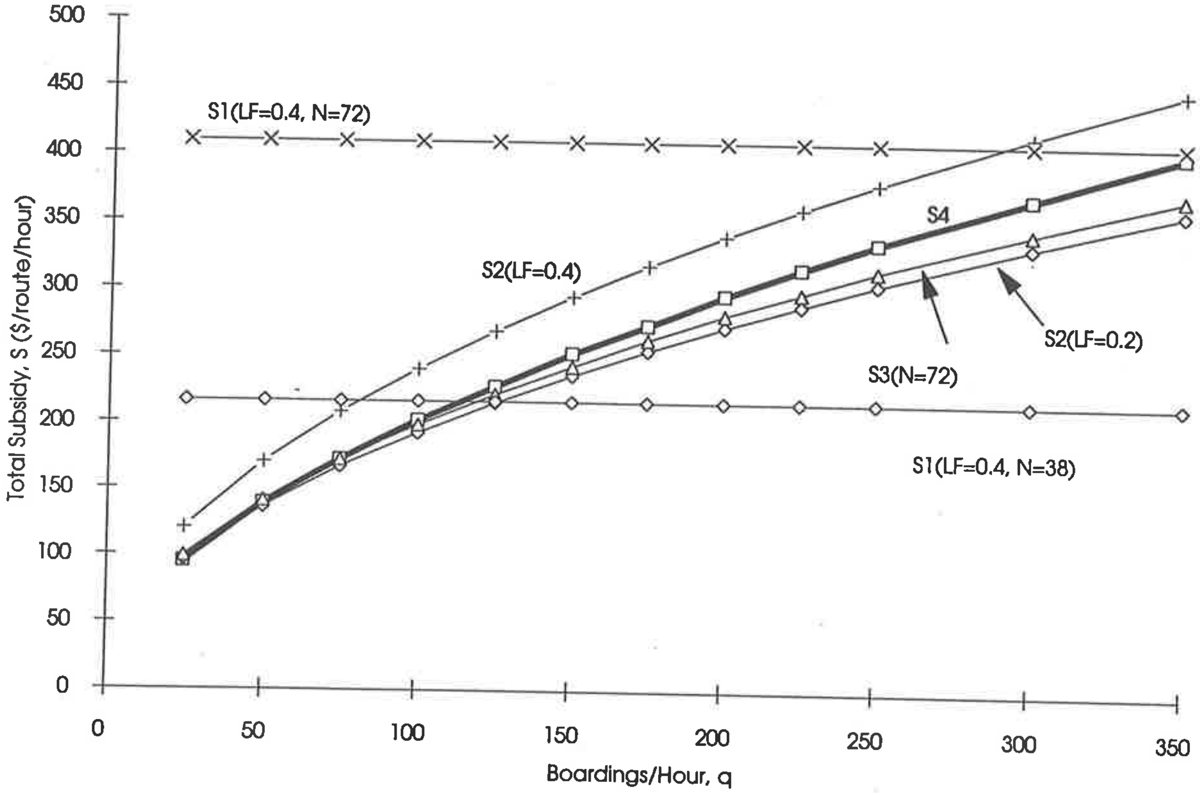
> ***RESULT 4 :*** *When bus size can be optimised, optimal bus size is less responsive to changes in patronage level when load factor can also be varied. Notwithstanding this, optimal bus size always increases with increases in patronage.*

The lower responsiveness occurs because, with an additional dimension of adjustment in case 4, additional patronage can be catered for by a joint increase in both bus size and load factor, rather than just bus size.

A number of points can also be noted about conventional urban bus sizes in developed countries, which for example in Adelaide is 78 passengers (seating plus standing, see section B.3.2 of appendix B). Quite clearly, at low demand levels, small buses have a role to play. At medium and higher demand levels, Figure 4.9 suggests that a case for smaller urban buses can only be made if a reasonably high target load factor is adopted (e.g. 0.4 or greater). If, on the other hand, *LF* can be varied (case 4), conventional buses may not be too big, and may in fact be too small in some high demand circumstances. As noted in section 2.5 of chapter 2, these outcomes are not inconsistent with the literature, in which the issue of optimal bus size remains somewhat unresolved.

Next consider the behaviour of *total* subsidy, *S*, which is reported in Figure 4.10. The figure plots a single *S* schedule for cases 3 and 4, two case 1 schedules ($LF = 0.4$, $N = 72$ and 38),

Figure 4.10 : Optimal Total Subsidy, All *LF/N* Cases



and two case 2 schedules (LF = 0.2 and 0.4). An interesting result is the constant total subsidy which occurs for any given case 1 situation,[30] although the constant level of S varies between case 1 situations. In contrast, total subsidy for all other cases grows as patronage increases. Note in particular the close similarity in S results between the unconstrained case 4 and case 3 with N= 72, a bus size close to the current average in Adelaide.

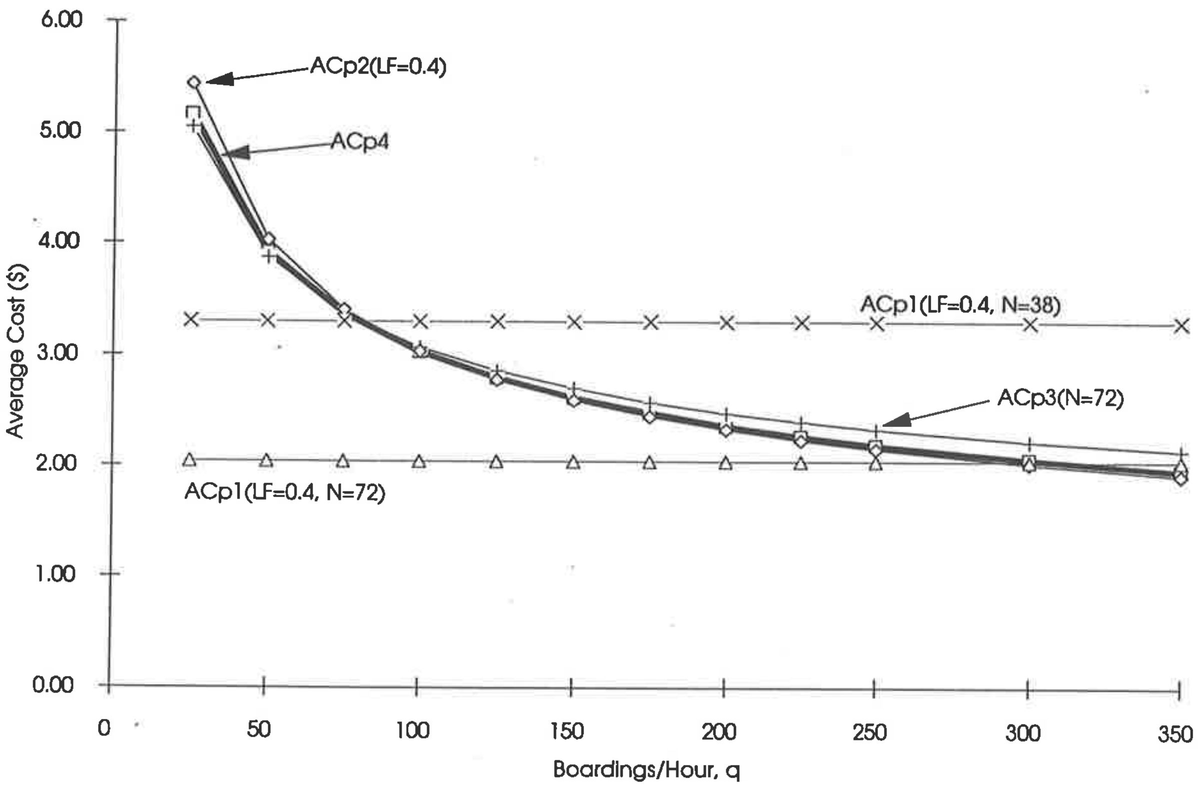
Finally, consider the composition of the economies of scale which underlie the bus subsidy argument and results. In all the LF/N fixity cases, as patronage increases, frequency also increases and thus headway falls. However, from chapter 3, a fall in headway leads in turn to average (frequency related) user cost, AC_F, falling, i.e. the Mohring effect. That is, there are user cost economies with respect to patronage. How, however, does average *producer* cost behave as patronage increases ? This is illustrated in Figure 4.11, which presents average producer cost curves for all four LF/N fixity cases. As discussed in section 4.4.2, when LF and N are given (case 1), $AC_p = MC_p$ and is constant, with AC_p and MC_p schedules being horizontal as in Figure 4.1. As a result, for case 1 situations, there are no producer cost economies with respect to patronage. Two corresponding case 1 AC_p curves are drawn in Figure 4.11, for N = 72 and 38. In contrast, the three other LF/N fixity cases (2, 3 and 4) each produce declining AC_p schedules, as illustrated in Figure 4.11.[31] This occurs because in each of these three cases, when q grows, F^* grows less than proportionally (unlike case 1 where $F^* \propto q$).

As a result, although the fall in average total cost (AC in Figures 4.3, 4.4 and 4.5) is due purely to user cost economies in case 1, for cases 2, 3 and 4, the economies of scale are due to declines in *both* AC_u *and* AC_p. That is, for case 2, 3 and 4 there are economies of scale *with respect to patronage* in *both* producer *and* user costs. Thus :

[30] This is a special feature of case 1 when the simple random model, with service unreliability $\sigma = 0$, is used. As will be seen in chapter 5, planned and logit user cost models yield a more conventional upward sloping S schedule for case 1.

[31] The declining nature of AC_p leads to a further point about Waters (1982a) diagrammatic presentation, on which Figure 4.1 here was based. Waters (1982b) argues that Figure 4.1, which is based on case 1 (LF and N given) and thus features a horizontal AC_p curve, also applies to more realistic cases where capacity (F) does *not* expand in direct proportion with q. However, Figure 4.11 suggests that each of the other three cases considered here conflict with this Waters' conclusion insofar as AC_p is no longer constant.

Figure 4.11 : Average Producer Cost, All *LF/N* Cases



***RESULT 5** : When both load factor and bus size are given, economies of scale exist in user costs only. When either or both load factor and bus size are allowed to vary, however, in addition to user cost economies, there are concurrently also producer economies of scale with respect to patronage, even though there are constant returns to producer costs with respect to vehicle-kms.*[32]

An interesting question is whether in this study, for cases 2, 3 and 4, the decline in average user and producer costs as patronage increases should be referred to as two separate types of economies. The approach adopted here is not to draw out this distinction, due to the fact that the declining nature of average producer cost is a result of the influence of user cost on frequency in the optimisation process. Consequently, the term *user economies of scale* will be used throughout to describe the overall economies of scale with respect to patronage of the combined total (user plus producer) cost.

Before concluding, it is important to consider how the results of this chapter might vary if, instead of using a random user cost model, a planned or logit user cost (see chapter 3) model had been used instead. This question is partly answered by the results of subsequent chapters, where these alternate models, particularly the logit model are considered, but some general points can be made here. The five key results of chapter 4 all flow from the family of average cost (*AC*) curves generated for the various *LF*/*N* fixity cases considered. If the random user model were replaced by a planned users model, the starting point for the chapter 4 analysis would be a new set of *AC* curves. The *general* shape of the *AC* curves would, however, be similar for both user cost models, and so one could expect the thrust of chapter 4 results to continue to hold.

Chapter 3 illustrated, however, that when the planned user cost model is used, user cost is less responsive to changes in service level (see Figure 3.8) than in the random users model. Therefore, one would expect the average cost schedules in the chapter 4 analysis to be flatter if a planned user cost model were used. With optimal unit subsidy being a function of the responsiveness of *AC*, unit subsidy would therefore be smaller and less responsive under a planned

[32] Note that it is the distinction highlighted in section 4.2.2 between intermediate services (veh-kms) and final services (trips, or patronage) (Small, 1992) that facilitates this result.

users model than those generated here in chapter 4. With the benefits of altering bus size also a function of the responsiveness of *AC*, optimal bus size may also be less responsive under a planned model.

On the whole, therefore, one could expect the general thrust of the results in this chapter to also apply under a planned users model, although the scale of the results is likely to be smaller. It is more difficult to draw firm conclusions about shifting to a logit user cost model. To the extent that the logit model yields user cost outcomes which are a combination of those of the random and planned models, the general thrust of the results in this chapter are also likely to hold under a logit user cost model. However, as will become evident in the next chapter, there can be some localised "folding" of average cost and unit subsidy schedules, and this may produce some localised folding in other variables and the results presented in chapter 4.

4.6 Chapter Summary and Conclusions

This chapter has had several aims. The main one was to set out and summarise the first best bus optimisation problem which generates user economies of scale, and its solution, thus providing a sound basis for the analytical chapters which follow in this study. A second aim was to consider and relate the user economies of scale concept for a number of load factor/bus size fixity cases, thus enabling various presentations in the literature to be linked. A final aim was to summarise and extend the graphical analysis of user economies of scale.

Four analytical frameworks were considered, differing with respect to the degree of constraint to which the bus optimisation problem is subject, where the degree of constraint is measured by the extent to which either or both load factor (*LF*) and bus size (*N*) are allowed to vary. The first best optimal pricing, service and subsidy optimisation formulation was derived and solved for each case. A quantitative analysis was then undertaken, using a simple user cost model for illustrative purposes, and the results for the four cases compared. The key findings are summarised below.

Result 1 : It was demonstrated that, from a diagrammatic perspective, the various *LF*/*N* fixity cases can be linked through the use of average total cost envelope curves, similar to the way short run and long run average cost curves are linked in conventional cost analysis. It was shown that the average cost curve of a less constrained optimisation case forms an envelope to the family of average cost curves for more constrained cases. Importantly, the enveloping property of average cost curves then allows the diagrammatic presentations of user economies of scale subsidy currently found in the literature to be integrated into a broader diagrammatic framework.

Result 2 : A well known and important rule in the literature on user economies of scale subsidy, is that optimal unit subsidy declines with the level of patronage. Although this rule continued to hold here, the rate of decline was found to vary with the degree of optimisation constraint. The less constrained was the bus optimisation problem (i.e. the more "long run" it is), the less pronounced was the rate of decline in unit subsidy as patronage increases. This result is due to unit subsidy being a function of the slope of the average cost vs patronage schedule, plus the fact that the slope of the average cost schedule varies less the less constrained is the *LF/N* case being considered (i.e. an implication of the envelope property of result 1). Result 2 is important because, given the important role of the negative relationship between optimal unit subsidy and patronage for relating optimal subsidy for routes of different demand density, result 2 indicates that the strength of this relationship varies depending on the degree of constraint which applies in the optimisation problem.

Result 3 : Although the ideal analytical approach is to undertake an unconstrained optimisation (provided the level of complexity is not excessive, and provided all policy variables can actually be varied), it was found that introducing a single optimisation constraint (either a given load factor or a given bus size) does not greatly distort optimal results away from those of an unconstrained analysis. A single constraint optimisation (with either load factor or bus size fixed) is therefore a reasonable approximation of the unconstrained problem. On the other hand, introduction of a second constraint has a much more substantial distorting impact on optimal outcomes.

Result 4 : Optimal bus size was found, in all cases, to increase with patronage. However, the rate of change of optimal bus size was found to be smaller when both load factor and bus size can be simultaneously optimised compared to when only bus size is optimised. The lower responsiveness occurs because marginal optimisation conditions can be better met by catering for the additional patronage by increasing both load factor and bus size concurrently, rather than just bus size.

Result 5 : The composition of economies of scale with respect to patronage varies depending on the extent to which the optimisation is constrained. In all cases, average *user* cost declines as patronage increases, that is there are user cost economies. On the other hand, the behaviour of average *producer* cost varies between *LF/N* fixity cases. When both load factor and bus size are given, average producer cost is constant and thus invariant to changes in patronage. However, if either or both load factor and/or bus size can be optimised, then average producer cost declines with increases in patronage, thus, in addition to user cost economies, there are also producer cost economies with respect to patronage (even though there are constant returns to scale in producer costs with respect to vehicle-kms). Notwithstanding this, the term *user economies of scale* will be used throughout the study to refer to the economies of scale in combined total (user plus producer) cost with respect to patronage.

Chapter Appendix :

4A.1 First Order Condition Derivation, Case 1 (*LF* and *N* both given)

The FOC $\frac{\partial ES}{\partial P} = 0$ is derived as follows. Setting $j = P$ in (4.19) and noting (4.22)

$$\frac{\partial ES}{\partial P} = \frac{\partial CS}{\partial g}\frac{\partial g}{\partial P} - \frac{\partial C_p}{\partial q}\frac{\partial q}{\partial P} + q + P\frac{\partial q}{\partial P} \qquad (4A.1)$$

Now, $$CS = \int_g^{\infty} q.dg \qquad (4A.2)$$

thus $$\frac{\partial CS}{\partial g} = -q \qquad (4A.3)$$

From (4.21) $$\frac{\partial g}{\partial P} = \left.\frac{\partial g}{\partial P}\right|_{\bar{q}} + \frac{\partial u}{\partial q}\frac{\partial q}{\partial P} + \frac{\partial v}{\partial q}\frac{\partial q}{\partial P}$$

$$= 1 + \frac{\partial (u+v)}{\partial q}\frac{\partial q}{\partial P} = 1 + \frac{\partial AC_F}{\partial q}\frac{\partial q}{\partial P} \qquad (4A.4)$$

Substituting (4A.3) and (4A.4) into (4A.1) :

$$\frac{\partial ES}{\partial P} = -q\left(1 + \frac{\partial AC_F}{\partial q}\frac{\partial q}{\partial P}\right) - \frac{\partial C_p}{\partial q}\frac{\partial q}{\partial P} + q + P\frac{\partial q}{\partial P}$$

$$= \left(-q\frac{\partial AC_F}{\partial q} - \frac{\partial C_p}{\partial q} + P\right)\frac{\partial q}{\partial P} \qquad (4A.5)$$

Setting $\frac{\partial ES}{\partial P} = 0$ yields the FOC result :

$$P = \frac{\partial C_p}{\partial q} + q\frac{\partial AC_F}{\partial q} \qquad (4A.6)$$

4A.2 First Order Condition Derivation, Case 2 (*LF* given, *N* variable)

The FOC $\frac{\partial ES}{\partial N} = 0$ is derived as follows. Setting $j = N$ in (4.19) and noting (4.38)

$$\frac{\partial ES}{\partial N} = \frac{\partial CS}{\partial g}\frac{\partial g}{\partial N} - \left.\frac{\partial C_p}{\partial N}\right|_{\bar{q}} - \frac{\partial C_p}{\partial q}\frac{\partial q}{\partial N} + P\frac{\partial q}{\partial N} \qquad (4A.7)$$

The third term on the RHS of the above line arises because if *N* changes, so does *u*(*q*/*N*) and *v*(*q*/*N*) and thus *g*, and thus *q*. Next, from (4.37) :

$$\frac{\partial g}{\partial N} = \left.\frac{\partial u}{\partial N}\right|_{\bar{q}} + \frac{\partial u}{\partial q}\frac{\partial q}{\partial N} + \left.\frac{\partial v}{\partial N}\right|_{\bar{q}} + \frac{\partial v}{\partial q}\frac{\partial q}{\partial N}$$

$$= \left.\frac{\partial (u+v)}{\partial N}\right|_{\bar{q}} + \frac{\partial (u+v)}{\partial q}\frac{\partial q}{\partial N}$$

$$= \left.\frac{\partial AC_F}{\partial N}\right|_{\bar{q}} + \frac{\partial AC_F}{\partial q}\frac{\partial q}{\partial N} \tag{4A.8}$$

Substituting (4A.3), (4A.6) and (4A.8) into (4A.7) :

$$\frac{\partial ES}{\partial N} = -q\left(\left.\frac{\partial AC_F}{\partial N}\right|_{\bar{q}} + \frac{\partial AC_F}{\partial q}\frac{\partial q}{\partial N}\right) - \left.\frac{\partial C_p}{\partial N}\right|_{\bar{q}} - \frac{\partial C_p}{\partial q}\frac{\partial q}{\partial N} + \left(\frac{\partial C_p}{\partial q} + q\frac{\partial AC_F}{\partial q}\right)\frac{\partial q}{\partial N}$$

$$= -q\left.\frac{\partial AC_F}{\partial N}\right|_{\bar{q}} - \left.\frac{\partial C_p}{\partial N}\right|_{\bar{q}} \tag{4A.9}$$

Setting $\frac{\partial ES}{\partial N} = 0$ yields the FOC result :

$$-q\left.\frac{\partial AC_F}{\partial N}\right|_{\bar{q}} = \left.\frac{\partial C_p}{\partial N}\right|_{\bar{q}} \tag{4A.10}$$

4A.3 First Order Condition Derivation, Case 3 (*N* given, *LF* variable)

(a) The FOC $\frac{\partial ES}{\partial P} = 0$ is derived as follows. The general forms of g and C_p given by (4.13) and (4.3) now apply. Setting $j = P$ in (4.19), and noting that $\frac{\partial C_p}{\partial P} = \frac{\partial C_p}{\partial q}\frac{\partial q}{\partial P}$, and that from (4.3) $\frac{\partial C_p}{\partial q} = 0$, then (4.19) reduces to :

$$\frac{\partial ES}{\partial P} = \frac{\partial CS}{\partial g}\frac{\partial g}{\partial P} + q + P\frac{\partial q}{\partial P} \tag{4A.11}$$

Now, from (4.13)

$$\frac{\partial g}{\partial P} = \left.\frac{\partial g}{\partial P}\right|_{\bar{q}} + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial P}$$

$$= 1 + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial P} \tag{4A.12}$$

Substituting (4A.3) and (4A.12) into (4A.11) yields :

$$\frac{\partial ES}{\partial P} = -q\left(1 + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial P}\right) + q + P\frac{\partial q}{\partial P}$$

$$= \left(-q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} + P\right)\frac{\partial q}{\partial P} \tag{4A.13}$$

Setting $\frac{\partial ES}{\partial P} = 0$ yields the FOC result :

$$P = q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} \tag{4A.14}$$

(b) The FOC $\frac{\partial ES}{\partial F} = 0$ is derived as follows. Setting $j = F$ in (4.19) and evaluating,

$$\frac{\partial ES}{\partial F}=\frac{\partial CS}{\partial g}\frac{\partial g}{\partial F}-\frac{\partial C_P}{\partial F}+P\frac{\partial q}{\partial F} \tag{4A.15}$$

Now, from (4.13) $$\frac{\partial g}{\partial F}=\frac{\partial u}{\partial F}+\frac{\partial v}{\partial LF}\left(\frac{\partial LF}{\partial q}\frac{\partial q}{\partial F}+\left.\frac{\partial LF}{\partial F}\right|_{\bar q}\right)+\left.\frac{\partial v}{\partial F}\right|_{\overline{LF},\bar q} \tag{4A.16}$$

Substituting (4A.3), (4A.14) and (4A.16) into (4A.15) :

$$\frac{\partial ES}{\partial F}=-q\left(\frac{\partial u}{\partial F}+\frac{\partial v}{\partial LF}\left(\frac{\partial LF}{\partial q}\frac{\partial q}{\partial F}+\left.\frac{\partial LF}{\partial F}\right|_{\bar q}\right)+\left.\frac{\partial v}{\partial F}\right|_{\overline{LF},\bar q}\right)-\frac{\partial C_P}{\partial F}+q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial F}$$

$$=-q\left(\frac{\partial u}{\partial F}+\frac{\partial v}{\partial LF}\left.\frac{\partial LF}{\partial F}\right|_{\bar q}+\left.\frac{\partial v}{\partial F}\right|_{\overline{LF},\bar q}\right)-\frac{\partial C_P}{\partial F} \tag{4A.17}$$

Setting $\frac{\partial ES}{\partial F}=0$ yields the FOC result :

$$-q\left(\frac{\partial u}{\partial F}+\frac{\partial v}{\partial LF}\left.\frac{\partial LF}{\partial F}\right|_{\bar q}+\left.\frac{\partial v}{\partial F}\right|_{\overline{LF},\bar q}\right)=\frac{\partial C_P}{\partial F} \tag{4A.18}$$

4A.4 First Order Condition Derivation, Case 4 (*LF* and *N* both variable)

The FOC $\frac{\partial ES}{\partial N}=0$ is derived as follows. The general forms of g and C_p given by (4.13) and (4.3) still apply. Setting $j = N$ in (4.19) and evaluating,

$$\frac{\partial ES}{\partial N}=\frac{\partial CS}{\partial g}\frac{\partial g}{\partial N}-\frac{\partial C_P}{\partial N}+P\frac{\partial q}{\partial N} \tag{4A.19}$$

Now, from (4.13) $$\frac{\partial g}{\partial N}=\frac{\partial v}{\partial LF}\left.\frac{\partial LF}{\partial N}\right|_{\bar q}+\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial N} \tag{4A.20}$$

Substituting (4A.3), (4A.14) and (4A.20) into (4A.19) :

$$\frac{\partial ES}{\partial N}=-q\left(\frac{\partial v}{\partial LF}\left.\frac{\partial LF}{\partial N}\right|_{\bar q}+\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial N}\right)-\frac{\partial C_P}{\partial N}+q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial N}$$

$$=-q\frac{\partial v}{\partial LF}\left.\frac{\partial LF}{\partial N}\right|_{\bar q}-\frac{\partial C_P}{\partial N} \tag{4A.21}$$

Setting $\frac{\partial ES}{\partial N}=0$ yields the FOC result :

$$-q\frac{\partial v}{\partial LF}\left.\frac{\partial LF}{\partial N}\right|_{\bar q}=\frac{\partial C_P}{\partial N} \tag{4A.22}$$

Chapter 5
OPTIMAL SUBSIDY AND CROSS SUBSIDY WITH A LOGIT MODEL

5.1 Introduction

This chapter focuses on the behaviour of optimal subsidy when the logit probabilistic choice model, or simply the *logit* choice model, the working model developed in chapter 3, is used to predict outcomes in user choice between random and planned behaviour. One reason for this focus is to extend the analysis of optimal subsidy undertaken in chapter 4 (which assumed the simple random user model) to account for a more realistic interpretation of user cost. In addition, a further aim is to address some important new bus subsidy results reported by Jansson, K. (1993) and Tisato (1990), outcomes which result directly from the use of a random vs planned bi-modal user behavioural model.

As discussed in section 2.6 of chapter 2, the work by Jansson and by Tisato generated two interesting new results :

(a) multiple local optima can occur in the bus optimisation problem, one optimum each for random and planned user behaviour; and

(b) there is a sudden and substantial increase in optimal unit subsidy when switching occurs between random and planned behaviour.

These new outcomes create difficulties for the analyst and the policy maker. First, the existence of multiple solutions creates additional complexity for the analyst since true optimisation requires the identification of the global optimum from multiple local optima. Second, the sudden increase in unit subsidy when behavioural switching occurs can distort the conventional negative relationship between optimal unit subsidy and route patronage reported in the literature (Jansson, 1980; Waters, 1982a; Gwilliam *et al*, 1985; Nash, 1988), i.e. that unit subsidy is stronger the thinner, or less patronised, the route.[1] The conventional negative unit subsidy/patronage relationship has provided a simple and important rule of thumb for describing how subsidy per trip varies between bus routes of different demand density, and has been a useful mechanism for explaining a key policy outcome of user economies of scale. Thus its potential demise is significant.

The *a priori* concern of this study was, however, that the new outcomes were being driven by the simplistic manner in which choice between random and planned behaviour is characterised in the Jansson/Tisato model. In that model, random vs planned choice outcomes are predicted using a purely *deterministic* framework[2] which can predict very sudden, or knife-edge, behavioural changes across the population of users for small changes in service levels (see discussion in chapter 3).

The aim of this chapter is to assess the robustness of these new results when random vs planned choice is modelled using the more realistic *logit* choice model. The chapter addresses the multiple local optima problem first. It explains the nature of the problem, tests its robustness, and assesses the magnitude and severity of the problem and thus whether ignoring the problem will lead to serious errors in optimal outcomes. Next the behaviour of unit subsidy, and the implications for its relationship with patronage level, are considered. The behaviour of total subsidy is then shown to point to an important implication for subsidy policy implementation in practice. Finally, the issue of optimal cross-subsidy between routes, which results from having a breakeven constraint (Gwilliam *et al*, 1985), is considered. The case for optimal cross-subsidy is related to the inverse optimal unit subsidy/patronage relationship. Thus, with the behaviour of unit subsidy being complicated by the nature of user cost, this may also have implications for cross-subsidy. Very little work has been previously done on optimal cross-subsidy between routes.

Both Jansson (1993) and Tisato (1990) present analyses for the case of a variable load factor (*LF*) and given bus size (*N*) (*LF*/*N* fixity case 3 of chapter 4). For consistency, and

[1] Jansson does discuss the implications of the sudden jump in optimal unit subsidy, but does so in the context of high vs low frequency routes, rather than high vs low patronage routes.

[2] The Jansson/Tisato model will be referred throughout simply as the *deterministic* choice model.

practicality, the analysis in this chapter was undertaken for the same case.[3] In quantitative work, the average bus size for Adelaide buses, $N = 78$ (see section B.3.2 of appendix B) was used.

5.2 The Scope for Multiple Local Optima

5.2.1 Deterministic Choice

The possibility of occurrence of multiple local optima (MLO) arises from the 2nd first order condition (FOC) of case 3 in chapter 4 (i.e. expression (4.41)). The LHS of (4.41) is the marginal direct[4] benefit of additional frequency, which is denoted here as *MBF*. The RHS of (4.41) is the marginal cost of additional frequency, which is denoted here as *MCF*. Thus optimal frequency exists when the marginal benefit and marginal cost of frequency enhancement are equated.

Consider *MBF* and *MCF* plotted against *F*. As shown in chapter 4, *MCF* (i.e. $\partial C_p/\partial F$) is constant for a given bus size, thus the *MCF* schedule will be horizontal. On the other hand, *MBF* declines as *F* increases. The reason for this is that, for any given patronage (q) level, as marginal increases occur in *F*, user cost falls by progressively smaller amounts, thus *MBF*, declines. In any given situation, there are two *MBF* schedules, MBF_r and MBF_p, one corresponding to each user behavioural mode (random and planned).[5] The three functions MBF_r, MBF_p and *MCF* are plotted against frequency, *F* in Figure 5.1 for a given q value. The MBF_r schedule lies above the MBF_p schedule since, as shown in chapter 3, when headway (and thus frequency) varies, user cost is more responsive under random than planned user behaviour.

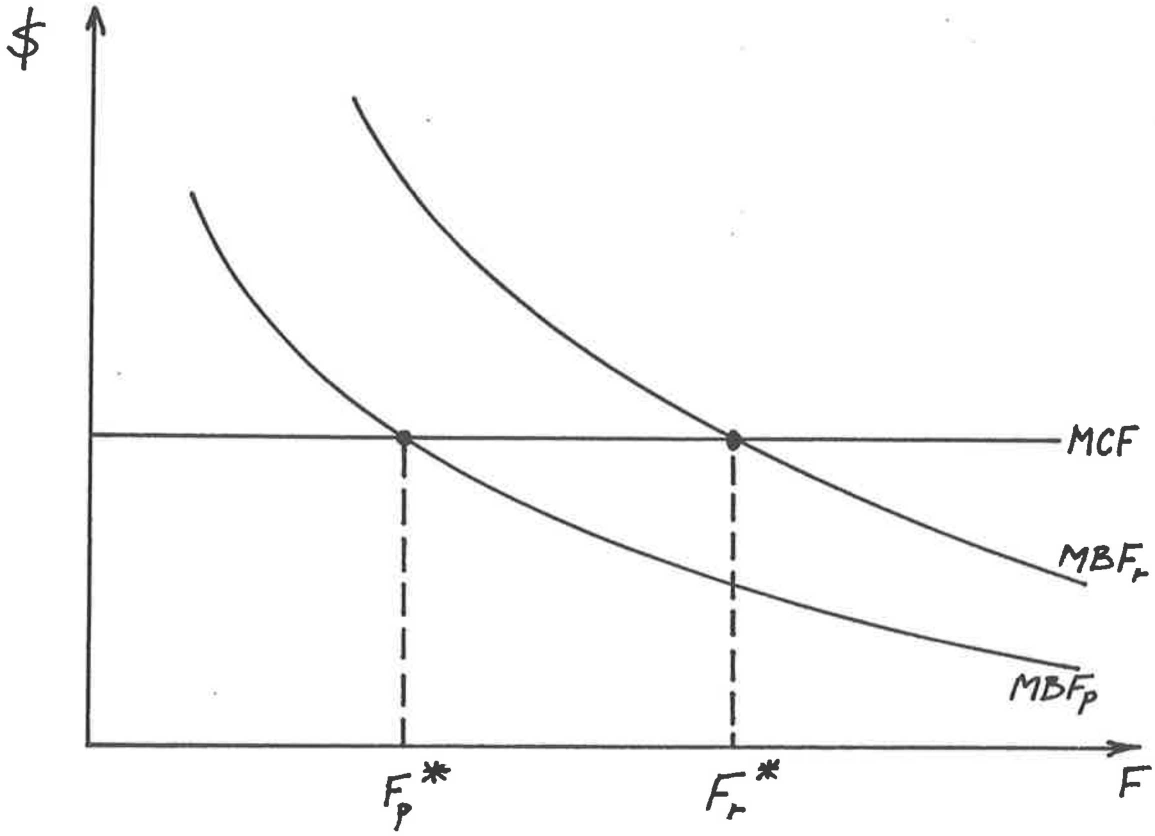
As with any FOC, economic surplus reaches a peak with respect to variation in the relevant optimising variable (in this case *F*) whenever the FOC holds, which here is when $MBF = MCF$. It is clear from Figure 5.1 that, *for a given behavioural mode*, this holds at only a *single F* value, and thus there is only a single peak in economic surplus. Denote the *F* values where this occurs (i.e.

[3] Jansson's analysis is complex, mainly due to the fact that the analysis is generalised by making the values of time savings a function of the length of delay experienced by users. The analysis here was restricted to the case where the values of time savings are independent of the length of delay, a case also considered by Jansson.

[4] See footnote 23 of chapter 4.

[5] The subscripts *r* and *p* continue to refer to random and planned behaviour.

Figure 5.1 : Optimal Frequency Determination, Random and Planned User Cases



where *MBF* = *MCF*) as F_r* and F_p* for the two modes of behaviour respectively. Therefore, if users always behaved in accordance with a single behavioural mode, *either* random *or* planned, then there would be *no* scope for MLO, with a single optimum occurring at F_r* or F_p* respectively.

As argued in chapter 3, however, users tend to choose a different behavioural mode in different situations : random behaviour tends to occur at higher frequency levels, and planned behaviour at lower frequency levels. In the *deterministic* choice model, all users act in a random fashion when $F > F_c$[6], in a planned fashion when $F < F_c$, and are indifferent when $F = F_c$, with all switching between modes occurring when $F = F_c$. *MBF* therefore corresponds with either MBF_r or MBF_p depending on the size of *F*. The effective *MBF* schedule, denoted MBF_e, is :

$$MBF_e = \begin{cases} MBF_p & \text{if } F \leq F_c \\ MBF_r & \text{if } F \geq F_c \end{cases} \qquad (5.1)$$

The MBF_e schedule is illustrated in Figure 5.2, with its main feature being a sudden vertical discontinuity in MBF_e at F_c resulting from the sudden switching of behavioural modes.

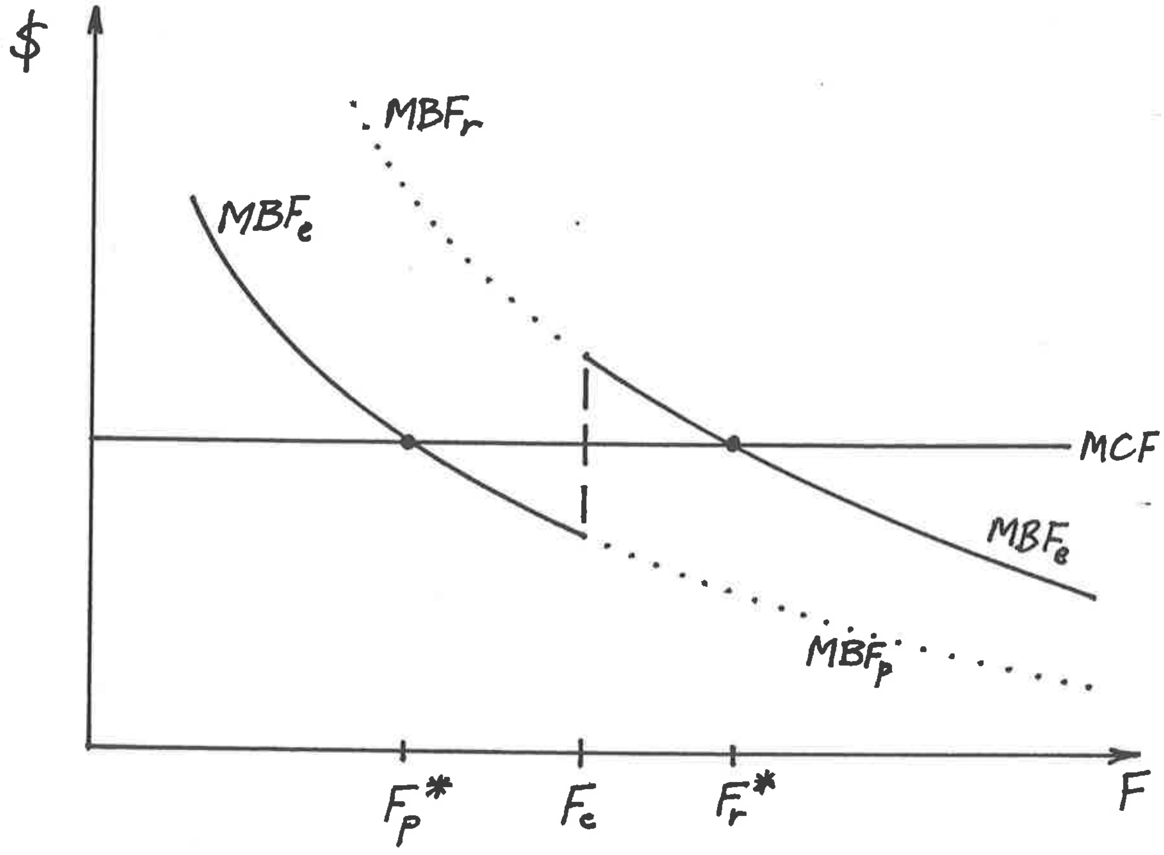
The possibility of MLO now arises because of the existence of this upward vertical discontinuity. If *MCF* does not pass through the discontinuity, then MBF_e and *MCF* will continue to cross only once, resulting in a single local (and thus global) optimum. However, if *MCF* passes through the vertical discontinuity in MBF_e, as is the case for the situation drawn in Figure 5.2, MBF_e and *MCF* will cut twice, and so twin local optima will exist *concurrently* at F_r* and F_p* (Tisato, 1990).

5.2.2 Probabilistic Choice : The Logit Model

Now consider the more general case in which choice between random and planned behaviour is modelled by the *logit* model (defined by expression (3.66)). With respect to the MLO issue, the framework of the problem is essentially the same as for the deterministic choice model.

[6] F_c is the critical frequency, the frequency equivalent of the critical headway, H_c, i.e. $F_c = 60/H_c$. Recall from chapter 3 that H_c is the headway at which random and planned deterministic user cost are equal, with users therefore being indifferent between behavioural modes.

Figure 5.2 : Derivation of *MBF_e* Schedule, Deterministic Choice Case



The difference, however, is in the way switching occurs between behavioural modes and the impact this has on MBF_e.

Once choice outcomes are predicted by the logit model, rather than all users suddenly switching modes at F_c, the change in behavioural mode across the population of users is more gradual, occurring over a range of F values. This was demonstrated in Figure 3.9 which showed the probability of random behaviour $P(r)$, and thus the proportion of users acting randomly R, changing over a spectrum of H, and thus F, values. The implication for the MBF_e schedule is that, rather than there being a sudden vertical discontinuity at F_c (as in Figure 5.2) with MBF_e switching completely from MBF_r to MBF_p at that point, there is a more gradual transition over an F range. MBF_e will be the weighted average of MBF for the two behavioural modes as follows :

$$MBF_e = R \, . \, MBF_r + P \, . \, MBF_p \qquad (5.2)$$

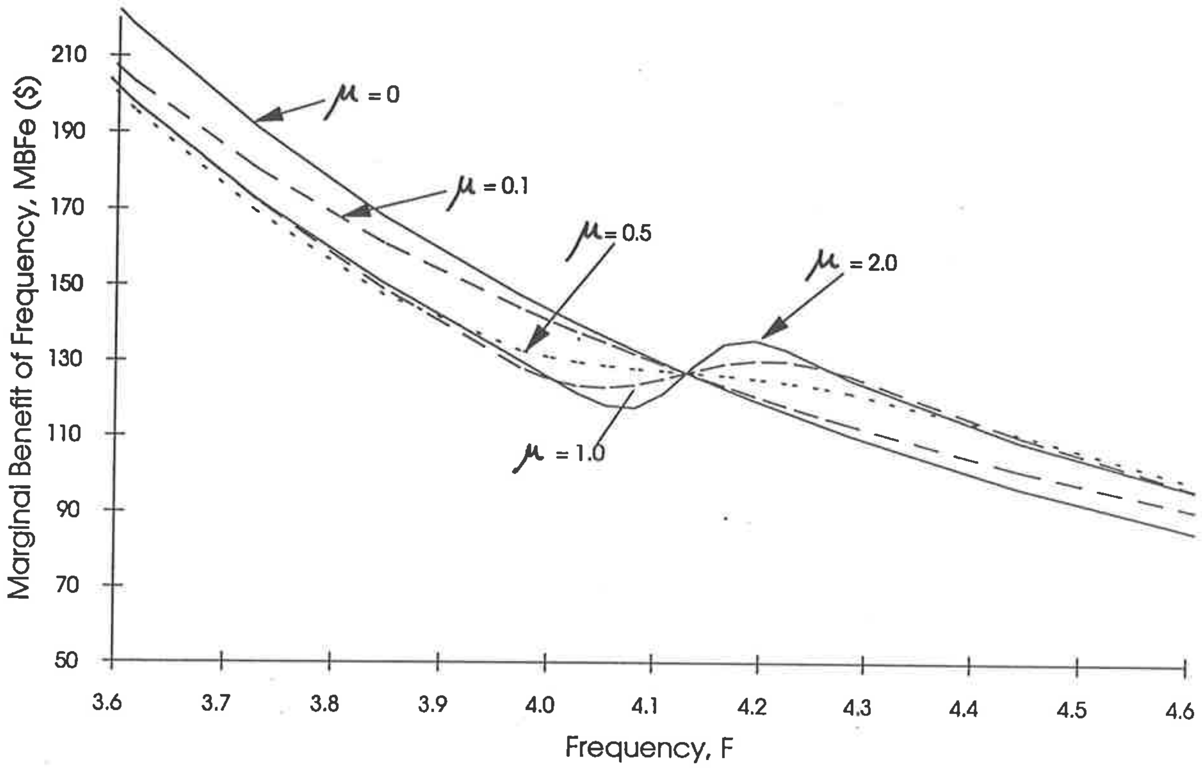
where R and P (= 1-R) are the proportions of random and planned users

and $R = P(r)$ is given by expression (3.66)

The resulting MBF_e function is graphed against F in Figure 5.3 for a number of values of the logit scale parameter μ, the parameter which is an indicator of the rate at which users switch from random to planned behaviour as F approaches F_c (see discussion of μ in section 3.8 of chapter 3).[7] The higher is μ, the more rapid will be the switching. The deterministic choice model, where all switching occurs when $F = F_c$ (thus the vertical upward discontinuity in MBF_e), coincides with having $\mu = \infty$ in the logit model. As μ declines below this extreme polar value, the sudden vertical discontinuity in MBF_e disappears. Figure 5.3 shows that, if μ remains relatively high (e.g. $\mu = 2$), the F range over which switching between modes occurs, and thus over which MBF_e switches between MBF_r and MBF_p, is relatively small. With switching occurring over such a small F range, the $\mu = 2$ case displays a rapid vertical rise in MBF_e in the neighbourhood of F_c, but the rise is now *continuous* rather than discontinuous (as in Figure 5.2), with MBF_e therefore having a positive slope over a range of F values.

[7] Figure 5.3 was derived using the middle parameter value set (PV Set 2) in Table B.6 in appendix B, and q = 220, a patronage level which yields MLO for the resulting curves.

Figure 5.3 : MBF_e Schedules for Various μ Values



As the size of μ continues to decline in Figure 5.3, switching between behavioural modes occurs more gradually over a progressively greater F range, with the positively sloped portion of MBF_e becoming increasingly flatter. For $\mu = 1.0$, MBF_e continues to have a positively sloped segment, but as μ declines further, for example, $\mu = 0.5$ (and all smaller μ values), the positively sloped portion of MBF_e eventually disappears, with the slope becoming negative throughout.

Provided cases such as $\mu = 1$ and $\mu = 2$ exist, where the slope of MBF_e changes from negative, to positive, and back to negative as F declines, i.e. there is "folding" in the MBF_e schedule,[8] the *possibility* of MCF cutting the MBF_e schedule more than once, thus resulting in MLO, will exist.[9] Note, however, that as μ declines, and the extent of MBF_e folding diminishes, so too does the scope for MLO also diminish. Further, once the slope of MBF_e becomes negative throughout (e.g. for $\mu = 0.5$), MLO are no longer possible since a horizontal MCF schedule can then only ever cut MBF once. If the critical μ value at which the slope of MBF_e just commences to become negative throughout is denoted as μ_c, it can then be concluded that :

(a) *MLO are possible <u>only</u> if* $\mu > \mu_c$ *; and,*

(b) *when* $\mu > \mu_c$*, the closer* μ *gets to* μ_c *the smaller the scope for the possible existence of MLO.*

Finally, note also that, to the extent that the deterministic choice model generates the greatest degree of "folding" of MBF_e, then it also provides the greatest scope for MLO.

5.3 The Significance of the Multiple Local Optima Problem

It is quite clear from the above discussion that MLO are possible when using both the deterministic and logit choice models. How important and significant, however, is the MLO problem? This question is addressed in this section through two quantitative assessments :

[8] The upward vertical discontinuity in the MBF_e schedule at F_c in the deterministic choice case can be thought of as an extreme case of "folding" in the MBF_e schedule.

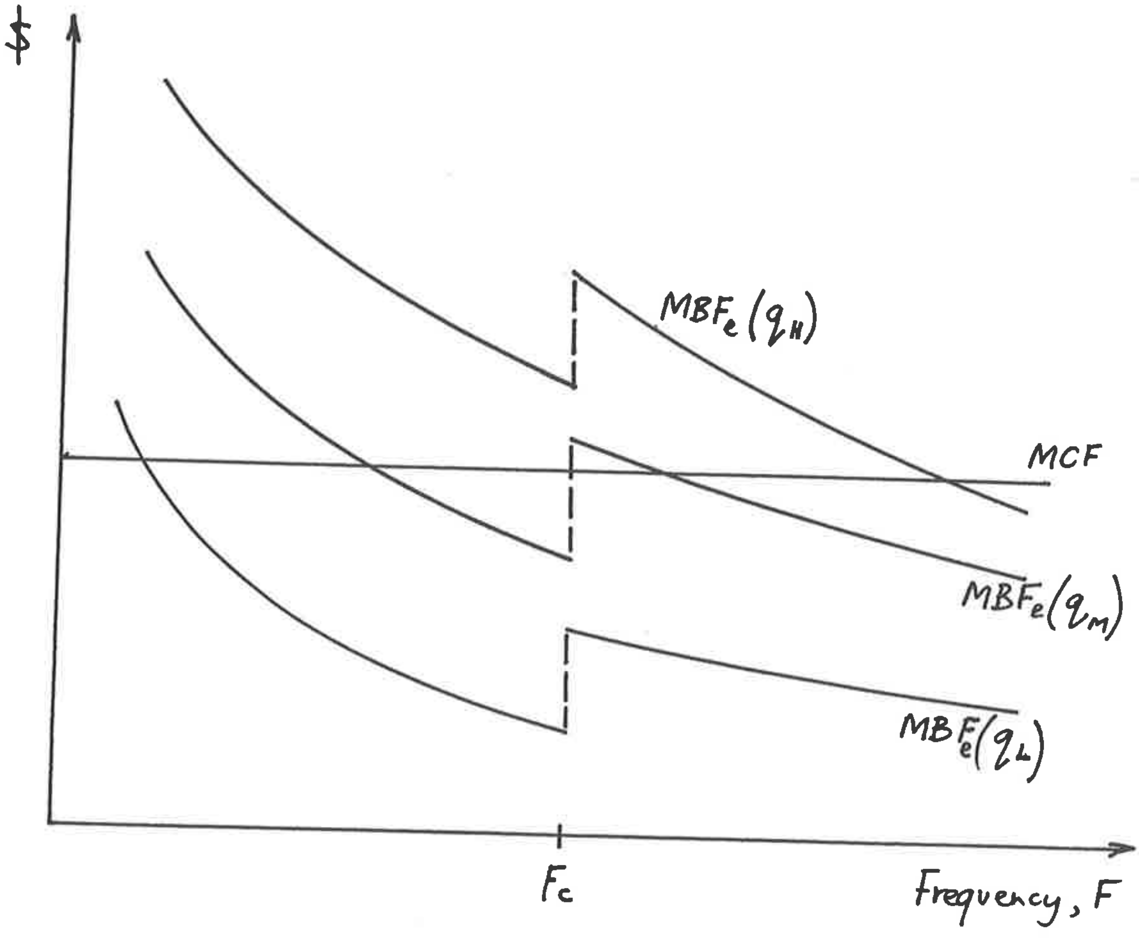
[9] In Figure 5.3, when MLO do occur, the MCF schedule will cut MBF_e three times : twice where MBF_e has a negative slope, and once where MBF is upward sloping. The first two local optima are points of economic surplus maxima, but the latter is a point of economic surplus minima since, as F increases towards the local optimum, marginal benefit is below marginal cost.

- In the first assessment, a comparison is made of the likely size of μ in practice with μ_c, the critical value of μ below which the slope of MBF_e is negative throughout thus eliminating the scope for MLO. The likely magnitude of μ in practice was discussed in section 3.8.3 of chapter 3, where a range of μ values were reported in Table 3.6 for different parameter value sets and different rates of switching between random and planned behaviour. The first assessment approach estimates the value for μ_c, enabling a comparison with the μ values in Table 3.6.

- In the second assessment, the range of patronage (q) levels for which MLO occur are estimated. This approach is useful because it provides an indicator of the proportion of demand situations in which MLO are possible. The assessment is undertaken using the polar deterministic choice model. Although, it has been argued that this is an unrealistic characterisation of user choice, since the scope for MLO is greatest in such a model, the assessment will provide an upper estimate of the MLO-generating patronage range. Figure 5.4 illustrates the basis and rationale for the assessment. The figure presents MBF_e schedules for three arbitrary patronage levels : low (q_L), medium (q_M) and high (q_H). The higher the patronage, the higher the position of the MBF_e schedule since the greater is the number of users benefiting from an increase in frequency (F). For the lower and upper MBF_e curves, with MBF_e and MCF intersecting only once, yielding a single optimum, patronage at these levels is therefore respectively too low and too high to yield MLO. On the other hand, the middle MBF_e curve in the figure is an example of a case where MLO do exist. As patronage rises or falls from this medium level, the MBF_e schedule is shifted up or down. Clearly, therefore, only a limited range of patronage levels clustered around q_M would actually yield MLO outcomes. The second assessment approach estimates the size of this MLO-generating patronage range, with the smallest and largest patronage levels of the range being denoted q_{min} and q_{max}.

These two assessments were quantified using the three parameter value sets presented in Table B.6 of appendix B, and for the four bus sizes discussed in Appendix B (see section B.3.2). The results are presented in Table 5.1. The table reports, for each PV set/bus size combination, two results. First, q_{min} and q_{max}, which define the range of MLO-generating patronage levels, are estimated, with the size of the range (q_{max} - q_{min}) then expressed as a percentage of the range's midpoint, $\frac{(q_{\min}+q_{\max})}{2}$. The second result reported is the critical μ value, μ_c, which results when q

Figure 5.4 : MBF_e Schedules for Low, Medium and High Patronage Cases



$= \frac{(q_{\min} + q_{\max})}{2}$, i.e. the μ_c value when patronage is in the middle of the MLO-generating range of demand conditions.

Table 5.1 : Summary of Multiple Local Optima Investigation Results

- *Parameter Value Set 1*

	Mini	*Midi*	*Rigid*	*Artic*
q_{min}	49	77	120	154
q_{min}	51	82	133	175
% diff	3.4	6.1	10.1	12.5
μ_c	1.1	0.6	0.4	0.33

- *Parameter Value Set 2*

	Mini	*Midi*	*Rigid*	*Artic*
q_{min}	75	122	196	254
q_{min}	77	129	212	281
% diff	2.6	4.9	8.0	10.1
μ_c	2	1	0.6	0.5

- *Parameter Value Set 3*

	Mini	*Midi*	*Rigid*	*Artic*
q_{min}	97	160	259	339
q_{min}	99	167	278	370
% diff	2.1	4.0	6.9	8.8
μ_c	3	1.4	0.8	0.65

All results were determined by trial and error. The μ_c values were identified by inspection from a plot like Figure 5.3, with an input μ value being varied until the slope of the MBF_e schedule just becomes negative throughout. The μ value at which this occurred was judged by visual inspection of the plot, and as a result, the μ_c values reported are approximate only, but sufficient for considerations here.

Consider first the logit model μ_c results. The key feature of the results in Table 5.1 is that, for all PV set/bus size combinations considered, μ_c is greater than the potential "in practice" μ values reported in Table 3.6 of chapter 3, and usually considerably greater. In contrast, it was established in section 5.2 that $\mu > \mu_c$ is required in order for MLO to be possible. It can be concluded, therefore, that if a logit model is used to predict choice by users between random and planned behaviour, as it is in this study, then it will only be in the rarest of cases that there will be even a possibility of MLO occurring. As a result, the MLO problem can effectively be ignored in the quantitative work in the remainder of this study.

The MLO-generating patronage range results in Table 5.1 are also interesting. The meagre size of the MLO-generating patronage range, q_{max} - q_{min}, particularly when expressed as a percentage of the range's midpoint, indicates that, even if random vs planned user choice was characterised by simple deterministic choice, thus maximising the chances of MLO existing, the range of patronage conditions which would generate MLO would be rather small. In other words, referring back to Figure 5.4, only a small variation in q away from q_M is required to shift the MBF_e schedule away from an MLO situation.

As a final point, note that, although MLO situations are likely to be uncommon, if such a situation did arise, the number of local maxima is likely to be small. In the case of the deterministic choice model, two local maxima would occur. Using the logit choice model, the number of local maxima would also be two. Two maxima would also result if other probabilistic choice models were used, provided they contained a distribution for the disturbance terms $f(\varepsilon_n)$ (see section 3.8.1 of chapter 3) which was uni-modal (as it is in the logit model, and in other often used probabilistic choice models, e.g. the linear probability model, and the probit model (Ben-Akiva and Lerman, 1985)). Therefore, even if an MLO situation did arise, with only two maxima likely, the task for the analyst trying to identify the global optimum will be relatively manageable.

The over-riding conclusion from the analysis here is that the general possibility of MLO existing in the bus optimisation problem is low. If user choice between random and planned behaviour is predicted by a probabilistic choice model, such as a the logit model used in this study, then the nature of such a model in practice suggests that there is unlikely to be any scope for MLO. As a result, no further consideration need be given to this issue in subsequent chapters.

5.4 Optimal Unit Subsidy

5.4.1 Deterministic Choice

The second key new result reported by Jansson (1993) involves the behaviour of optimal unit subsidy, s^*. Jansson notes that for any given level of frequency, F, the corresponding s^* will always be greater for random behaviour than planned. As a result, when frequency reaches a level at which users switch from planned to random behaviour, there will be a concurrent sudden jump in the size of s^*. Although not spelt out explicitly by Jansson, the interesting thing about this outcome is that it has important implications for the conventional negative relationship between unit subsidy and patronage level.

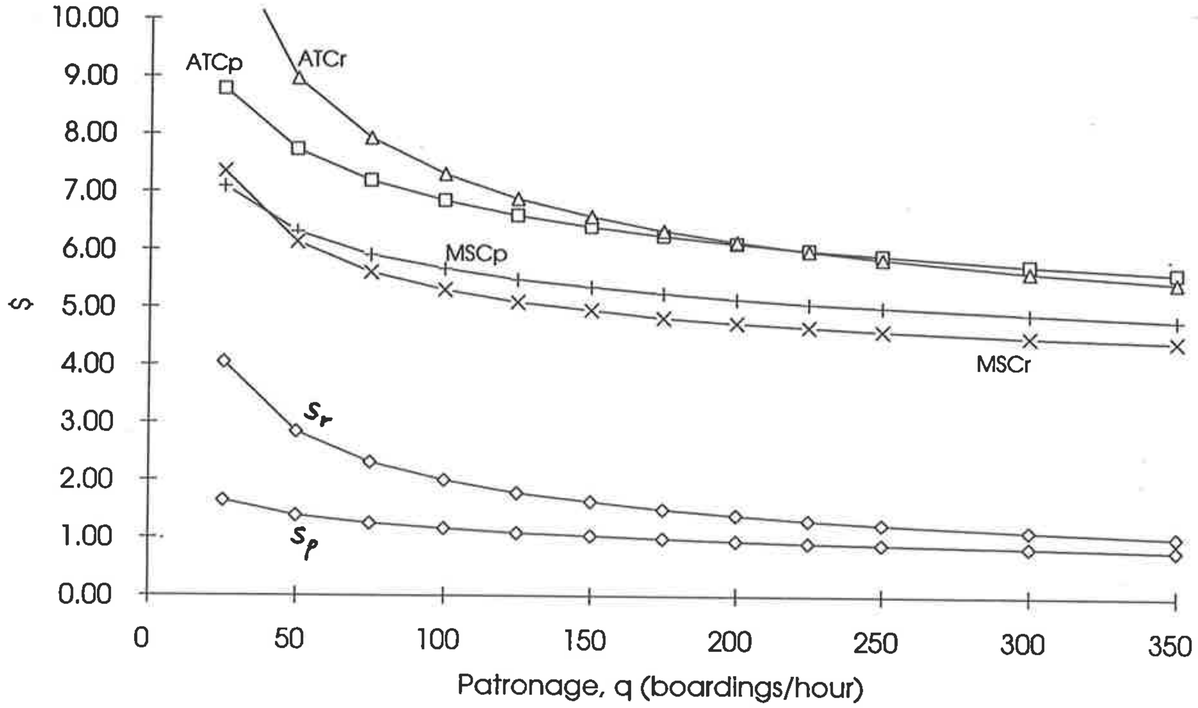
To commence with, consider the relationship between optimal unit subsidy and patronage for random and planned behaviour separately. Consider Figure 5.5 which plots average total cost (*ATC*), marginal social cost (*MSC*) and optimal unit subsidy (s^*) for the two behavioural modes.[10] *ATC* and *MSC* are plotted because s^* is derived from them. Recall from chapter 4 that *ATC* declines as patronage increases because of the positive externality to all users associated with the additional frequency required to cater for the extra demand, although the rate of decline is dampened by the negative passenger congestion externality arising from the additional users (see discussion in section 4.4.4 of chapter 4). As a result, *MSC* lies below *ATC*, with the gap between *ATC* and *MSC* being a measure of the net positive externality at the margin. Optimal unit subsidy, s^*, was shown to be also exactly equal to the marginal net externality, and thus the gap between *ATC* and *MSC*.

The main feature to note from Figure 5.5 is the fact that the entire s^* schedule for random behaviour lies above that for planned behaviour. This is due to the gap between *ATC* and *MSC* curves, and thus s^*, being a function of the *slope* of *ATC*. The steeper the *ATC* curve, the bigger the gap (the marginal net positive externality), and thus the bigger is s^*. Therefore, since the ATC_r schedule is steeper than the ATC_p schedule, at any given patronage level, then s^* is always greater for random behaviour than planned.

The continuous negative slope of the unit subsidy schedules in Figure 5.5 illustrates the important conventional negative relationship between optimal unit subsidy and patronage level, i.e. optimal unit subsidy is greater the thinner or less patronised the route (e.g. Waters, 1982a; Gwilliam *et al*, 1985). The relationship is explained as follows. The lower the patronage level, the lower will

[10] All quantitative work and resulting figures in the remainder of the chapter, including Figure 5.5, are based on parameter value set 2 in Table B.6 of appendix B.

Figure 5.5 : Average Total Cost, Marginal Social Cost and Optimal Unit Subsidy Schedules for Random and Planned Behaviour



be the frequency, F, required to accommodate it. The lower F is, however, the greater the marginal positive externality from additional frequency, and thus the greater the optimal unit subsidy.

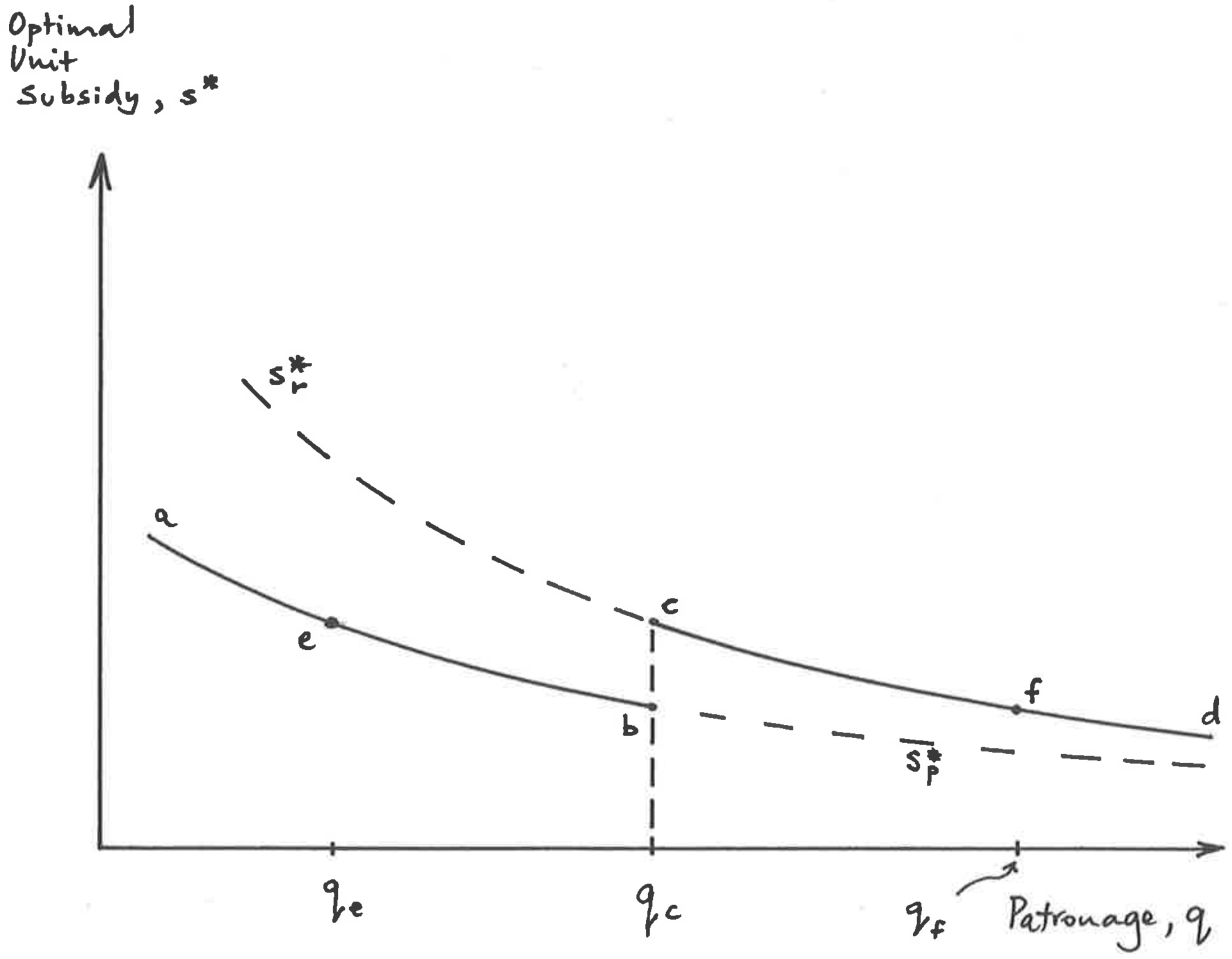
Now consider optimal unit subsidy when the *deterministic* choice model predicts behavioural mode choice by users, as it does in the Jansson model. When patronage, q, is low, optimal frequency, F^*, will also be low, yielding planned behaviour on the part of all users, and optimal unit subsidy in accordance with the s^*_p schedule. As q increases, so too does F^*. Eventually, q and F increase sufficiently that all users switch at exactly the same time from planned to random behaviour, as predicted by the deterministic model (see chapter 3). Accordingly, optimal unit subsidy will suddenly increase to coincide with the s^*_r schedule, as suggested by Jansson. If the patronage level at which switching from planned to random behaviour occurs is denoted as the critical patronage, q_c, the "effective" s^* schedule will then consist of the composite curve *abcd* illustrated in Figure 5.6, following the s^*_p schedule when $q < q_c$, and the s^*_r schedule when $q > q_c$. The sudden vertical jump in the size of s^* at q_c reflects the bigger gap (at any given q) between *ATC* and *MSC* under random behaviour than planned.

The sudden increase in optimal unit subsidy at q_c, resulting in the composite unit subsidy schedule *abcd* in Figure 5.6, is an important result because it challenges the conventional negative relationship between optimal unit subsidy and patronage. In this setting, *it is no longer possible to claim that a lower patronage level will have an associated higher optimal unit subsidy, s^**. To assist in illustrating cases in which the relationship is broken, denote points b and c in Figure 5.6 as the planned and random unit subsidy levels at q_c where switching occurs. Then, denote point e as the planned unit subsidy exactly equal to random unit subsidy at point c, and point f as the random unit subsidy exactly equal to the planned unit subsidy at point b.

Using these reference points, a number of observations can be made :

- The negative relationship applies when the patronage levels being compared are either both bigger, or both smaller, than q_c.
- For comparisons of patronage levels either side of q_c, i.e. one bigger and one smaller than q_c, the negative relationship still applies provided there is a large enough gap between the

Figure 5.6 : Optimal Unit Subsidy Schedule for Deterministic Choice Model



patronage levels being compared. For example, unit subsidy at all point to the left of point *e* is greater than unit subsidy at all points to the right of point *c*. The same is true of comparisons of points to the left of point *b* and to the right of point *f*.

- There are many comparisons, however, of points in the range *eb* with points in the range *cf* for which the negative relationship breaks down. For example, unit subsidy at a point half way between *b* and *e* has lower unit subsidy than a point just to the right of *c*.

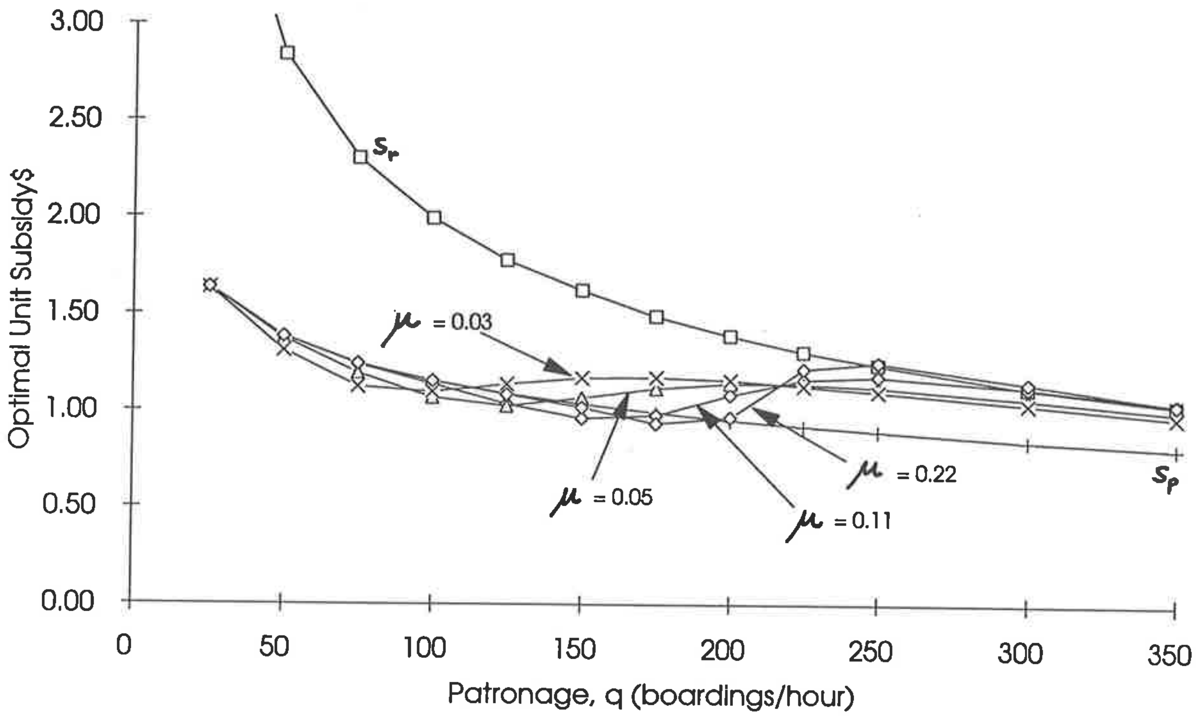
5.4.2 The Logit Choice Model

The key question to now address is whether a similar distortion of the conventional optimal unit subsidy/patronage negative relationship also holds when a logit choice model is used.

As discussed in chapter 3, compared to the deterministic choice model, the logit model generates more gradual switching between planned and random behaviour as frequency, F, increases. With F^* increasing gradually as patronage (q) grows, the transition between planned and random behaviour, and the transition of s^* from s^*_p to s^*_r also occurs gradually over a range of patronage levels, rather than at q_c as in Figure 5.6. This is evident in Figure 5.7 which presents the s^* curves generated by the logit choice model for the range of μ values discussed in section 3.8.3 of chapter 3, plus the s^*_p and s^*_r schedules from Figure 5.5 as reference schedules. The vertical discontinuity in s^* has now disappeared, and has been replaced, in each μ case, by a smooth continuous curve which approximates s_p at low demand levels, and gradually approaches s_r as patronage level (and F) increases. All of the logit unit subsidy curves in Figure 5.7 display a gradual transition of s^* from s^*_p to s^*_r. The lower is μ, the more gradual is the switching from planned to random behaviour, and thus the more gradual is the transition of s^* from s^*_p to s^*_r.

Figure 5.7 illustrates that, even when a logit choice model is used (rather than the simpler deterministic choice model), the *potential* still exists for the conventional negative relationship between optimal unit subsidy, s^*, and patronage, q, to be broken. In Figure 5.7, the relationship is broken in each of the μ cases considered, evidenced by the existence of upward sloping segments in the s^* schedules, over which s^* *increases* with q. However, as the value of μ falls, and thus switching becomes more gradual, two patterns emerge. First, the patronage range over which the s^* schedule has a positive slope gets smaller. Second, the positively sloped portion of the s^*

Figure 5.7 : Optimal Unit Subsidy for Logit Choice Model



schedule becomes flatter (and for small enough μ values, would revert to having a negative slope throughout). Both these patterns suggest that, the smaller μ is, the fewer the number of patronage comparison cases where the conventional negative relationship would be broken. A final result to note is that, in some circumstances, for example the $\mu = 0.03$ case, for optimal unit subsidy to be relatively constant over which a wide range of patronage levels.

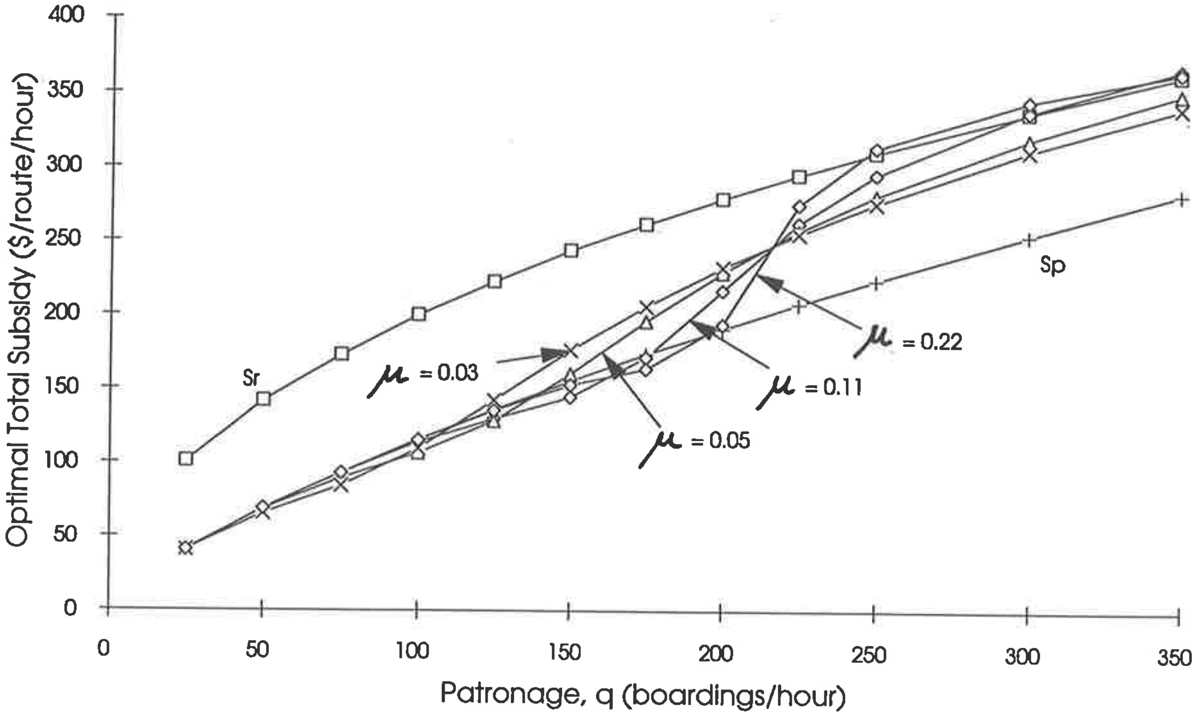
In summary, Jansson's analysis, and the behaviour of optimal unit subsidy in Figure 5.7 here, suggest that the conventional negative relationship between optimal unit subsidy and patronage level *does not hold as a general rule.* The introduction of bi-modal user behavioural choice raises the possibility for the conventional relationship to be broken, making it difficult in speculate, *a priori,* the direction of the relationship for any given pair of patronage levels. However, the more gradual switching between planned and random behaviour occurs, the fewer the cases where the conventional relationship is broken.

5.5 Optimal Total Subsidy

Consider next the behaviour of optimal total subsidy, S^*. Figure 5.8 plots S^* for the range of logit μ values considered in Figure 5.7, plus S^* for the random and planned cases (S^*_r and S^*_p) as reference schedules. S^* rises steadily with patronage level in all situations. With a logit model, S^* displays a gradual transition between S^*_p and S^*_r as q rises, in line with similar behaviour of unit subsidy in Figure 5.7. The larger the logit scale parameter μ, the more rapid the switching between modes, and the more rapidly S^* grows over a progressively shorter q range whilst switching occurs. In the polar case of $\mu = \infty$ (the deterministic choice model) there would be a sudden vertical jump in S from S_p to S_r at the patronage level at which switching occurs. Such an outcome is best approximated in Figure 5.8 by the $\mu = 0.22$ case.

The behaviour of S^* is an interesting issue from a policy perspective. First, policy makers concerned with implementing optimal outcomes which lead to economic efficiency must be prepared (based on Figure 5.8) to continually increase total subsidy as patronage rises. This is not

Figure 5.8 : Optimal Total Subsidy for Logit Choice Model (Case 3, *LF* variable, *N* = 78)



an unusual outcome, however, and need not cause problems[11]. Secondly, and more importantly, as patronage grows through the range over which behavioural mode switching occurs, successive increments in patronage require greater increases in subsidy to maintain optimality. This is suggested by the steepening in Figure 5.8 of the S^* schedule over this range. Further, the bigger the logit scale parameter, the greater the S^* increments, and the smaller the q range over which they occur. Consequently, in cases of relatively large μ values (implying rapid switching behaviour), S^* will grow quite rapidly for relatively small increments in demand.

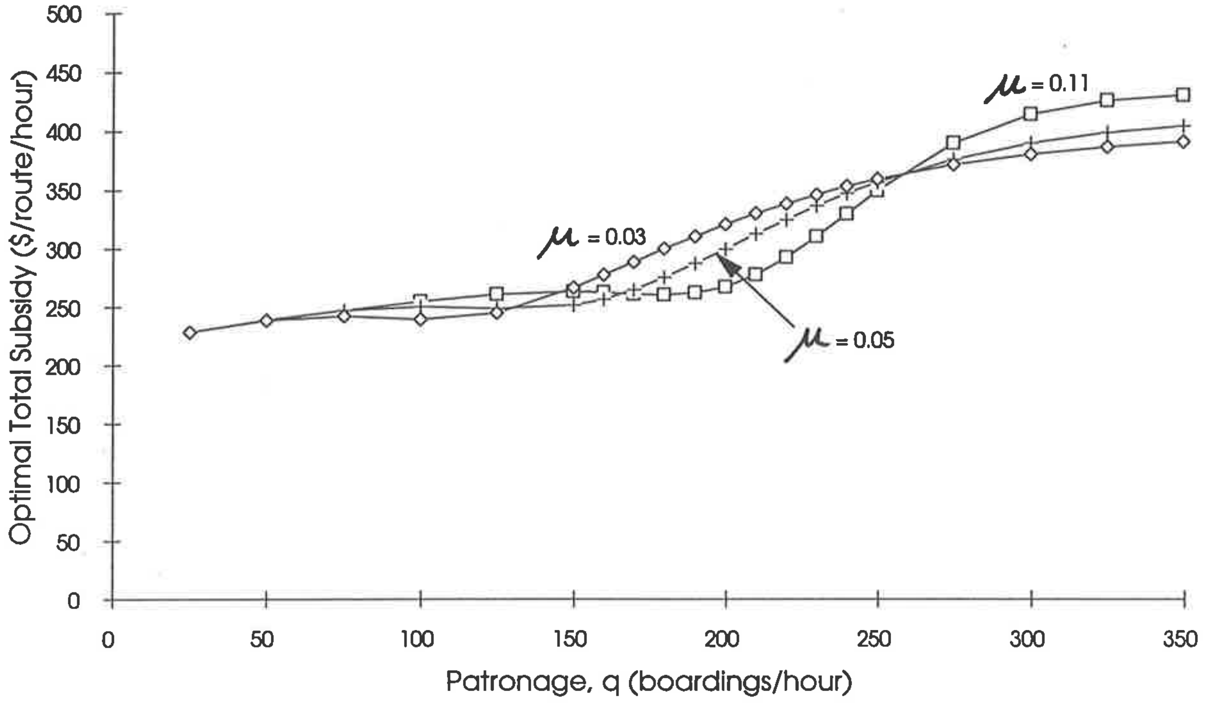
This suggests that, in some circumstances, policy makers interested in maintaining optimal economic outcomes will be faced with a need to implement quite significant jumps in S^* over possibly quite short periods of time as growth in patronage occurs. This would be the case when patronage levels approach the range over which mode switching occurs, and when patronage is growing rapidly. In the real world, there are often political and financial constraints which may limit the rate at which S^* can be increased. If this is the case, policy-makers will need to at least be aware of projected S^* values, and if possible attempt to manage the situation, for example by earmarking funds on a more gradual basis in advance. Of course this in itself would have associated opportunity costs, requiring careful overall consideration of what might constitute optimal policy.

Figure 4.10 in chapter 4 showed that total subsidy behaved in a fairly consistent manner across the load factor (*LF*)/bus size (*N*) fixity cases 2, 3 and 4 considered in that chapter. One could therefore expect the above analysis, undertaken for *LF*/*N* case 3 (*LF* variable, *N* fixed), to also apply fairly well to cases 2 and 4. Figure 4.10 showed quite different total subsidy behaviour, however, for *LF*/*N* case 1 (*LF* and *N* both fixed), and so it is worth briefly focusing on this before moving on.

Figure 5.9 plots, for *LF*/*N* case 1, S^* for the random and planned cases plus for three logit μ value cases. Although S^*_p and S^*_r do display some degree of response to changes in patronage (q),

[11] Although there may *perceived* limits to how high total subsidy may rise, as has increasingly been the case in many cities in recent times, including Adelaide, due to public finance constraints and scope to reduce costs through productivity improvements (see discussion in sections 2.1 and 2.2 of chapter 2).

Figure 5.9 : Optimal Total Subsidy for Logit Choice Model (Case 1, *LF* = 0.4, *N* = 78)



the responsiveness is rather mild,[12] thus S^* is relatively stable for both individual mode cases. This is certainly true when one compares the responsiveness of S^*_r and S^*_p here with that for *LF/N* cases 2, 3 and 4 in Figure 4.10. In the logit cases in Figure 5.9, there is also considerable stability at patronage levels outside the range over which switching occurs, where planned and random behaviour are the norm respectively. Within the switching range, however, there is a much more pronounced response. Although the rates of response during switching are similar in magnitude to those for *LF/N* case 3 in Figure 5.8, there is now a greater contrast with the otherwise relatively stable unit subsidy levels. One could expect, therefore, that implementation problems faced by policy makers could be much more pronounced in *LF/N* case 1.

5.6 Optimal Cross-Subsidy

The final area of focus of this chapter is the question of *optimal* cross subsidy between routes when the operator is required to meet a *financial breakeven constraint* across a collection of routes. The case for having positive route cross-subsidies[13] in the optimum has been previously established and discussed in the literature (Nash, 1982; Gwilliam *et al*, 1985; Evans, 1987). It is an important argument because cross-subsidies are usually associated with undertakings and enterprises in a non-optimal context, whereas here the outcome is part of the optimal solution. The aim of the analysis here is to assess the impact on cross-subsidy result of the introduction of the logit model of user behavioural choice. The *a priori* suspicion was that, with cross-subsidy being closely related to first best unit subsidy, the fluctuating movements in unit subsidy discussed in section 5.4 may produce similar movements in cross-subsidy.

[12] In chapter 4, the S^* vs q schedule (see Figure 4.10) was horizontal for *LF/N* fixity case 1. This was due to the assumption that buses were perfectly reliable ($\sigma = 0$). In contrast, Figure 5.9 was constructed with $\sigma > 0$, producing non-constant S^* (- the influence of service unreliability on subsidy is addressed in detail in chapter 6).

[13] Cross-subsidy is defined (Gwilliam *et al*, 1985) as having economic profits on some routes financing economic losses on other routes.

5.6.1 Second Best Financially Constrained Optimisation

Optimal cross-subsidies between routes are the outcome of the bus optimisation problem when it is subject to an overall financial breakeven constraint across a group of routes (Gwilliam *et al*, 1985). With the requirement to meet a subsidy constraint, the bus optimisation problem is now a second best problem, i.e. given that the first best optimum is not achievable, what is the best outcome possible subject to meeting the constraint.

To illustrate the case for optimal cross-subsidies, it is necessary to reformulate the first best optimisation problem presented in Chapter 4 (see sections 4.4.1 and 4.4.4). A general expression for the second best constrained subsidy bus optimisation problem is :

$$\max_{\forall j \text{ policy variables}} \quad ES$$

$$\text{subject to } S \leq S_T$$

where S is total subsidy, and S_T denotes a limiting value that S cannot exceed, with $S_T \geq 0$.

This problem can be solved by the Kuhn-Tucker non-linear optimisation technique. In this chapter, however, the focus will be limited to the case of meeting a *financial breakeven constraint.* The above inequality constraint is thus replaced by an equality constraint, plus S_T is set equal to zero given the need to break even, i.e.

$$\max_{\forall j \text{ policy variables}} \quad ES$$

$$\text{subject to } S = 0 \; (= S_T)$$

The Langrangean approach then transforms the problem into the following simple unconstrained optimisation problem :

$$\max_{\forall j \text{ policy variables}} \quad L$$

where the Lagrangean, $L = ES + \lambda\,(S_T - S)$

and λ is the Lagrange multiplier, the marginal value of relaxing the subsidy constraint by one unit (i.e. one dollar), where $\lambda \geq 0$.

The problem can be solved for each route individually, resulting, by definition, in each route breaking even, and thus no cross-subsidies between routes. Gwilliam *et al* (1985) have illustrated, however, that social welfare can be further improved by maximising *combined ES subject to an overall subsidy constraint, expression (5.3) below (- a breakeven constraint in this case -), applying across routes in aggregate, with cross-subsidy between routes* i.e.

$$\sum_i S_i = S_T \tag{5.3}$$

Consider the case of two routes, denoted 1 and 2.[14] The Lagrangean is then :

$$L = ES_1 + ES_2 + \lambda(S_T - S_1 - S_2) \tag{5.4}$$

Substituting for *ES* and *S* from (4.16) and (4.18) yields :

$$L = CS_1 + CS_2 + \lambda S_T - (1+\lambda)\ (C_{p1} - P_1 q_1 + C_{p2} - P_2 q_2) \tag{5.5}$$

where $S_T = 0$ in the break even case.

As discussed in section 4.4.4, for the *LF/N* fixity case 3 being considered here, *P* and *F* are the policy variables to be optimised. There are thus five first order conditions (FOCs) : one for each of the two prices, P_1 and P_2, one for each of the two frequencies, F_1 and F_2, and one with respect to the Lagrange multiplier, λ. The FOC's $\frac{\partial L}{\partial P_i} = 0$, $\frac{\partial L}{\partial F_i} = 0$ and $\frac{\partial L}{\partial \lambda} = 0$ (which are evaluated in the chapter appendix) yield respectively :

$$P_i = q_i \frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i} - \frac{\lambda}{(1+\lambda)}\frac{q_i}{\partial q_i / \partial g_i} \tag{5.6}$$

$$-q_i\left(\frac{\partial u}{\partial F_i} + \frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial F_i}\bigg|_{\bar{q}} + \frac{\partial v}{\partial F_i}\bigg|_{\overline{LF},\bar{q}}\right) = \frac{\partial C_{pi}}{\partial F_i} \tag{5.7}$$

and the budget constraint $C_{p1} - P_1 q_1 + C_{p2} - P_2 q_2 = S_T (= 0)$ (5.8)

Expression (5.7) from the FOC with respect to F_i yields the same result as the first best problem in chapter 4, i.e. expression (4.41). Expression (5.6), from the FOC with respect to P_i, is the second best price, which in turn has two components. The first component is the marginal negative passenger congestion externality, the first best price in chapter 4 (see expression 4.40)). With chapter 4 showing that first best subsidies are optimal at all patronage levels, component two of the second best price (5.6) is a markup above marginal social cost (i.e. above first best price)

[14] This is the case considered throughout this section.

required to raise revenue so that the subsidy constraint can be met. The "constrained markup" above first best price is denoted CMU_i, i.e.

$$CMU_i = -\frac{\lambda}{(1+\lambda)}\frac{q_i}{\partial q_i/\partial g_i} \tag{5.9}$$

The minus sign is required because $\frac{\partial q_i}{\partial g_i} < 0$.

As Gwilliam *et al* (1985) explain, other things equal, the markup will be greater where :

(a) patronage, q_i, is greater; and

(b) when patronage is less responsive to generalised cost, i.e. the smaller $\partial q_i/\partial g_i$ is, or alternatively, the greater $\partial g_i/\partial q_i$ is, i.e. the steeper the generalised cost inverse demand curve.

These two effects can be integrated, however, via generalised cost elasticity of demand by manipulating (5.6) as follows. Generalised cost elasticity of demand, ε_g is given by :

$$\varepsilon_g = \frac{\partial q}{\partial g}\frac{g}{q} \tag{5.10}$$

thus (5.6) becomes $$P_i = q_i\frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i} - \frac{\lambda}{(1+\lambda)}\frac{g_i}{\varepsilon_{gi}} \tag{5.11}$$

Now, $q_i\frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i}$ is the marginal passenger congestion externality, i.e. MC_F - AC_F. Substituting this into (5.11), and (5.11) into $g = P + AC_u = P + AC_F + AC_o$ (from (4.5) and (4.12)), and noting that $MC_o = AC_o$ since AC_o is constant, yields :

$$g_i = MSC_i - \frac{\lambda}{(1+\lambda)}\frac{g_i}{\varepsilon_{gi}} \tag{5.12}$$

where $MSC = MC_F + MC_o$ (which excludes an MC_p term since from (4.3) $MC_p = 0$).

Finally, rearranging (5.12) yields :

$$\frac{g_i - MSC_i}{g_i} = -\frac{\lambda}{(1+\lambda)\varepsilon_{gi}} \tag{5.13}$$

This relationship is the conventional inverse elasticity rule found in optimal taxation literature (Ramsey, 1927) and optimal pricing literature (e.g. see Brown and Sibley (1986) on Ramsey prices). It indicates that the percentage markup of generalised cost above *MSC* (i.e. generalised cost in a first best setting) is inversely proportional to the elasticity of demand.

Under these optimal outcomes, for there to be no cross-subsidy between routes, each route would have to (simultaneously) exactly cover its average producer costs. That is the markup on each route would have to be exactly equal to the first best unit subsidy on each route. If this exact outcome does not eventuate, then, given the overall breakeven requirement, one route must be over-recovering costs and the other route under-recovering, with :

- the profit making route *cross-subsidising* the loss making route; and
- the loss making route *being cross-subsidised* by the profit making route.

The breakeven constraint requires that total subsidy, S, be identical in size, but opposite in sign, for the two routes, i.e. $S_1 = -S_2$. Thus optimal cross-subsidy, CRS, is the absolute value of S on either route :

$$CRS = |S_i|$$

Denoting CRS_{12} as the cross-subsidy from route 1 to route 2, then :

$$CRS_{12} = -S_1 = S_2 = -CRS_{21} \qquad (5.14)$$

5.6.2 Generating a Constrained Optimal Solution

An optimal solution can be generated using the five FOC's ((5.6) to (5.8)) to solve for the five variables P_1, P_2, F_1, F_2 and λ, for any given pair of patronage levels q_1 and q_2. First, optimal frequencies F_1^* and F_2^* can be determined by solving (5.7) for each route, given q_1 and q_2. Then P_1^*, P_2^* and λ^* can be determined by solving the remaining three FOC's. This is done by substituting (5.6) for P_1 and P_2 in (5.8) and solving to yield λ^*. Thus (5.8) becomes :

$$S_T = X - Y + \frac{\lambda}{(1+\lambda)} Z \qquad (5.15)$$

where
$$X = \sum_i C_{pi} \qquad (5.16)$$

$$Y = \sum_i \left(q_i^2 \frac{\partial v}{\partial LF_i} \frac{\partial LF_i}{\partial q_i} \right) \qquad (5.17)$$

and
$$Z = \sum_i \left(\frac{q_i^2}{\partial q_i / \partial g_i} \right) \qquad (5.18)$$

and rearranging (5.15) yields :

$$\lambda^* = -\frac{S_T + Y - X}{S_T + Y - X - Z} \qquad (5.19)$$

λ^* can then be substituted into (5.6) to yield P_1^* and P_2^*.

5.6.3 Simulation Results

To illustrate the case for cross-subsidies, and more importantly to investigate the impact on cross-subsidy of using the logit model to predict choice between random and planned behaviour, a number of simulation runs were undertaken using the exponential demand function given by expression (4.14) in chapter 4. All the simulation results were constructed by setting q_1 at some selected level, then observing how CRS_{12} changed as q_2 was varied. The results are summarised in Figures 5.10 to 5.12.

For the exponential demand model, from (4.14), $\frac{\partial q_i}{\partial g_i} = -\beta q_i$, thus (5.9) becomes :

$$CMU_i = \frac{\lambda}{(1+\lambda)\beta} \tag{5.20}$$

That is, for any pair of patronage levels q_1 and q_2, the markup will be the same on both routes.[15] Then, from (5.6), optimal second best price becomes :

$$P_i = q_i \frac{\partial v}{\partial LF_i} \frac{\partial LF_i}{\partial q_i} + \frac{\lambda}{(1+\lambda)\beta} \tag{5.21}$$

Thus, when comparing two routes, with an identical markup for both routes, the direction of cross-subsidy will depend completely on the relative size of first-best unit subsidies :

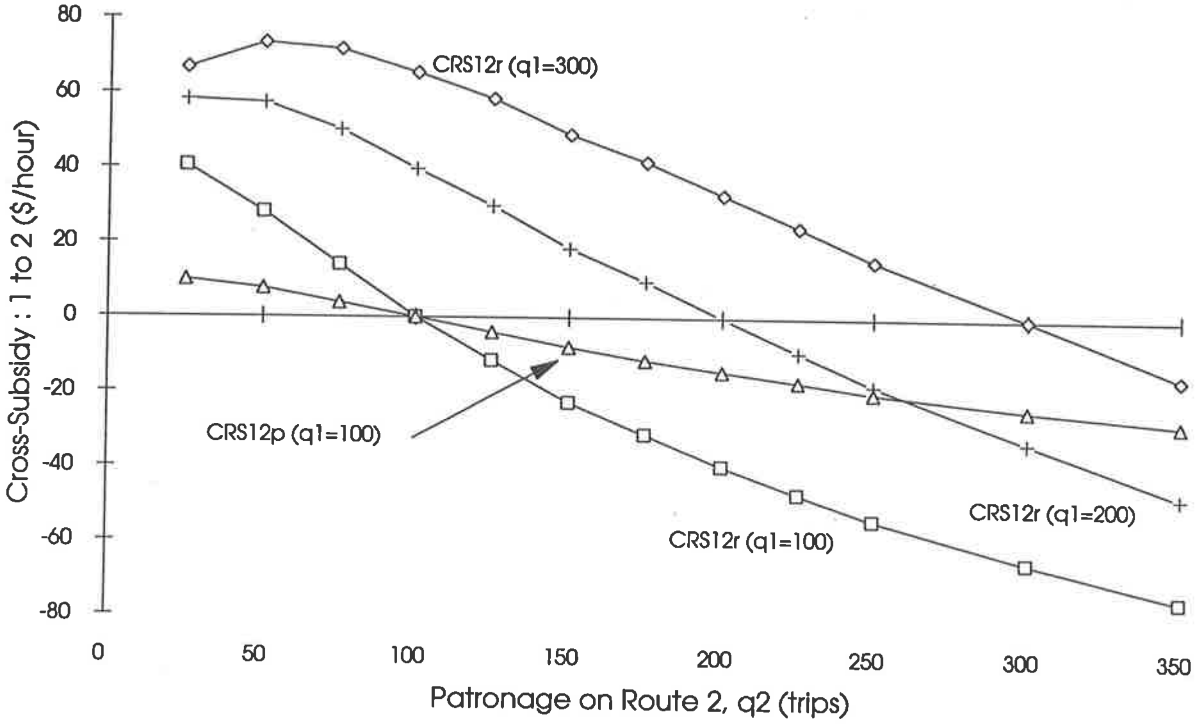
***Result 1** : for an overall breakeven outcome, it must be the case that the route with the lower first-best unit subsidy will be cross-subsidising the route with the higher first best unit subsidy.*

This is an important result which forms the underlying basis of the assessment of cases below.

Consider first the situation where users act in either a random or planned manner. Figure 5.10 plots four single behavioural mode schedules of CRS_{12} (the cross-subsidy from route 1 to route 2) : three schedules are for random user behaviour, one for each of three q_1 values (q_1 = 100, 200 and 300), and the fourth schedule is for planned user behaviour for the q_1 = 200 case. In each case, $CRS_{12} > 0$ when $q_1 > q_2$, and $CRS_{12} < 0$ when $q_1 < q_2$, that is, the higher patronage route always cross-subsidises the lower patronage route, consistent with findings in the literature (Gwilliam *et al*, 1985; Evans, 1987). This occurs because, as shown in Figure 5.5, for single behavioural mode cases (random and planned), first best unit subsidy is always lower on higher patronised routes (i.e.

[15] But for each different patronage pair, λ^* will change, and thus so will the markup.

Figure 5.10 : Optimal Cross Subsidy From Route 1 to Route 2 - Random and Planned Behaviour Cases



the conventional negative relationship discussed in section 5.5). It then follows from result 1 that a higher patronage route will cross-subsidise one of lower patronage.

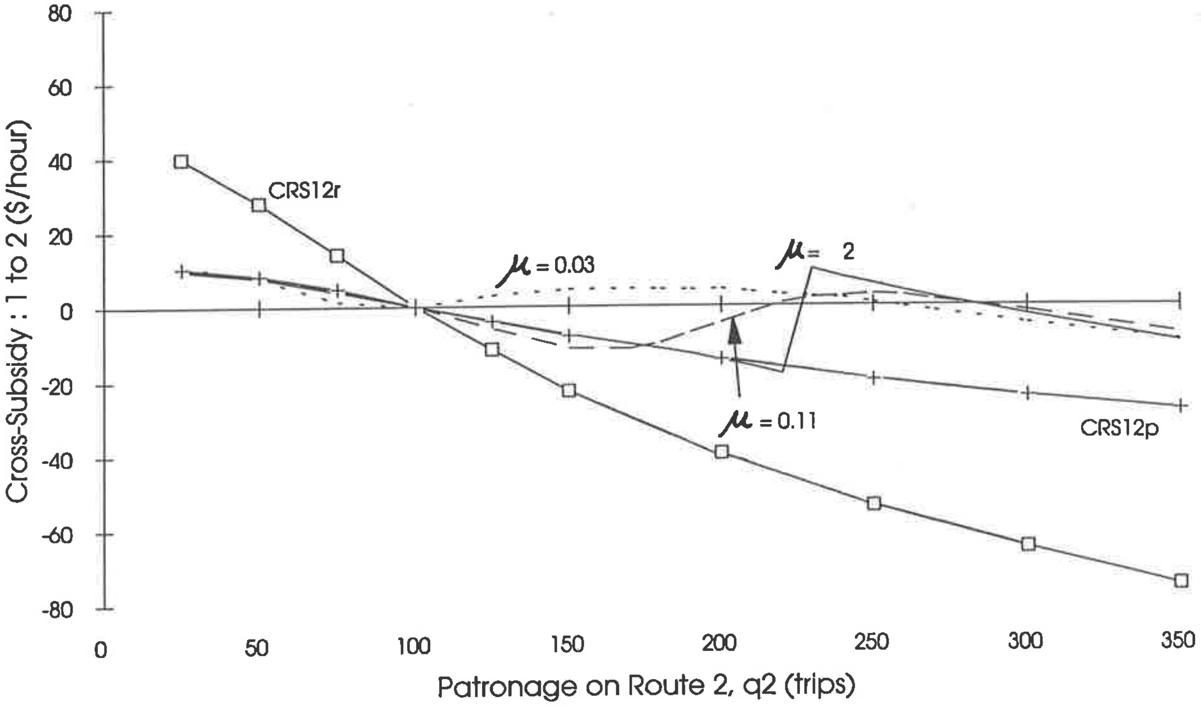
Note also in Figure 5.10 that $CRS_{12} = 0$ when $q_1 = q_2$ since both the second best markup, and the first best unit subsidy, will be the same on both routes. Further, the bigger the difference between q_1 and q_2, the greater the magnitude of cross-subsidy. Finally, note that the planned behaviour model generates lower cross-subsidies, and thus a flatter cross-subsidy schedule, than does the random behaviour model. This is due to the first best unit subsidy schedule being flatter under planned behaviour (see Figure 5.5), and thus the differences in first best unit subsidy between a given pair of patronage levels, and correspondingly the level of cross-subsidy, will be lower.

Now consider the more interesting question of how cross-subsidy behaves when planned and random behaviour are considered simultaneously in a discrete choice framework, specifically, the logit choice model. The resulting cross-subsidy outcomes are summarised in Figures 5.11 and 5.12 for two cases, where $q_1 = 100$ and 300 respectively.[16] In each case, cross-subsidy results are presented for a range of logit μ values, plus the random and planned schedules (CRS_{12p} and CS_{12r}) are also plotted for comparative purposes.

In Figure 5.11, consider first the case of $\mu = 2$, in which switching occurs very quickly (and therefore approximates the simple deterministic choice model). For q_2 values below q_c, CRS_{12} is almost identical to that which occurs under pure planned behaviour since that is in fact the behavioural mode at both patronage levels. Once q_2 approaches q_c, however, there is a rapid switch to random behaviour on route 2, but importantly, with q_1 still $< q_c$, behaviour on route 1 is still planned. We therefore now have different behavioural modes on the two routes. With a switch to random behaviour on route 2, there is a corresponding sudden increase (see Figure 5.7) in first best unit subsidy on that route. With first best unit subsidy on route 1 unchanged, result 1 suggests that there should be a rapid relative shift in revenue raising from route 2 to route 1, resulting in the

[16] With $q_c \approx 220$ boardings/hour, these two q_1 values were chosen to yield a different mode of behaviour on route 1 in each case (planned behaviour when $q_1 = 100$, and random when $q_1 = 300$). In contrast, the mode of behaviour on route 2 will vary as q_2 varies, starting with planned behaviour when q_2 is low, and gradually switching to random as q_c is approached and exceeded.

Figure 5.11 : Optimal Cross Subsidy From Route 1 to Route 2 - Logit Model, q_1 = 100 Case



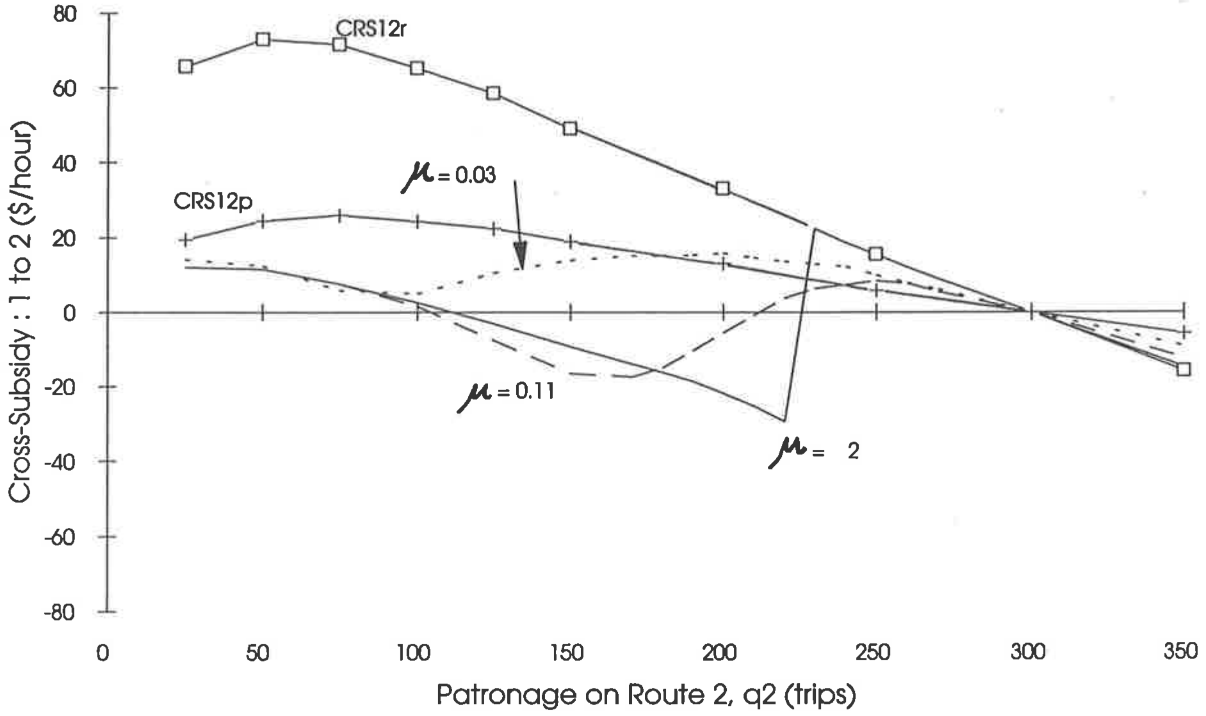
sudden rise in the cross-subsidy schedule evident in Figure 5.11. The rise in first best unit subsidy on route 2 is so great that it now exceeds that on route 1, leading to CRS_{12} moving from negative to positive. As q_2 increases further, first best unit subsidy on route 2 declines gradually and thus so does cross-subsidy.

The other logit curves in Figure 5.11 reflect less rapid switching from planned to random behaviour on route 2 as q_2 approaches q_c. As a result, the need to increase the relative proportion of revenue raising from route 2 to route 1 changes more gradually, requiring a more gradual change in cross-subsidy compared to the rapid change experienced for $\mu = 2$. Note the close relationship between the cross-subsidy schedules here and the corresponding first best unit subsidy schedules in Figure 5.7. The more gradual the change in unit subsidy in Figure 5.7, the more gradual the change in cross-subsidy in Figure 5.11.

A similar, but almost symmetrical pattern of outcomes occur in Figure 5.12 for the case of $q_1 = 300$. In this case, $q_1 > q_c$, thus there is always random behaviour on route 1. When q_2 also exceeds q_c, behaviour on route 2 is also random , and thus the logit schedules lie close to the pure random cross-subsidy schedule. As q_2 declines towards q_c, there is switching of behaviour on route 2 to planned. This causes route 2 first best unit subsidy to decline, implying the need to shift relative revenue raising efforts from route 1 to route 2, resulting in a corresponding decline in the cross-subsidy from route 1 to route 2. If μ is large, e.g. μ =2, switching is very rapid, and the corresponding decline in cross-subsidy is also rapid. As μ gets smaller, and so switching is more gradual, so too the rate of change in cross-subsidy is more gradual.

The behaviour of cross-subsidy depicted in Figures 5.11 and 5.12 under a logit model differs in several respects from that reported in the literature and from the single behavioural mode cases in Figure 5.10. First, it is no longer necessarily the case that high patronage routes always cross subsidise low patronage routes. In some circumstances the reverse may occur, with high patronage routes being cross subsidised by low patronage routes. Consider for example, the case in Figure 5.12 of $\mu = 0.11$ when q_2 lies approximately between 105 to 215. Over this q_2 range, CRS_{12} is negative and thus route 1 is being cross subsidised, yet route 1 has the higher patronage (300).

Figure 5.12 : Optimal Cross Subsidy From Route 1 to Route 2 - Logit Model, q_1 = 300 Case



Second, it is no longer necessarily the case that the magnitude of cross-subsidy increases when the difference in patronage between routes gets larger. For example, in Figure 5.12 when μ = 0.11, as q_2 falls over the approximate range 165 to to 105, the magnitude of cross-subsidy declines, yet the patronage difference has grown.

Third, as the difference in patronage between routes gets larger, it is no longer necessarily the case that cross-subsidy changes uniformly. It can now change rapidly in some circumstances (e.g. the μ = 2 example) and slowly in others (e.g. when μ = 0.03, there is quite a wide range of q_2 values over which cross-subsidy remains relatively stable, especially in Figure 5.11).

Fourth, in the single behavioural mode cases in Figure 5.10, the only time when $CRS_{12} = 0$ was when $q_1 = q_2$. However, once a logit model is used to predict choice between modes, CRS_{12} can also equal zero even when $q_1 \neq q_2$. In the μ = 2 and 0.11 cases in both Figures 5.11 and 5.12, $CRS_{12} = 0$ at two other q_2 values. This occurs because, for these μ values, switching occurs rapidly enough to cause a change in CRS_{12} which is large enough for CRS_{12} to change sign. In contrast, for the μ = 0.03 case, switching does not occur rapidly enough for this to occur, and so $q_1 = q_2$ is the only location where $CRS_{12} = 0$.

Overall, introduction of the logit model has had a significant impact on optimal cross-subsidy analysis, making it more difficult to predict *a priori* how cross subsidy will behave.

5.7 Chapter Summary and Conclusions

The aim of this chapter was to investigate the influence on bus subsidy analysis of using a probabilistic *logit* model to predict the outcome of the discrete user choice between random and planned behaviour. One reason for doing so was to extend the analysis of optimal subsidy beyond the single behavioural mode analysis of chapter 4. In addition, there was a need to test the robustness of recent new results in the literature (which were derived using a simpler, purely deterministic, behavioural mode choice framework) : the existence of multiple local optima in bus optimisation; and sudden increases in optimal unit subsidy as behavioural mode switching occurs. The chapter also considered optimal total subsidy, and optimal cross-subsidy, an aspect of user economies of scale subsidy analysis which has received limited attention to date.

The chapter demonstrates that scope exists for *multiple local optima* (MLO) in both deterministic and probabilistic (logit) discrete choice frameworks. When a logit model is used, the scope for MLO depends on the rate at which users switch between random and planned behaviour, the indicator of which is the logit model scale parameter μ. The more gradual the rate at which switching occurs (i.e. the smaller the μ value), the smaller will be the scope for MLO. In fact, for MLO to be feasible at all, μ must exceed a minimum critical value (which varies between situations). The analysis here found, however, that μ values in practice will be below these required minimum values. As a result, use of a logit model with realistic μ values is *unlikely* to generate MLO situations.

The scope for MLO is greater, and in fact greatest, when choice between random and planned behaviour is described by a simple deterministic choice model. Even if such a model is used, however, the range of patronage levels which are necessary to generate MLO is likely to be quite small. Finally, even if, in the unlikely situation, MLO actually arose, providing the probability density function underlying the choice model is uni-modal, as is the case with the deterministic model, the logit model and other commonly used probabilistic choice models, then the number of local maxima will be two, making the task of identifying the global optimum relatively manageable.

The overriding conclusion that can be drawn is, therefore, that the possibility, in practice, of MLO existing in the bus optimisation problem is quite low, particularly if a logit model is used to model user behavioural mode choice. As a result, since the logit model forms an integral part of the analysis in this study, the question of MLO need not be further addressed in subsequent chapters.

The considerations of *unit subsidy* centred on recent work by Jansson (1993) in which optimal unit subsidy was found to change suddenly and substantially as users switched between random and planned behaviour. An implication of this result is that the conventional negative relationship between unit subsidy and patronage level breaks down. The analysis of this chapter has demonstrated that it is possible for the conventional unit subsidy/patronage relationship to also break down even when a logit choice model is used (rather than the deterministic model used by Jansson), making it difficult to speculate *a priori* the direction of the relationship between unit subsidy and patronage. However, the range of patronage levels over which the relationship falters, and the severity of the contradiction, decreases the more gradual is mode switching.

The key result in the analysis undertaken of *total subsidy* was that the growth of optimal total subsidy (in response to increases in patronage) is greater as mode switching occurs than it is before or after switching. Further, the more rapidly switching occurs, the greater is the relative growth of total subsidy during switching. Policy makers concerned with maintaining optimal outcomes in terms of economic efficiency may therefore be faced with the prospect of implementing quite significant jumps in total subsidy levels in potentially short periods of time as growth in patronage occurs in the presence of switching.

The literature has previously established that it is optimal to have *cross-subsidies* between bus routes when an overall breakeven constraint is imposed across a group of, or all, routes. The main focus here was to consider how optimal cross-subsidy results were affected by the introduction of the logit choice model. Random user behaviour was found to generate higher cross-subsidy levels than planned behaviour. In addition, when either random or planned behaviour applied throughout, the result previously reported in the literature, that a high patronage route should cross-subsidise a low patronage route, was confirmed. Further, cross-subsidy varied uniformly as patronage differences between routes varied.

Use of the logit model, however, was found to generate quite different cross-subsidy outcomes. With a logit model, it is now also possible for low patronage routes to cross-subsidise high patronage routes. In addition, as the patronage difference between routes grows, the cross subsidy no longer necessarily increases, nor does it necessarily change gradually : the cross-subsidy can now both increase and decrease, and can change both quickly and slowly, depending on the rate of switching between random and planned behaviour. Finally, although in a single behavioural mode model cross-subsidy is equal to zero only when route patronage levels are equal, in a logit model an optimal cross subsidy of zero is also possible in situations where route patronages differ. On the whole, it becomes more difficult to predict *a priori* how cross subsidy will behave.

Chapter Appendix : Optimal Cross-Subsidy Analysis, Derivation of First Order Conditions

Following the analysis in section 4A.3 in the appendix to chapter 4 with L (given by (5.5)) as the objective function, the FOC with respect to P_i, $\frac{\partial L}{\partial P_i} = 0$, is evaluated as follows.

$$\frac{\partial L}{\partial P_i} = \frac{\partial CS_i}{\partial g_i}\frac{\partial g_i}{\partial P_i} - (1+\lambda)\left(-q_i - P_i\frac{\partial q_i}{\partial P_i}\right) \tag{5A.1}$$

Note that there is no C_p term in (5A.1) because since $\frac{\partial C_{pi}}{\partial P_i} = \frac{\partial C_{pi}}{\partial q_i}\frac{\partial q_i}{\partial P_i}$, and from (4.3) $\frac{\partial C_{pi}}{\partial q_i} = 0$, then $\frac{\partial C_{pi}}{\partial P_i} = 0$. Substituting (4A.3) and (4A.12) into (5A.1) yields :

$$\frac{\partial L}{\partial P_i} = -q_i\left(1 + \frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i}\frac{\partial q_i}{\partial P_i}\right) + (1+\lambda)P_i\frac{\partial q_i}{\partial P_i} + q_i + \lambda q_i \tag{5A.2}$$

Setting $\frac{\partial L}{\partial P_i} = 0$ and rearranging :

$$P_i(1+\lambda)\frac{\partial q_i}{\partial P_i} = q_i\frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i}\frac{\partial q_i}{\partial P_i} - \lambda q_i$$

thus

$$P_i = \frac{1}{(1+\lambda)}q_i\frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i} - \frac{\lambda}{(1+\lambda)}\frac{q_i}{\partial q_i/\partial P_i} \tag{5A.3}$$

Now,

$$\frac{\partial q_i}{\partial P_i} = \frac{\partial q_i}{\partial g_i}\frac{\partial g_i}{\partial P_i} \tag{5A.4}$$

Then substituting (4A.12) into (5A.4) and rearranging yields :

$$\frac{\partial q_i}{\partial P_i} = \frac{\partial q_i}{\partial g_i}\Bigg/\left(1 - \frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i}\frac{\partial q_i}{\partial g_i}\right) \tag{5A.5}$$

Substituting (5A.5) for $\frac{\partial q_i}{\partial P_i}$ in (5A.3), the third component of (5A.3) reduces to :

$$-\frac{\lambda}{(1+\lambda)}q_i\frac{\left(1 - \frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i}\frac{\partial q_i}{\partial g_i}\right)}{\partial q_i/\partial g_i}$$

$$= -\frac{\lambda}{(1+\lambda)}\frac{q_i}{\partial q_i/\partial g_i} + \frac{\lambda}{(1+\lambda)}q_i\frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i} \tag{5A.6}$$

causing (5A.3) to reduce to :

$$P_i = q_i\frac{\partial v}{\partial LF_i}\frac{\partial LF_i}{\partial q_i} - \frac{\lambda}{(1+\lambda)}\frac{q_i}{\partial q_i/\partial g_i} \tag{5A.7}$$

The FOC with respect to F_i, $\dfrac{\partial L}{\partial F_i} = 0$, is complex to evaluate. Recall from chapter 4, that the FOC in cost minimisation $\left.\dfrac{\partial TC}{\partial F_i}\right|_{\bar{q}} = \left.\dfrac{\partial (C_p + C_u)}{\partial F_i}\right|_{\bar{q}} = 0$ is equivalent, and easier to evaluate. Now:

$$\left.\frac{\partial (C_{pi} + C_{ui})}{\partial F_i}\right|_{\bar{q}} = \frac{\partial C_{pi}}{\partial F_i} + \frac{\partial C_{ui}}{\partial F_i} \tag{5A.8}$$

Noting (4.4) and (4.5), and noting that since AC_o is constant then $\partial AC_o/\partial F = 0$, then (5A.8) reduces to :

$$\left.\frac{\partial (C_{pi} + C_{ui})}{\partial F_i}\right|_{\bar{q}} = \frac{\partial C_{pi}}{\partial F_i} + q_i \frac{\partial AC_{Fi}}{\partial F_i} \tag{5A.9}$$

Then with $AC_{Fi} = u(F_i) + v(LF(q_i, F_i, N), F_i)$, (5A.9) evaluates to :

$$\left.\frac{\partial (C_{pi} + C_{ui})}{\partial F_i}\right|_{\bar{q}} = \frac{\partial C_{pi}}{\partial F_i} + q_i \left(\frac{\partial u}{\partial F_i} + \frac{\partial v}{\partial LF_i} \left.\frac{\partial LF_i}{\partial F_i}\right|_{\bar{q}} + \left.\frac{\partial v}{\partial F_i}\right|_{\overline{LF},\bar{q}} \right) \tag{5A.10}$$

Setting $\left.\dfrac{\partial (C_p + C_u)}{\partial F_i}\right|_{\bar{q}} = 0$, (5A.10) reduces to :

$$-q_i \left(\frac{\partial u}{\partial F_i} + \frac{\partial v}{\partial LF_i} \left.\frac{\partial LF_i}{\partial F_i}\right|_{\bar{q}} + \left.\frac{\partial v}{\partial F_i}\right|_{\overline{LF},\bar{q}} \right) = \frac{\partial C_{pi}}{\partial F_i} \tag{5A.11}$$

which is identical to (4A.18) the result when there is no subsidy constraint. Although not presented here because of its complexity, the full derivation of $\dfrac{\partial L}{\partial F_i} = 0$ generated exactly the same result.

As usual, the FOC with respect to the Lagrange multiplier λ, $\dfrac{\partial L}{\partial \lambda} = 0$, yields the budget constraint

$$C_{p1} - P_1 q_1 + C_{p2} - P_2 q_2 = S_T (= 0) \tag{5A.12}$$

Chapter 6
SERVICE UNRELIABILITY AND SUBSIDY

6.1 Introduction

There is an interesting link between the degree of unreliability of bus services and the level of optimal bus subsidy which arises from using the user cost models of chapter 3 in subsidy analysis. In chapter 3, unreliability was established as a key determinant of user cost through the user cost component stochastic supply delay, the delay experienced by users if services fail to depart at their scheduled time. With service unreliability influencing user cost, and user cost being a key determinant of optimal bus subsidy, it follows that there is a potential link between service unreliability and optimal subsidy. This relationship has not been previously explored in the literature,[1] and the aim of this chapter is to therefore investigate the nature and implications of the relationship.

Exploring the nature of the link is important for two reasons. First, to the extent that service unreliability exists in bus networks, it is usually seen as good management practice to reduce the level of unreliability (provided of course that the marginal benefit of doing so exceeds the marginal cost). If unreliability is to be reduced, however, it is useful to be aware that this will impact on optimal subsidy, and to understand the scale of this impact. In fact, it will be argued in the chapter that a case can be made for policies on unreliability management and subsidy setting to be made jointly. Second, as will be shown in the chapter, there are circumstances where an increase in unreliability leads to an increase in optimal subsidy. In these circumstances, if subsidy is to be

[1] The only exception is Tisato (1990, 1992). However, although unreliability did influence user cost and in turn subsidy in that work, unreliability played a largely incidental role, thus the nature and magnitude of the subsidy/unreliability link was not investigated.

increased, it will be important to ensure that adequate complementary policies are in place to directly address the problem of unreliability. Otherwise the community is likely to question the logic of increasing bus subsidy in response to increases in service unreliability.

The chapter proceeds as follows. First, the nature of the subsidy/unreliability link is considered, including an assessment of the direction and magnitude of the impact of changes in unreliability on subsidy, and any related policy implications. Second, the question of how unreliability and subsidy policy should be managed in light of the fact that the two are inter-related is considered. Third, the role of road congestion in the first best analysis of bus subsidy[2] is briefly explored. Such a role becomes possible here because road congestion variability is an important source of bus unreliability.

Integral throughout the analysis in the chapter is a recognition of the bi-modal nature of user behavioural choice discussed in previous chapters, with the logit choice model developed in chapter 3 used to predict choice between random and planned behaviour. As in chapter 5, the analysis here is undertaken for the situation where load factor (*LF*) is variable and bus size (*N*) is given (*LF*/*N* fixity case 3 of chapter 4). The base analysis is undertaken for $N = 78$, the average bus size (seating plus standing passengers per bus) in Adelaide (see section B.3.2 of appendix B), with the sensitivity of the results to a smaller bus size assessed later.

6.2 The Impact of Unreliability on Subsidy

The aim of this section is to establish the direction and magnitude of the impact that changes in unreliability have on optimal total bus subsidy. The nature of the impact of unreliability on subsidy was established through a series of simulation runs for various levels of unreliability. The indicator of the level of unreliability continues to be σ, the standard deviation of bus departure time from the scheduled time. Section B.3.1 of appendix B indicated that a range of $\sigma = 1$ to 4 mins was an appropriate range of unreliability levels to consider.[3]

[2] In contrast to the more common linking of road congestion and bus subsidy in a second best context (see discussion in section 2.4.2 of chapter 2).

[3] Other parameter values used in the simulation runs were those in appendix B, including $f = 3$ corresponding to the median parameter value set in Table B.6 of appendix B.

It is useful to initially focus on the two single user behavioural mode cases, random and planned behaviour, as a progressive step towards considering the more realistic and complex case of logit discrete choice between these two behavioural modes.[4]

6.2.1 Random and Planned Behaviour Cases

Figure 4.2 of chapter 4, which illustrated a generalised optimal situation for *LF/N* case 3, continues to provide the relevant framework for the analysis being considered here. In Figure 4.2, optimal patronage, q^*, is determined by the intersection of the marginal benefit (*MB*) and marginal social cost (*MSC*) curves. With average total cost (*ATC*) declining, $MSC < ATC$, and so a subsidy is required to generate the optimal outcome. Optimal unit subsidy, s^*, equals the gap between *ATC* and *MSC* at q^*, with optimal total subsidy, S^*, being then given by the product of s^* and q^*.

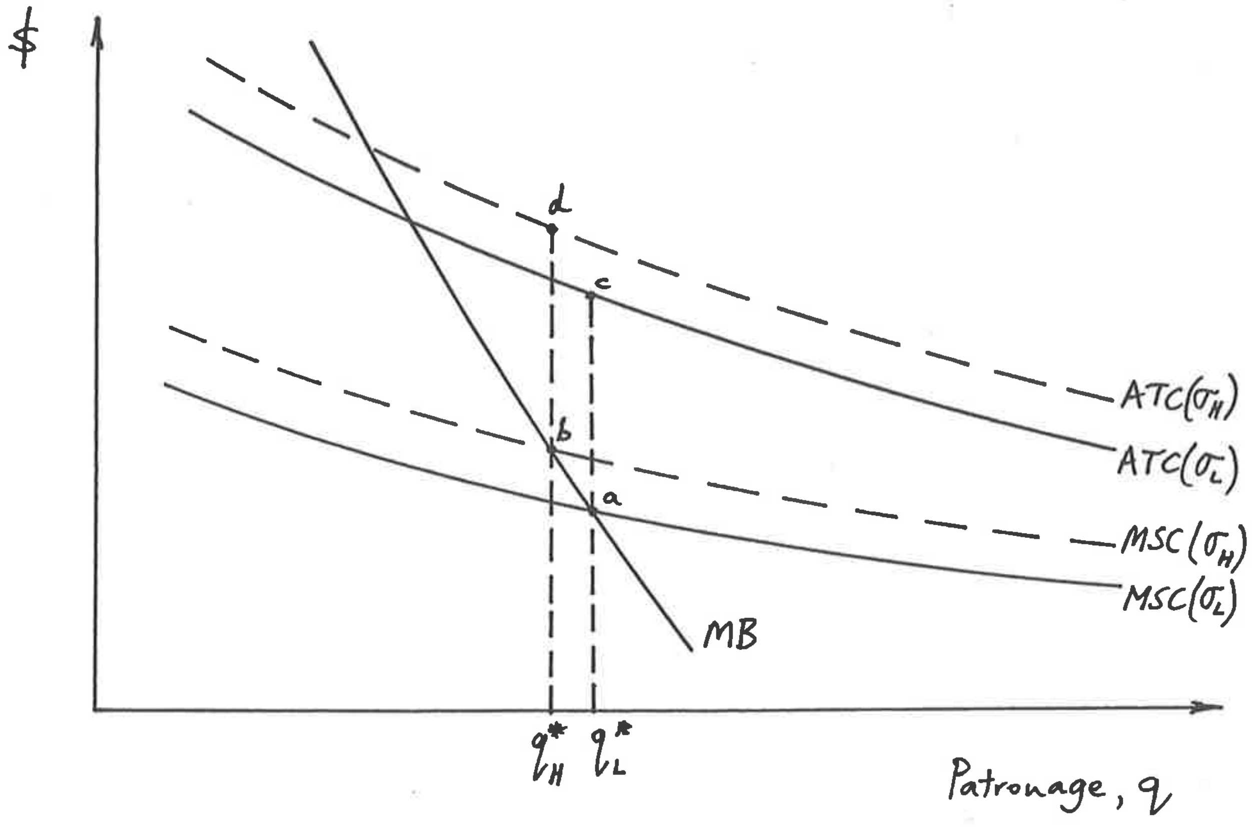
The demand curve is unaffected by changes in the level of unreliability (σ). Any demand response to changes in σ will consist of movements along the demand curve due to resulting changes in generalised cost faced by the user. To the extent that σ has any impact on subsidy outcomes, it is therefore due to shifts in the *ATC* and *MSC* curves. Two impacts are possible. First, relative shifts in the *ATC* and *MSC* schedules impact on optimal unit subsidy, s^*, which (as was established in chapter 4) is measured by the gap between the *ATC* and *MSC* schedules. Second, shifts in the *MSC* schedule shifts the intersection of *MSC* and *MB* and thus alters optimal patronage, q^*. Simulation runs for three unit changes in σ (1 to 2, 2 to 3, and 3 to 4) revealed a consistent pattern of impacts on *ATC* and *MSC*. In all cases, *ATC* and *MSC* increased at all patronage levels, *i.e. increasing σ always lifts the ATC and MSC schedules vertically.* These shifts of the *ATC* and *MSC* schedules are explained in the chapter appendix.

With the *ATC* and *MSC* schedules being lifted by an increase in unreliability (σ), the subsidy situation to be analysed here is therefore the one presented in Figure 6.1. The figure presents the optimal situation for two different levels of unreliability, σ_L and σ_H, where $\sigma_H > \sigma_L$. When unreliability increases from σ_L to σ_H, the following results can be noted :

- the optimum position (where *MB* and *MSC* intersect) moves from point *a* to point *b*;

[4] The subscripts $_r$ and $_p$ continue to signify random and planned user behaviour.

Figure 6.1 : Shifts in *ATC* and *MSC* Schedules Due To an Increase in Service Unreliability (σ)



- optimal patronage *declines* from q^*_L to q^*_H. This effect is denoted throughout as the *patronage effect*;
- optimal unit subsidy changes from s^*_L (= distance *ac*) to s^*_H (= distance *bd*). This effect is denoted throughout as the *unit subsidy effect*. With both *ATC* and *MSC* schedules rising, plus a change in q^*, the direction of change in unit subsidy from point *a* to point *b* is therefore unclear *a priori*; and
- optimal total subsidy, the product of optimal unit subsidy, s^*, and optimal patronage, q^*, changes from S^*_L to S^*_H. With the direction of change in s^* unclear, the direction of change in S^* also cannot be determined *a priori*.

The patronage effect is unambiguously negative. However, simulation runs were required to determine the sign of the unit subsidy effect. Once again, the runs were undertaken for a number of unit changes in σ (1 to 2, 2 to 3, and 3 to 4), and for a range of demand levels.[5] The simulation runs revealed consistent results for all unreliability (σ) and demand (α) levels considered. For random behaviour, an increase in σ results in a fall in unit subsidy, whilst the reverse is true for planned behaviour.

Table 6.1 provides an overall summary of the signs of the patronage and unit subsidy effects, and the total subsidy effect. A negative/positive sign (-/+) denotes a negative/positive effect, i.e. a rise in unreliability (σ) results in a fall/rise in the relevant variable. In the case of random behaviour, the patronage and unit subsidy effects are both negative throughout, resulting in a clear reduction in total subsidy. For the case of planned behaviour, the patronage and unit subsidy effects are opposite in sign, the former being negative and the latter positive. The relative strengths of these determine the overall total subsidy effect. The patronage effect dominates when $\alpha <\approx 700$, and thus the total subsidy effect is -ve, whilst the reverse is true when $\alpha >\approx 700$. Overall, therefore, the direction of impact of unreliability on subsidy varies between and within the user modes of behaviour.

[5] With patronage affected by σ, the indicator of demand level used here is the level of potential demand, α, the patronage level which would occur if generalised cost faced by the user, g, were set to zero. This is the patronage level where the demand curve cuts the horizontal (patronage) axis in Figure 6.1.

Table 6.1 : Direction of Patronage and Unit Subsidy Effects

Behavioural Mode	*Patronage Effect*	*Unit Subsidy Effect*	*Total Subsidy Effect*
Random	-	-	-
Planned	-	+	- if $\alpha <\approx 700$ + if $\alpha >\approx 700$

Note : *- implies a decline in the variable when unreliability, σ, increases.*
+ implies an increase in the variable when σ increases.

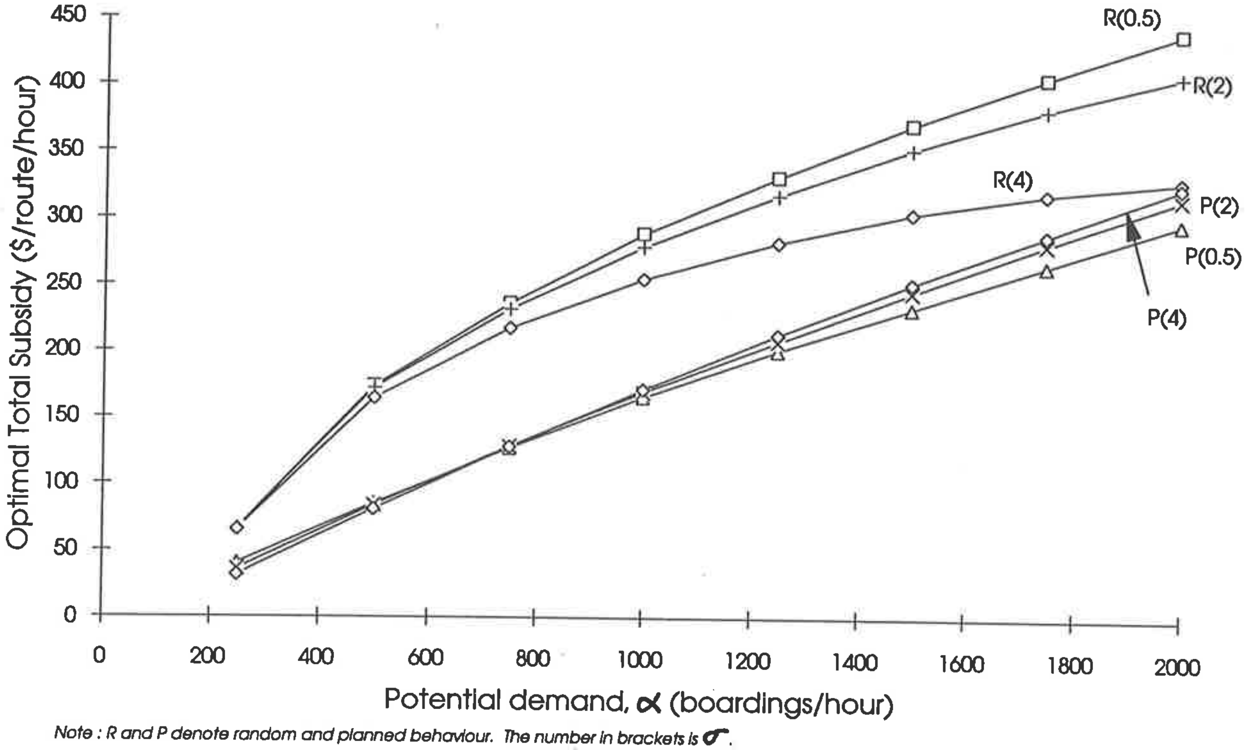
The results of Table 6.1 are confirmed in Figure 6.2 which plots optimal total subsidy (S^*) for three unreliability (σ) levels (0.5, 2 and 4) against values of demand level (α) : an increase in σ unambiguously decreases S^*_r, and either decreases or increases S^*_p depending on whether $\alpha <\approx$ or $>\approx 700$. Inspection of the percentage changes in subsidy in Figure 6.2 reveal that they are generally reasonably modest in magnitude. For planned behaviour, the impacts are particularly small. The changes are larger in the random behaviour case, but even there the percentage change in subsidy for any unit change in unreliability (σ) is generally less than 10%.

6.2.2 The Logit Choice Case

Now consider the more general, and important, case where users make a discrete choice between random and planned behaviour, with the logit model developed in chapter 3 predicting choice outcomes. The random and planned behaviour analysis of the last section provides a useful foundation for considerations here since logit subsidy results will tend towards those of the planned and random behaviour cases when these modes of behaviour dominate, namely at low and high demand levels respectively. However, the effects of behavioural mode switching, an integral part of the logit model, on optimal subsidy must now also be considered.

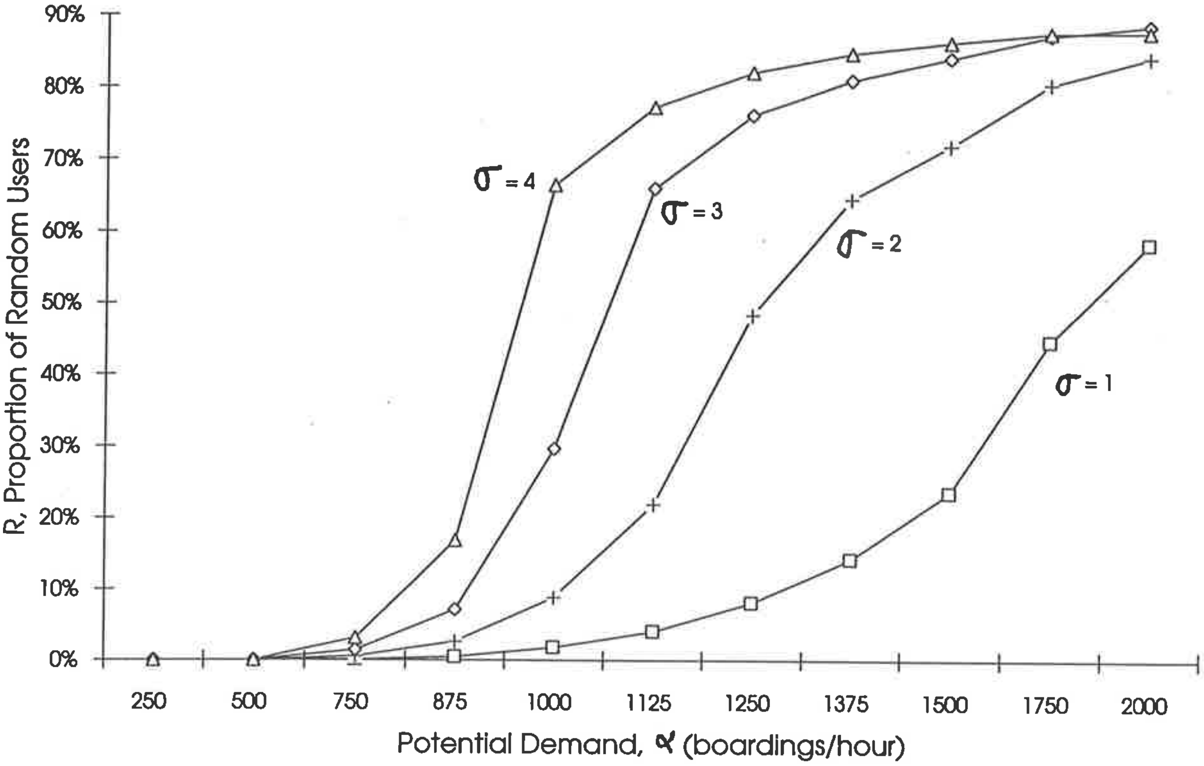
Figure 6.3 summarises the behavioural mode switching patterns which emerge for different levels of unreliability (σ), with the logit scale parameter, μ, set at 0.11 (- the value of μ will be varied in due course). The figure plots R, the proportion of random users, where an increase in R indicates that switching is occurring from planned to random behaviour. A consistent pattern exists for each curve. At low demand (α) levels, planned behaviour dominates. As demand level

Figure 6.2 : Optimal Total Subsidy Under Random and Planned Behaviour



Note : R and P denote random and planned behaviour. The number in brackets is σ.

Figure 6.3 : Proportion of Random Users In The Logit Model



increases, switching to random behaviour occurs progressively, with the commencement and rate of switching varying with σ, in accordance with the following two switching effects :

- the *switching commencement effect* - the bigger is σ, the lower the α level at which switching from planned to random behaviour commences; and
- the *switching rate effect* - the bigger is σ, the more rapid is the switching from planned to random behaviour as demand (α) increases.

To explain these two effects, reference must once again be made to the behaviour of the user cost component u. Chapter 3 demonstrated that, in a logit model, the probability of random behaviour, and thus the proportion of users acting in a random manner, was a function of two things : the difference u_r-u_p, which in turn is function of service headway, H; and the logit scale parameter, μ, which determines how quickly mode switching occurs as the critical headway, H_c (the headway at which u_r= u_p and thus random and planned behaviour are equally likely), is approached. Figure 6A.1 illustrated how this choice framework is affected by an increase in unreliability (σ) : the u_r schedule is raised and becomes flatter, whilst the u_p schedule is raised and becomes steeper. The key resulting impact in relation to the logit model is that there is a corresponding increase in H_c, with Table 6.2 showing that H_c increases progressively with increases in σ, although at a diminishing rate. This is intuitively reasonable : as σ increases, the benefits to the user of acting in a planned fashion diminish, thus increasing the proportion of situations in which the user is likely to act in a random manner.

Table 6.2 : Influence of Unreliability on Critical Headway, H_c

Service Unreliability, σ	*Critical Headway, H_C*
1	10.2
2	14.5
3	17.7
4	20.1

Note : All units are in minutes.

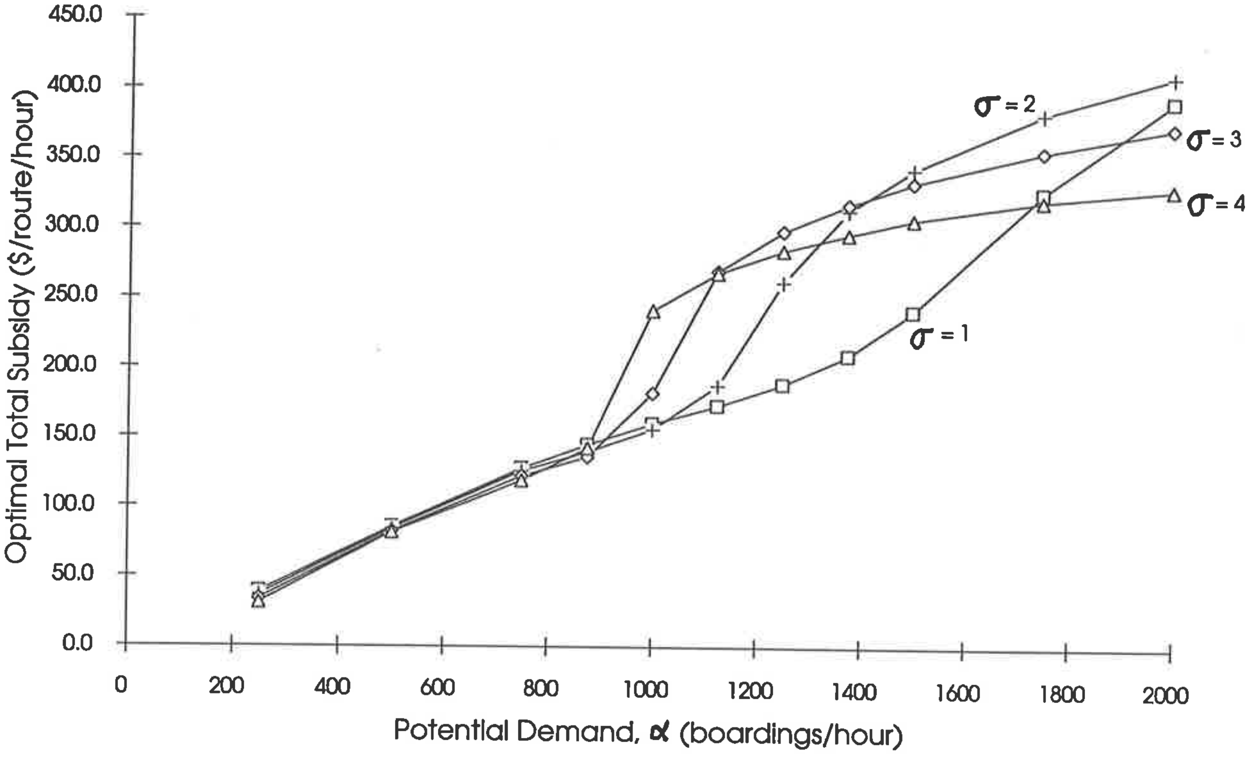
The increase in H_c provides an explanation of the *switching commencement effect*. At low demand (α) levels, service frequency is low and therefore headway (H) is high, thus planned behaviour occurs. As demand grows, frequency increases and H decreases. The fact that the higher σ is the higher is H_c then means that, as α grows and H falls, H_c is reached more quickly and therefore the lower the α level at which switching from planned to random behaviour commences (i.e. the *switching commencement effect*).

The growth in H_c is also critical to explaining the *switching rate effect*. Compare for example the $\sigma = 1$ and 4 cases. There are a number of steps to the argument. First, compared to $\sigma = 1$, the $\sigma = 4$ case is characterised by switching occurring at higher H values. Second, with frequency increasing relatively proportionally with increases in demand level (α), and given $H = 60/F$, a unit change in α will lead to a bigger change in H the bigger H is. Third, the bigger the change in H, *ceteribus paribus*, the more rapid the rate of switching. As a result, switching occurs more rapidly in the $\sigma = 4$ case than when $\sigma = 1$. The argument generalises to marginal changes in unreliability : the greater is σ, the more rapidly switching will occur from planned to random as demand increases (i.e. the *switching rate effect*).

The switching patterns observed in Figure 6.3 facilitate clearer understanding of the behaviour of optimal subsidy, S^*, which is plotted against demand level (α) in Figure 6.4 for four unreliability levels : $\sigma = 1$, 2, 3 and 4. Each curve consists of three stages, coinciding with the three stages of the R curves in Figure 6.3. To illustrate, consider for example the $\sigma = 4$ total subsidy curve. In the first stage, commencing at very low demand levels, planned behaviour dominates, with S corresponding closely with the relevant S_p schedule in Figure 6.2, growing steadily as α increases. The second stage coincides with the range of demand levels over which the bulk of the switching from planned to random behaviour occurs. In the $\sigma = 4$ case, this occurs approximately from $\alpha = 850$ to 1000. In the second stage, as the proportion of random users grows, S grows more rapidly as S is increasingly weighted towards the higher subsidy levels of the random subsidy schedule S_r observed in Figure 6.3.[6] As α grows further, when the majority of switching has occurred, the third stage is reached where random behaviour now dominates, and S corresponds increasingly closely with the random S schedule, S_r, in Figure 6.2. In the $\sigma = 4$ case, the third stage coincides with $\alpha >\approx 1000$. With switching occurring increasingly earlier the higher is σ (the *switching commencement effect*), the second stage is reached at lower demand levels. Further, with

[6] This type of pattern of more rapid growth in S whilst behavioural mode switching is occurring was similarly observed in figure 5.8 in chapter 5. There, however, S was plotted against optimal patronage rather than the potential demand level, α.

Figure 6.4 : Optimal Total Subsidy With Logit User Behavioural Choice



mode switching occurring more slowly the lower σ is (the *switching rate effect*), the more gradual is the rise in S during the second stage.

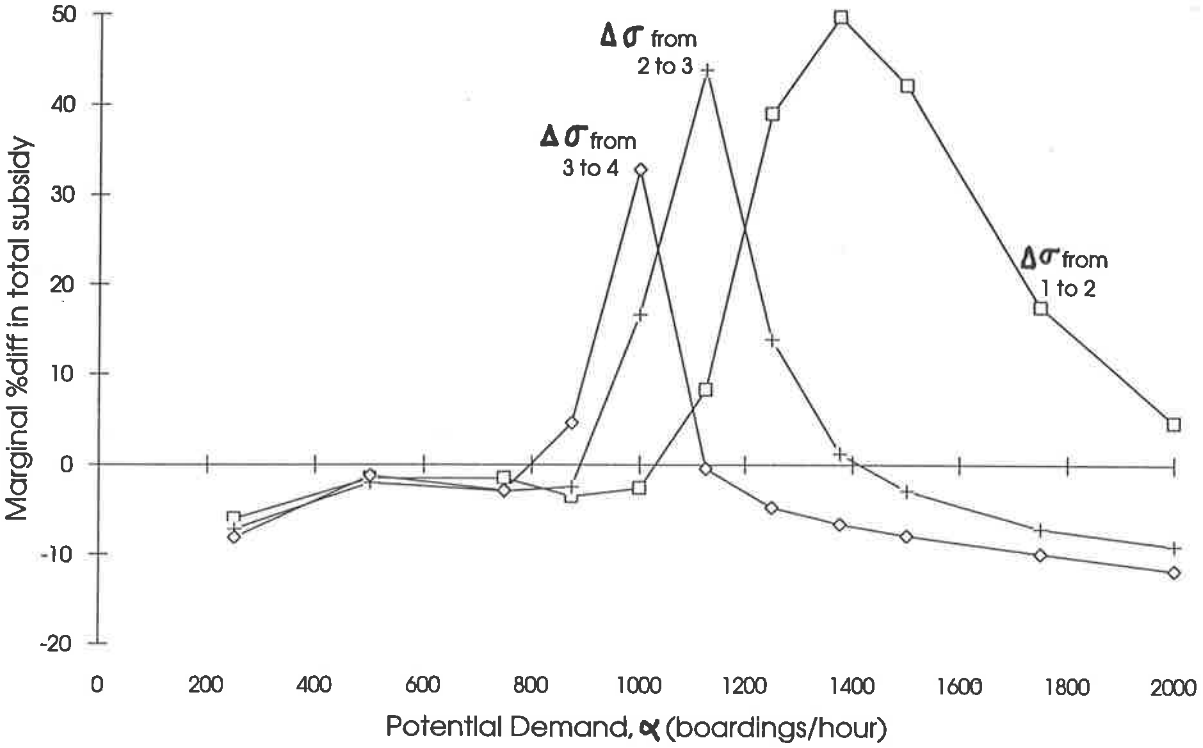
Marginal Impacts

Figure 6.4 has clearly illustrated that varying σ has a significant impact on optimal bus subsidy. One way of *summarising* the scale of the impact is through the % change which occurs in total subsidy when σ changes. The % change in S^* of unit changes in σ (i.e. from $\sigma = 1$ to 2, from 2 to 3, and from 3 to 4), denoted here as the *marginal* impact, is summarised in Figure 6.5 for the case where $\mu = 0.11$ and bus size (N) = 78. The main feature of the marginal impact curves in Figure 6.5 is the substantial peak which each curve displays over the range of demand values which coincide with users switching from planned to random behaviour. Within the peak, the marginal impact is positive (i.e. the increase in σ results in an *increase* in S^*), and of quite substantial magnitude, with the largest % impact being 33%, 47% and 55% respectively for the three unit changes in σ considered. Outside the peak, the marginal impact is negative throughout, i.e. a unit increase in σ yields a *decrease* in S^*, although the magnitude of impact is far more modest than within the peak, generally below 10%.[7]

The peak in each marginal impact curve occurs because of the different switching patterns which occur for the different unreliability (σ) cases. For example, consider a unit increase in σ from 2 to 3. The *switching commencement effect* suggests that, as demand increases, switching in the $\sigma = 2$ case lags behind switching in the $\sigma = 3$ case. Accordingly, the rise in S^* in Figure 6.4 for $\sigma = 2$ lags behind the rise in S^* for the $\sigma = 3$ case. As a result, for the (approximate) demand range $\alpha = 900$ to 1400, S^* is greater for $\sigma = 3$ than $\sigma = 2$, creating a *gap* between the S^* curves and causing the peak in the marginal impact schedule in Figure 6.5. In essence, at any given demand level in this range, the higher is unreliability, the greater will be the proportion of users acting in a random manner, and thus (given the higher subsidies that occur under random behaviour (see Figure 6.2)) the greater will be optimal subsidy.

[7] Before and after the peak, behaviour is exclusively planned and random respectively. Therefore, the impact of unreliability on subsidy coincides with that reported at the end of section 6.2.1 in discussion of Figure 6.2.

Figure 6.5 : Impact on Total Subsidy of Unit Changes in Service Unreliability (σ)



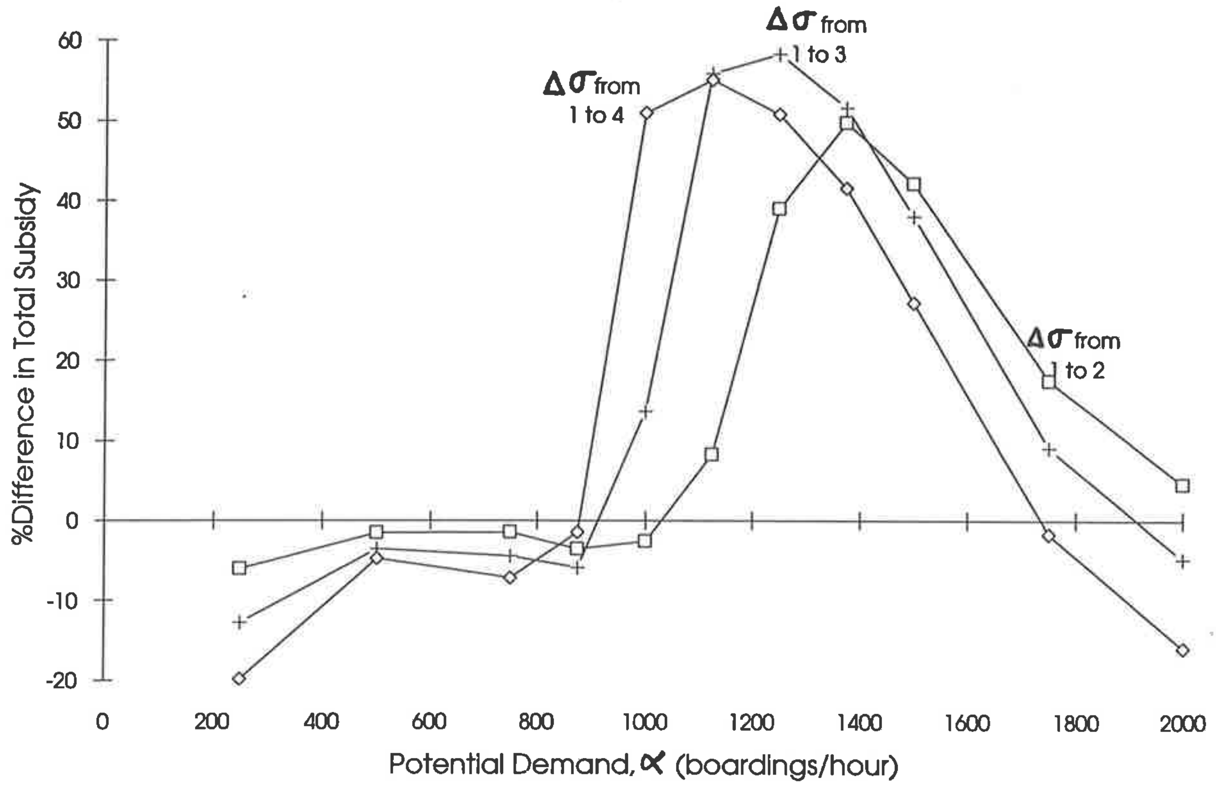
A further feature of the peak in the marginal impact curves is that, the smaller is the level of unreliability, the more pronounced is the peak in two respects : the largest % change is greater, and the peak persists over a greater range of demand levels. The peak for the unit change in σ from 2 to 3 is therefore higher and broader than the peak for the unit change in σ from 3 to 4, and the peak for the unit change in σ from 1 to 2 is in turn higher and broader than the peak for the unit change in σ from 2 to 3. This result is due to the *switching rate effect*, which causes switching to occur more gradually the lower that σ is. As a result, the S^* schedule in Figure 6.4 becomes progressively flatter as σ declines, increasing the size and spread of the gap between consecutive S^* curves.

The marginal impacts reported above suggest an interesting implication for policy makers and analysts. The above analysis suggests that, if a policy maker wishes to pursue the objective of economic efficiency, then there are circumstances where they should increase subsidy in response to an increase in unreliability. This would be the case at those demand levels coinciding with the peak of the marginal impact curves in Figure 6.5. Not only should subsidy be increased in response to an increase in unreliability, but the required increase in subsidy can be quite substantial. How, however, would the community react to such a policy recommendation? Although this study has no evidence of a definite response, it is not unreasonable to postulate a possible negative response from the community, for the simple reason that a subsidy might be seen as a form of financial assistance being paid in a situation of worsening performance. Although the policy maker would be making such recommendations on sound economic efficiency grounds, they may find a clash with public perceptions. It may be difficult therefore for policy makers to bring about these subsidy increases, even though they would be justified in doing so based on economic efficiency grounds.

Aggregate Impacts

It is also interesting to observe the impact of larger changes in unreliability (σ), which will be denoted as *aggregate* impact, in contrast to the impacts of unit (*marginal*) changes discussed above. Figure 6.6 plots the % change in optimal total subsidy for three cases, in which σ changes from a base value of σ = 1 to an end value of 2, 3 and 4, i.e. the three changes in σ are 1 to 2, 1 to 3, and 1 to 4. Once again, the main feature of the figure is the peak in each curve, coinciding with relative mode switching effects. On the whole, the shape and scale of the peak does not vary significantly between the three cases considered. The peak is, however, shifted gradually leftwards

Figure 6.6 : Impact on Total Subsidy of Larger Changes in Service Unreliability (σ)



as the change in σ becomes progressively bigger, due to the fact that switching is commencing at lower demand levels the higher is the end value of σ to which σ moves. The figure therefore suggests that the scale of impact on total subsidy is fairly robust to changes in σ of different size, not varying significantly between cases of small and large changes in σ.

6.2.3 Sensitivity Analysis

The sensitivity of the above results to changes in key parameters was also assessed. This was done by observing the sensitivity of the marginal impact curve in Figure 6.5 for the case where σ changes from 2 to 3. Sensitivity to changes in two parameters were considered to be of major importance : the logit scale parameter, μ, which reflects how quickly behavioural mode switching occurs as H approaches H_c; and, bus size, N.

The influence of variation in μ is illustrated in Figure 6.7. The marginal impact curve does appear to be quite sensitive to the choice of μ. The bigger μ is, that is the more quickly mode switching occurs, the narrower is the peak in the marginal impact curve, but the greater is the peak's height. For the more rapid switching cases, $\mu = 0.22$ and $\mu = 0.11$, the scale of the peak is quite significant. Although the % changes in the $\mu = 0.05$ case are not negligible, they are still substantial, and they persist over a greater range of demand levels.

Figure 6.8 then illustrates the influence on the marginal impact curves of varying bus size. In this case, the scale of the marginal impact peak appears to not vary significantly as bus size changes, although the width of the peak increases somewhat as bus size increases. The main feature is that the peak moves leftward as bus size decreases, i.e. the lower the bus size, the lower the demand range over which the peak occurs. This is the case because a smaller bus size results in higher frequency and thus lower headway, and thus the lower will be the demand (α) level at which mode switching will occur. Overall, varying bus size appears to influence the range of demand levels over which changes in unreliability have an impact, but does not greatly alter the scale of the impact.

The over-riding conclusion that can be drawn from the analysis in this section is that service unreliability can have quite a significant influence on subsidy analysis and on the level of optimal subsidy, particularly when a logit model is used to predict user choice between random and planned

Figure 6.7 : Sensitivity of Marginal Impact to Changes in μ

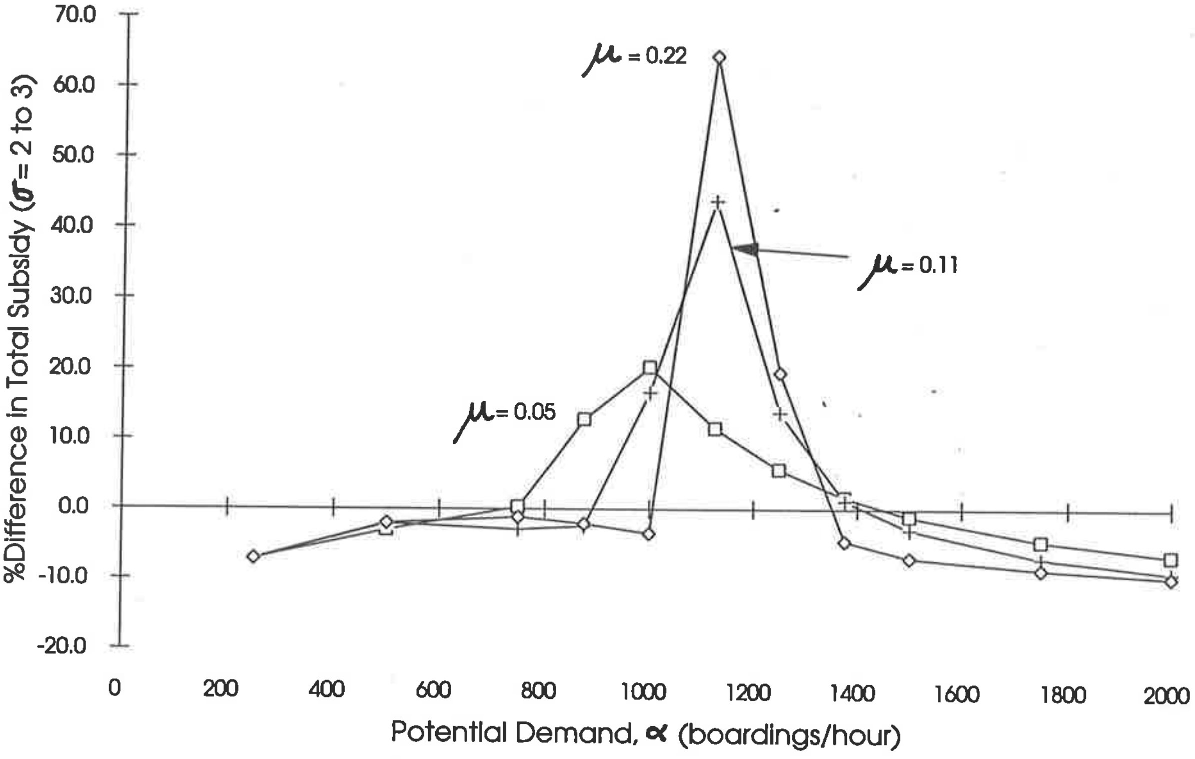
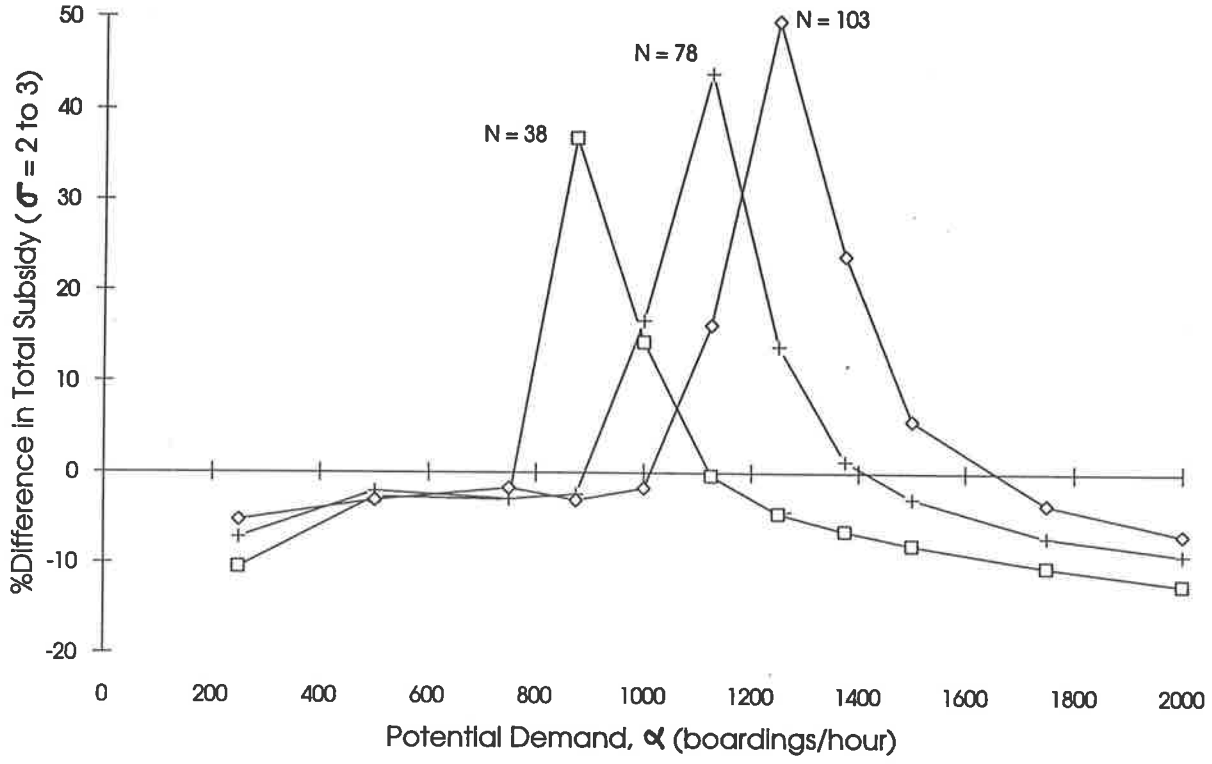


Figure 6.8 : Sensitivity of Marginal Impact to Changes in Bus Size, *N*



behaviour. It is essential, therefore, for service unreliability to play a role in the analysis and estimation of optimal bus subsidy.

6.3 The Role of Service Unreliability in Subsidy Policy and Analysis

6.3.1 Unreliability Management and Optimal Subsidy Policy

The last section has demonstrated that there is an important link between service unreliability and subsidy which can significantly influence optimal subsidy results. One problem has already been identified, however, in terms of putting the link into practice, namely, the potential difficulties associated with trying to increase subsidy in response to an increase in unreliability. This in turn raises the broader issue of what is the appropriate level of unreliability which should be used in setting subsidy policy in situations where scope exists for unreliability to be reduced through appropriate measures? Should the current level of unreliability be used, or should some lower level be used which accounts for the potential for improving service reliability?

In the *medium to longer term*, it is clearly desirable, and intuitively sensible, for unreliability to be directly reduced through appropriate measures, and for subsidy determination to be based on the *improved* level of reliability. In the *short term,* the appropriate action is less clear. One could argue that given the economic efficiency link between optimal subsidy and unreliability established above, the greatest gains in economic surplus can be realised by recognising all components of unreliability, no matter what their source, when setting subsidy. Based on this logic, subsidy should be based on the *actual* level of service unreliability, irrespective of whether scope exists for reducing that unreliability. On the other hand, setting subsidy in this way may act as a disincentive to the implementation of unreliability reducing measures, preventing the long run from being reached, which suggests basing subsidy on a level of unreliability below actual levels may be more appropriate.

Bus service unreliability is caused by a range of different factors. Strathman and Hopper (1993) and Adebisi (1986) identify a number of important sources, including *inter alia* : the number of alighting passengers, the length of experience of the bus driver, whether the bus driver was employed part-time or full-time, and variability in passenger demand and road congestion from day to day. Discussions with the bus operator in Adelaide (Seaman, 1994) revealed that union imposed work practices can also influence driver behaviour in ways which are detrimental to service unreliability. For the sources of service unreliability that are within the control of the bus operator, better management and operating practices can lead to more reliable services. It is clearly good management practice to pursue these improvements (provided the benefits of doing so exceed the costs).

Even if the sources of unreliability within the control of the operator can be minimised, variation in user demand and road congestion still remain as causes of service unreliability. Although the occurrence of these factors are outside the control of the operator, their impact can be reduced through measures such as planned non-running times between bus runs, reducing the number of bus stops, exclusive bus lanes, etc. These measures, however, clearly generate producer and user costs of their own. As a result, reducing these sources of unreliability therefore involves a trade-off between the associated benefits and the costs, with a positive level of unreliability always likely to be the optimal outcome (Strathman and Hopper, 1993).

Such a trade-off suggests that the level of unreliability can itself be optimised. In addition, with optimal subsidy being a function of unreliability, a case can also be made for both the optimal level of unreliability and the optimal level of subsidy being jointly determined in a combined optimisation exercise. In such an optimisation, in addition to economic surplus being a function of the variables discussed in chapter 4, it will now also be a function of the benefits and costs of reducing unreliability through measures such as bus lanes, non-running times, etc. This joint optimisation was not undertaken in this study however, and remains a topic for future research.

6.3.2 Road Congestion in First Best Bus Subsidy Analysis

As discussed above, variation in road congestion from day to day is one of the many sources of bus service unreliability. There is an important reason, however, to focus on this separately from other sources of unreliability. As pointed out in chapter 2 (see section 2.4.2), road congestion has played a major role in the past in the analysis of bus subsidy, being the focus of one of the key arguments used to justify subsidy. The argument has been that subsidy is justified on the grounds that it is a second-best instrument for managing road congestion in a world where roads are unpriced, as they have almost universally tended to be.

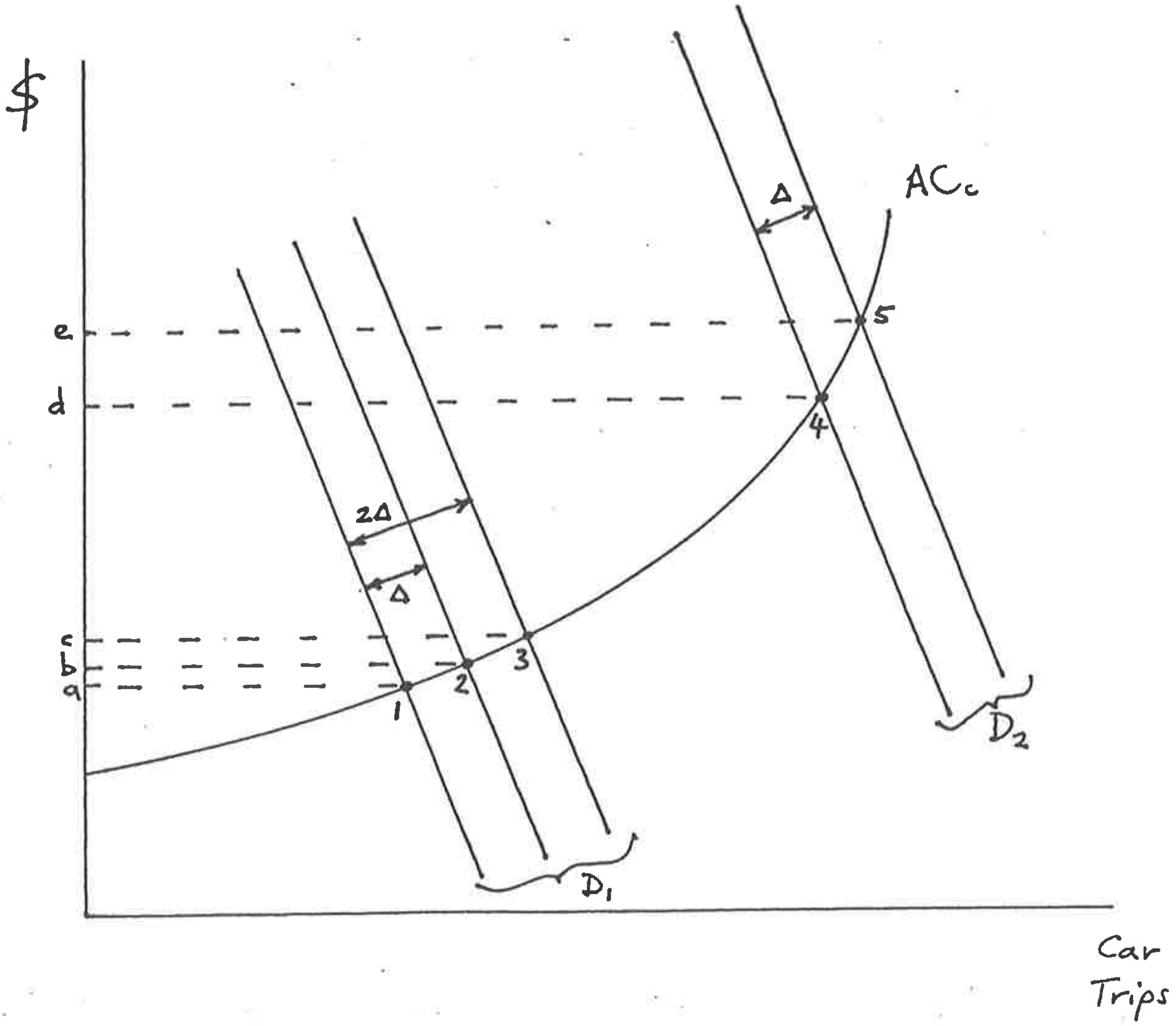
The argument has received considerable criticism, however, in recent times. The criticism has been that, although the argument is theoretically sound, with the cross price elasticity between public transport and car travel being very small, significant levels of congestion are required to obtain any significant reduction in road congestion from subsidised public transport fares. As a result, it will only be in the world's most highly congested cities that this argument is likely to play an important role in subsidy justification. The argument is likely to play a secondary role in a city like Adelaide where congestion levels are relatively modest.

With the importance of the second best subsidy argument seemingly diminished, it may appear that road congestion would then play a limited role only in optimal subsidy analysis. This is not the case, however, since road congestion can also play a role in a *first best* setting. Two such cases can be established. First, Kerin (1990) has argued that, as bus size increases, the case for user economies of scale subsidy is at least partially offset by the deleterious effect that larger buses have on road congestion. Second, road congestion also influences subsidy once one accounts for the fact that road congestion is one of the factors contributing to unreliability,[8] which in turn influences optimal subsidy (as illustrated in this chapter). The greater the variability in road congestion, the greater the degree of service unreliability, which in turn affects the optimal bus subsidy as illustrated in Figure 6.4. Road congestion therefore still has a role to play in subsidy analysis, but now in a first best, rather than a second best, context.

Two important factors which determine the extent of road congestion variability are the degree of car travel demand fluctuation, and the current level of road congestion. The greater these two variables, the greater will be the level of unreliability. To demonstrate this, consider Figure 6.9.

[8] Road congestion affects unreliability through its influence on bus speed variability. Congested road conditions are not on their own sufficient to result in service unreliability. If daily road congestion patterns were repeated from day to day, operators could allow for this in service scheduling. However, once road congestion, and thus bus speed, varies from day to day, it becomes increasingly difficult to keep buses running to schedule. As discussed earlier, the impact of road congestion variability can be partly managed with appropriate measures such as bus lanes. In addition, the bus driver arguably has some capacity, through speed changes, to try to keep to a schedule when faced with changing road congestion conditions. This is clearly the case when congestion is lighter than expected. When congestion is heavier than expected, however, there is usually limited scope for speed to be increased due to the greater than expected congestion, and ultimately speed restrictions (Seaman, 1994).

Figure 6.9 : Road Congestion Variability and Service Unreliability



The figure shows demand fluctuation occurring at two different levels of car travel demand, *D1* and *D2*. The system equilibrium, which is at the intersection of the *D* curve and the car travel time cost curve, AC_c, is an indicator of the level of road congestion. Fluctuation in this equilibrium thus indicates fluctuation in road congestion and service unreliability.

For demand level *D1*, the figure shows two levels of demand fluctuation, Δ and 2Δ, with the equilibrium moving from point 1 to point 2, and from point 1 to point 3, respectively in the two cases. The resulting change in the level of road congestion ($a \rightarrow\ b$ vs $a \rightarrow\ c$) is clearly greater for the case of the larger demand fluctuation. Road congestion variation, and thus service unreliability, will therefore be greater the greater is the degree of demand fluctuation.

Next compare a uniform demand fluctuation, of size Δ, for the two demand situations *D1* and *D2*. The equilibrium changes from point 1 to point 2, and from point 4 to point 5, respectively in the two cases. The resulting change in the level of road congestion ($a \rightarrow\ b$ vs $d \rightarrow\ e$) is clearly greater for the higher demand level *D2*, due to the non-linear nature of AC_c (which is caused by the non-linear nature of road travel time functions (Akcelik, 1978)). Road congestion variation, and thus service unreliability, will therefore be greater the greater the current level of road congestion.

6.4 Chapter Summary and Conclusions

One of the deficiencies in user economies of scale subsidy analyses undertaken to date has been the almost complete lack of recognition of, and attention to, the influence of service unreliability on optimal bus subsidy. A link exists between service unreliability and optimal bus subsidy because subsidy is strongly influenced by user cost, which is in turn affected by service unreliability. The aim of this chapter has been to explore the nature of this relationship between service unreliability and subsidy. The impact on optimal subsidy of changes in the level of unreliability was assessed, the issue of how unreliability and subsidy policy should be managed in light of the fact that the two are inter-related was considered, and a role for road congestion (through its influence on unreliability) in first-best subsidy analysis was briefly established.

The over riding conclusion that can be drawn from the analysis of the impact on subsidy of changes in service unreliability is that unreliability can have quite a significant influence on subsidy analysis and on the level of optimal subsidy. The impact is rather small in cases where a single mode of user behaviour occurs, either random or planned. In contrast, the impact is much more significant when a logit model is used to predict mode choice. The most critical contributing factor to the logit model results is the influence that changes in unreliability have on the timing and nature of behavioural mode switching patterns. Changes in unreliability can lead to sudden changes in these switching patterns, which can in turn result in quite substantial changes in optimal subsidy levels. Percentage impacts of 50% (and greater in some cases) were found at demand levels over which users switch from planned to random behaviour. Overall, one could conclude that service unreliability forms an important part of optimal subsidy analysis and estimation. Neglecting unreliability, or incorrectly measuring its scale, can lead to significant errors in optimal subsidy estimates.

Notwithstanding the validity of the unreliability/optimal subsidy relationship on economic efficiency grounds, some potential problems and issues arise with respect to putting the link into practice. A problem which was identified related to the result that subsidy should be *increased* in response to an increase in service unreliability in certain circumstances, namely, when the level of demand falls within the range of values over which users switch from planned to random behaviour. However, the community may oppose such a policy recommendation for the simple reason that additional subsidy might be seen as a form of financial assistance being paid in a situation of worsening performance. It may be difficult therefore for policy makers to bring about such subsidy increases, even though they would be justified in doing so on economic efficiency grounds.

A broader issue is the question of what is the appropriate level of unreliability which should be used in setting subsidy policy in situations where scope exists for unreliability to be reduced through appropriate measures? Should the current level of unreliability be used, or should some lower level be used which accounts for the potential for improving service reliability? In the *medium to longer term*, it is clearly desirable for unreliability to be directly reduced through appropriate measures, and for subsidy determination to be based on the *improved* level of reliability. In the *short term*, the appropriate action is less clear. The greatest gains in economic surplus can be made by basing subsidy policy on *actual* levels of service unreliability, but to the extent that higher subsidy may act as a disincentive to implementing unreliability reducing measures, basing subsidy on a level of unreliability below actual levels may be more appropriate.

Improvements in operating and management practices by the bus operator will reduce unreliability. In addition, measures which reduce the impact of variation in user demand and road congestion, such as bus lanes and planned non-running times, also have an important role to play. However, a trade-off exists between the costs and benefits of such measures, suggesting that a positive level of unreliability is optimal. Further, with service unreliability and optimal subsidy closely related, for an overall optimum, these two should be jointly optimised.

Finally, it was argued that, with road congestion variability being a determinant of service unreliability, road congestion has a role to play in the first best analysis of bus subsidy. This is in contrast to past experience where road congestion has mainly played a second best role in subsidy justification as an instrument for managing road congestion in a world of unpriced roads.

Chapter Appendix :

In section 6.2.1, it was reported that an increase in unreliability (σ) leads to the *ATC* and *MSC* vs *q* schedules being lifted vertically. This can be explained as follows. From (4.10) and (4.35), $ATC = AC_p + AC_F(= u + v) + AC_o$. Then, with AC_o constant, and noting (4.3), (4.9) and (3.49), it is clear that the only element of *ATC* which is directly influenced by σ is *u*, that part of frequency related user cost which is unrelated to passenger congestion effects. In addition, (4.25) and (4.26) indicate that *MSC* is a function of both *ATC* and its slope. Therefore, the only influence of unreliability (σ) on *ATC* and *MSC* is through its influence on *u*, and as such the behaviour of *u* is thus the key to understanding the impact of σ.

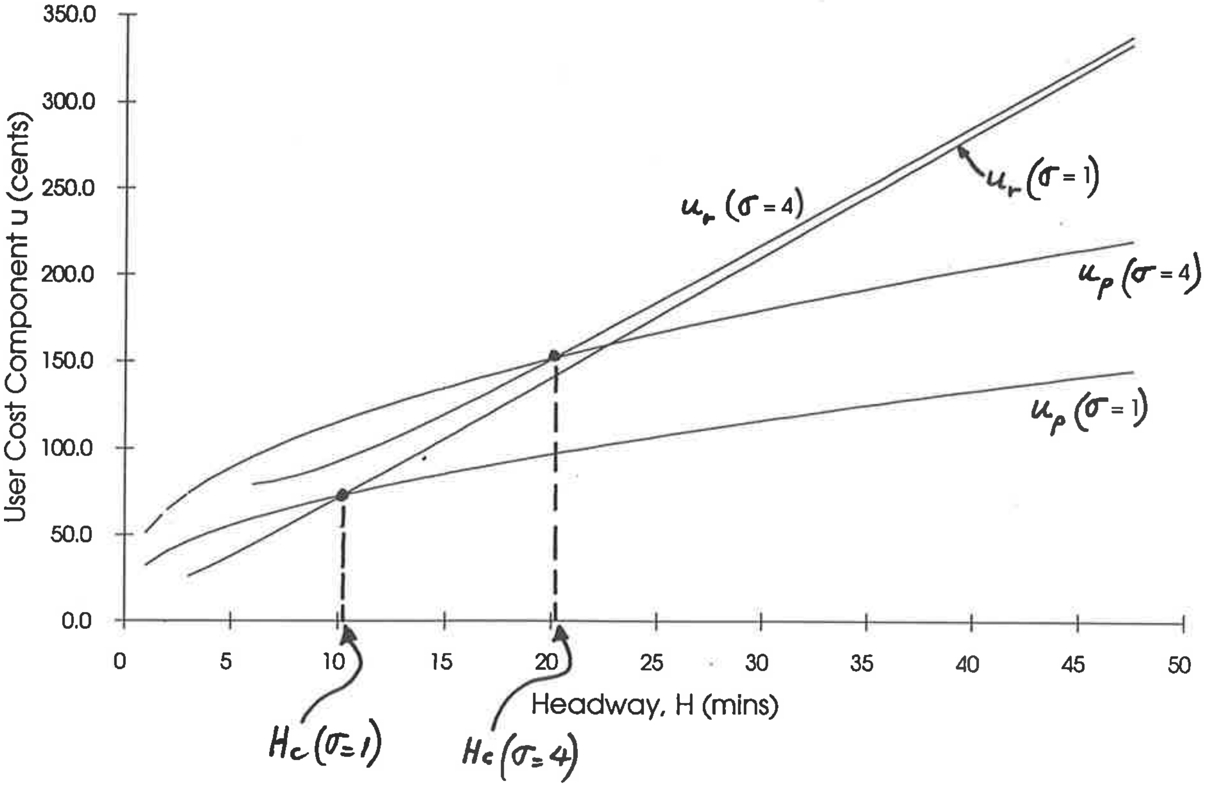
Figure 6A.1[9] demonstrates the influence of an increase in σ on *u* for random and planned behaviour (u_r and u_p). There are two changes to note. First, for both behavioural modes, *u* increases at all *H* values. Second, the slope of the *u* schedule changes : the u_r schedule becomes flatter, whilst the u_p schedule becomes steeper.[10] The fact that *u* increases throughout, explains why the *ATC* schedule rose for both behavioural modes in the simulation runs. The rise in the *MSC* vs *q* schedule in the simulation runs is, however, a bit more complicated to explain. As discussed above, *MSC* is a function of both *ATC* and its slope : *MSC* rises when *ATC* rises, and falls when the slope of *ATC* rises. For random behaviour, a rise in σ caused *u* (and thus *ATC*) to rise, but its slope fell, thus *MSC* must unambiguously rise. On the other hand, for planned behaviour, a rise in σ caused *u* (and thus *ATC*) and its slope to both rise, with the net impact on *MSC* therefore being unclear. Inspection of the simulation runs revealed that the net effect was for the *MSC* vs *q* schedule to rise.

[9] Which is a repetition of Figure 3.8, but drawn for two σ values. The two σ values chosen were 1 and 4 to illustrate the range of possible movement in the *u* schedules as σ changes.

[10] Chapter 3 has already provided explanations for these effects :

- For random behaviour, recall from chapter 3 (section 3.5.2) that unreliability generates a distribution of actual headways distributed around the scheduled headway, with a greater proportion of users therefore arriving at the bus stop in the longer headways, thus biasing upwards the expected wait, *u*, i.e. an upward shift in the *u* vs *H* schedule. It was also explained that this upward biasing was stronger the smaller the scheduled headway, thus causing the *u* vs *H* schedule to become flatter as σ increased.
- When user behaviour is planned, recall from chapter 3 (section 3.6.3) that an increase in σ results in the SSD_p vs *H* schedule (and thus the *u* vs *H* schedule) rising and becoming steeper. It rises because if buses are more unreliable, the user will miss a bus on more occasions because it arrives early, leading to a longer expected wait. The schedule is steeper because, with more unreliable buses causing the expected delay from missing a bus to rise, the greater is the reduction in the expected delay brought about by a unit reduction in *H*.

Figure 6A.1 : The Influence of Unreliability on User Cost Component *u*



Chapter 7
USER ECONOMIES OF SCALE SUBSIDY IN ADELAIDE

7.1 Introduction

The previous chapters in this study have been concerned with presenting, and extending in a number of ways, the general analysis of optimal pricing, frequency and subsidy for urban buses, with user economies of scale (*UES*) as the central focus. The contributions made so far have been of a general nature, relevant to public transport analysis across different cities. In this chapter, the focus now shifts to application of the user economies of scale concept to Adelaide and estimation of the subsidy levels which can be justified on these grounds. The purpose of doing so is to broaden the set of information available to policy makers in Adelaide who have the responsibility of administering and generating pricing, service level and subsidy policy.

Several qualifications apply. First, this study is limited to urban bus transport only, the dominant mode of public transport in Adelaide (and most other Australian cities).[1] Second, there are a range of subsidy arguments which have been advanced in the past in favour of subsidy (see the discussion in section 2.4 of chapter 2). By focusing on *UES*, the subsidy results presented here must be seen as part only of a broader analysis of subsidy which gives due weight to those alternative subsidy arguments. Notwithstanding this, as argued in chapter 2 (see section 2.4), the *UES* argument is an important argument in favour of subsidy, especially in a relatively low road congestion city like Adelaide where the popular second best road congestion management argument for subsidy is likely to play a secondary role only. The *UES* subsidy argument should therefore carry a fair degree of weight in subsidy assessment, as it has in overseas subsidy studies (e.g. Mohring, 1972; Jansson, 1979; Glaister, 1982; 1987; Bly and Oldfield, 1987).

[1] Recall footnote 1 of chapter 1.

The assessment in this chapter makes a number of contributions beyond previous Adelaide work (Dodgson, 1985, 1986; Chalmers, 1990; Kerin, 1990 : section 2.6 of chapter 2 identified limitations of these studies), consisting of :

- A more comprehensive disaggregated analysis - Dodgson's work was undertaken at the highly aggregated entire network/all day average level, whilst Kerin only considered a single route in the peak period. Although Chalmers considered both peak and off-peak, his work was limited to only a couple of bus routes. In contrast, this chapter estimates subsidy for the majority of the network for both peak and off-peak, and does so by working upwards from a considerable level of disaggregation.
- A closer assessment of the relationship between patronage level and subsidy - the earlier Adelaide studies cast little attention on the relationship between patronage level and subsidy, yet the literature (e.g. Gwilliam *et al*), and earlier chapters of this study (see chapter 5 in particular) indicate that this is an important component of *UES* subsidy assessment.
- The use of an improved user behavioural choice model - the Chalmers work, the main piece of previous *UES* subsidy estimation in Adelaide, used the simple random arrivals model to represent user behaviour. As chapter 3 suggested, however, this is a rather limited and unrealistic model of user behaviour. Dodgson and Kerin use a more realistic model, but one which lacks strong theoretical underpinnings. The sensitivity of subsidy results to the nature of user behavioural assumptions and models (see Tisato, 1992; and chapters 5 and 6 above) suggests reestimation for Adelaide using the superior user cost model developed in chapter 3 is justified.[2]
- Optimisation with explicit recognition of two market segments, full fare paying vs concession users. This contrasts with previous work which modelled a single average bus user.

The chapter also tests the off-peak robustness of Kerin's (1990) peak period conclusion that *UES* subsidy will be small.

[2] The analysis here incorporates some modifications to the general user cost model presented in chapter 3 to make it more suitable for the analysis of Adelaide buses (see section 7A.1 of the appendix to this chapter for details).

7.2 Framework for the Analysis

7.2.1 The Merits and Need For Disaggregation

The aim of the study is to identify optimal policy settings for the Adelaide bus system. As with all system studies, an important decision to be made is the choice of the level of disaggregation for the analysis, that is, the way the system should be split into separate parts (if at all) for separate analysis. There are both benefits (advantages) and costs (disadvantages) of disaggregation, with the optimal level of disaggregation being determined by the nature of the trade-off between these.

The *benefits* of disaggregation flow from being able to explicitly recognise in subsequent analysis (and policy recommendations which flow from it) the variations that exist across the system in both demand (preferences) and supply (costs), thus improving the scope for increasing allocative efficiency. In a study where all relationships are linear, disaggregation has limited benefits, with the simple approach of modelling the entire system as one representative element (in this case a representative route) with average parameter values used to quantify optimal outcomes likely to yield satisfactory results. On the other hand, when system relationships are non-linear, a simple averaging approach may yield significant errors, and the validity of such an approach could therefore be questioned.

In Adelaide, as in most cities, bus system relationships tend to have non-linear characteristics, implying tangible benefits from disaggregation. Two dimensions of non-linearity exist in this case study. First, as shown in chapter 3, delays vary non-linearly with patronage. Thus variations in demand between different parts of the network result in significantly different delays. It is important therefore that demand variation between areas be modelled. A second example of non-linearity is the pattern of bus use, with two peak periods in each weekday, and demand on weekends being lower than on weekdays. Peak load economics suggests that bus capital costs are determined exclusively by peak needs, so costs for peak and off-peak will be significantly different, requiring these periods to be modelled separately.

Although disaggregation has merits, this does not necessarily imply that one should adopt the highest level of disaggregation possible, since there are also costs associated with disaggregation. Disaggregation involves two types of *costs*, or *disadvantages*. First, the more disaggregated the analysis, the greater the volume of data required and the more detailed is that data. Therefore, the greater the disaggregation, the greater the commitment of resources and time required for data collection, collation and modelling. A second disadvantage of disaggregation can be that the more one disaggregates, the less implementable are the analytical policy recommendations for the simple reason that in practice, for reasons of feasibility or political acceptability, there may be limitations to the extent that policy settings can be varied. In that case, high levels of disaggregation may be out of step with practical policy making considerations, with the analyst thus running the risk of their analysis being of little interest to policy makers. Although highly disaggregated analysis can still have merits in its own right, if the analysis to be undertaken is to be of use in policy formulation, then it is sensible to link the level of disaggregation to feasible levels of policy setting implementation in practice.

For example, the analytical approach in this study could consist of analysis of every route in the network, with a unique set of policy settings being determined for each route. In reality, however, one of the aims in policy design is *simplicity*, in order to ensure understanding and transparency. For example, in practice, considerable weight tends to be given to the simplicity objective when designing a pricing system, so pricing systems are rarely, if ever, designed on an individual route basis. Basing analysis on an individual route level is therefore judged to be too detailed, with a coarser level of disaggregation being appropriate.

Overall, a trade-off therefore exists when deciding on an appropriate level of disaggregation for analysis. On the one hand, greater disaggregation allows for greater variation in preferences and costs to be reflected in analysis and the policy recommendations which flow from it, thus improving allocative efficiency. On the other hand, the cost of data collection, and the need for simplicity and transparency, demands some restriction to the degree of disaggregation used.

7.2.2 The Disaggregation Approach Adopted

In settling on a level of disaggregation for this study, a pragmatic approach was adopted to making this trade-off. First, account was taken of the fact that current policy settings have a very high degree of uniformity across the system, and that economic efficiency could probably be improved if some of this uniformity was relaxed.[3] In particular, with the focus in this study on user economies of scale (*UES*), a less uniform system would allow variations in costs (including user costs, and thus *UES*) and preferences that exist in the network to be better reflected in policy settings. Second, the availability of existing data was a strong influential factor on the level of disaggregation adopted.

Fortunately, a comprehensive disaggregated data base on the Adelaide bus system was already available, with the level of disaggregation considered appropriate for use here (with some modifications). This system, the Routes and Services Information System (ROSIS) developed by the STA (now TransAdelaide), models the bus network as a series of service corridors, where a service corridor is a collection of routes which serve a similar geographical area. This system was considered appropriate for a number of reasons. First, it economises on data collection needs. Second, the ROSIS system offers a reasonable degree of disaggregation compared to the simple whole network average approach, and is more manageable than individual route analysis. Third, the transport agencies within the state government (TransAdelaide, the Passenger Transport Board, and the Transport Policy Unit) indicated that results at this level of reporting would be useful, with further disaggregation yielding small additional benefits for policy purposes.[4,5] The ROSIS system

[3] Some uniformity and coordination across the bus system is highly beneficial, for example in ticketing and information systems, a key finding from British bus deregulation (Evans, 1990). In other respects, uniformity is harder to justify on economic grounds. For example, the Industry Commission (1994) recently argued that prices should vary with distance travelled, and between peak and off-peak, to reflect associated cost variations. Although both elements have been reflected in pricing systems in Adelaide over time to some degree, there have also been periods where pricing structures have reverted back away from these ideals. As footnotes 13 and 14 of chapter 2 note, the current situation in Adelaide is one where prices do not vary with distance travelled (other than a lower fare for trips < 3.2 km), but there is a price differential between peak and off-peak. Other than this, the pricing structure is uniform across the entire bus system (except for social justice price discounts which are discussed below in section 7.2.4).

[4] In fact, having policy settings which vary by ROSIS corridor may also be considered by some to be too complicated a policy system.

[5] An alternative disaggregation option was also initially considered, namely, modelling a limited number of route types. Given the importance of demand level in UES determination, a sensible example might be to model three route demand categories, high demand, medium demand, and low demand routes, and then assign each actual route in the network to one of these stylised route types. Neither this approach, nor the ROSIS system, is ideal, with each compromising some descriptive aspect, as do all aggregation approaches. In the stylised route type approach, although routes within a route group have similar demand, they may be geographically dispersed, and so the associated costs and user characteristics may differ considerably. Under the ROSIS approach, on the other hand, demand variation within a group may be greater, but costs and users characteristics are likely to display less variation. In the end, the ROSIS system was adopted because of its ready availability, and the fact that it is already being used by, and familiar to, the transport agencies.

of bus service corridors was therefore adopted as the framework for disaggregated analysis in this study.

The adopted disaggregation approach has two dimensions : temporal and spatial.

(i) *Temporal*

Two time periods are modelled, Peak (*PK*) and Off-Peak (*OP*), defined as follows :

- *PK* : 6am-9 am, plus 3pm-6pm on working weekdays
- *OP* : the interpeak period 9am-3pm on working day weekdays, evenings (6pm-12am), weekends, and public holidays.

This approach reflects the key differences that exist in demand characteristics, whilst at the same time limiting (for simplicity) the number of time periods and maintaining consistency with the ROSIS system (the time boundaries being drawn directly from ROSIS).[6] As per ROSIS, *PK* represents an average of the am and pm peaks since the difference in demand between the two peaks is not sufficient to justify separate modelling.

Adelaide buses run according to two separate service networks. The "Day" network operates for *PK* plus for part of *OP*, i.e. the interpeak period, 6pm - 7pm on Monday to Thursday, 6pm - 10pm on Friday, and until 7pm on Saturday. The "Night" network, which is essentially a reduction from the Day network in terms of both route coverage and frequency of service, runs for the remaining operating hours of *OP*. *PK* is therefore fully reflected by the Day network. *OP*, on the other hand, is a mix of the Day and Night networks. Network characteristics for *OP* (i.e. number of routes, and route length) were therefore derived by weighting the Day and Night networks by the number of hours that each network operates (see section B.2.2.1 of Appendix B for details).

[6] A single off-peak period was adopted to keep the number of analytical cases to be considered at a manageable level. Some may also argue that from a practical fare setting viewpoint a single differentiation between peak and off-peak is sufficient. Further investigations, particularly more detailed corridor case studies, could usefully consider two separate off-peak periods, one being the interpeak, and the other evenings, weekends and public holidays combined. Some testing of the implications of such a distinction is made at relevant points in the analysis that follows.

(ii) *Spatial*

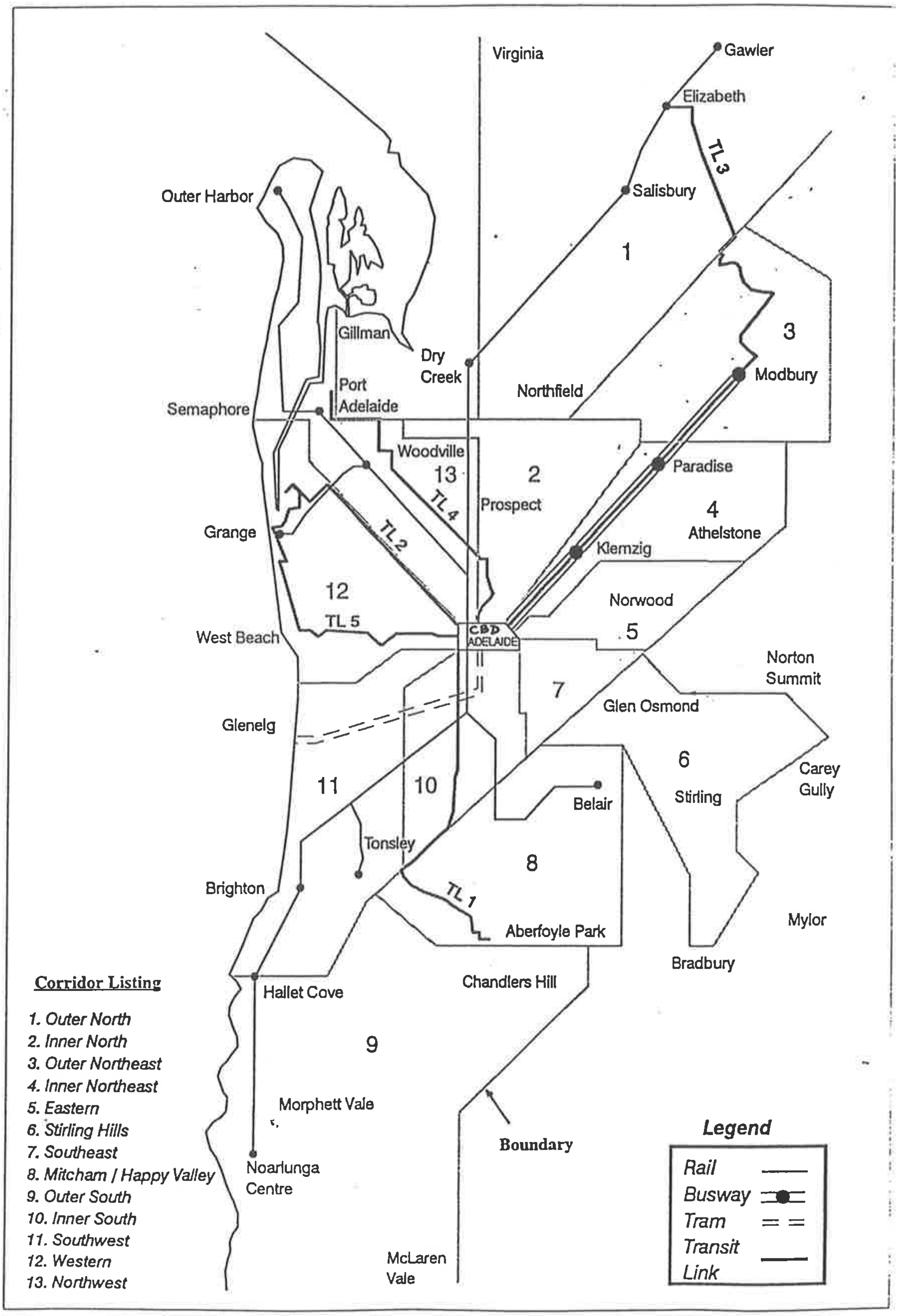
The ROSIS system divides the metropolitan bus service into 18 corridors. For corridors 1-13, which are listed and illustrated in Figure 7.1, all routes in a corridor serve a similar area, and all routes are timetabled. The remaining ROSIS "corridors" consist of : corridor 14, a collection of cross-suburban routes scattered throughout the metropolitan area, the main one being a ring route service, the Circle Line, which runs around the CBD at a distance of about 5 to 7 kms; corridor 15, the "Beeline", a short free bus route running within the CBD; corridor 16, school runs; corridor 17, a small number of midi bus services; and corridor 18, other non-route services. To keep the analysis manageable, only the true corridors, corridors 1-13, which in 1992/93 accounted for 89% of boardings and 91% of vehicle-kms (STA, 1993a), are considered in this study.

Some of the corridors, namely the outer corridors, run radial services which pass through inner corridors on their way to/from the CBD. It is important to note, however, that for most of the time they do *not* also serve the inner corridors when passing through them. This is true for most weekday daytime services, where the majority of buses are either express or limited stop buses when passing through other corridors. The former do not stop in inner suburbs, whilst the latter only drop off in inner suburbs in the am peak, and only pickup in inner suburbs in the pm peak. For the night network, the situation is slightly different, where some outer services also service inner areas (thus allowing some inner corridor routes to not be run during these times periods). On the whole, it is reasonable to model the corridors as a collection of essentially independent markets.

Several corridors (in particular corridors 1, 3, 9, 11 and 13) are complicated by the existence of a mix of radial and feeder/regional type of routes. Figure 7.2 illustrates the proportion of radial routes in each corridor.[7] The different service frequencies and trip lengths associated with

[7] A convention adopted throughout this chapter in the presentation of figures which compare results across corridors, is to plot the corridors on the horizontal-axis in a manner which relates to their relative location in the metropolitan area. A useful depiction for this is to think of routes as falling into one of the following aggregated categories : Inner Suburban, Middle Suburban, and Outer Suburban. Following the allocation of corridors to these categories suggested in STA (1992a) (i.e. Inner (corridors 2,5,7,10,12), Middle (3,4,11,13), and Outer (1,6,8,9)), the corridors are plotted from left to right in accordance with this classification. The Inner Suburban corridors are at the left end, the Outer Suburban corridors are at the right end, with the Middle Suburban corridors in between.

Figure 7.1 : ROSIS System Corridor Boundaries



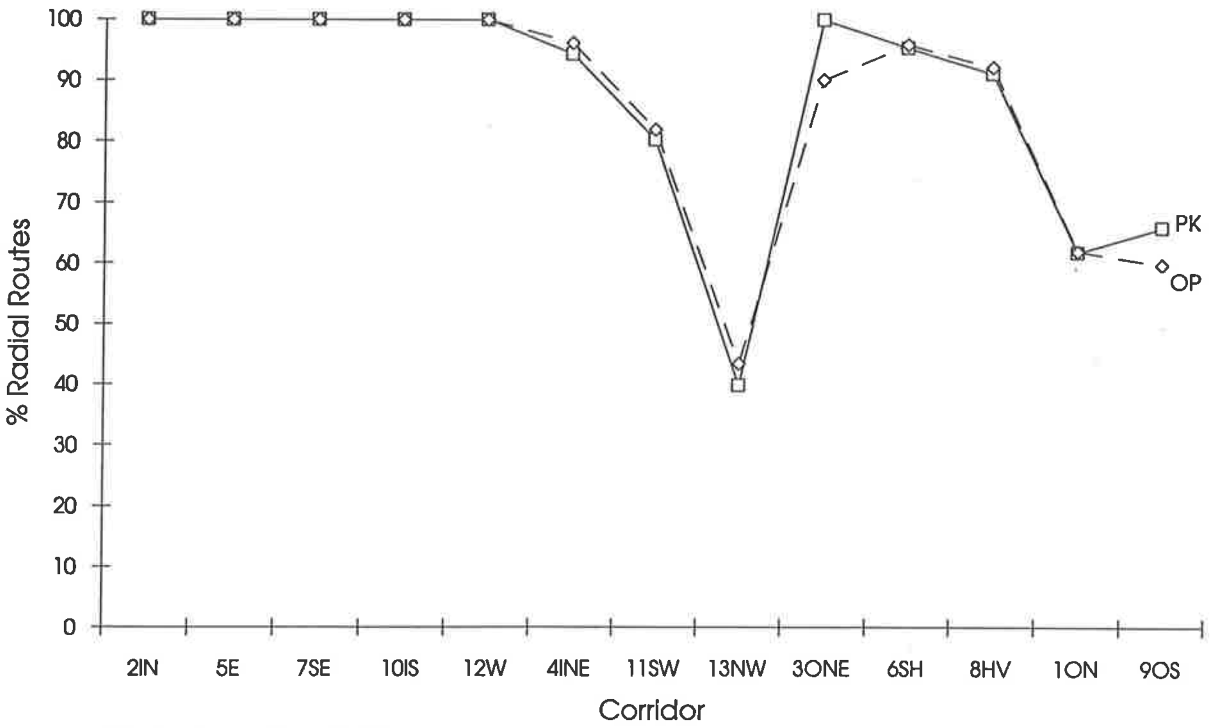
Corridor Listing

1. *Outer North*
2. *Inner North*
3. *Outer Northeast*
4. *Inner Northeast*
5. *Eastern*
6. *Stirling Hills*
7. *Southeast*
8. *Mitcham / Happy Valley*
9. *Outer South*
10. *Inner South*
11. *Southwest*
12. *Western*
13. *Northwest*

Legend

Rail	───
Busway	─●─
Tram	= =
Transit Link	───

Figure 7.2 : Proportion of Radial Routes



Note : See Figure 7.1 for corridor listing by name.

these two service types suggests a further level of disaggregation, namely *within* the corridor, with radials and feeders modelled separately, may be beneficial. This further disaggregation is *not* undertaken here, mainly because data was not readily available by the radial vs feeder distinction. In addition, this would introduce the need for a further fare category (radial vs feeder travel), adding to the complication of the pricing system. Instead, corridor parameters will reflect the average of both route types within each corridor, with a single fare for all travel within a corridor (notwithstanding concession discounts, see section 7.2.4 below). Some investigation of the radial vs feeder problem is warranted, however, outside of this study.

To enable meaningful comparison of results between corridors, the unit level of analysis within each corridor is the representative route, with representative route parameters for each corridor being derived by averaging across all routes within the corridor. Aggregate results for each corridor are subsequently derived by factoring up accordingly. Appendix B (see section B.2) summarises the derivation of representative route data.

7.2.3 Pricing per Journey vs per Boarding ?

An important question to consider at the outset is whether the pricing system should operate on a *per journey* or *per boarding* basis.[8] A journey is defined as travel from an origin (e.g. home) to a destination (e.g. work). A boarding, on the other hand, is defined as the act of travelling on a particular mode of transport (e.g. a bus). The distinction between journeys and boardings becomes clear when one consider a multi-boarding journey (or trip). For example, travel to work from home may require the traveller to travel by bus to the city centre (boarding 1) and then transfer to another service (e.g. rail) and travel to the work destination which is in the suburbs (boarding 2). This single home-based journey to work thus consists of two boardings (bus, then rail) with a transfer in between.

[8] A further important aspect of designing a pricing policy is the relationship between fare and distance travelled. To enable the focus to remain on *UES* effects, fares vary with distance in only a simplistic manner, varying only to reflect the variation in route length *between* corridors. There is *no* fare distance relationship, however, *within* a corridor, with trips of differing lengths facing the same fare. Distance based fares are subject to current investigation in Adelaide by the Passenger Transport Board, and have been studied in considerable detail elsewhere (see footnote 13 of chapter 2).

Pricing on a per boarding basis allows one to set a different price for each corridor, and thus allows prices to vary significantly across the network. This way, it is possible to reflect through pricing the different user costs (the determinant of user economies of scale (*UES*)) which may occur in different corridors throughout the network.

Once one moves to a per journey pricing system (the current system in Adelaide), it is still technically possible to have prices which reflect cost variations between corridors. There would need to be a price, however, for each journey type, where journey type is defined here by the combination and location of boardings which make up the journey. For example, consider the two boarding (bus-rail) journey discussed above. A different price could be determined for every feasible corridor origin/corridor destination pair. If there were χ corridors in the network, this would yield a 2-dimensional matrix of prices of size χ by χ. For journeys with three boardings, a 3-dimensional price matrix of size χ by χ by χ would be required. Thus a price would be struck for each journey depending on its corridor origin-destination characteristics.

Such a per journey pricing system would require a ticketing system that could identify the nature of each journey and charge accordingly. Although such an approach might be feasible in a system with a relatively small number of corridors, it would probably be judged to be too complex for the 14 corridor system being considered here.[9] For the purpose of this study, this is assumed to be the case. Therefore, given the aim here of having prices which reflect user cost and demand variations between corridors, pricing *per boarding* is assumed.

7.2.4 Social Justice Policy

A feature of the analysis in this chapter is that the role of social justice or social equity policy is explicitly recognised and incorporated in modelling. The Government of South Australia currently delivers assistance on social justice grounds through user side subsidies which finance a system of concession fares[10] for certain groups of travellers (students, the aged, the unemployed,

[9] Although new technologies like smartcard and geographical information systems may in future provide workable approaches for situations with many corridors.

[10] A case of third degree price discrimination.

etc).[11] The importance and influence this has on public transport in Adelaide is illustrated by Figure 7.3, which reports the proportion of concession bus users in Adelaide. Two things can be noted. First, in nearly all cases, concession users dominate the bus travel market, particularly in the off-peak. Second, on the whole, the proportion of concession users is fairly uniform across corridors within each period. The main exceptions are the NorthWest and Outer NorthEast corridors which have the highest and lowest proportion of concession users respectively.

In this study, the thrust of the Government's current social justice policy is retained by incorporating the concession fare system into the analytical and quantitative optimisation and modelling.[12] This could be done in a number of ways : maintaining the existing percentage difference between full and concession fares; maintaining the existing absolute difference; putting a ceiling on concession fares; or maintaining existing total concession fare subsidy. The first option is adopted for use here. The other options have a number of problems. The third and fourth options appear to introduce undue rigidity into the system. In addition, the last option would lead to lower per unit concession subsidies as the number of concession users grows, and the third option may lead to excessive levels of assistance if full fares grow. The second option, whilst more flexible, worsens the relative position of concession groups when the overall level of bus prices rise, which could lead to equity objections in some circumstances. Maintaining the existing percentage difference would appear to be the option which best maintains the current equity relativities between concession and non-concession users.

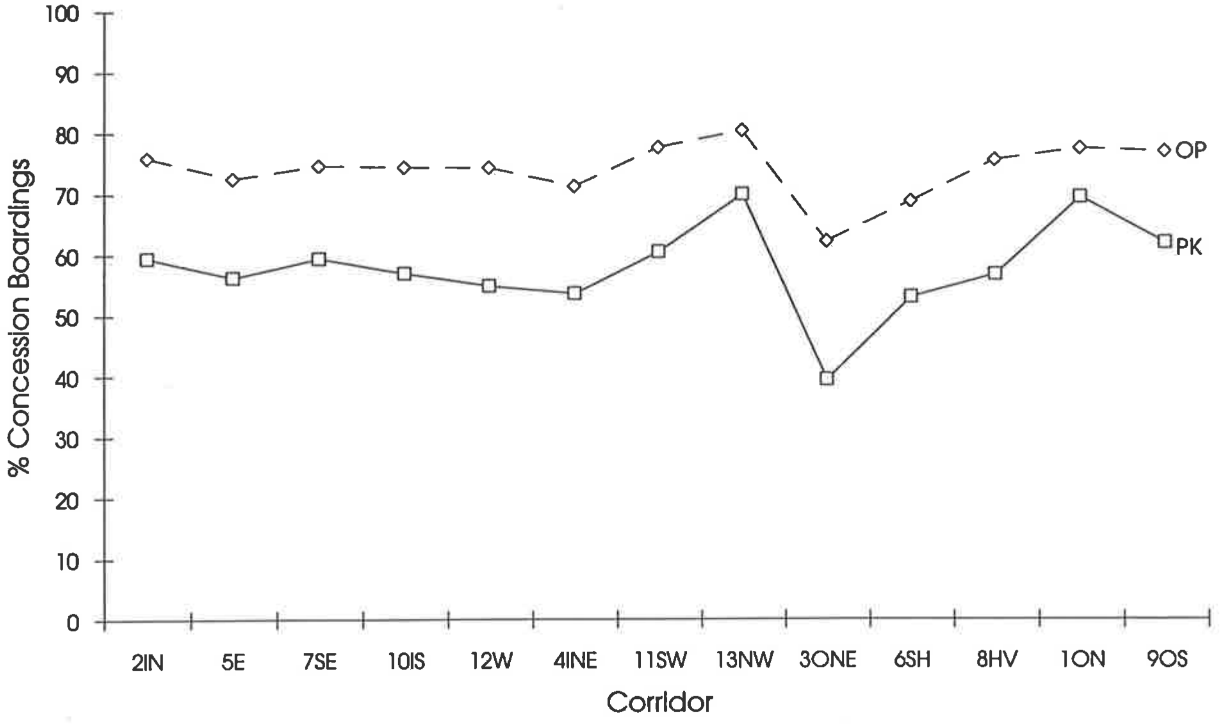
7.3 The Optimisation Problem

The optimisation problem is identical in principle to that of chapter 4, however it is extended here in two ways : it is subject to the social justice policy constraint that has been imposed, namely,

[11] In addition, production side subsidies also assist the needy by keeping the general level of fares below what they would otherwise be. Such an approach has at times been advocated as an appropriate tool for achieving equity goals, although from an economic efficiency perspective it does so in a less than fully effective manner (this issue was also discussed in section 2.3 of chapter 2).

[12] A broader approach not considered here, is to address from first principles the delivery of financial assistance to those who are judged to be in need of it, assessing in particular the effectiveness of the concession fares approach for delivering equity goals, and identifying superior policy instruments (if they exist).

Figure 7.3 : Proportion of Concession Boardings



Note : See Figure 7.1 for corridor listing by name.

that the current level of concessions, expressed in percentage terms, be maintained; and, it is cast as a peak load problem.

7.3.1 Definitions and Assumptions

Users are modelled as two users groups : *regular* full fare paying non-concession users, and *concession* users. For notational convenience, these two groups will be denoted throughout by the subscripts $_1$ and $_2$. Denoting P_1 and P_2 as the prices paid by the two groups of users respectively, the Social Justice Price Discount, *SJPD*, received by concession users is :

$$SJPD = P_1 - P_2 \tag{7.1}$$

whilst the Social Justice Discount Factor, *SJDF*, expressed as a percentage, is given by :

$$SJDF = \frac{SJPD}{P_1} = \frac{P_1 - P_2}{P_1} \tag{7.2}$$

Rearranging (7.2) then yields :

$$P_2 = (1 - SJDF)\, P_1 \tag{7.3}$$

The key to the analysis of peak load problems (Steiner, 1957; Williamson, 1966) is the separation of peak and off-peak demand, and the treatment of capital costs. The standard approach is to allocate capital costs (in this case bus capital costs) to the peak (Kerin, 1989). Costs incurred in the peak are thus peak operating costs *plus* capital costs, whilst only operating costs are incurred in the off-peak. The reason for this is straightforward. It is peak demand which influences the need for bus capacity. A growth or decline in peak demand translates into a rise or fall in bus capacity requirements. On the other hand, changes in off-peak demand have no influence on bus capital requirements. A fall in off-peak demand does not alter the need for the current capacity to cater for the peak. If off-peak demand were to grow, then, given there is already spare capacity in the form of unused buses that are required for the peak but not the off-peak, no additional capital will be required. This approach is the basis of bus costing in Adelaide (STA, 1994), with the resulting unit costs for the two time periods reported in section B.3.3 of appendix B.

A feature of the peak load problem analysed here is that, in the off-peak, although there is spare bus capacity, not all that capacity needs to be used. Service levels in the peak and off-peak can be made to vary independently by having different peak and off-peak bus frequencies. For simplicity, it is also assumed that demand cross elasticities between time periods are zero. Overall therefore, peak and off-peak periods can be analysed as separate optimisations, identifying optimal policy settings for each period.

In chapter 4, a number of load factor (*LF*) and bus size (*N*) fixity/flexibility cases were considered in a taxonomic analysis. The possibility of *LF* being fixed was considered there for several reasons : for analytical completeness, because such an assumption has been used in previous analyses, and because it can act as a simple operational policy. For the analysis of optimal bus subsidy in Adelaide in this chapter, however, *LF* has been allowed to vary, and thus be optimised. On the other hand, fixity of *N* is likely to be a realistic constraint in the short (and possibly medium) term. Further, result 3 in chapter 4 showed that when *LF* is allowed to vary, marginal changes in *N* from its current level led to only relatively small changes in subsidy results. The analysis here is therefore undertaken with *N* fixed at the current bus size, $N = 78$, and the analysis is thus an extension of the variable load factor/fixed bus size analysis presented in chapter 4 (*LF/N* fixity case 3).

7.3.2 First Best Optimisation

Consider the representative route within any given corridor for any given time period. The optimisation problem is :

$$\max_{P_1, P_2, F} \; ES \tag{7.4}$$

$$\text{where } ES = CS_1 + CS_2 - S \tag{7.5}$$

$$\text{and} \quad S = C_p - P_1 q_1 - P_2 q_2 \tag{7.6}$$

The fact that P_2 is related to P_1, however, means that P_2 can be eliminated as an optimisation variable, since the optimal value of P_1 automatically implies the optimal value of P_2 for any given social justice price discount (*SJPD*) or social justice discount factor (*SJDF*). Initial attempts to set up and solve this optimisation problem were based on using expression (7.3) where P_2 is a direct function of *SJDF*. Unfortunately, attempting to solve the problem in this form proved extremely complex analytically. A more satisfactory approach, for which tractable solutions were able to be generated relatively easily, was to express P_2 as a function of *SJPD* as in (7.1), and then solve repeatedly with *SJPD* iterating until the target *SJDF* resulted. The final problem is thus :

$$\max_{P_1, F} \; ES \tag{7.7}$$

where $ES = CS_1 + CS_2(P_1) - S$ (7.8)

and $S = C_p - P_1q_1 - (P_1 - SJPD)q_2$ (7.9)

Solving the first order condition $\frac{\partial ES}{\partial P_1} = 0$ (see section 7A.2 of the chapter appendix) yields the following expression for optimal regular (non-concession) fare :

$$P_1^* = q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} + SJPD\frac{q_2}{q} \tag{7.10}$$

which section 7A.3 of the chapter appendix shows is consistent with the equating of marginal benefit and marginal social cost, the optimal outcome.

As discussed in chapter 4, optimal frequency F^* can be determined in two ways : either through the first order condition $\frac{\partial ES}{\partial F} = 0$ of the above maximisation problem, or from the first order condition $\frac{\partial TC}{\partial F} = 0$ of the dual 2-stage cost minimisation/welfare maximisation problem (see discussion in section 4.3 of chapter 4). The first approach proved analytically complex, so the simpler second approach was used. Solving $\frac{\partial TC}{\partial F} = 0$ yields (see section 7A.4 of the chapter appendix) :

$$-q\left(\frac{\partial u}{\partial F} + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial F}\bigg|_{\bar{q}} + \frac{\partial v}{\partial F}\bigg|_{\overline{LF},\bar{q}}\right) = \frac{\partial C_p}{\partial F} \tag{7.11}$$

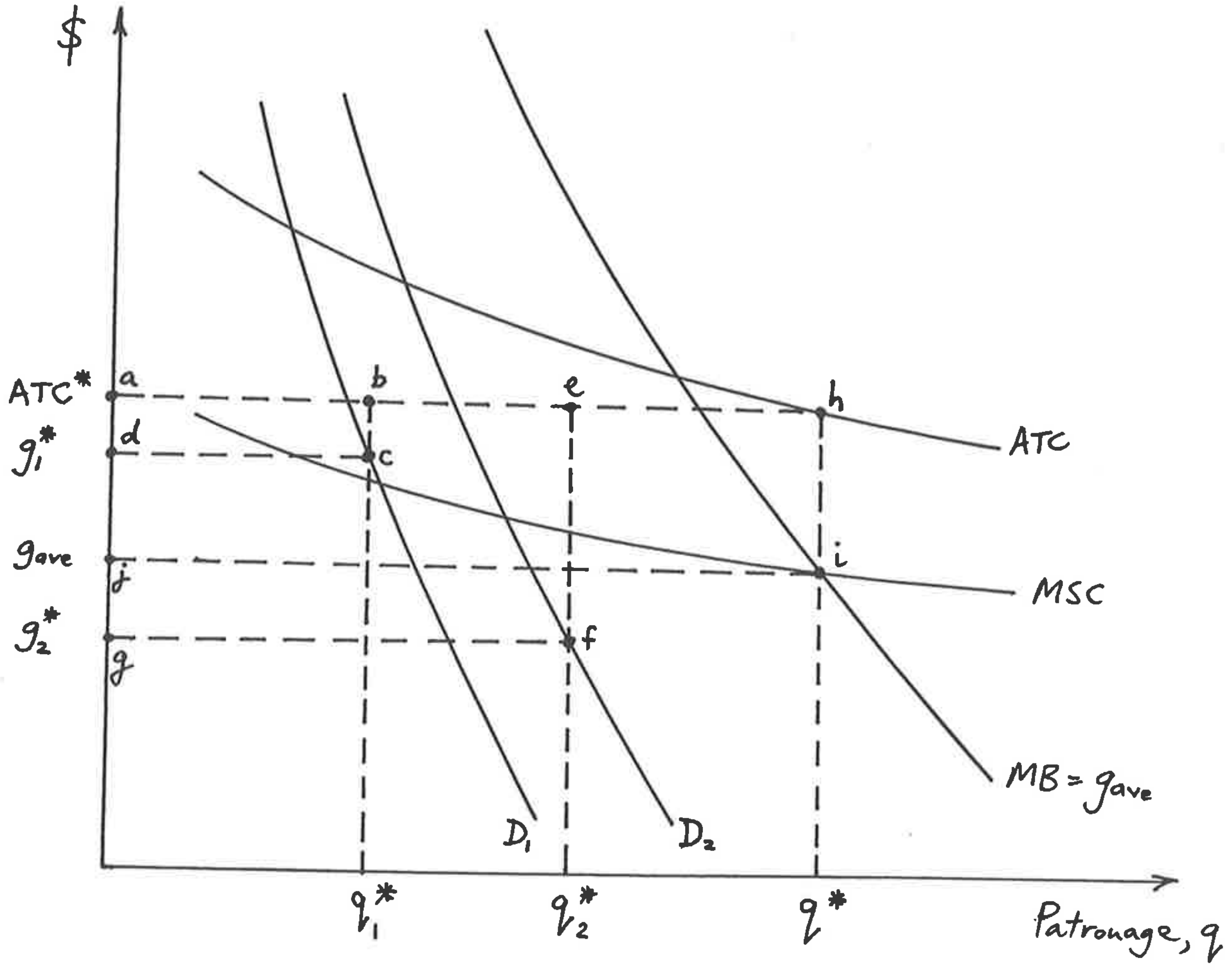
the same optimal condition which applied when optimising F in chapter 4, i.e. at the margin, the net benefit and cost of an additional unit of frequency must be equated.

7.3.3 A Diagrammatic Presentation of the First Best Framework

The optimisation problem can also be presented in the familiar marginal benefit (*MB*)/long run marginal social cost (*MSC*) diagrammatic framework for any given time period, as in Figure 7.4. Curves D_1 and D_2 are the generalised cost demand curves for regular (full) fare, and concession fare, users respectively. Curve *MB* measures the overall marginal benefit (as derived in section 7A.3 of the chapter appendix), whilst curves *MSC* and *ATC* are the minimum cost curves achieved by optimising F at all q values according to (7.11).

It is *important* to note from the outset that, for the exponential demand model being used here (expression (4.14)), the financial position can be inferred by the difference between the *MB* and

Figure 7.4 : Single Period Analysis Framework



ATC curves at any patronage, *q*, level. This rule was established in chapter 4 for the analysis without concession fares, and also holds in the analysis here with concession fares. This can be established as follows. First note that $MB = g_{ave}$ (see section 7A.5 in the chapter appendix). Then $MB - ATC = g_{ave} - ATC$, and noting that $g = P + AC_u$ and $ATC = AC_u + AC_p$, yields $MB - ATC = P_{ave} - AC_p$, i.e. the unit financial surplus/deficit. It then follows that :

- at the intersection point of *MB* and *ATC*, $g = ATC$, and thus $P_{ave} = AC_p$, a breakeven outcome;
- at lower *q* values, $MB > ATC$, thus $P_{ave} > AC_p$, i.e. a surplus exists; and
- at higher *q* values, $ATC > MB$, thus $AC_p > P_{ave}$, i.e. a deficit results, and thus a need for subsidy.

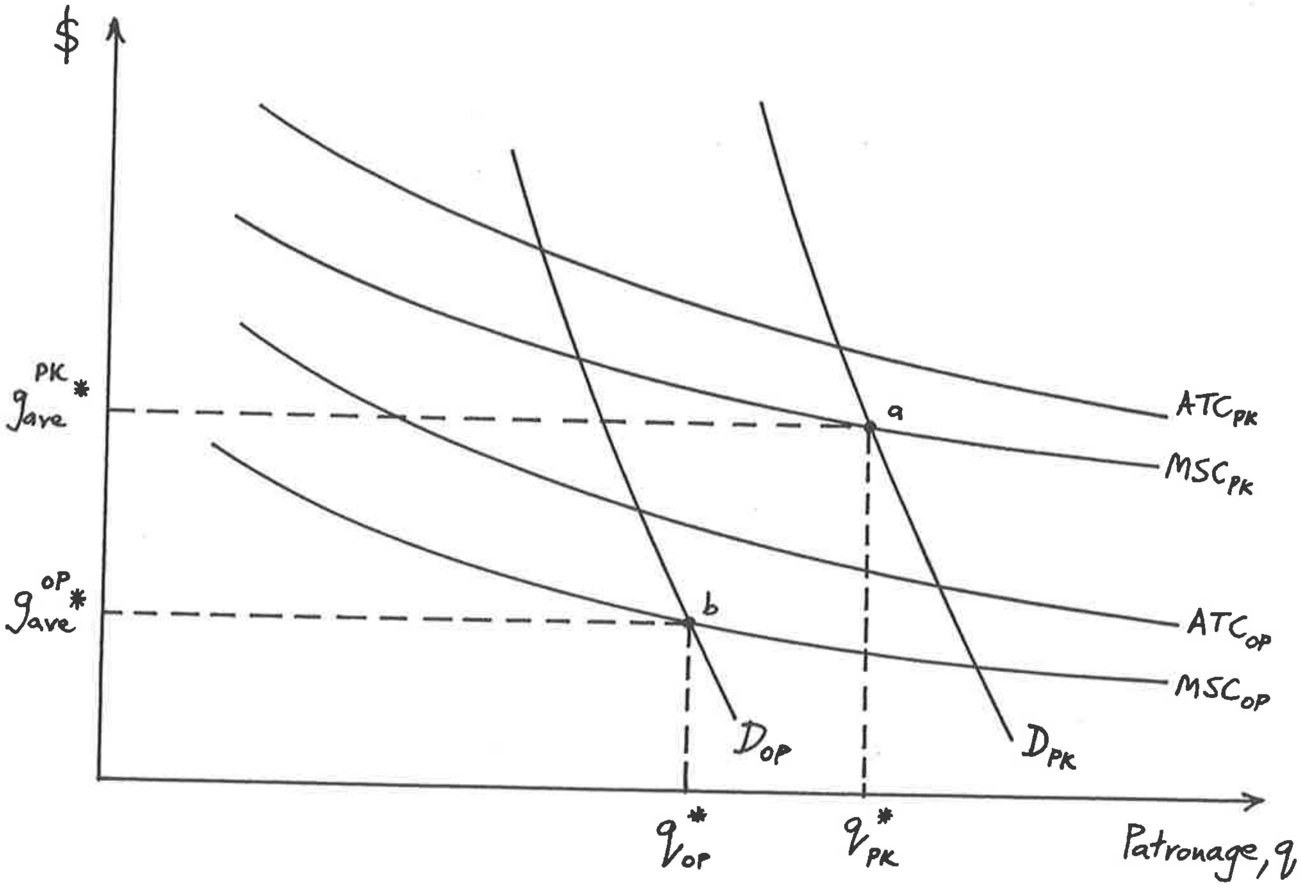
Therefore, the gap between the MB curve and the ATC curve measures the unit financial surplus/deficit outcome.

The optimal solution in Figure 7.4 lies where $MB = MSC$, at point *i* and output level q^*. With *ATC* declining throughout due to user economies of scale (*UES*), and thus $MSC < ATC$ throughout, the optimum will always involve a subsidy, with average unit subsidy across the two market segments, s_{ave}, being the distance *hi* and total subsidy *S* equal to area *ahij*. For market segment *j*, the difference between the optimal unit cost $ATC(q^*)$ (or simply ATC^*) and g_j^* measures the unit subsidy delivered to each user, i.e. $ATC^* - g_1^*$ measures unit subsidy received by full fare paying users, s_1, whilst $ATC^* - g_2^*$ indicates unit subsidy per concession user, s_2. Total subsidy for each group, S_1 and S_2, are then areas *abcd* and *aefg* in Figure 7.4 respectively (with the sum of these two areas being equal in size to area *ahij*, the total subsidy derived from the average unit subsidy).

Figure 7.4, and the above discussion associated with it, applies for both of the time periods. It is useful, however, to be able to bring together diagrammatically the nature of the overall problem across both time periods. This is done in Figure 7.5. To avoid cluttering the diagram, the market segment demand curves (regular vs concession users) are not shown here, limiting the presentation to *MB*, *MSC* and *ATC* curves for each period.

On the demand side, the relative position of the respective *MB* curves reflects the underlying level of demand in the two time periods, with demand level ranking being peak, off-peak. On the cost side, both producer and user costs exist in each period. For convenience, unit user cost parameters and user cost functions are assumed to be the same in both periods. As explained

Figure 7.5 : Multiple Period Analysis Framework



earlier, unit producer costs are higher in the peak. With similar user cost functions throughout, overall (producer and user) unit costs are also higher in the peak. This is illustrated by the vertical positioning of the two sets of cost curves in Figure 7.5. The optimal outcomes are at points *a* and *b* respectively. With a lower cost structure, and a lower demand level, g^*_{ave} is likely to be lower for off-peak than peak.

7.3.4 Second Best Optimisation : The Opportunity Cost of Public Funds

The second best case considered here is where the financing of optimal subsidy from public funds is not neutral from an efficiency perspective.[13] That is, there are efficiency, or deadweight costs, involved with such financing. As a result, the opportunity cost of raising a $1 of finance is greater than $1. As mentioned in Chapter 2, this is also called the shadow price of public funds (Dodgson and Topham, 1987). We denote the marginal opportunity cost of public funds as *MOCPF*, and the associated marginal financing deadweight cost as κ, where

$$MOCPF = 1 + \kappa \tag{7.12}$$

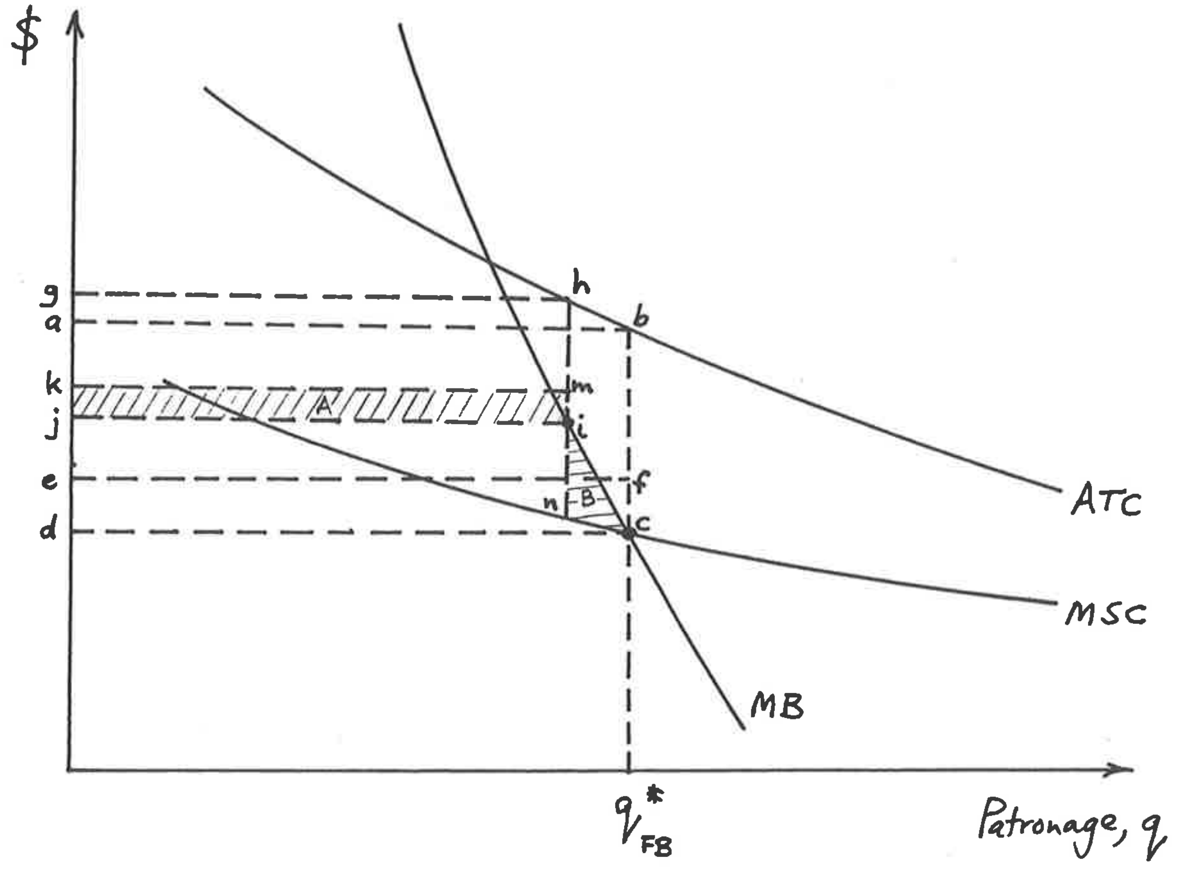
Findlay and Jones (1982) estimate that, in Australia, the *MOCPF* lies in the range $1.23 to $1.65.[14] Freebairn (1995) has recently reconsidered the opportunity cost of public funds in Australia, adjusting earlier studies to account for sticky wages and unemployment, confirming significant efficiency costs of public fund raising. Freebairn found considerable sensitivity in marginal distortionary costs to model assumptions, yielding *MOCPF* values as high as $1.73. For the majority of the analysis here, a κ value of 0.4, an approximate midpoint value in the Findlay and Jones range is adopted, with sensitivity testing of lower values.

This framework is illustrated in Figure 7.6 which shows the first best diagram modified to take account of costly finance. Figure 7.6 schematically shows the *MB*, *MSC* and *ATC* curves for a given time period. The first best optimum is at point *c* coinciding with *MB* = *MSC*, with first best subsidy given by area *abcd*. Once we move to a second best framework, the existence of subsidy generates a financing deadweight loss, *FDWL*, where :

[13] The other second best case (not considered here) is where, because of a lack of road pricing, public transport subsidy is used as an instrument for managing road congestion (see discussion in section 2.4.2 of chapter 2).

[14] Overseas studies have found lower values, for example in the UK (Dodgson and Topham, 1987).

Figure 7.6 : Second Best Analysis When Public Finance is Costly To Raise



$$FDWL = \kappa S \qquad (7.13)$$

In a second best world, what was previously a first best situation at point *c* now generates a *FDWL* equal to area *cdef*.[15] The bigger κ is, the bigger is the *FDWL* area.

From the old first best position, *ES* can be improved by moving up along *MB*, which reduces the size of total subsidy (*S*)[16] and thus also reduces the size of the *FDWL*. This reduction in *FDWL* is achieved, however, at the expense of introducing a new deadweight loss due to the introduction of a divergence between *MB* and *MSC*. If, for example, we move up to point *i* on the *MB* curve, *S* reduces to area *ghij*, *FDWL* reduces to area *ijkm* (shaded area *A*), whilst the new divergence deadweight loss (*DDWL*) is area *cin* (shaded area *B*). The optimal solution consists of moving up the *MB* curve until the ***marginal*** increase in the *DDWL* (the ***increase*** in area *B*) is exactly ***equal*** to the ***marginal*** reduction in *FDWL* (the ***reduction*** in area *A*), thus resulting in the combined deadweight loss, the sum of *FDWL* and *DDWL* (area *A* plus area *B*) being ***minimised.***

To undertake such second best analysis, the first best subsidy formulation presented in section 7.3.2 must be modified accordingly. The optimisation now has an additional cost component, financing deadweight loss *FDWL*, which enters the economic surplus (*ES*) expression. Thus (7.8) can be rewritten as follows :

$$ES = CS - S - FDWL \qquad (7.14)$$

$$\text{where } FDWL = \kappa S$$

Thus

$$ES = CS - (1+\kappa)S \qquad (7.15)$$

Solving the optimisation problem (see section 7A.6 in the chapter appendix) yields :

$$P_1 = q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} + SJPD\frac{q_2}{q} + \frac{\kappa}{(1+\kappa)\beta} \qquad (7.16)$$

The first two components of (7.16) are the first best regular fare (see expression (7.10). The third component is the second best markup which is constant (for any given κ and β) for the demand case considered, the case of an exponential demand function. Note the similarity with the breakeven optimisation result in Chapter 7 where markups were also required. Here the markup is required

[15] The diagram is constructed on the illustrative assumption that κ = 0.25, i.e. the *FDWL* area equals one quarter of the subsidy area.

[16] Since both patronage (*q*) and unit subsidy (*ATC* - *MB*) fall.

for the following simple reason. With subsidy now being costly to finance, it is desirable to raise bus prices, and thus reduce q, in order to limit the size of subsidy and thus the financing deadweight loss. A trade-off situation arises, however, since in order to limit financing opportunity cost, deadweight losses are incurred from the divergence of prices away from their first best optima. The overall optimum occurs when, at the margin, the increase in divergence deadweight loss balances the saving in financing deadweight loss.

7.4 Optimal Results

To commence with, consider the first best (*FB*) and second best (*SB*) results before the introduction of competitive tendering (*CT*), scheduled to occur in the near future in Adelaide (see the discussion in section 2.2 of chapter 2). The results are generated using the general solutions presented in the last section. In each case considered, generation of solutions consisted of the following steps :

1. Select an initial patronage value, q_i.
2. Given q_i, optimise frequency (*F*) through (7.11).
3. Iterate the regular fare (P_1) and the social justice price discount (*SJPD*) until the current social justice discount factor ($SJDF_o$), and q_i, are achieved.
4. Compare the resulting values of *MB* and *MSC*.
5. If $MB = MSC$, the optimal solution has been reached. If $MB \neq MSC$, repeat steps 1 to 4, with alternative values of q_i, until $MB = MSC$ is reached. When iterating q_i, increase (decrease) q_i if $MB >$ (<) MSC.

The key results are summarised in Table 7.1.

There are a number of general patterns in the first best and second best results presented in Table 7.1 which should be noted. These are discussed in the sub-sections that follow.

7.4.1 <u>First Best Results</u>

- In a first best world, ES_{PK} is substantially greater than ES_{OP}, highlighting peak periods as the prime market for service delivery from an economic efficiency perspective. In the peak, net benefits

Table 7.1 : Representative Route Optimal Results Before Competitive Tendering

Corridor	*q* *(boardings/hr)*	*F* *buses/hr*	*LF*	*R* *%*	P_1 *cents*	s_{ave} *cents*	s_1 *cents*	s_2 *cents*	*S* *$/hr*	*ES* *$/hr*
Inner North										
PK - current	215	3.5	0.27	27	109	119	85	142	256	513(410)[2]
- FB	193	3.0	0.28	21	144	90	45	120	175	516
-SB, κ = 0.4	130	2.2	0.25	11	331	-6	-123	51	-8	475
OP - current	102	2.0	0.16	9	99	81	37	95	82	85(52)
- FB	99	1.9	0.16	9	103	79	33	93	78	85
- SB, κ = 0.4	62	1.4	0.13	8	207	56	-44	77	34	59
Eastern										
PK - current	238	5.2	0.20	44	108	147	115	172	351	498(357)
- FB	202	3.3	0.26	24	122	81	45	109	164	557
- FB(R)	225	3.9	0.24	100	84	120	96	140	271	533
-SB, κ = 0.4	137	2.4	0.24	13	298	-17	-115	42	-23	521
OP - current	81	2.4	0.11	12	100	125	84	141	102	31(-10)
- FB	72	1.6	0.14	8	88	92	57	106	66	51
- FB(R)	134	3.2	0.13	100	40	128	113	135	172	48
- SB, κ = 0.4	42	1.2	0.11	8	184	75	-5	98	32	24
SouthEast										
PK - current	136	3.4	0.17	26	107	148	115	170	201	285(204)
- FB	116	2.2	0.22	11	127	87	48	113	101	314
-SB, κ = 0.4	81	1.7	0.2	9	297	-1	-103	52	-1	291
OP - current	51	1.3	0.12	8	98	88	46	102	44	38(21)
- FB	60	1.5	0.12	8	81	98	64	111	59	39
- SB, κ = 0.4	35	1.1	0.10	8	176	83	3	103	29	26
Inner South										
PK - current	180	2.9	0.27	19	108	123	90	148	221	421(333)
- FB	160	2.6	0.27	15	156	93	44	128	149	422
-SB, κ = 0.4	110	2.0	0.24	9	331	6	-109	69	7	384
OP - current	74	1.6	0.14	8	98	120	78	134	88	32(-3)
- FB	61	1.3	0.14	8	122	109	55	126	66	34
- SB, κ = 0.4	30	0.9	0.11	8	234	107	-4	131	33	4
Western										
PK - current	186	3.2	0.25	24	109	178	146	205	331	332(200)
- FB	143	2.3	0.27	12	182	107	50	148	153	359
-SB, κ = 0.4	98	1.7	0.24	9	356	24	-97	93	24	317
OP - current	75	1.4	0.16	8	102	114	70	129	85	37(4)
- FB	67	1.3	0.15	8	138	108	47	127	72	38
- SB, κ = 0.4	33	0.9	0.12	8	253	108	-14	132	36	8
Inner NorthEast										
PK - current	150	3.2	0.22	23	110	186	155	213	280	256(144)
- FB	116	2.1	0.26	10	172	109	58	148	126	289
-SB, κ = 0.4	79	1.6	0.23	8	342	26	-85	94	21	255
OP - current	62	1.5	0.15	8	101	130	88	147	81	21(-11)
- FB	46	1.2	0.14	8	143	118	56	140	54	21
- SB, κ = 0.4	20	0.7	0.10	8	262	137	12	165	27	-5

Table continued next page

Table 7.1(cont) : Optimal Results Before Competitive Tendering

Corridor	*q* *(boardings/hr)*	*F* *buses/hr*	*LF*	*R* *%*	P_1 *cents*	s_{ave} *cents*	s_1 *cents*	s_2 *cents*	*S* *$/hr*	*ES* *$/hr*
SouthWest										
PK - current	169	2.8	0.29	17	109	207	173	229	350	252(111)
- FB	119	2.0	0.28	10	243	129	47	172	154	273
-SB, κ = 0.4	80	1.5	0.25	8	430	51	-106	116	41	227
OP - current	85	1.7	0.18	8	99	146	101	159	124	15(-34)
- FB	51	1.2	0.16	8	193	125	31	144	64	20
- SB, κ = 0.4	21	0.64	0.12	8	336	148	-27	170	31	-9
NorthWest										
PK - current	96	2.3	0.20	12	108	188	148	205	179	162(90)
- FB	74	1.6	0.22	8	189	119	47	147	88	176
-SB, κ = 0.4	49	1.2	0.19	8	393	33	-129	79	16	152
OP - current	50	1.5	0.12	8	104	121	72	133	60	22(-2)
- FB	44	1.3	0.12	8	115	112	57	125	49	23
- SB, κ = 0.4	24	0.88	0.10	8	224	108	-5	126	26	3
Outer NorthEast										
PK - current	349	4.1	0.21	35	109	204	182	238	710	535(250)
- FB	264	2.4	0.27	13	154	101	69	148	267	675
-SB, κ = 0.4	183	1.8	0.24	9	301	17	-54	100	32	608
OP - current	99	1.6	0.12	8	105	171	135	193	169	-6(-73)
- FB	64	1.0	0.13	8	129	139	93	164	89	16
- SB, κ = 0.4	18	0.44	0.08	8	240	221	125	258	39	-26
Stirling Hills										
PK - current	126	1.6	0.34	8	106	222	194	247	280	170(58)
- FB	100	1.5	0.28	8	287	167	82	225	167	191
-SB, κ = 0.4	61	1.1	0.24	8	487	92	-68	176	56	139
OP - current	37	0.62	0.21	8	102	173	135	190	65	-4(-30)
- FB	50	1.0	0.18	8	229	148	53	177	74	8
- SB, κ = 0.4	**[3]									
H Valley/Mitcham										
PK - current	162	2.3	0.30	12	103	267	237	290	432	146(-27)
- FB	97	1.5	0.28	8	299	167	70	223	162	184
-SB, κ = 0.4	60	1.0	0.24	8	504	96	-85	174	57	134
OP - current	42	.91	0.16	8	100	275	234	289	115	-47(-92)
- FB	**									
- SB, κ = 0.4	**									
Outer North										
PK - current	148	2.4	0.26	13	104	226	189	243	336	194(59)
- FB	101	1.7	0.26	8	254	140	44	174	142	220
-SB, κ = 0.4	66	1.2	0.23	8	461	61	-127	110	40	179
OP - current	56	1.4	0.14	8	105	214	168	227	120	-29(-77)
- FB	**									
- SB, κ = 0.4	**									

Table continued next page

Table 7.1 (cont) : Optimal Results Before Competitive Tendering

Corridor	*q* *(boardings/hr)*	*F* *buses/ hr*	*LF*	*R* *%*	P_1 *cents*	s_{ave} *cents*	s_1 *cents*	s_2 *cents*	*S* *$/hr*	*ES* *$/hr*
Outer South										
PK - current	144	1.9	0.32	9	104	220	187	241	317	197(70)
- FB	105	1.6	0.28	8	288	155	55	203	162	211
-SB, $\kappa = 0.4$	65	1.1	0.24	8	497	80	-109	146	52	161
OP - current	44	1.2	0.13	8	103	257	212	271	113	-41(-87)
- FB	**									
- SB, $\kappa = 0.4$	**									

Notes :
1. Abbreviations : FB - first best; FB(R) - first best with random user model (rather than the logit model); SB - second best; subscript 1 denotes regular (full fare paying) users; subscript 2 denotes concession users; κ *is the marginal public financing deadweight loss.*
2. The number in brackets is current ES in a SB world of distortionary public finance collection, and the preceding number out of brackets is ES in a FB world of costless fund raising. This notation applies in describing current ES in all corridors.
*3. ** denotes that ES was < 0 for all possible q values in this corridor/time period, so the optimum involves zero service provision.*

(i.e. *ES* > 0) can be attained from running services in all corridors. This is also the case in the majority of corridors in the off-peak.

There are three corridors, however, Happy Valley/Mitcham, Outer North and Outer South, where all service delivery options yield $ES_{OP} < 0$. The main reason for this result is the high cost of service delivery to these corridors due to their long route lengths. If decisions about operating in a corridor rested purely on economic efficiency grounds, off-peak services in these three corridors would, therefore, need to be withdrawn. Three qualifications are required however. First, especially in the Outer North and Outer South corridors, radial and feeder routes have not been modelled separately. If differences exist in the level of returns from these two route types, then collective modelling is likely to disguise them. Therefore, before any closure decisions could be made, these two route types would need to be modelled in their own right.

Second, for reasons of analytical manageability, we are considering here only a single aggregated off-peak period, consisting of the time between am and pm peaks (i.e. the interpeak period), evenings, weekends and public holidays. Interpeak demand is substantially greater than demand in the remaining off-peak time periods (with average boardings per hour being 12500 vs 4500 respectively (STA, 1993a)). Thus the need to withdraw services in a corridor is likely to apply only to the very low demand periods, and not the interpeak period. To test this point, the first best results for the Outer North corridor, one of the three critical corridors where $ES_{OP} < 0$ in Table 7.1, were rerun for two separate off-peak periods : *OP1*, the interpeak period; and *OP2*, evenings, weekends and public holidays combined. In *OP1*, the first best optimum yielded $ES^* =$ \$55/route/hour, an economically sustainable outcome. In contrast, in *OP2*, ES was < 0 at all possible q values. Thus, in the case of the Outer North corridor, the unsustainable economic performance in the off-peak reported in Table 7.1 is reflecting the very low demand levels in evening, weekend and public holiday periods (*OP2*), with the interpeak (*OP1*) yielding positive net benefits.

Third, with public transport being such an important, and in some cases the only, source of mobility for many members of the community, particularly those that are transport disadvantaged, governments would no doubt be prepared, as they currently are, to deliver significant subsidies (over and above concession discounts) in low demand time periods on equity and social justice grounds to ensure a minimum level of mobility is provided at all times.[17]

- As one would expect, given the major difference in demand level between peak and off-peak, optimal patronage (q^*) is always greater in peak than off-peak.

- In all corridors, frequency (F) is greater in the peak, i.e. $F_{PK}{}^* > F_{OP}{}^*$. There are two effects underlying this result. In any given situation, F is optimised to ensure (7.11) holds, delivering an optimal trade-off at the margin between the time cost reduction benefits of additional F $\left(i.e. -q\left(\frac{\partial u}{\partial F}+\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial F}\Big|_{\bar{q}}+\frac{\partial v}{\partial F}\Big|_{\overline{LF},\bar{q}}\right)\right)$, and the associated additional producer costs $\left(i.e.\ \frac{\partial C_p}{\partial F}\right)$. The greater the marginal user benefits the greater is F^*, and the greater the marginal producer cost the smaller is F^*. With patronage being greater in the peak (i.e. $q^*_{PK} > q^*_{OP}$), the marginal benefit schedule is greater in the peak, tending to push F_{PK} above F_{OP}, *ceteris paribus*. On the other hand, there is an offsetting effect on the cost side, where $(C_p/VK)_{PK} > (C_p/VK)_{OP}$ means the peak marginal producer cost schedule will be above that in the off-peak, which will push,

[17] The question of whether providing a minimum level of service of public transport in all corridors is the most appropriate way of ensuring mobility for the transport disadvantaged is not addressed here.

ceteris paribus, F^*_{PK} below F^*_{OP}. With $F_{PK}^* > F_{OP}^*$ in all cases, the demand effect dominates the cost effect.

- Planned user behaviour dominates in all corridors and time periods. The greatest proportion of random users in any one situation is for the Eastern corridor in the peak where 24% of users act in a random fashion.

- The ranking of average unit subsidy, s_{ave}, in peak vs off-peak varies between corridors, i.e. $s_{PK} > s_{OP}$ in some cases, and $s_{PK} < s_{OP}$ in others. Although based on the conventional negative relationship between optimal unit subsidy and patronage reported in the literature this would seem an odd result, as chapter 5 showed (see section 5.4), the conventional relationship can break down over some patronage ranges. The cause of this occurring in chapter 5 was the switching which occurs from planned to random behaviour as patronage increases, which can result in optimal unit subsidy being higher for higher patronage. This is one explanation for why the $s_{PK} > s_{OP}$ result may occur.

Another reason contributing to having $s_{PK} > s_{OP}$ in some cases lies in the fact that the *ATC* curve is steeper, at any given patronage (q), in the peak than the off-peak. With the *ATC* schedule steeper, the gap between *ATC* and *MSC*[18], and thus optimal unit subsidy (= *ATC* - *MSC*), will in turn be greater for peak than off-peak at any given q value. In other words, the s_{PK} schedule lies above the s_{OP} schedule. There are two reasons why the *ATC* schedule is steeper at any given q value. First, at any given q value, with $(C_p/VK)_{PK} > (C_p/VK)_{OP}$, the AC_p schedule, and thus the *ATC* schedule, is steeper for peak than for off-peak. A second reinforcing effect is the fact that $(C_p/VK)_{PK} > (C_p/VK)_{OP}$ tends to push F_{PK} below F_{OP}, *ceteris paribus*, making random behaviour more likely, and thus making the AC_u (and thus *ATC*) schedule steeper in the peak than the off-peak.

The net result when one combines the conventional negative relationship between optimal unit subsidy and patronage with the above counteracting effects is that both ranking outcomes between s_{PK} and s_{OP} are possible depending on the difference between $(C_p/VK)_{PK}$ and $(C_p/VK)_{OP}$,

[18] Where $MSC = ATC + q\dfrac{\partial ATC}{\partial q}$ is a function of the slope of *ATC* (where $\dfrac{\partial ATC}{\partial q} < 0$).

the difference between peak and off-peak patronage levels, and the rate of switching between planned and random behaviour. Table 7.1 shows that $s_{OP} > s_{PK}$ in 5 corridors, whilst $s_{OP} < s_{PK}$ in 4 corridors.

- With respect to total subsidy (S), $S^*_{PK} > S^*_{OP}$ in all situations. Thus even in cases where $s^*_{OP} > s^*_{PK}$, the difference between q^*_{PK} and q^*_{OP} is great enough to ensure $S^*_{PK} > S^*_{OP}$. As a result, the ranking of total subsidy between time periods matches the ranking of patronage (q).

- Table 7.1 also shows that subsidy flows to both user groups, regular and concession (i.e. s_1 and s_2 are both > 0).

7.4.2 Second Best Results Compared To First Best

- As explained in section 7.3.4, moving to a second best optimum involves reducing patronage (q) below q_{FB}. Table 7.1 shows this to be the case in all corridors and time periods, along with a corresponding fall in frequency (F). Although F falling raises average user cost, AC_u, Table 7.1 suggests that notwithstanding this, fares (P_1 and P_2) need to rise to increase generalised cost, g ($= P + AC_u$), sufficiently to bring about the required fall in q.

- Not surprisingly, with a sudden introduction of a new major cost (i.e. the financing deadweight loss, *FDWL*), $ES^*_{SB} < ES^*_{FB}$ in all corridor/time period cases. As a result, a greater number of corridors fail to generate positive net benefits (i.e. $ES > 0$) in the off-peak than in the peak.

- In all corridors, total subsidy (S) in both periods is lower in the second best setting than the first best. This result is not surprising given the nature of adjustment required in moving from first best to second best discussed in section 7.3.4. The greater is the marginal financing deadweight loss, κ, the greater is the required adjustment, and the bigger the difference between S^*_{FB} and S^*_{SB}. The outcome which is unusual *a priori*, however, is the fact that in all corridors, as κ increases, the reduction in total subsidy is much greater in the peak than in the off-peak. So much so that in some cases S_{PK} reduces to, or below, zero, and in many cases $S_{OP} > S_{PK}$, in contrast to the first best situation where almost invariably the reverse result held. These are important results, but they require closer scrutiny to be fully explained. This will be done in section 7.4.3. One outcome of that further analysis is that the reduction in total subsidy and economic surplus outcomes from first best to second best varies in proportion with the size of κ. Thus the changes from first best to second best reported in Table 7.1 are relatively substantial because $\kappa = 0.4$ is likely to be towards the upper end of likely κ values.

- In contrast to the first best case where both user groups (regular (full fare) paying, and concession) receive a subsidy, in a second best setting many cases exist, particularly in the peak, where full fare paying users now make a net monetary contribution (i.e. $g_1 > ATC$) rather than receiving a subsidy, although concession users continue to be subsidised (i.e. $g_2 < ATC$).

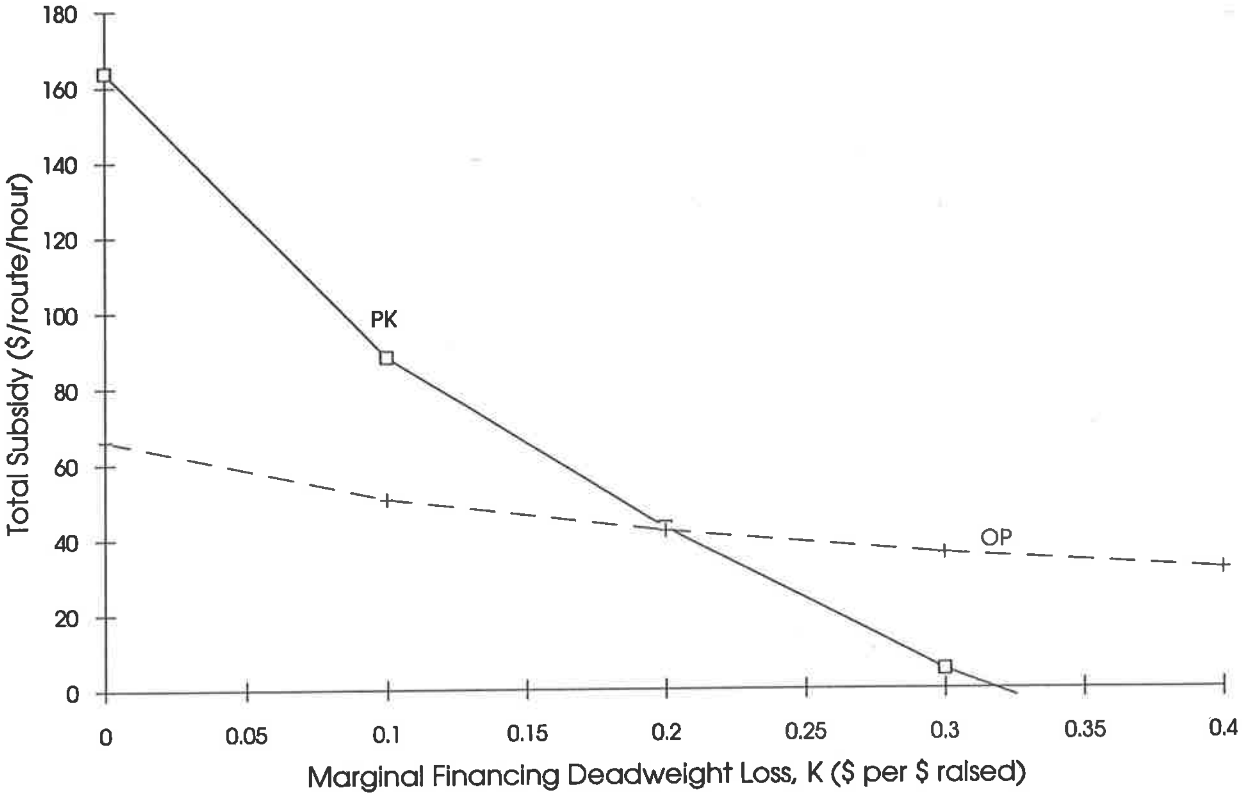
7.4.3 Detailed Analysis of Second Best Outcomes

Two questions remain regarding second best outcomes. First, given there is some uncertainty about the value the marginal financing deadweight loss, κ, actually takes (Findlay and Jones, 1982; Freebairn, 1995), how different are second best outcomes if κ takes on values other than the one used in the last section ($\kappa = 0.4$)? Second, what explanation can be offered for the unusual movements in total subsidy (S) between first best and second best settings observed in the last section?

These questions have been addressed here by studying in greater detail outcomes for one corridor, the Eastern corridor. Table 7.2 reports second best outcomes for the Eastern corridor for four values of κ : 0.1, 0.2, 0.3 and 0.4. Results for $\kappa = 0$, which are the first best results, are also reported as a base for comparison. Table 7.2 demonstrates a gradual reduction of patronage (q) through marking up of fares as κ increases, and a gradual reduction in economic surplus (ES) as public funds become increasingly costly to raise. The key results are the behaviour of unit and total subsidy (s and S).

Figure 7.7 summarises S^*_{SB} outcomes for the range of κ values considered. There are two interesting outcomes to note. First, in the case of the peak, if κ is large enough, $S^*_{SB,PK}$ eventually reduces to zero. This occurs when κ is approximately 0.32. Thus, in this case, removing even the last dollar of subsidy is worthwhile because the resulting saving of distortionary finance cost is greater than the gain in divergence deadweight loss from marking up g above MSC. If larger κ values apply, $S^*_{SB,\ PK}$ drops below zero, i.e. it is optimal to run a surplus rather than a subsidy. The reason for this is that if raising public finance is distortionary enough, then it is welfare improving to

Figure 7.7 : Second Best Optimal Subsidy, Eastern Corridor



raise public finance in the public transport market and at the same time relieve more costly revenue raising elsewhere in the economy. This is consistent with the general principles of optimal taxation (Ramsey, 1927) where taxes (or revenues) are raised in the least distortionary manner.

Table 7.2 : Second Best Outcomes Before Competitive Tendering, Eastern Corridor

Case	*q* *(boardings/hr)*	*F* *buses/ hr*	*LF*	*R* *%*	P_1 *cents*	s_{ave} *cents*	s_1 *cents*	s_2 *cents*	*S* *$/hr*	*ES* *$/hr*
(a) *PK*										
$\kappa = 0$	202	3.3	0.26	24	122	81	45	109	164	557
$\kappa = 0.1$	178	3.0	0.26	20	176	49	-5	88	88	540
$\kappa = 0.2$	163	2.8	0.25	17	217	26	-43	72	43	531
$\kappa = 0.3$	149	2.6	0.25	15	259	3	-81	56	5	525
$\kappa = 0.4$	137	2.4	0.24	13	298	-17	-115	42	-23	521
(b)*OP*										
$\kappa = 0$	72	1.6	0.14	8	88	92	57	106	66	51
$\kappa = 0.1$	60	1.5	0.13	8	121	85	34	103	51	42
$\kappa = 0.2$	52	1.4	0.12	8	145	80	18	101	42	35
$\kappa = 0.3$	46	1.3	0.11	8	166	77	5	99	36	29
$\kappa = 0.4$	42	1.2	0.11	8	184	75	-5	98	32	24

Notes :
1. Abbreviations : subscript 1 denotes regular (full fare paying) users; subscript 2 denotes concession users; κ *is the marginal public financing deadweight loss.*

The second, and more important general feature to note, is that the rate of reduction in S^*_{SB} in response to increments in the marginal financing deadweight loss, κ, is much smaller for off-peak than peak. So much so that, for $\kappa \geq 0.2$, $S^*_{SB,OP}$ becomes greater than $S^*_{SB,PK}$, which contrasts with the consistent first best outcome in all corridors where it was observed in Table 7.1 that $S_{FB,PK} > S_{FB,OP}$. Thus if public finance is costly enough to raise (i.e. if κ is big enough), then the strongest argument for subsidy is in the lower demand period, the off-peak. This is an important result since it contradicts the popular view often encountered that off-peak subsidies need to be justified on social justice and minimum service grounds rather than economic efficiency grounds. The discussion above illustrates that there is also a sound economic efficiency argument for subsidies flowing to a significant extent to lower demand periods.

The reason why (as κ increases) $S_{SB,OP}$ declines relatively moderately compared to $S_{SB,PK}$ is as follows. In the peak, as generalised cost, g, is marked up above *MSC*, the growth in *ATC* is relatively modest, and so *ATC* - *MB*, and thus s, declines fairly rapidly. As a result, a relatively small increase in g yields a fairly significant reduction in S, and thus also in financing deadweight loss (*FDWL*). Therefore, *FDWL* can be readily reduced with only modest divergences between *g* and *MSC*, and thus moderate divergence deadweight losses. Large reductions in *S* can thus be achieved prior to the optimal trade-off between reduction and gain of the two deadweight losses (financing and divergence).

In the off-peak, on the other hand, as *g* is marked up above *MSC*, *ATC* rises fairly steeply. Thus *ATC* - *MB*, and *s*, decline much more slowly than is the case in the peak. As a result, a relatively large increase in *g* is required to yield a significant reduction in *S* (and thus in *FDWL*). Thus, divergence deadweight losses are generated more quickly in the off-peak, with the optimal deadweight loss trade-off being reached with smaller reductions in *S*. In essence, in the off-peak, it is quite difficult to reduce *s* (and thus *S*) to any significant extent by marking up *g*. Thus, achieving any significant reduction in *S* in this way will also lead to substantial introduction of divergence deadweight losses.

The above explanation is confirmed by Figures 7.8 and 7.9 which plot, for the Eastern corridor, unit subsidy *s* (= *ATC* - *MB*) and total subsidy *S* (= *sq*) as patronage (*q*) is allowed to vary. It plots *s* and *S* for the current Eastern peak and off-peak demand curves, plus for levels of demand reduced and boosted by 50% in both peak and off-peak to test the sensitivity to demand level variation. For the peak, Figure 7.8 shows unit subsidy is positive at higher *q* values (where the first best optimum lies), gradually declining and becoming a unit *surplus* as *q* is reduced. The corresponding behaviour of *S* in Figure 7.9 shows *S* declining quickly towards a surplus situation as *q* reduces. Therefore, reductions in *S*, and thus reductions in *FDWL*, can be achieved with fairly moderate reductions in patronage (*q*).

In the off-peak, Figure 7.8 shows that whilst reducing patronage (*q*) does reduce *s*, the rate of decline is much more modest than in the peak, and if the reduction in *q* is large enough, *s* eventually begins to rise. This last outcome is due to the fact that at very low *q* values, *ATC* actually rises more quickly than *MB*.[19] The corresponding behaviour of *S* in Figure 7.9 sees *S* declining much more slowly in the off-peak than in the peak and never reaching a surplus situation in the off-

[19] This explains the result in Table 7.1 that, for some corridors, moving to a second best optimum saw *s* rise.

Figure 7.8 : Unit Subsidy, Eastern Corridor

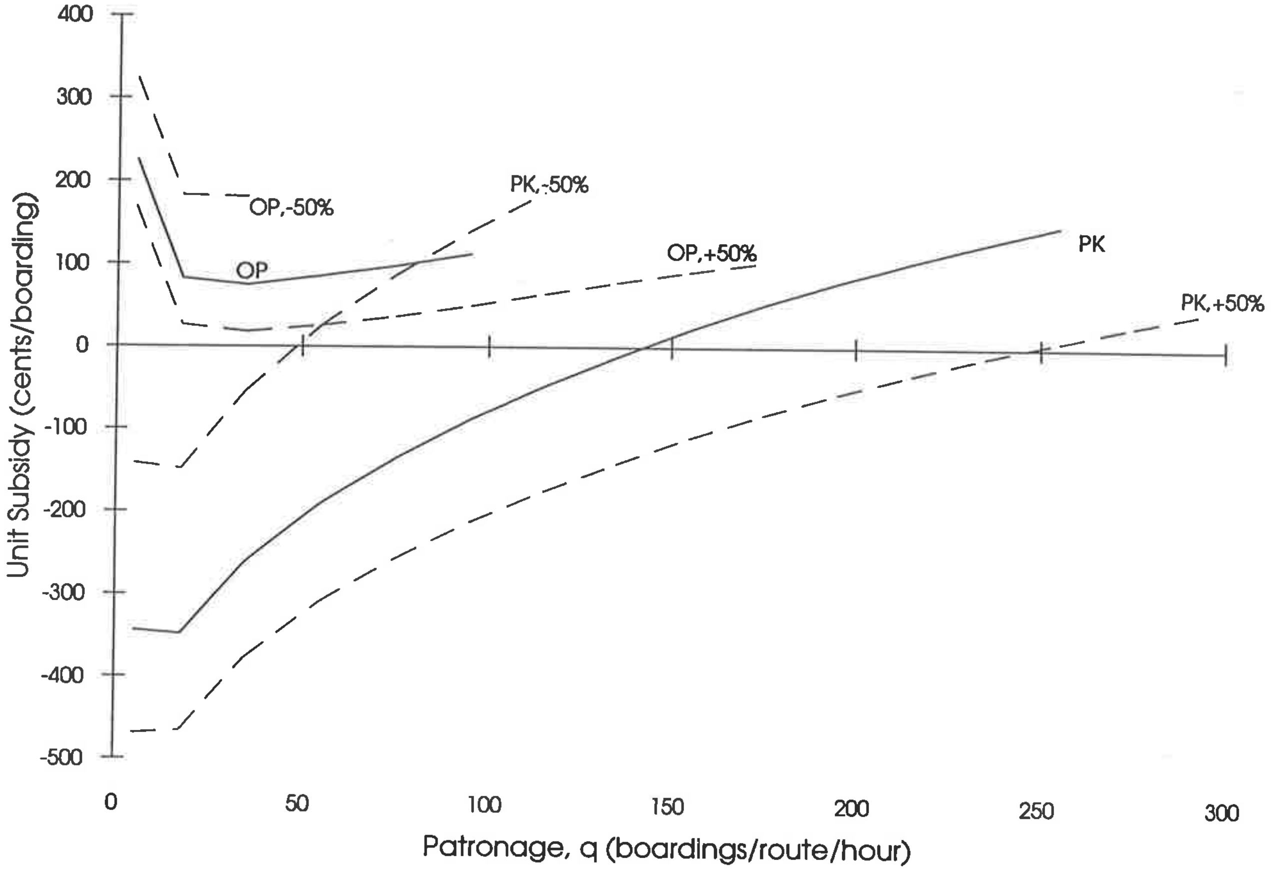
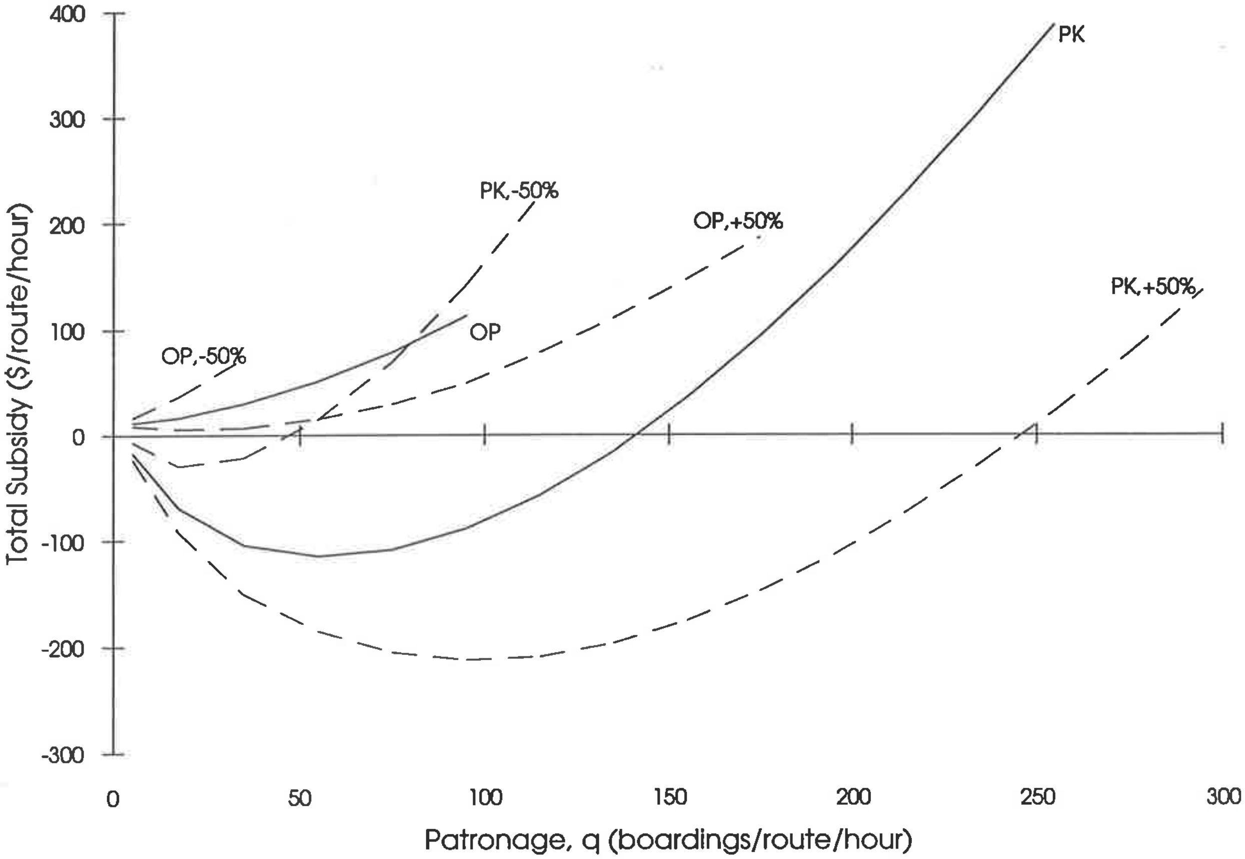


Figure 7.9 : Total Subsidy, Eastern Corridor



peak. Finally, note that the sensitivity plots in Figures 7.8 and 7.9, whilst altering the position and slopes of the curves, do not alter the relative difference between peak vs off-peak outcomes.

Overall, moving from first best to second best has a major impact on optimal outcomes. For the peak, optimal total subsidy (S*) reduces significantly. On the other hand, for the lower demand off-peak, the difference between first best and second best outcomes is nowhere near as marked, with second best subsidy still being quite substantial in size, particularly on a unit subsidy basis.

7.4.4 Results Under Random Behaviour

It is also useful to consider the first best results derived from using the *random* user behaviour model, rather than the logit model (on which the results in this chapter are based). Not only has the random model been widely used in most past analyses, but the most significant previous study of the Adelaide situation (Chalmers, 1990) assumed this type of user behaviour exclusively. Table 7.1, reports first best results using the random model for the Eastern corridor (the *FB*(*R*) lines for peak and off-peak in Table 7.1). The results show that using the random model has a significant impact in both peak and off-peak on optimal unit subsidy and total subsidy, yielding values well above those generated with the logit model.[20]

An important implication of this is that one of Chalmers' (1990) main conclusions should be treated with some scepticism. Chalmers finds that popular arguments, which claim that current subsidy levels in Adelaide are too high, are not supported by his optimal results. This conclusion is drawn, however, from his analysis using the random user behaviour model which tends to overestimate subsidy in cases where planned behaviour, or a mix of planned and random behaviour, might be expected. As indicated, in section 7A.1 of the appendix to this chapter, the anecdotal experience in Adelaide is that planned behaviour is the dominant form of user behaviour. Thus, Chalmers results may be significantly biased by the strong random user behaviour assumption built into his analysis. The same can be said of most past subsidy analyses (Tisato, 1992).

[20] Given the first best logit outcome in this corridor results in 24% random behaviour, the relative impact of using or not using the random model would be even greater if the base of comparison were closer to a pure planned situation rather than the logit outcome. This would be the case in corridors with lower demand levels where lower optimal frequency (F*) values make planned behaviour more likely.

7.5 Comparison of Optimal Results and Current Settings

With optimal outcomes now identified, how do these compare with the current policy settings and other outcomes in each corridor/time period, and thus what is the magnitude of the changes required to attain optimal positions ? The reference point for these considerations is once again Table 7.1.

Except for a small number of exceptions (the Eastern, South East and Stirling Hills corridors in the off-peak), a consistent pattern emerges across corridors and time periods. In these common cases, the changes required to move us from the current situation to optimal outcomes have the following characteristics :

- Reductions are required in frequency (*F*). This suggests that at the current frequencies, marginal time costs are generally not big enough to justify the higher marginal producer costs of increasing *F*. Economic surplus (*ES*) can therefore be increased by allowing *F* to decline which raises time costs by less than the resulting rise in producer cost.

- With *F* falling, the likelihood of random behaviour falls, and therefore so does the proportion of random users, *R*.

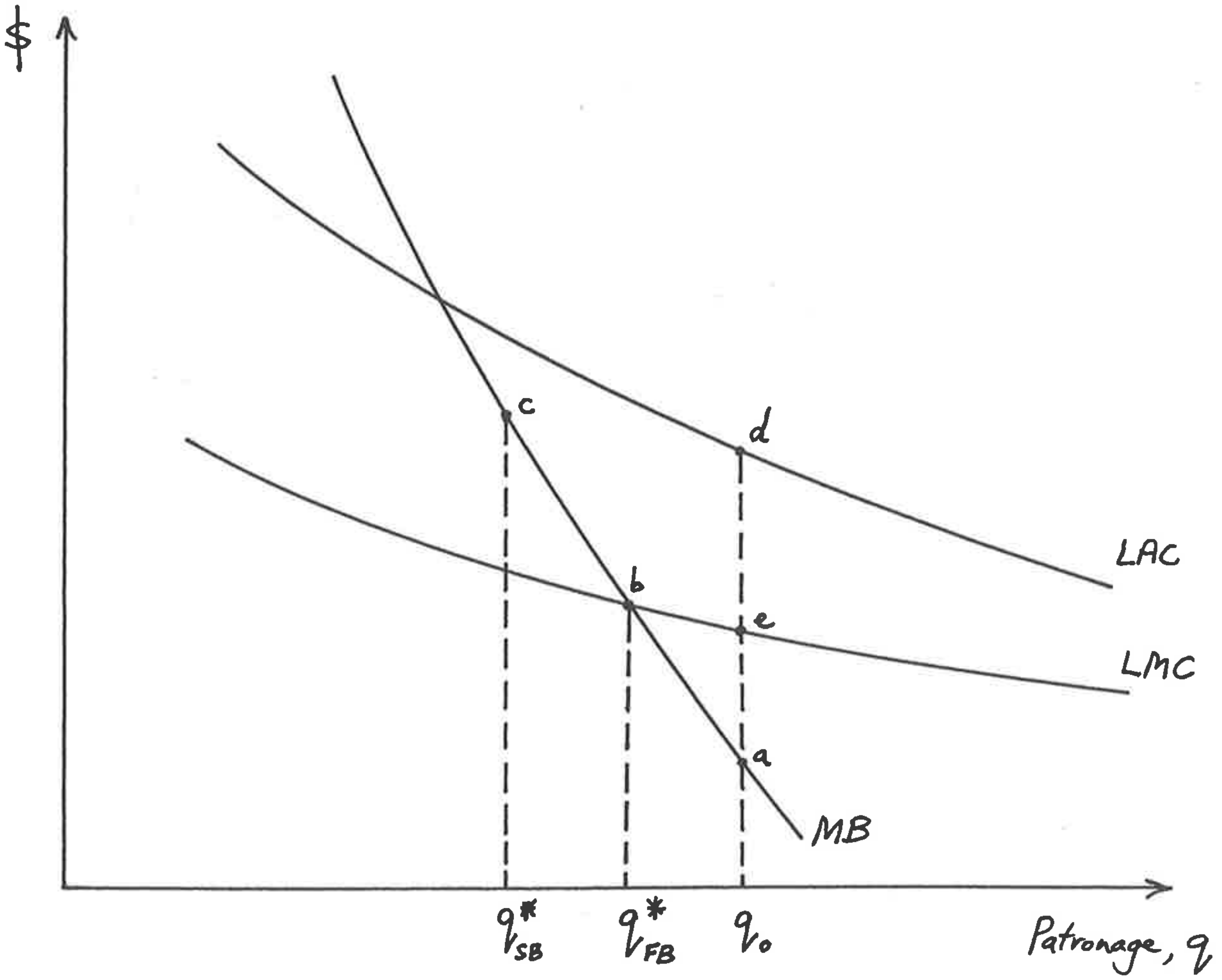
- Like *F*, moving to optimal outcomes also requires patronage, q, to fall.

- Optimal fares are higher than existing fares. Although *F* falling causes AC_u to rise, fares must also rise in order to achieve the required patronage reductions.

- Average unit subsidy, s_{ave}, falls. To understand why, consider Figure 7.10 which represents a typical situation in either period. With $q_o < q^*$, this suggests we are currently at a point like *a* in Figure 7.10. If we assume momentarily that q_o is being produced in the least costly manner, so that *ATC* at q_o is $LAC(q_o)$, then moving to point *b* would result in both *MB* and *AC* rising, but with the former by moreso, and thus s (= *ATC* - *MB*) falling. To the extent that production at q_o is less than fully cost efficient (i.e. frequency (*F*) is not being optimised at q_o), so that $AC(q_o) > LAC(q_o)$, this would merely reinforce the decline in s brought about from moving to the first best optimum at point *b*.

- Total subsidy (*S*) falls. The reduction in both s_{ave} and q explains this result.

Figure 7.10 : Comparison of Typical Current Situation With Optimal Outcomes



- In addition, the changes are greater in magnitude in the peak than the off-peak, and the changes required to move to a second best optimum (illustratively positioned at point c in Figure 7.10) are of significantly greater magnitude than those which required to reach a first best optimum.

The exceptions to the above general pattern of outcomes are the Eastern, SouthEast and Stirling Hills corridors in the off-peak. In the Eastern corridor, fares also fall along with q, F and s_{ave}. The reason for this is that the required fall in F (and thus associated rise in AC_u) is rather substantial, thus, with only a modest fall in q required (and thus a modest rise in g to achieve it), scope exists for P to fall. In the SouthEast corridor, optimising F requires it to be increased marginally. In addition, additional patronage can be accommodated. To achieve this, fares need to fall. Referring again to Figure 7.10, this type of situation would exist at a patronage (q) value below q_{FB}. The rightward move to the first best optimum results in MB and AC falling. The former must fall by more to attain the fall in s reported in Table 7.1. The Stirling Hills corridor reports rises in q, F and P, and a fall in s_{ave}. The feature of the base case for this corridor is the very low F and high H. Thus AC_u is disproportionally higher than that which occurs in the other corridor/time periods. The substantial rise in F in moving to first best therefore causes a very significant fall in AC_u. The size of this fall in AC_u is so significant that substantial fare rises are required to limit the size of the growth in q.

It is interesting to compare the changes suggested here for frequency and fares with those suggested by Dodgson (1985, 1986) as the chief conclusion to his study of urban public transport subsidy in Australia's major cities :

> *" ... this study indicated that there are benefits to be derived from a reduction in the level of public transport services in many of the Australian cities, and a switch of the subsidies saved to finance lower fare levels ... this conclusion was reached for both bus services and rail services ... "* (Dodgson, 1985, p.77)

The current study finds that, on the whole, frequency must fall and fares must rise. Thus there is consistency with respect to how frequency should move, but not fares. Although the two studies propose fares should move in opposite directions, there is no conflict in this result if one contrasts the objectives of the two studies. The Dodgson work considered the changes required in fares and frequency in order to maximise the return from subsidy *given existing subsidy levels were to be maintained.* Thus, his findings were not unconstrained first outcomes, but rather *sub-optimal* outcomes from an optimisation subject to a subsidy constraint in a first best world (i.e. costless public finance). With total subsidy given, a reduction in frequency, and thus a rise in AC_u, allows fares to fall. Alternatively, here the changes in fares and frequency move us to *optimal* outcomes, with subsidy reducing significantly rather than remaining fixed. With subsidy falling, fares needs to rise (even though AC_u falls with the fall in frequency).

7.6 The Sensitivity of Economic Surplus

The above comparison of current and optimal settings has not addressed one key area, namely, the impact on economic surplus (*ES*) of moving to optimal outcomes, and it is to this issue that the focus now turns. This is an important issue since *ES* is the prime indicator in economic terms of the net benefits to be gained from instituting policy reform.

Table 7.1 reports, for each corridor/time period, two current *ES* values : one assumes a first best world of costless fund raising; and the other (in brackets) assumes a second best world of costly financing where there are financing deadweight losses. Inspection of Table 7.1 reveals that all corridor/time situations are sub-optimal, that is, *ES* could be improved by moving toward the optimum. The *extent* of sub-optimality, and thus the degree to which *ES* can be improved by moving to an optimum, does vary, however, between corridor/time situations. This merely reflects the fact that some corridors are coincidentally currently closer to an optimum than others. For example, moving to a first best optimum in the Inner North corridor in the peak has a marginal impact only on *ES*, increasing it from $513/route/hour to $516. On the other hand, growth in *ES* is much more substantial in the Eastern corridor in the peak, growing from $498/route/hour to $557. The main reason for the difference is that the Eastern corridor has a much bigger gap between current and optimal frequency (*F*), which when corrected, leads to significant cost savings and thus improvement of *ES*.

Notwithstanding this heterogeneity, there is, however, an important consistent pattern across the *ES* results in Table 7.1, namely, that moving to an optimum in any given corridor/time setting brings about a much more substantial gain in *ES* in a second best world than in a first best world. Extending the example comparison used above, whilst the Inner North corridor in the peak saw little change in ES_{FB}, ES_{SB} alters more significantly, growing from $410/route/hour to $475. In addition, the Eastern corridor in the peak, which saw ES_{FB} rise by \$59/route/hour, now experiences a bigger rise in ES_{SB} which moves from \$357/route/hour to \$521, a change of \$164.

In several cases, the current situation yields negative net benefits (i.e. $ES < 0$), but optimisation allows it to become a service of net value (i.e. $ES > 0$). In a first best world this is the case for the Outer NorthEast and Stirling Hills corridors in the off-peak, whilst it is true of the Inner South and NorthWest corridors in the off-peak in a second best world.

Our understanding of the behaviour of *ES*, and the benefits of various policy changes, can be further enhanced by considering the move from a current situation to an optimal position as consisting of two changes :

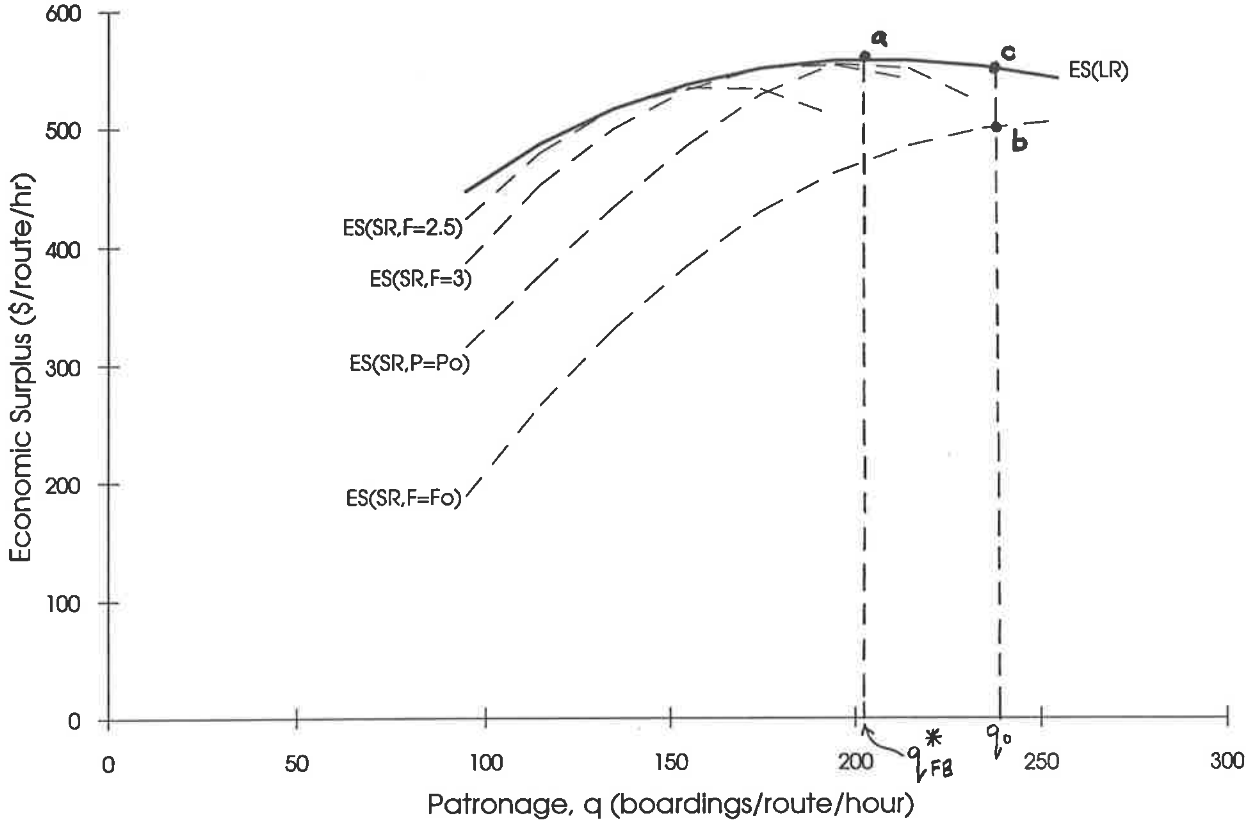
- Change 1 : This consists of optimising frequency (*F*) and fares (*P*), i.e. obtaining the optimal mix/balance between *F* and *P*, at the *current* level of *q*. That is, given the current patronage for a route, maximising the efficiency with which service is delivered to that clientele.
- Change 2 : With change 1 in place, the second change consists of moving to the optimal patronage level, q^*, whilst at the same time maintaining an optimal mix of *F* and *P* at all times.

In Figure 7.10, the current situation is represented by point *a* at $q = q_o$. If the optimal mix of *F* and *P* prevails, *AC* would coincide with point *d* on *LAC*. If, on the other hand, a sub-optimal mix of *F* and *P* prevails, *AC* will lie above point *d*, with production being less than fully cost efficient. Change 1 therefore consists of ensuring *AC* occurs at point *d* (and correspondingly *LMC* at point *e*) by achieving an efficient balance between *F* and *P*. Change 2 then consists of reducing *q* to q^* whilst ensuring that *AC* and *MC* continue to lie on *LAC* and *LMC* at all times.

Next, we adopt the distinction between short run (*SR*) and long run (*LR*) as used in cost analysis in microeconomics. In this present context, the *SR* consists of situations where *F* or *P* are fixed, whilst the *LR* is a situation where both can be varied. Using this distinction, another way of interpreting changes 1 and 2 is to see change 1 as ensuring that we are in a *LR* optimisation position, rather than a *SR* one, whilst change 2 ensures we reach an optimum in the *LR* setting.

The *SR*/*LR* distinction is also a useful adjunct for comparing economic surplus (*ES*) in optimal and non-optimal settings. To illustrate, consider once again the Eastern corridor in the peak. Figure 7.11 plots, for a first best world, a family of *ES* curves relating *ES* and patronage (*q*) for

Figure 7.11 : Economic Surplus, Eastern Peak Corridor, First Best World



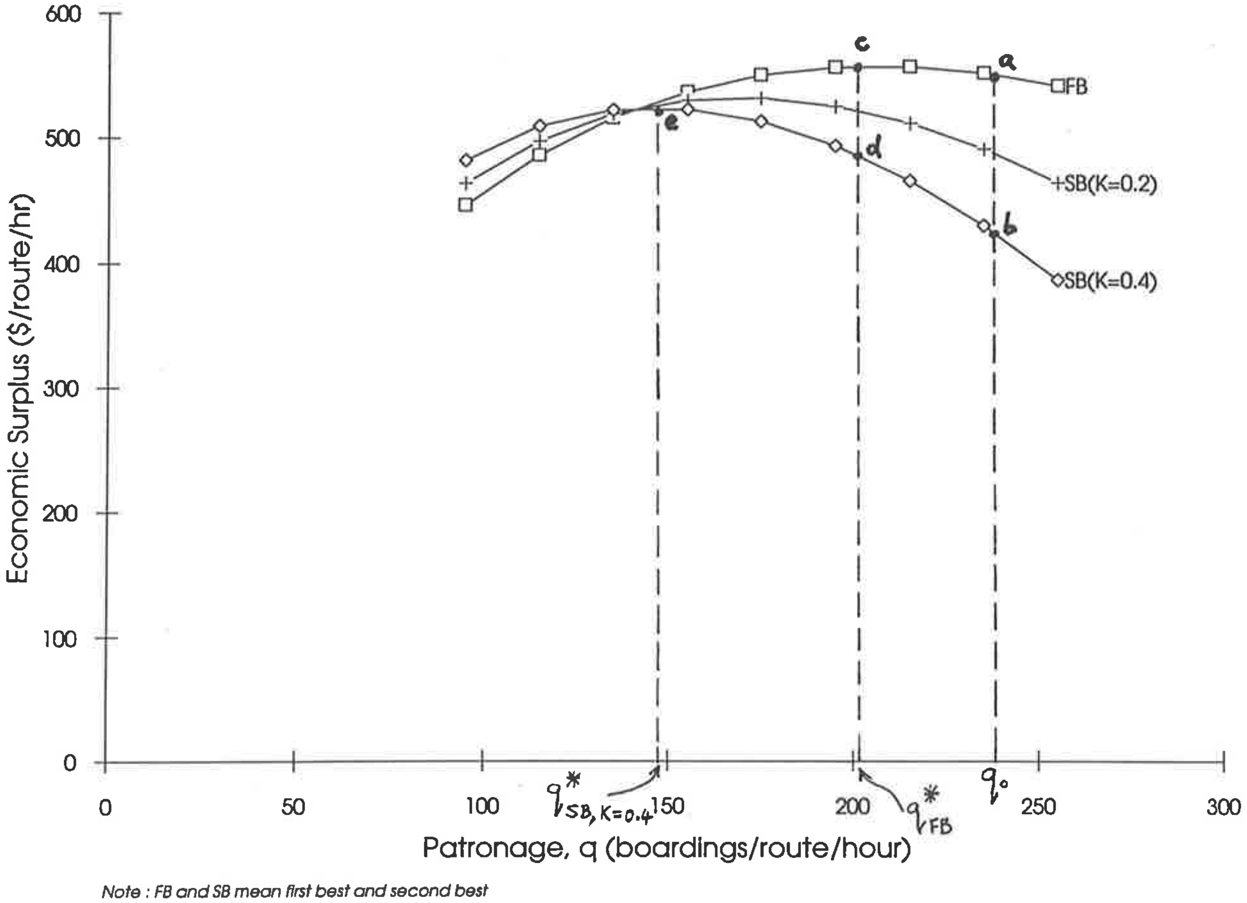
Note : SR and LR mean short run and long run various policy settings. The ES_{LR} curve in Figure 7.11 illustrates ES with P and F simultaneously optimised at any given q level. Each of the other curves in Figure 7.11 are ES_{SR} curves which report ES where one of the policy variables is fixed, but the other can be optimised. The ES_{LR} schedule forms an envelope to the entire collection of ES_{SR} curves (illustrated in the figure by drawing four ES_{SR} curves : the first three for $F = 2.5$, 3.0 and 5.2 (the current F value) respectively, with P optimised in each case; and the fourth with the current average fare fixed but F optimised). As one would expect, ES_{SR} is always less than ES_{LR} at any given q level. The overall first best optimum is at the peak of the ES_{LR} curve (point a in Figure 7.11), whilst all other points on the ES_{LR} and ES_{SR} curves are sub-optimal outcomes.

With both the P/F mix, and q, being currently sub-optimal in the Eastern peak corridor, we are presently situated at point b in Figure 7.11. Splitting the move from the current position to the overall first best optimum (i.e. the move from point b to point a) into its two change components, change 1 produces a move from point b to point c, whilst change 2 moves us from point c to point a. In this particular corridor, the current F is well above F^*, generating an ES_{SR} curve ($ES_{SR}(F_o)$) well below ES_{LR}. Combining this with the fact that the ES_{LR} schedule is quite flat, results in change 1 generating most of the gain in ES associated with moving to the optimal outcome (point a). Change 1 will also dominate in other corridor cases where the current situation is well away from the optimum. In cases where the margin between current and optimal situations is less pronounced, the gap between ES_{SR} and ES_{LR} at q_o will be smaller, with changes 1 and 2 then providing a more equal contribution to improving ES.

The flatness of the ES_{LR} schedule is an interesting outcome since it implies that the benefits of change 2 in a move to the overall optimum are always likely to be fairly modest. This suggests that pursuing optimisation of patronage (q), and associated reductions in the total level of subsidy, may not be a particularly worthwhile policy reform. To test the robustness of this result, the shape of the ES_{LR} schedule was analysed in a second best context, and also for several other corridors.

Initially, continue to consider the Eastern Peak corridor. Figure 7.12 plots for the Eastern Peak corridor a comparison of ES_{LR} schedules in first best and second best worlds, with second best

Figure 7.12 : *ES(LR)* Schedules, Eastern Peak Corridor, First and Second Best Situations



Note : FB and SB mean first best and second best curves being presented for two values for the marginal financing deadweight loss, κ, 0.2 and 0.4.[21] The key feature to note is that the rise in ES_{LR} which results from optimising q (i.e. reducing q from q_o to q^*) is much greater in a second best world than in a first best world, with the gain increasing in magnitude the higher is the marginal financing deadweight loss, κ. There are two factors contributing to this result :

(a) The first factor is that, as Figure 7.12 shows, the ES_{LR} schedule is more responsive to changes in q in a second best world, with the responsiveness increasing progressively as κ increases.[22] The implication of this is that, from the sub-optimal position at q_o (i.e. point a in a first best world, or point b in a second best ($\kappa = 0.4$) world), a given reduction in q, say q_o - q^*_{FB}, produces a much bigger rise in ES_{LR} in a second best setting. Also the greater is κ, the greater is the rise in *ES*. Thus, provided an optimal mix of *P* and *F* can always be implemented, reducing q is a more useful policy in a second best setting than in a first best one.[23]

(b) The second factor evident in Figure 7.12 causing the gains in *ES* associated with moving to an optimum to be greater the greater is κ, is that the fall in q required to reach the optimum from the sub-optimal q_o is greater the greater is κ. For example, in a first best world (i.e. $\kappa = 0$), moving to the optimum involves moving from point a to point c. In a second best world with $\kappa = 0.4$, however, a bigger reduction in q is required to move to the optimum, i.e. the move from point b to point e.

[21] The three curves cross at the same q level, the level coinciding with where subsidy moves from positive to negative (i.e. a move from deficit to surplus).

[22] To see why this is the case consider Figure 7.10 again. The current situation at point a generates a divergence deadweight loss triangular shape *abe* which is an efficiency cost in both first and second best worlds. Reducing q therefore reduces this deadweight loss and increases economic surplus (*ES*) in both worlds. Whilst the gains in first best world are limited to this, in a second best world, the reduction in q also results in a reduction in subsidy and thus a reduction in financing deadweight loss (which is a proportion κ of subsidy). Therefore, in a second best world, there is a second source of deadweight loss reduction which raises *ES*. In addition, the reduction in financing deadweight loss (a rectangular area) is likely to be of greater magnitude than the change in the triangular divergence deadweight loss, making the gain in *ES* even greater in a second best world than a first best world. Further, the larger is κ, the greater the initial financing deadweight loss, and thus the bigger the reduction in this, and the bigger the gain in *ES*, from a unit patronage reduction.

[23] A corollary of this is the reverse result, namely, that once an optimum has been reached, again provided *F* and *P* can be optimised throughout, maintaining q close to its optimal value is relatively more important the greater the marginal cost of public finance.

Next compare the above results with those from several other corridors. Table 7.3 compares the Eastern Peak corridor results just discussed with similar results for two other diverse corridors in the peak (SouthWest and Outer NorthEast), and with results for the Eastern corridor in the off-peak. The table reports the change in ES_{LR} resulting from a move along the ES_{LR} curve for a given change in q (i.e. a move in Figure 7.12 from point a to point c in a first best setting, and from point b to point d in a second best setting). Throughout Table 7.3, the rise in ES_{LR} (i.e. the ΔES column) is consistently larger in a second best setting than a first best setting. The size of the falls differ between corridors because the size of the change q_o - q^*_{FB} varies between corridors, with a positive relationship between the change in q_o - q^*_{FB} and ES.

Table 7.3 : Responsiveness of Economic Surplus

Corridor	$ES_{LR}(q_o)$ *\$/route/hr*	$ES_{LR}(q^*_{FB})$ *\$/route/hr*	Δ *ES* *\$/route/hr*
Eastern			
PK - FB	553	557	4
-SB, κ = 0.4	434	491	57
OP - FB	49	51	2
- SB, κ = 0.4	16	24	8
SouthWest			
PK - FB	256	273	17
-SB, κ = 0.4	117	212	95
Outer NorthEast			
PK - FB	648	674	26
-SB, κ = 0.4	409	568	159

Notes :
Abbreviations : ES is economic surplus; κ is the marginal financing deadweight loss.

The sensitivity testing just undertaken reveals that the merits of moving along the ES_{LR} curve, i.e. optimising patronage once the *P/F* mix has been optimised, varies significantly between first and second best settings. The gains from such a move are relatively modest in a first best world, but more substantial in a second best world, with the size of the second best gains increasing as public finance becomes increasingly costly.

Several general conclusions can be drawn from the analysis in this section :[24]

- In a *first best* world, once the mix between fares and frequency has been optimised, the gain from then optimising patronage (which also optimises subsidy) is likely to be relatively modest. This in turn suggests that, in a first best world, the gains in moving to the optimum can only be substantial if fares and frequency are significantly out of balance. A corollary of this is that ensuring that fares and frequency are balanced is the essential component of a strategy for improving *ES* in a first best world.[25]
- In a *second best* world, a policy of maintaining a balance between fares and frequency is also important. However now, once the *P/F* mix is optimised, a policy of optimising patronage (and thus subsidy) will yield greater gains (and thus play a more important role) than in a first best world, with the size of the gains increasing progressively as public finance becomes increasingly costly to raise at the margin.

7.7 The Impact of Competitive Tendering

The analysis so far has been based on the current producer cost structure. The focus now turns to considering how the above results vary if producer costs can be lowered. Reductions in producer costs are expected to flow from recent reforms in Adelaide, which have led to a planned move away from monopoly bus service provision to the pending introduction of competitive tendering. Based on experience overseas, unit operating cost reductions of the order of 20% to 30% are feasible through the introduction of competitive tendering (Stanford, 1992). The SA government has based its justification for the introduction of competitive tendering on the potential for cost savings, claiming competitive tendering will result in annual cost savings of $M 34 within 5 years across the public transport system (a reduction of approximately 23%) (The Advertiser, 1994). Discussions with TransAdelaide (Willis, 1995) revealed that a 20% reduction in operating

[24] The overall thrust of the results here are similar to those from the more general analysis by Akerlof and Yellen (1985) who show that movements from an optimum are often of "second order" magnitudes in a first best type setting, but of "first order" magnitudes in other settings where there are disturbances in the system, such as taxes or externalities.

[25] The need to maintain an optimal mix of *P* and *F* was also found to be of critical importance in the Dodgson (1985) study (as have other studies, e.g.Nash, 1982; Hensher, 1989a; Glaister, 1987).

costs could be considered a realistic expectation in the case of Adelaide buses, and has therefore been used as a working figure in the analysis that follows.

7.7.1 Impact on Subsidy in the Short Run

The immediate impact of competitive tendering depends on whether cost reductions are passed on to consumers or whether they are absorbed to reduce subsidy. The general current thrust of public sector reform in South Australia, with its focus on improving the financial position of the state, suggests that the latter strategy, subsidy reduction, may be favoured. For the Adelaide bus system as a whole, a 20% decline in operating costs would result in a fall in costs, and thus subsidy, of around \$28M per annum. For the 13 ROSIS corridors being considered in this study, system operating cost is \$98.7M, with competitive tendering therefore yielding a cost, and subsidy, reduction of \$17.7M.

7.7.2 Impact on Optimal Outcomes in the Longer Run

In the longer run, efficiency can be improved by moving to optimal outcomes (as discussed in sections 7.5 and 7.6). It is important to note, therefore, that the introduction of competitive tendering also influences the nature of both first best and second best optimal outcomes. This occurs due to the fact that, as a result of introducing competitive tendering, both the *MSC* and *ATC* curves shift downwards.[26] The previous first best and second best outcomes immediately become sub-optimal, requiring adjustments to be made to policy settings to ensure moving to the new optimal outcomes.

[26] The lowering of the *ATC* and *MSC* schedules due to competitive tendering cannot be unambiguously explained *a priori*. There are a number of conflicting impacts. At any given q, introducing competitive tendering lowers C_p/VK. This means that the cost of additional F falls, so F^* will increase, which will unambiguously cause AC_u to fall. The net impact on AC_p is unclear however. Whilst the increase in F causes AC_p to rise, the initial fall in C_p/VK causes AC_p to fall. The net impact on AC_p, and thus ATC $(= AC_p + AC_u)$ is thus ambiguous. In addition, with MSC $(= ATC + \frac{\partial ATC}{\partial q})$, the uncertainty about how *ATC* will change also makes the impact on *MSC* ambiguous.

Simulation runs of the Eastern corridor, however, generated a fall in both *ATC* and *MSC* schedules resulting from the introduction of competitive tendering. For example, for the peak period, the percentage reduction in *ATC* and *MSC* proved to be consistently around 6%-7% over the entire q range. Although formal simulation runs were not undertaken for other corridors, the impact of competitive tendering on policy variables and other outcomes to be discussed below shows considerable consistency, suggesting that the *MSC* and *ATC* schedules also fall in a similar manner in other corridors/time periods.

With the *MSC* curve falling, there will be a new set of first best and second best outcomes for each period. The situation is described in Figure 7.13, where the notation *BCT* and *ACT* refers to before and after competitive tendering respectively. With the *MSC* schedule falling, the first best optimum shifts from point *a* to point *b*, resulting in q^* increasing, and g^* falling.

Table 7.4 reports first best and second best results after competitive tendering for three illustrative corridors : the Eastern corridor (which is considered for reasons of continuity given its illustrative use earlier), the important Outer NorthEast corridor (which includes the NorthEast Busway), and the SouthEast corridor (one of the corridors which gave off-peak outcomes in Table 7.1 which were out of step with the general trends in the table). Comparison of Tables 7.1 and 7.4 reveals that the relativities between second best and first best results after competitive tendering are identical to those before competitive tendering and so are not repeated here. The impact of competitive tendering on optimal outcomes is therefore gauged by comparing first best results before competitive tendering and after competitive tendering. Comparing Tables 7.1 and 7.4, the following points, which apply to all three corridors in both time periods, can be noted :

- with optimal patronage (q^*) increasing and C_p/VK decreasing (which respectively increase the marginal benefit, and decrease the marginal cost, of additional *F*), there is a resulting rise in optimal frequency (F^*);
- notwithstanding the reduction in average user cost (AC_u) resulting from the increase in F^*, the fall in g_{ave} due to q^* increasing is sufficient to allow optimal fares, P_1^* and P_2^*, to decline;
- optimal unit subsidy (s^*) falls for two reasons : first, with C_p/VK falling, the *ATC* curve becomes flatter, reducing the gap between the *ATC* and *MSC* schedules, and thus reducing s_{FB} at all *q* values; second, as *q* increases, the gap between *ATC* and *MSC* curves also reduces;
- optimal total subsidy (S^*) remains relatively *steady* : it rises marginally in some cases, and falls in others. This stability is due to the fact that s^* and q^* change in opposite directions, thus restricting the size of the net change in S^* ($=s^*q^*$);
- as one would expect, the reduction in cost structure resulting from the introduction of competitive tendering results in a direct enhancement of economic surplus (*ES*).

Figure 7.13 : Impact of Competitive Tendering On First Best Optimum

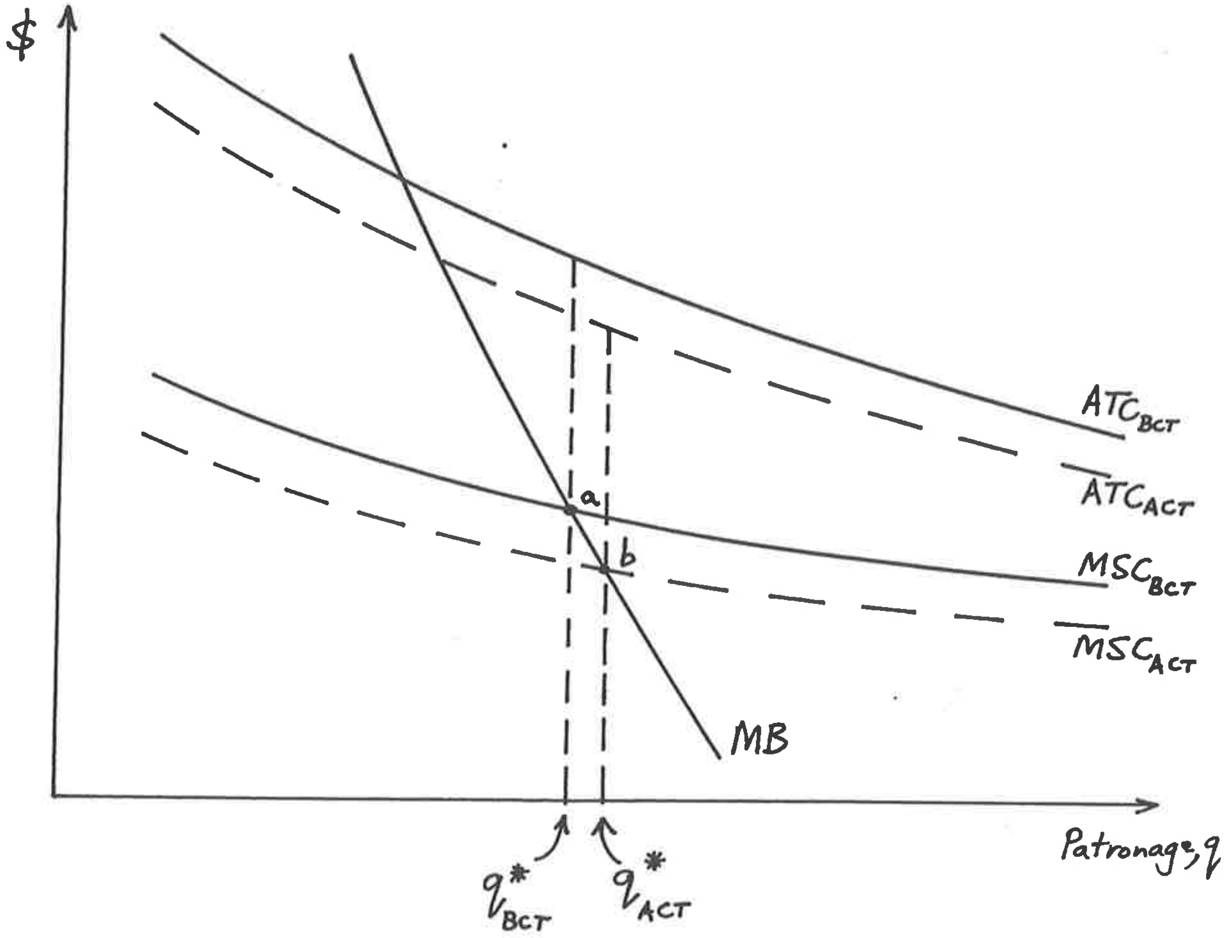


Table 7.4 : Representative Route Optimal Results After Competitive Tendering

Corridor	q *(boardings/hr)*	F *buses/hr*	*LF*	*R* *%*	P_1 *cents*	s_{ave} *cents*	s_1 *cents*	s_2 *cents*	*S* *$/hr*	*ES* *$/hr*
Eastern										
PK - FB	220	3.7	0.25	30	97	80	51	102	176	609
-SB, κ = 0.4	147	2.7	0.24	16	275	-24	-115	31	-36	577
OP - FB	84	2.0	0.13	9	70	77	49	88	64	73
- SB, κ = 0.4	54	1.5	0.11	8	158	54	-14	75	29	48
SouthWest										
PK - FB	134	2.3	0.28	11	208	112	43	151	149	327
-SB, κ = 0.4	91	1.7	0.25	8	387	31	-107	92	29	285
OP - FB	72	1.6	0.17	8	148	96	27	113	69	49
- SB, κ = 0.4	41	1.1	0.14	8	264	79	-54	100	33	22
Outer NorthEast										
PK - FB	287	2.6	0.26	16	132	91	64	131	260	765
-SB, κ = 0.4	200	2.0	0.24	10	278	2	-62	80	5	707
OP - FB	93	1.4	0.13	8	95	108	76	128	100	52
- SB, κ = 0.4	48	0.9	0.10	8	187	103	32	135	49	10

Notes :
1. Abbreviations : subscript 1 denotes regular (full fare paying) users; subscript 2 denotes concession users;
κ *is the marginal public financing deadweight loss.*

Importantly, the introduction of competitive tendering *softens* the impact of moving from current policy settings to optimal outcomes. Table 7.1 showed that, before competitive tendering, substantial changes were required in order to move to optimal outcomes : F and q must fall, whilst P must rise, reform proposals which are likely to be met with social and political resistance. In contrast, the introduction of competitive tendering causes F^* and q^* to rise, and P^* to fall, thus making the size of the adjustments required in moving to optimal outcomes smaller after competitive tendering than before competitive tendering. In a first best world, this makes the adjustments required to reach an optimum quite moderate in some cases. For example, in the Eastern corridor, moving from the current situation to a first best optimum after competitive tendering requires a 7% reduction in q (compared to 15% reduction before competitive tendering), a slightly smaller reduction in F than is required before competitive tendering, and a 10% reduction in fares (compared to the 13% increase required before competitive tendering). In a second best (κ = 0.4) world, however, even after competitive tendering has been introduced, major policy changes are still required to reach an optimum. Further, the adjustments increase in size as the marginal deadweight loss from raising public finance (κ) increases.

7.8 Aggregated Results for the Adelaide Bus System

An important aspect of the analysis in this chapter has been its disaggregated focus since it has ensured that demand and cost variations across the bus system can be reflected in optimal outcomes. Bringing these disaggregated results together, what are the overall implications for bus subsidy (*S*) and economic surplus (*ES*) for the system in aggregate?

Figure 7.14 presents a summary of aggregated *S* and *ES* results for the combined 13 bus corridors considered in this chapter. The Figure is presented in two parts : part (a) plots subsidy, whilst part (b) plots economic surplus. In each part of the Figure, results are reported for a number of cases. There are four groupings of cases along the horizontal axis (denoted by the numbers 1 to 4) covering the four possible combinations of first and second best worlds,[27] and before and after competitive tendering, as summarised in Table 7.5.

Table 7.5 : Summary of Case Groupings for Aggregate Result Reporting

	Before Competitive Tendering	*After Competitive Tendering*
First Best World	Grouping 1	Grouping 2
Second Best World	Grouping 3	Grouping 4

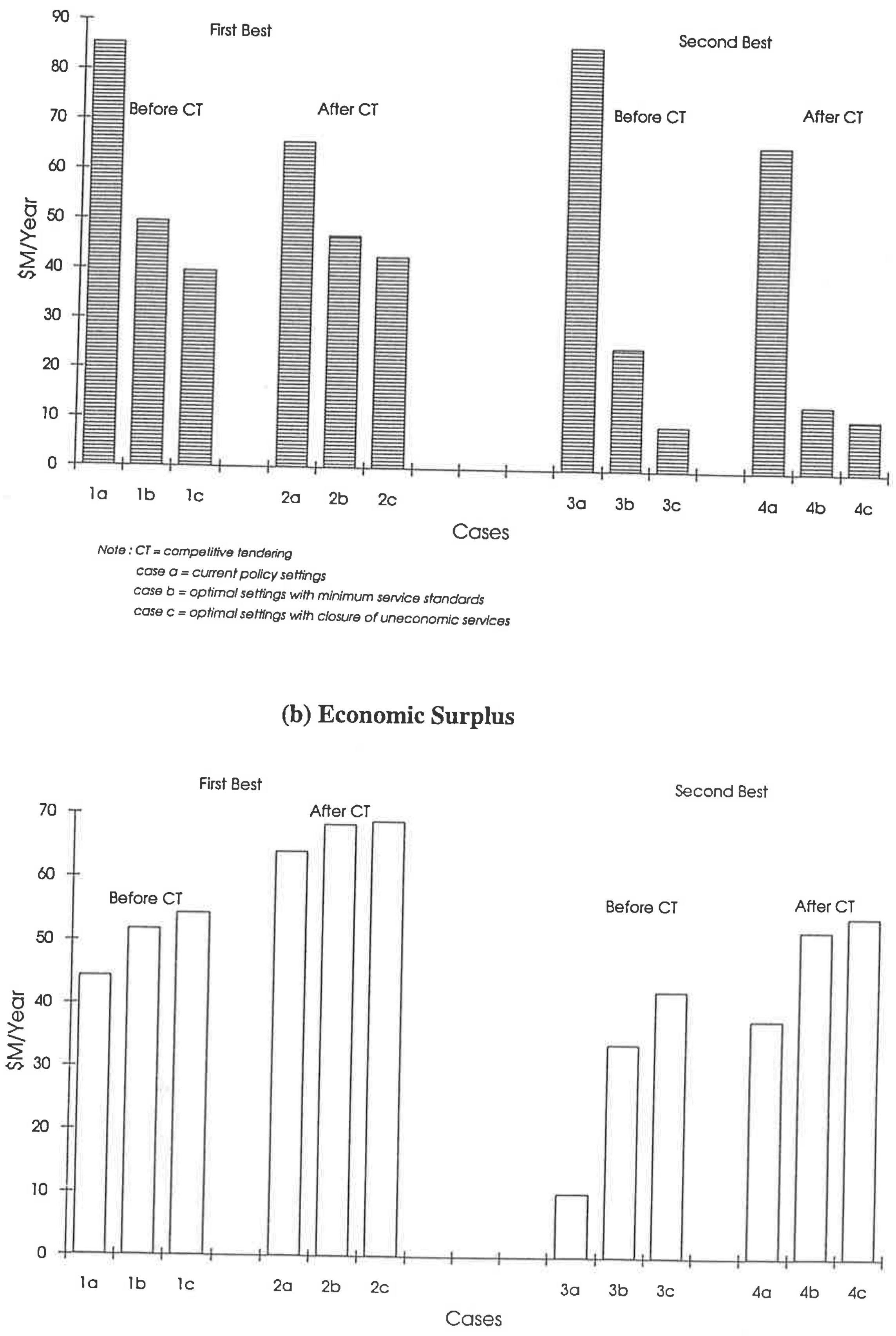
In addition, within each grouping, three cases are considered, denoted as follows :

- ***a*** : Outcomes under current policy settings;
- ***b*** : Optimal outcomes in corridors where $ES^* > 0$, but in corridors where $ES^* < 0$, results reported are based on a minimum service level of $F = 1$ and fares which deliver patronage reductions of similar scale to those occurring in the optimised corridors (i.e. where $ES^* > 0$); and
- *c* : Optimal outcomes in corridors where $ES^* > 0$, and withdrawal of all services on corridors where $ES^* < 0$.

Cases *a* to *c* therefore reflect different positions the government may take. Case *a* reflects a continuation of existing policies, case *c* reflects fully optimal policies (from an economic efficiency

[27] The second best results are presented for a marginal financing deadweight loss, κ, = 0.4.

Figure 7.14 : Aggregated Results for Adelaide Buses : Comparison of Current and Optimal Outcomes



(a) Subsidy

Note : CT = competitive tendering
case a = current policy settings
case b = optimal settings with minimum service standards
case c = optimal settings with closure of uneconomic services

(b) Economic Surplus

perspective) including service withdrawal of uneconomic routes where net benefits (*ES*) are negative, and case *b* is an intermediate case where, rather than shut down uneconomic routes, they continue to operate, delivering a minimum service level.

The before competitive tendering results in Figure 7.14 indicate that current subsidy (column 1 a) is well above the optimal user economies of scale subsidies (cases *b* and *c*). In a first best world, moving to optimal subsidy case *b* (column 1b) would require a 41% reduction in subsidy, whilst achieving the more radical optimal subsidy case *c* (column 1c) would result in a 53% reduction in subsidy. In a second best world before competitive tendering, substantially greater subsidy reductions are warranted due to the fact that reducing subsidy not only improves allocative efficiency in the bus market (as it does in a first best setting), but it also reduces the distortionary costs elsewhere in the economy associated with the raising of public finance. Moving to second best optimal subsidy cases *b* and *c* (columns 3b and 3c) would require subsidy reductions of 71% and 89% respectively. These second best reductions are for a marginal deadweight loss of public fund raising, κ, = 0.4. The subsidy reduction would be less (more) pronounced if public fund raising was less (more) distortionary at the margin.

The introduction of competitive tendering results in lower operating costs, and thus presents an immediate opportunity to reduce subsidy levels. This has a substantial immediate impact on both subsidy and economic surplus (compare columns 1a and 2a). Even after competitive tendering has been introduced, however, current policy settings remain sub-optimal, with further subsidy reductions required in order to reach optimal outcomes. In a first best world, moving to optimal subsidy cases *b* and *c* (columns 2b and 2c) would require a further subsidy reduction (from column 2a) of 23% and 27% respectively. In a second best world, even larger reductions are required, with a move to optimal subsidy cases *b* and *c* (columns 4b and 4c) requiring a further subsidy reduction (from column 4a) of 79% and 83% respectively.

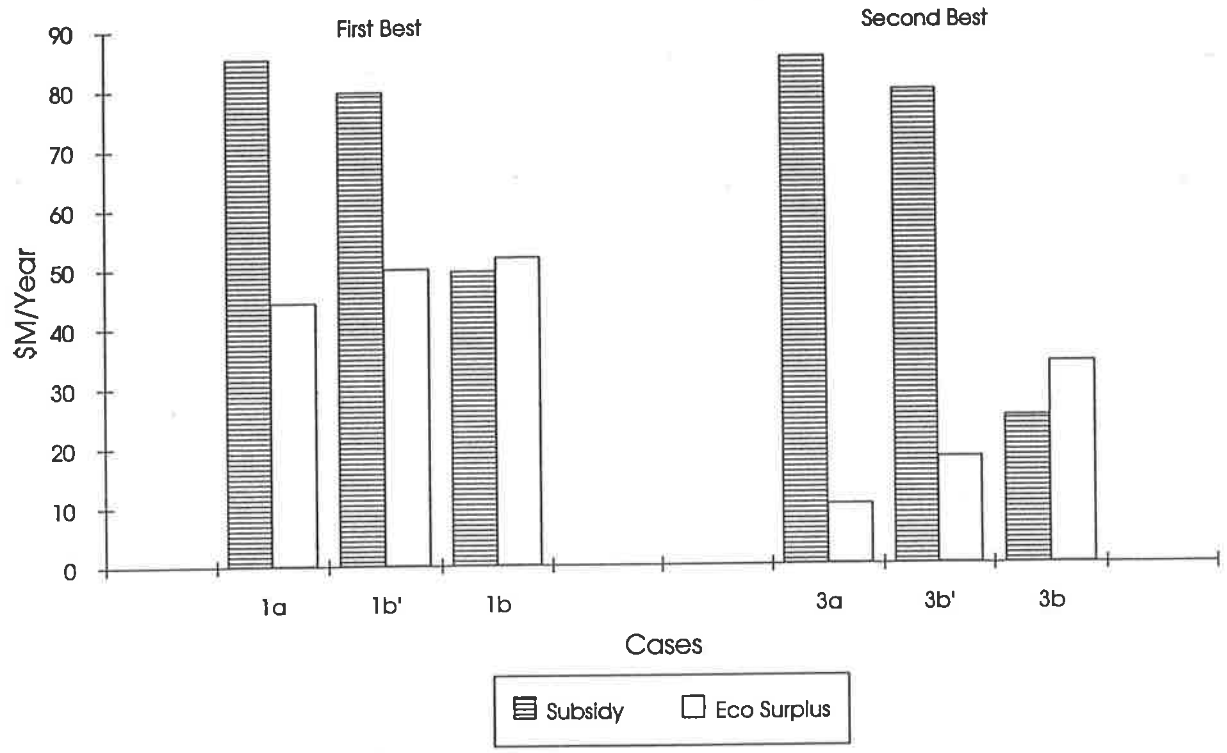
Overall, substantial subsidy reductions can be justified from current levels, with the required reductions being greater in a second best world, and with the scale of reductions required in a second best world varying in proportion with the efficiency cost of raising public finance. It is important to note, however, that in no case is zero subsidy justified. The lowest optimal subsidy in the cases considered was $A 10m, although if the marginal efficiency cost of public finance continued to grow, optimal subsidy would eventually be driven to zero.

The results in Figure 7.14 illustrate the adjustments required, and the gains to be made, in moving to first and second best optimal outcomes from the current sub-optimal situation. As noted in section 7.6 at the corridor level, however, a considerable proportion of the gains from moving to an optimum can come from ensuring that the mix of fares (*P*) and frequency (*F*) is optimised (or at least improved). Figure 7.15 (which reports before competitive tendering results only) illustrates that this result also holds at an aggregated level for the Adelaide bus system. For both first best and second best settings (presented on the left and right hand sides of the figure respectively), Figure 7.15 presents subsidy and economic surplus results for each of three cases. Cases *a* and *b* and their results (column pairs 1a, 1b, 3a and 3b) are those used in Figure 7.14, i.e. current policy settings, and optimal settings with a minimum service standard, respectively. The third case, *b*' (column pairs 1b' and 3b'), is an intermediate case in which the balance between *P* and *F* is optimised whilst the existing patronage, q_o, is maintained.

Figure 7.15 shows that, in a first best world, focusing on optimising the mix between *P* and *F* without also altering *q* (i.e. moving from case *a* to case *b*') yields a majority of the gains in economic surplus (*ES*) that come from also optimising (reducing) *q* (i.e. moving from case *a* to case *b*). Notice from Figure 7.15 that optimising *q* also leads to a substantial reduction in subsidy. In a second best world, the gains from optimising *q* become more important, in fact in Figure 7.15, where the second best results are based on a marginal efficiency cost of public funds, κ, = 0.4, the gains from optimising *q* are actually greater than the gains from optimising the mix of *P* and *F*. As explained in section 7.6, this is due to two factors which are stronger in a second best setting than a first best setting : the greater responsiveness of *ES* to changes in *q*, and the greater reduction in *q* required to move to an optimum.

The conclusion one can draw in a first best setting is that, if policy makers are reluctant to reduce patronage levels to their optimum levels, a likely situation given the concern of the current state government about declining public transport useage, concentrating on improving the mix between *P* and *F* will be a highly beneficial partial response. In a second best setting where public

Figure 7.15 : Aggregated Results for Adelaide Buses : Moving To Optimal Outcomes With and Without Patronage Optimised



Note : case a = current policy settings
case b = optimal settings with minimum service standards
case c = optimal settings with closure of uneconomic services

fund raising is distortionary, however, obtaining the right balance between fares and frequency is still important, but now reducing patronage to optimal levels (and associated reductions in subsidy) become increasingly important as distortionary effects of public fund raising become more pronounced.

7.9 Summary and Conclusions

This chapter has presented a disaggregated analysis of optimal outcomes in the Adelaide bus system. The benefit of a disaggregated approach has been the ability to determine policy settings which vary across the system to reflect variations in demand and cost differences, including differences in user economies of scale, the key focus of this study.

From a general analytical perspective, the study has revealed a number of key results. In a first best world, the simple rule, found in the literature, which claims that unit subsidy is inversely related to the level of patronage, and thus that unit subsidy is greater in off-peak than peak, falls down in a peak/off-peak model. The rule holds in some corridors, but breaks down in others. The relationship between total subsidy and patronage was consistent, however, with conventional models : the greater is patronage, the greater is total subsidy, with peak subsidy exceeding off-peak subsidy . Moving to a second best world, where public fund raising is distortionary and therefore costly in efficiency terms, sees total subsidy fall more quickly in the higher demand peak period than the lower demand off-peak. So much so that, if public fund raising is distortionary enough, optimal total subsidy in the off-peak can *exceed* that in the peak.

The other set of general results relate to the gains in economic surplus of moving from non-optimal to optimal outcomes. In a first best world, the gains from reducing sub-optimal levels of patronage (and associated reductions in subsidy) are quite moderate. Greater gains can be attained by being less concerned about optimising patronage levels, and more concerned about optimising the balance between fares and frequency levels at any point in time. In a second best world, this generalisation is not as strong. With public fund raising being costly, reductions in patronage (and thus subsidy) generate increasingly greater net benefits as the cost of public finance grows at the margin. Obtaining an optimal balance between fares and frequency is still important, but now so too is attaining optimal patronage. The bigger the marginal distortionary effects of public fund raising, the greater the gains from attention to the level of patronage and therefore subsidy.

In the Adelaide context, the main results are as follows. First, a number of corridors are economically unsustainable in the off-peak, although the state government's broader social policy objectives will almost certainly require minimum services to continue to be delivered in these cases. Second, the random/planned logit choice user cost model predicted a high proportion of planned user behaviour (which is consistent with local anecdotal evidence), and subsidy results well below those of an earlier disaggregated study of Adelaide buses which was based on the simple random user behaviour model. To the extent that the logit model provides a more realistic representation of user behaviour, the subsidy results which it generates are more likely to be indicative of correct optimal subsidy levels.

Third, compared to optimal outcomes, the current Adelaide bus situation consists of excessive frequency and patronage levels, insufficient fare levels, and excessive subsidy levels. This result was found in over 80% of all corridor/time period combinations analysed. In addition, the size of these discrepancies was found to be greater in the peak than the off-peak, and of considerably greater size in a second best world compared to a first best world.

Even after the introduction of competitive tendering (and the potential cost and subsidy reductions which should follow), current policy settings will remain sub-optimal. On the positive side, however, the introduction of competitive tendering reduces the gap between current and optimal policy settings, and thus reduces the scale, and impact, of policy changes required to move to optimal outcomes. In a first best world, the changes required would be moderate in some corridors, but the required changes remain substantial throughout in a second best world.

For the Adelaide bus system in total, current subsidy levels are significantly greater than those which can be justified on user economies of scale grounds. The greater the cost of public finance, the greater the size of this discrepancy. Even after the benefits of competitive tendering have been reaped, a case can still be made for reducing the level of bus subsidy. Notwithstanding this, optimal subsidy need not necessarily drop to zero. In a first best world, optimal subsidy still exceeds approximately $A40m in all cases considered. Only is a second best world does optimal subsidy approach, and eventually drop to, zero as public finance becomes increasingly expensive to raise.

The final consideration is the off-peak robustness of Kerin's peak conclusion that *UES* subsidy will be small. It is essential to note that Kerin draws his conclusion based on analysis in a second best world, with the marginal distortionary cost of public finance the same as that considered here, $0.40 for each $1 raised. The analysis here found that in such a second best world, total peak plus off-peak subsidy, is also rather small. Therefore, within a second best world, Kerin's conclusion is robust in both the off-peak, and also on an overall basis. Two important points need to be stressed however. First, the overall size of optimal subsidy, and Kerin's conclusion, is strongly dependent on the cost of public finance. Improvements in the way public finance is raised to make it less distortionary will increase the level of optimal bus subsidy. Secondly, as discussed above, in a *relative* sense, off-peak subsidy becomes increasingly important the more distortionary is the raising of public finance.

Chapter Appendix

7A.1 User Cost Model Modifications for Application to Analysis of Adelaide Buses

Throughout this study, the logit model developed in chapter 3 has been used to predict user choice between random and planned behaviour when accessing a bus. This model continues to be used in this chapter, but with two modifications to make it more suitable for the analysis of Adelaide buses.

First, there are cases where users may follow a different choice rule. One example is "transfer" boardings to a second bus on a trip. Transfers can be split into three categories. Category one consists of cases where users actually make two separate trips within two hours, e.g. travel from home to shopping, and then a return journey home. However, the ticketing system used in Adelaide defines all bus usage within a two hour period as a single trip, and so records the return home from shopping as a transfer boarding. For this type of situation, the logit model would seem to be an appropriate choice model since the user continues to have the opportunity to catch the second bus in either a random or planned fashion.

Categories two and three consist of cases where users are making a single trip, but are forced to transfer modes during the trip. Category two consists of those cases where the transfer occurs at a major transport interchange, in which case the arrival and departure of the two modes are usually closely coordinated by the public transport operator. With close coordination, the transfer delays are likely to be similar to those experienced in the planned user model. Category three consists of those cases where the arrival and departure times of the two modes involved in the transfer cannot easily be coordinated. This is the case, for example, for transfers that occur in the CBD, and is also the case for transfers at regional shopping centres which do not contain a transport interchange, and transfers onto cross-suburban routes. In this case, modal arrival and departure times are essentially randomly related, implying that user transfer delays would tend to follow the random model.

Discussions with TransAdelaide (Willis, 1995) and the Passenger Transport Board (Wilson, 1995) revealed that : the proportion of category 1 transfers was unknown, but was unlikely to exceed 20%; of the remaining transfers, between 25% and 50% are believed to be category 3 transfers, with a figure of 40% adopted for analytical purposes. With transfers constituting (on average) about 25% of all boardings, about 12% and 8% of all boardings are category 2 and 3 transfers respectively, for which (based on the above discussion) planned and random user behaviour is respectively assumed.[28]

A second modification to the logit choice model was required because some problems were encountered when calibrating the model for Adelaide. Although data for a detailed calibration of the model was unavailable, discussions with the Passenger Transport Board (Gargett, 1994) revealed that its anecdotal experience in Adelaide is that people switch between random and planned behaviour at around H_c = 10 minutes, which is highly consistent with experience elsewhere (e.g. Seddon and Day, 1974).[29] However, the parameter values derived in appendix *B* yield H_c values above 10 mins (see Table B.6 of appendix *B*), with the model thus predicting a higher proportion of random user arrivals than local experience would suggest. Further sensitivity testing with other user cost parameter values revealed H_c results consistently above 10 mins. One reason why the model predicts high H_c values may be that the user choice model being used here assumes risk neutral users. In reality, a significant number of users are likely to be risk averse, something which would tend to make users more prone to using a timetable, i.e. acting in a planned manner, thus lowering H_c.

In order to make the H_c value in the random vs planned user choice model more consistent with local experience, a uniform scaling cost was applied to random costs in the model to achieve H_c = 10 mins.

[28] A further extension of the model would consist of recognising that users who regularly travel at the same time are more likely to act in a planned fashion. No information was available to estimate the size of this effect, and so it has been excluded from the modelling here. The proportion of random users predicted by the model used in this chapter is therefore likely to be an upper estimate.

[29] As a result, with most headways in Adelaide being above 10 mins, users are currently more likely to arrive at bus stops in a planned fashion than in a random fashion (except for transfers as discussed above).

7A.2 Derivation of First Order Condition With Respect To P_1, First Best World

The first order condition $\frac{\partial ES}{\partial P_1} = 0$ is derived as follows. From (7.8) and (7.9) :

$$\frac{\partial ES}{\partial P_1} = \frac{\partial CS_1}{\partial g_1}\frac{\partial g_1}{\partial P_1} + \frac{\partial CS_2}{\partial g_2}\frac{\partial g_2}{\partial P_1} - \frac{\partial C_p}{\partial P_1} + q_1 + P_1\frac{\partial q_1}{\partial P_1} + q_2 + P_1\frac{\partial q_2}{\partial P_1} - SJPD\frac{\partial q_2}{\partial P_1} \qquad (7A.1)$$

Now, $g_i = P_i + AC_{ui}$

but with F being the same for both concession and non-concession (regular) users, and LF being a function of q, where $q = q_1 + q_2$ (7A.2)

then $AC_{ui}(LF, q) = AC_{uj}(LF, q) = AC_u$ (7A.2a)

thus, $g_i = P_i + AC_u$

i.e. $g_1 = P_1 + AC_u$ (7A.3)

and noting (7.1) $g_2 = P_1 - SJPD + AC_u = g_1 - SJPD$ (7A.4)

Thus, noting (4.5) and (4.10), and noting that AC_o is constant,

$$\frac{\partial g_i}{\partial P_1} = \frac{\partial g_j}{\partial P_1} = \left(1 + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial P_1}\right) \qquad (7A.5)$$

$$\text{where } \frac{\partial q}{\partial P_1} = \frac{\partial q_1}{\partial P_1} + \frac{\partial q_2}{\partial P_1} \qquad (7A.6)$$

Recall from (4A.3) that $\frac{\partial CS_i}{\partial g_i} = -q_i$ (7A.7)

and that, as in chapter 4, for LF/N case 3 being considered here, $\frac{\partial C_p}{\partial q} = 0$, thus

$$\frac{\partial C_p}{\partial P_1} = \frac{\partial C_p}{\partial q}\frac{\partial q}{\partial P_1} = 0 \qquad (7A.8)$$

Then, substituting (7A.5), (7A.7) and (7A.8) into (7A.1) yields :

$$\frac{\partial ES}{\partial P_1} = -q_1\left(1 + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial P_1}\right) - q_2\left(1 + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial P_1}\right) + q_1 + P_1\frac{\partial q_1}{\partial P_1} + q_2 + P_1\frac{\partial q_2}{\partial P_1} - SJPD\frac{\partial q_2}{\partial P_1}$$

$$= P_1\frac{\partial q}{\partial P_1} - \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\frac{\partial q}{\partial P_1}q - SJPD\frac{\partial q_2}{\partial P_1} \qquad (7A.9)$$

Setting $\frac{\partial ES}{\partial P_1} = 0$ yields :

$$P_1 = q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} + SJPD\frac{\partial q_2/\partial P_1}{\partial q/\partial P_1} \qquad (7A.10)$$

Then, noting (7A.6), expressing $\frac{\partial q_i}{\partial P_1}$ as $\frac{\partial q_i}{\partial g_i}\frac{\partial g_i}{\partial P_1}$, and noting from (7A.5) that $\frac{\partial g_i}{\partial P_1} = \frac{\partial g_j}{\partial P_1}$, (7A.10) reduces to :

$$P_1 = q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} + SJPD\frac{\partial q_2/\partial g_2}{\frac{\partial q_1}{\partial g_1} + \frac{\partial q_2}{\partial g_2}} \qquad (7A.11)$$

This general expression for P_1^* reduces to a simpler expression when the exponential demand function (4.14) is used, as it is in this study. From (4.14) :

$$\frac{\partial q_i}{\partial g_i} = -\beta q_i \qquad (7A.12)$$

Thus, (7A.11) reduces to :

$$P_1 = q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} + SJPD\frac{q_2}{q} \qquad (7A.13)$$

7A.3 Alternative Optimisation Framework : Marginal Benefit and Marginal Cost

Adding and subtracting C_u ($= AC_u q = AC_u(q_1 + q_2)$) to (7.8) and rearranging yields an alternative formulation of the optimisation problem, the familiar formulation of *ES* in terms of gross benefit (*GB*) and total cost (*TC*), i.e.

$$\begin{aligned} ES &= (CS_1 + P_1 q_1 + AC_u q_1) + (CS_2 + P_2 q_2 + AC_u q_2) - (C_p + AC_u q) \\ &= (CS_1 + g_1 q_1) + (CS_2 + g_2 q_2) - TC \\ &= GB_1 + GB_2 - TC \\ &= GB - TC \qquad (7A.14) \end{aligned}$$

Optimising patronage (q) yields a familiar marginal benefit (*MB*) and marginal cost (*MSC*) framework,[30] i.e.

$$\frac{\partial ES}{\partial q} = \frac{\partial GB}{\partial q} - \frac{\partial TC}{\partial q} = 0$$

i.e.

$$MB \; (= \frac{\partial GB}{\partial q}) = MSC \; (= \frac{\partial TC}{\partial q}) \qquad (7A.15)$$

To evaluate *MB*, note that $\frac{\partial GB}{\partial P_1} = \frac{\partial GB}{\partial q}\frac{\partial q}{\partial P_1}$ and thus

$$MB = \frac{\partial GB}{\partial q} = \frac{\partial GB/\partial P_1}{\partial q/\partial P_1} \qquad (7A.16)$$

But,

$$GB = GB_1(q_1) + GB_2(q_2)$$

thus

$$\frac{\partial GB}{\partial P_1} = \frac{\partial GB_1}{\partial q_1}\frac{\partial q_1}{\partial P_1} + \frac{\partial GB_2}{\partial q_2}\frac{\partial q_2}{\partial P_1}$$

and with

$$MB_i = \frac{\partial GB_i}{\partial q_i} = g_i$$

then

$$\frac{\partial GB}{\partial P_1} = g_1\frac{\partial q_1}{\partial P_1} + g_2\frac{\partial q_2}{\partial P_1} \qquad (7A.17)$$

[30] In a *short run* setting, frequency (*F*) is given and q is optimised via adjustments in P_1 and P_2. In a *long run* setting, *F* is first optimised for any given q, with optimal prices then set to yield the optimum q.

Substituting for g_2 from (7A.4), (7A.17) becomes

$$\frac{\partial GB}{\partial P_1} = g_1\left(\frac{\partial q_1}{\partial P_1} + \frac{\partial q_2}{\partial P_1}\right) - SJPD\frac{\partial q_2}{\partial P_1}$$

and then noting (7A.6) yields :

$$\frac{\partial GB}{\partial P_1} = g_1\frac{\partial q}{\partial P_1} - SJPD\frac{\partial q_2}{\partial P_1} \qquad (7A.18)$$

Substituting (7A.18) into (7A.16) yields :

$$MB = g_1 - SJPD\frac{\partial q_2/\partial P_1}{\partial q/\partial P_1} \qquad (7A.19)$$

and the derivation from (7A.10) to (7A.11) then allows (7A.19) to reduce to :

$$MB = g_1 - SJPD\frac{\partial q_2/\partial g_2}{\left(\frac{\partial q_1}{\partial g_1} + \frac{\partial q_2}{\partial g_2}\right)} \qquad (7A.20)$$

Next evaluate *MSC*.

$$MSC = \frac{\partial TC}{\partial q} = \frac{\partial C_p}{\partial q} + \frac{\partial C_u}{\partial q} \qquad (7A.21)$$

As mentioned in section 7A.2, $\frac{\partial C_p}{\partial q}$ = 0, thus (7A.21) becomes :

$$MSC = AC_u + q\frac{\partial AC_u}{\partial q} \qquad (7A.22)$$

Then, substituting (7A.20) and (7A.22) into (7A.15), and recalling (7A.3) yields :

$$P_1 + AC_u - SJPD\frac{\partial q_2/\partial g_2}{\left(\frac{\partial q_1}{\partial g_1} + \frac{\partial q_2}{\partial g_2}\right)} = AC_u + q\frac{\partial AC_u}{\partial q}$$

thus $$P_1{}^* = q\frac{\partial AC_u}{\partial q} + SJPD\frac{\partial q_2/\partial g_2}{\left(\frac{\partial q_1}{\partial g_1} + \frac{\partial q_2}{\partial g_2}\right)}$$

which matches (7A.11) since from (4.5) and (4.10) $AC_u = u + v + \overline{AC_o}$, and thus $\frac{\partial AC_u}{\partial q} = \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}$. Thus the optimal prices derived in section 7A.2 are consistent with attaining *MB* = *MSC*, the optimal outcome.

Using the exponential demand model, (7A.12) allows (7A.20) to be further simplified to :

$$MB = g_1 - SJPD\frac{q_2}{q} \qquad (7A.23)$$

7A.4 Derivation of First Order Condition With Respect To *F*, First Best World

Total cost (*TC*) is $$TC = C_p + C_{u1} + C_{u2} \qquad (7A.24)$$

Recalling from p. 4 -19 of chapter 4 that $C_p = \frac{\partial C_p}{\partial F} F$, (7A.24) becomes :

$$TC = \frac{\partial C_p}{\partial F} F + AC_{u1}\, q_1 + AC_{u2}\, q_2 \qquad (7A.25)$$

Then, noting (7A.2) and (7A.2a), (7.25) becomes :

$$TC = \frac{\partial C_p}{\partial F} F + AC_u\, q \qquad (7A.26)$$

This expression is identical to the expression which applies in the optimisations in chapter 4, thus the optimality condition which applies there will also apply here, i.e.

$$-q\left(\frac{\partial u}{\partial F} + \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial F}\bigg|_{\bar{q}} + \frac{\partial v}{\partial F}\bigg|_{\overline{LF},\bar{q}}\right) = \frac{\partial C_p}{\partial F} \qquad (7A.27)$$

7A.5 Equivalence of Marginal Benefit and Average Generalised Cost

The average generalised cost, g_{ave}, is given by :

$$g_{ave} = P_{ave} + AC_u \qquad (7A.28)$$

$$\text{where } P_{ave} = \frac{P_1 q_1 + P_2 q_2}{q}$$

Substituting for P_2 from (7.1) and rearranging :

$$P_{ave} = P_1 \text{ - } SJPD\frac{q_2}{q}$$

Substituting into (7A.28) and noting (7A.3) :

$$g_{ave} = g_1 \text{ - } SJPD\frac{q_2}{q} \qquad (7A.29)$$

which matches with expression (7A.23) for *MB*.

7A.6 Derivation of First Order Condition With Respect To P_1, Second Best World

Repeating the first best analysis from (7A.1) to (7A.10) with the new expression (7.15) for *ES* yields :

$$P_1 = \frac{1}{(1+\kappa)} q \frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q} + SJPD\frac{\partial q_2/\partial P_1}{\partial q/\partial P_1} - \frac{\kappa}{(1+\kappa)}\frac{q}{\partial q/\partial P_1} \qquad (7A.30)$$

Two steps are required to expand this expression. First, the derivation from (7A.10) to (7A.11) indicates that :

$$\frac{\partial q_2/\partial P_1}{\partial q/\partial P_1} = \frac{\partial q_2/\partial g_2}{\frac{\partial q_1}{\partial g_1} + \frac{\partial q_2}{\partial g_2}} \qquad (7A.31)$$

Second, an expression for $\partial q/\partial P_1$ is required. Expanding (7A.6),

$$\frac{\partial q}{\partial P_1}=\frac{\partial q_1}{\partial g_1}\frac{\partial g_1}{\partial P_1}+\frac{\partial q_2}{\partial g_2}\frac{\partial g_2}{\partial P_1}$$

which, recalling from (7A.5) that $\frac{\partial g_i}{\partial P_1}=\frac{\partial g_j}{\partial P_1}$, reduces to :

$$\frac{\partial q}{\partial P_1}=\left(\frac{\partial q_1}{\partial g_1}+\frac{\partial q_2}{\partial g_2}\right)\frac{\partial g_1}{\partial P_1} \qquad (7A.32)$$

Substituting (7A.5) for $\frac{\partial g_1}{\partial P_1}$ in (7A.32) and gathering all $\frac{\partial q}{\partial P_1}$ terms to one side yields :

$$\frac{\partial q}{\partial P_1}=\left(\frac{\partial q_1}{\partial g_1}+\frac{\partial q_2}{\partial g_2}\right)\Bigg/\left(1-\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}\left(\frac{\partial q_1}{\partial g_1}+\frac{\partial q_2}{\partial g_2}\right)\right) \qquad (7A.33)$$

Substituting (7A.31) and (7A.33) into (7A.30) and rearranging and simplifying yields :

$$P_1=q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}+SJPD\frac{\partial q_2/\partial g_2}{\left(\frac{\partial q_1}{\partial g_1}+\frac{\partial q_2}{\partial g_2}\right)}-\frac{\kappa}{(1+\kappa)}\frac{q}{\left(\frac{\partial q_1}{\partial g_1}+\frac{\partial q_2}{\partial g_2}\right)} \qquad (7A.34)$$

The first two components of (7A.34) are the first best regular fare (see expression (7A.11). The third component is the second best markup.

When the exponential demand model (4.14) is used, (7A.34) simplifies. Noting (7A.12), (7A.34) reduces to :

$$P_1=q\frac{\partial v}{\partial LF}\frac{\partial LF}{\partial q}+SJPD\frac{q_2}{q}+\frac{\kappa}{(1+\kappa)\beta} \qquad (7A.35)$$

i.e. a constant markup (for any given κ and β) above first best price.

Chapter 8
SUMMARY AND CONCLUSIONS

Overall Nature Of The Study

This study has examined a particular argument for public transport subsidy, known as the *User Economies of Scale* (*UES*) subsidy argument. The bus system in metropolitan Adelaide, South Australia, was used as a case study.

The motivation for examining the issue of urban bus subsidy lies in the recent focus in Australia on microeconomic reform, an important element of which has been a reassessment of the financial performance of the public sector. All areas of public finance, including public transport subsidy, are being subjected to closer scrutiny, with urban bus subsidy therefore being a worthwhile and relevant topic for research and investigation.

The focus was limited to urban buses for two reasons. First, in Adelaide, like most other major Australian cities, bus transport is the dominant form of public transport. Second, the pending introduction of competitive tendering in the delivery of bus services in Adelaide is expected to lead to cost reductions, which will allow financial performance to be improved through subsidy reduction. This raises the questions, however, of how subsidy should be managed once the gains from competitive tendering have been reaped, and what level of bus subsidy can be justified in Adelaide on economic efficiency grounds?

Economic efficiency grounds have been interpreted in this study as the *UES* argument for subsidy (first identified by Mohring). The basis of the argument is that economies of scale exist in frequency related user (time) costs associated with bus travel which justify a subsidy under standard efficient marginal social cost pricing rules, even though the delivery of bus services tend to occur under constant returns to scale in producer costs. Other arguments which exist in favour of a subsidy have not been considered here for several reasons. The argument which has been most commonly advanced in Australia in favour of a subsidy is that it is a second best policy for managing road congestion when roads are unpriced, but this tends to be of secondary importance in low congestion cities like Adelaide. Other subsidy arguments often have little basis, and their objectives can often be achieved through more appropriate and effective policies. Third, the *UES* argument has played an important role in subsidy analysis overseas, yet has received limited attention in Australia, making it ripe for investigation in the Australian context.

The Scope Of The Research

A review of the literature identified a number of unresolved issues, or issues requiring further attention.

First, a range of perceptions exist in the literature about the size of subsidies which can be justified under the *UES* argument. Some (e.g. Jansson, 1979) have suggested that "massive" user economies of scale exist, justifying substantial optimal bus subsidies. Others (e.g. Walters, 1982) have argued that optimal subsidy is much smaller, particularly if the commonly used assumptions of monopoly service provision and conventional bus size are relaxed. In an important review, Kerin (1990, 1992) argues that optimal subsidy results are highly sensitive to the assumptions on which they are based, and that variation in assumptions accounts for the range of interpretations reported in the literature regarding the appropriate scale of optimal bus subsidy. After relaxing many commonly made assumptions, Kerin draws the tentative conclusion that only small subsidies are likely to be justified on *UES* grounds.

Whilst Kerin's critique is an important one, his tentative conclusion needs to be qualified. His conclusion partly relies on subsidy generating production inefficiencies through a leakage of the subsidy into producer costs. However, in the case of Adelaide buses, the planned introduction of competitive tendering should overcome this concern. In addition, Kerin's analysis focused on the peak period only, so an important question to ask is whether Kerin's conclusion is robust in the off-peak. This is a particularly important question to ask since the literature suggests that the *UES* effect is stronger (on a unit subsidy basis) the thinner, or less patronised, the route. There is some justification therefore for further research to test Kerin's conclusion in the off-peak.

Second, justification was found for focusing further on user cost modelling (a key input into optimal subsidy determination), and its impact on subsidy analysis. Most past subsidy studies have modelled user costs on the assumption that users arrive at bus stops in a *random* manner. Empirical evidence shows however, that whilst random user arrivals are likely to occur when service frequency is high, as frequency declines users are increasingly likely to act in a *planned* manner, using timetable information to coordinate their arrival time with bus departure times. Recent subsidy studies have started to recognise, and account for, this variation in user behaviour in their user cost models. One approach has been to use location specific empirically based user cost relationships. An alternative approach, and the one adopted here, has been to model user behaviour as the outcome of a discrete choice between random and planned behaviour, with the user choosing the behavioural mode which minimises their user cost, resulting in users switching between modes as service frequency varies.

The discrete user choice between random and planned behaviour has to date been modelled in a very simplistic deterministic manner. An implication is that the model predicts the possibility of very sudden, or knife-edge, behavioural changes across the population of users. This in turn produces two interesting new subsidy results (Jansson, K., 1993). First, multiple local optima become possible in the bus optimisation, complicating subsidy analysis for the analyst. Second, sudden increases can occur in optimal unit subsidy when behavioural mode switching occurs. This result is particularly interesting because it suggests that the conventional negative relationship reported in the literature between optimal unit subsidy and the level of patronage, a useful rule of thumb for explaining how optimal subsidy should vary between routes of different patronage levels, may break down in some circumstances. There was some justification therefore for investigating the extent to which these new results are a product of the simple deterministic choice framework used to date. The approach adopted was to reconsider Jansson's results using the more realistic framework of probabilistic choice, a framework widely used in modelling discrete choices in transport and other areas of analysis. A logit model, a commonly used probabilistic choice model, was used in this study.

A further topic related to user cost modelling which was considered worthy of investigation was the potential link between optimal subsidy and service unreliability. A relationship exists because the subsidy is strongly influenced by user cost, and service unreliability is a determinant of user cost. User economies of scale subsidy analyses to date have not investigated the nature of this relationship.

Third, given the limited attention the *UES* argument for subsidy has received in Adelaide (and Australia) to date, there was justification for estimating optimal *UES* subsidy levels for Adelaide buses. The *UES* concept has played a role in a small number of previous studies in which Adelaide was considered (Dodgson, 1985, 1986; Chalmers, 1990; Kerin, 1990), however, scope exists for extending these analyses further. Dodgson did not address the question of optimal subsidy levels, focusing instead on the optimal use of a given (existing) subsidy. In addition, the Dodgson study was undertaken at a highly aggregated (whole network daily average) level, and thus did not consider the important relationship between unit subsidy and patronage level, which would be reflected in how optimal subsidy might vary between peak and off-peak, and between routes of different demand levels.

Although Chalmers undertook a more disaggregated analysis, a serious problem with that work was the use of the simple assumption of random user arrivals at loading points, which significantly biases optimal subsidy results upwards. In addition, neither Dodgson nor Chalmers considered the efficiency costs associated with the raising of public finance to fund subsidies, a critique which applies to many other subsidy studies. Finally, although the study by Kerin used Adelaide data for model calibration, the analysis was undertaken at a general level, and did not generate optimal results for Adelaide.

The conclusion drawn from the literature review was that there was a case for undertaking a study of *UES* bus subsidy which :

- develops an improved user cost model, and investigates the link between service unreliability and subsidy;
- tests the robustness of the new results recently reported by Jansson, namely, the possible existence of multiple local optima in the bus optimisation problem, and sudden increases in optimal unit subsidy as users switch between behavioural modes;
- tests the robustness in the off-peak of the tentative peak period conclusion drawn by Kerin that only small subsidies are likely to be justified on *UES* grounds; and

- estimates optimal *UES* subsidy at a disaggregated level for Adelaide buses, and which is cognizant of the efficiency costs associated with public fund raising, and also the pending introduction of competitive tendering of bus services in Adelaide.

This study has attempted to address these shortcomings. Chapters 3 to 6 undertook a general (i.e. not purely Adelaide specific) analysis of *UES* subsidy. Chapter 3 developed an improved model of user cost, a *logit* model. Chapter 4 formally set out, solved and presented diagrammatically the first best *UES* subsidy argument, including an overview of previously used formulations. Chapter 5 assessed the impact of the introduction of the logit model on optimal unit and total subsidy, plus optimal cross-subsidy, and considered the robustness of Jansson's new results. Chapter 6 addressed the link between optimal subsidy and service unreliability. Finally, chapter 7 reported the Adelaide case study. Optimal levels of subsidy were estimated by bus corridors, for peak and off-peak time periods, before and after the introduction of competitive tendering, and with and without public finance being costly to raise. The off-peak robustness of Kerin's peak period conclusion that optimal *UES* subsidy is small was also tested.

Overview Of Findings And Contributions

An overview of the main findings and contributions of the study is presented below. Detailed explanations underlying the results (found in the body of each chapter) are not reproduced here.

User Cost Modelling (chapter 3)

The major contribution of the chapter was to extend the random vs planned user behaviour discrete choice model from its existing purely deterministic context, to one of probabilistic choice based on random utility theory. A theoretical binary *logit* choice model was adopted as the working model for predicting random vs planned choice outcomes across the population of users. The benefit of this development is that it allows switching between random and planned behaviour to occur gradually over a range of service frequency levels as frequency varies, rather than the sudden knife-edge switching at one specific frequency value which occurs in the simpler deterministic choice model.

A number of other model improvements were made. First, an expanded and more consistent set of user cost definitions was developed, including a new term, *service delay*, to describe the delay caused by the scheduling of bus services (in place of the term schedule delay, which can be more usefully reserved for other contexts), and an expanded concept of stochastic delay which encompasses delay caused by both stochastic user demand (stochastic *demand* delay) and stochastic service departures (stochastic *supply* delay). Second, it was shown that although existing models of random user cost err in their prediction of the time at which users arrive at a bus stop, this error does not bias their prediction of user cost. Third, the complex model of stochastic supply delay under planned behaviour which currently exists in the literature was replaced by a simpler fitted model which is easier to use in subsidy policy analysis.

Optimal Subsidy Formulations (chapter 4)

This chapter set out and solved the first best bus optimisation problem from which the *UES* subsidy argument arises. The main purpose of doing so was to provide a sound theoretical foundation for the subsidy analysis of subsequent chapters.

In addition, by adopting a taxonomic framework, consisting of the four possible combinations of load factor (*LF*) and bus size (*N*) being either given or variable, it was possible to better integrate and relate previously used optimisation frameworks and diagrammatic presentations. From a diagrammatic perspective, the various *LF*/*N* fixity cases can be linked through the use of average total cost envelope curves, similar to the way short run and long run average cost curves are linked in conventional cost analysis. The enveloping property of average cost curves then allows the existing diagrammatic presentations of user economies of scale subsidy currently found in the literature to be integrated into a broader diagrammatic framework.

A well known and important rule in the literature on user economies of scale subsidy, is that optimal unit subsidy declines with the level of patronage. The analysis here showed that the more "long run" the optimisation, i.e. the smaller the degree to which load factor and bus size are constrained, the less pronounced was the rate of decline in unit subsidy as patronage increases. This is an important result because, given the important role of the negative relationship between optimal unit subsidy and patronage for relating optimal subsidy for routes of different demand density, the result indicates that the strength of this relationship varies depending on the degree of constraint which applies in the optimisation problem.

Optimal bus size was found, in all cases, to increase with patronage. However, the rate of change of optimal bus size was found to be smaller when both load factor and bus size can be simultaneously optimised compared to when only bus size is optimised. The lower responsiveness occurs because total costs can only be minimised if the additional patronage is catered for by increasing both load factor and bus size concurrently, rather than just bus size.

The Impact of a Logit Model on Optimal Subsidy and Cross-Subsidy (chapter 5)

The introduction of a logit model to predict user choice between random and planned behaviour was found to have a significant impact on optimal subsidy analysis. One implication of using a logit model was that the problem of multiple local optima recently identified by Jansson largely disappears, for a number of reasons. First, switching in practice between random and planned behaviour is likely to occur more gradually than the very rapid switching rates which were found to be required to generate scope for multiple local optima to exist. Second, even if switching were to occur rapidly enough for multiple local optima to be a possibility, only a small range of patronage levels are likely to generate multiple local optima. Finally, even if, in the unlikely situation, multiple local optima actually arose, the deterministic choice model, the logit choice model, and other commonly used probabilistic choice models will yield only two local maxima, making the task of identifying the global optimum relatively manageable.

Jansson's other new result was the sudden increase in unit subsidy when behavioural mode switching occurs, thus breaking down the conventional negative relationship between unit subsidy and patronage level. When a logit model is used to predict mode choice, the conventional negative unit subsidy/patronage relationship may also break down, making it difficult to speculate *a priori* the direction of the relationship. However, the more gradual the rate at which mode switching occurs between random and planned behaviour, the smaller will be the range of patronage levels over which the relationship falters, and the less severe will be the nature of the breakdown in the relationship.

Use of the logit choice model also produces an interesting policy implication with respect to total subsidy. The growth in optimal total subsidy (in response to increases in patronage) was found to be greater whilst mode switching is occurring than it is before or after switching. Further, the more rapidly switching occurs, the greater is the relative growth of total subsidy during switching. Policy makers concerned with maintaining optimal outcomes in terms of economic efficiency may therefore be faced with the prospect of implementing quite significant jumps in total subsidy levels in potentially short periods of time as growth in patronage occurs in the presence of mode switching.

The literature has previously established that it is optimal to have cross-subsidies between bus routes when an overall breakeven constraint is imposed across a group of routes. Previous studies have found that : high patronage routes should cross-subsidise low patronage routes; cross-subsidy increases uniformly as the patronage difference between routes grows; and cross-subsidy is zero only when routes have the same patronage. The analysis here generated a number of variations to these conventional results. First, random user behaviour generates higher cross-subsidy levels than planned behaviour. Second, when a logit model is introduced, it is now also possible for low patronage routes to cross-subsidise high patronage routes. In addition, as the patronage difference between routes grows, the cross subsidy no longer necessarily increases, nor does it necessarily change gradually : the cross-subsidy can now both increase and decrease, and can change both quickly and slowly, depending on the rate of switching between random and planned behaviour. Further, an optimal cross subsidy of zero is now possible in situations where route patronages differ. On the whole, it becomes more difficult to predict *a priori* how cross subsidy will behave.

Service Unreliability and Subsidy (chapter 6)

Service unreliability was found to have a significant influence on subsidy analysis, and thus plays an important role in optimal subsidy estimation. The impact is relatively small in cases where a single mode of user behaviour occurs, either random or planned. In contrast, the impact is much more pronounced when a logit model is used to predict mode choice, due to the influence that changing unreliability has on the timing and nature of mode switching patterns. Changes in unreliability can lead to sudden changes in these switching patterns, and in turn quite substantial changes in optimal subsidy levels. Percentage impacts of 50% (and greater in some cases) were found at demand levels coinciding with behavioural mode switching.

Notwithstanding the economic validity of the unreliability/optimal subsidy relationship, some potential problems and issues arise with respect to putting the link into practice. One problem relates to the result that subsidy should be *increased* in response to an increase in service unreliability in certain circumstances, namely, over the demand levels coinciding with users switching from planned to random behaviour. However, the community may object to such a policy recommendation on the grounds that additional subsidy might be seen as a form of financial assistance being paid in a situation of worsening performance. It may be difficult therefore for policy makers to bring about such subsidy increases, even though they would be justified in doing so on economic efficiency grounds.

A broader issue is the question of the appropriate level of unreliability which should be used in setting subsidy policy in situations where scope exists for unreliability to be reduced through appropriate measures? It was argued here that, in the *medium to longer term*, it is clearly desirable for unreliability to be directly reduced through appropriate measures, and for subsidy determination to be based on the *improved* level of reliability. In the *short term*, the appropriate action is less clear. The greatest gains in economic surplus can be made by basing subsidy policy on *actual* levels of service unreliability, but to the extent that higher subsidy may act as a disincentive to implementing unreliability reducing measures, basing subsidy on a level of unreliability below actual levels may be more appropriate.

Improvements in operating and management practices by the bus operator will reduce unreliability. In addition, measures which reduce the impact of variation in user demand and road congestion, such as bus lanes and planned non-running times, also have an important role to play. However, a trade-off exists between the costs and benefits of such measures, suggesting that a positive level of unreliability is optimal. Further, with service unreliability and optimal subsidy closely related, for an overall optimum, these two should be jointly optimised.

Finally, since variability in road congestion is a determinant of service unreliability, road congestion has a role to play in the first best analysis of bus subsidy. This is in contrast to past experience where road congestion has mainly played a second best role in subsidy justification as an instrument for managing road congestion in a world of unpriced roads.

User Economies of Scale Bus Subsidy In Adelaide (chapter 7)

Two sets of important general results arose from the disaggregated analysis of Adelaide buses. The first relates to the relative size of total subsidy in peak and off-peak periods. In a first best world, the peak subsidy always exceeds the off-peak subsidy. In a second best world, public fund raising is distortionary, and therefore costly in efficiency terms. As the marginal efficiency cost of public finance increases, although optimal subsidy falls in both periods, it declines more rapidly in the peak, reducing the margin between peak and off-peak subsidy, so much so that if public fund raising is distortionary enough, optimal peak subsidy can drop below off-peak subsidy, with off-peak subsidy then *exceeding* peak subsidy.

The second set of general results relate to the gains in economic surplus of moving from non-optimal to optimal outcomes. In a first best world, the gains from reducing patronage and associated reductions in subsidy are quite moderate. Greater gains can be attained by being less concerned about optimising patronage levels, and more concerned about optimising the balance between fares and frequency levels at any point in time. In a second best world, this generalisation is not as strong. With public fund raising being costly, reductions in patronage (and thus subsidy) generate increasingly greater net benefits as the cost of finance grows at the margin. Obtaining an optimal balance between fares and frequency is still important, but now so too is attaining optimal patronage. The bigger the marginal distortionary effects of public fund raising, the greater the gains from attention to the level of patronage and therefore subsidy.

A number of important conclusions can be drawn from the Adelaide optimisations. First, a number of corridors are economically unsustainable in the off-peak. Broader state government social policy objectives will, however, almost certainly require minimum services to continue to be delivered in these corridors.

Second, the random/planned logit choice user cost model predicted a high proportion of planned user behaviour (which is consistent with local anecdotal evidence), and subsidy results well below those of an earlier disaggregated study of Adelaide buses which was based on the simple random user behaviour model. To the extent that the logit model provides a more realistic representation of user behaviour, the subsidy results which it generates are more likely to be indicative of correct optimal subsidy levels.

Third, compared to optimal outcomes, the current Adelaide bus situation consists of excessive frequency and patronage levels, insufficient fare levels, and excessive subsidy levels. This result was found in over 80% of all corridor/time period combinations analysed. In addition, the size of these discrepancies was found to be greater in the peak than the off-peak, and of considerably greater size in a second best world compared to a first best world.

Even after the introduction of competitive tendering (and the potential cost and subsidy reductions which should follow), current policy settings will remain sub-optimal. On the positive side, however, the introduction of competitive tendering reduces the gap between current and optimal policy settings, and thus reduces the scale, and impact, of policy changes required to move to optimal outcomes. In a first best world, the changes required would be moderate in some corridors, but the required changes remain substantial throughout in a second best world.

For the Adelaide bus system in total, current subsidy levels are significantly greater than those which can be justified on user economies of scale grounds. The greater the cost of public finance, the greater the size of this discrepancy. Even after the benefits of competitive tendering have been reaped, a case can still be made for reducing the level of bus subsidy. Notwithstanding this, optimal subsidy need not necessarily drop to zero. In a first best world, optimal subsidy still exceeds approximately $A40m in all cases considered. Only is a second best world does optimal subsidy approach, and eventually drop to, zero as public finance becomes increasingly expensive to raise.

The final consideration is the off-peak robustness of Kerin's peak conclusion that *UES* subsidy will be small. It is essential to note that Kerin draws his conclusion based on analysis in a second best world, with the marginal distortionary cost of public finance the same as that considered here, $0.40 for each $1 raised. The analysis here found that in such a second best world, total peak plus off-peak subsidy, is also rather small. Therefore, within a second best world, Kerin's conclusion is robust in both the off-peak, and also on an overall basis. Two important points need to be stressed however. First, the overall size of optimal subsidy, and Kerin's conclusion, is strongly dependent on the cost of public finance. Improvements in the way public finance is raised to make it less distortionary will increase the level of optimal bus subsidy. Secondly, as discussed above, in a *relative* sense, off-peak subsidy becomes increasingly important the more distortionary is the raising of public finance.

Future Research

There are a number of directions in which further research could proceed. Empirical evaluation of the parameters of user choice between random and planned behaviour would be worthwhile. Such studies would enable the logit model to be refined to reflect more accurately user behaviour for given case studies, including calibration of the logit scale parameter μ (which reflects the speed of switching between behavioural modes), and testing of alternative aggregation methods.

A second area suitable for further research is the trade-off between the benefits and costs of reducing service unreliability, and the joint optimisation of unreliability and subsidy. This would require estimation of benefit and cost functions associated with measures which reduce unreliability, incorporation of these functions into the optimisation framework outlined in this study, and joint estimation of optimal levels of service unreliability and subsidy.

Finally, it is important to remember that user economies of scale, whilst a central argument for subsidy, is not the only basis for subsidy justification. It is important to keep track of other subsidy arguments, and continue to reassess their relative importance. For example, the second best road congestion argument for subsidy was not considered here because of the relatively low level of road congestion in Adelaide. As conditions change over time, however, this argument may grow in importance, making it necessary to assess optimal subsidy on both user economies of scale and road congestion management grounds.

Appendix A
SUMMARY OF NOTATION

Subscripts (most common use of)

r	random user behaviour
p	planned user behaviour, and producer
u	user
B	backward activity rescheduling
F	frequency related, and forward activity rescheduling
L	the first bus before t_p
R	the first bus after t_p
AD	already departed
NYD	not yet departed
FB	first best
SB	second best
PK	peak period
OP	off-peak period

Acronyms, Variables and Parameters

Acronym or Parameter	*Definition*
A	constant in LF equation, and constant in SSD_p equation
ABS	Australian Bureau of Statistics
AC	average cost
AC_F	average frequency related user cost
AC_o	average non frequency related user cost
AC_p	average producer cost
AC_u	average user cost
ACT	after competitive tendering
AFC	average fixed cost
ATC	average total cost
AVC	average variable cost
b_1, b_2	constants in SDD function
BCT	before competitive tendering
BT	Bowman and Turnquist planned waiting time model
BTE	Bureau of Transport Economics
c_1, c_2	parameters in C_p/VK vs N relationship
C_p	total producer cost
C_u	total user cost
C_n	choice set of person n
CBD	central business district
CRS_{12}	cross subsidy from route 1 to route 2
CS	consumer surplus
CT	competitive tendering
CUC	choice user cost
d	proportion of total boardings in linehaul direction
du_{rp}	$u_r - u_p$
D	total user delay
DUC	deterministic user cost
e	exponential
$E(.)$	expected value of (.)
E_{max}	maximum minutes that any bus departs before its scheduled time
ES	economic surplus
f	unit frequency delay cost
$f(\varepsilon_n)$	density function of $\varepsilon_n$distribution
$f(t_a)$	density function of t_a distribution
F	service frequency
FB	first best
F_c	critical frequency
FD	frequency delay
FDC	frequency delay cost
$FDWL$	financing deadweight loss
FOC	first order condition
g	generalised cost of travel
GB	gross benefit
GS	government surplus
H	headway (mins) between consecutive services
H_c	critical headway
I	planned information cost
IC	Industry Commission
L_d	duplicated route-kms in a corridor's service areas
L_1	route-kms in a corridor's non-service areas
L_2	total route-kms in a corridor's service areas
L_N	network route-kilometres
L_r	route length (kms)
L_t	trip length (kms)
LAC	long run average cost
LF	load factor
LF_T	target load factor
$LFmult$	load factor multiple
L_{max}	maximum minutes that any bus departs after its scheduled time
LMC	long run marginal cost
LR	long run
MAL	maximum allowable load
MB	marginal benefit
MBF	marginal net benefit of frequency enhancement
MBF_e	effective MBF
MC	marginal cost
MCF	marginal producer cost of frequency enhancement
MC_p	marginal producer cost
MC_u	marginal user cost
MC_F	marginal frequency related user cost
MC_o	marginal non frequency related user cost
MLO	multiple local optima
MSC	marginal social cost
N	bus size
N_r	actual number of routes in a corridor
N_{re}	effective number of routes in a corridor
OH_y	number of hours per annum over which bus service is provided
OP	off-peak period
OP_1	interpeak part of off-peak
OP_2	evening/weekend/public holiday part of off-peak
$p(t)$	probability density function for distribution of t
$P(t)$	cumulative probability function of $p(t)$
P	fare charged, and proportion of planned users
P_1	regular (non-concession) full fare
P_2	concession fare
$P_n(r)$	probability that person n will act randomly
$P_n(p)$	probability that person n will act in a planned fashion
PK	passenger kilometres, and peak period
PTB	Passenger Transport Board
PV	parameter value
q	patronage (boardings/hour)
q_c	critical patronage level coinciding with switching between random and planned behaviour
R	proportion of random users
$ROSIS$	routes and services information system
s	unit subsidy
s_1	unit subsidy received by regular fare passengers
s_2	unit subsidy received by concession fare passengers
S	total subsidy
S_T	target subsidy constraint
SA	South Australia
SB	second best
SDD	stochastic demand delay
$SDDC$	stochastic demand delay cost
$SJPD$	social justice price discount
$SJDF$	social justice discount factor
SK	seat kilometres
SR	short run
SSD	stochastic supply delay
$SSDC$	stochastic supply delay cost
STA	State Transport Authority
t	time
t_a	time that user arrives at the bus stop
t_V	time spent travelling in-vehicle
t_L	departure time of first bus before t_p
t_p	the time at which the user would prefer a bus to depart
t_R	departure time of first bus after t_p
TC	total cost
TM	Travers Morgan Pty Ltd
u	component of frequency related user cost not influenced by LF effects
UC	user cost
UES	user economies of scale
v	component of frequency related user cost influenced by LF effects
v_V	value of in-vehicle travel time saving
v_w	unit waiting time cost
VK	vehicle-kilometres of service
wk	walk time
W	waiting time
$W(t_a)$	wait incurred by user arriving at bus stop at time t_a
α	level of potential passenger demand in exponential demand model
β	constant in exponential demand model
ε	price elasticity of demand
ε_g	generalised cost elasticity of demand
ε_n	discrete choice model stochastic disturbance term
κ	distortionary cost of a dollar of public fund raising
σ	standard deviation of bus departure times
σ_H	standard deviation of bus headway
ϕ, γ	constants in SSD_p function
μ	logit model scale parameter
μ_c	critical μ value
λ	Lagrange multiplier in optimisation with breakeven constraint over group of routes
%dir	directional split

Appendix B
PARAMETER VALUE DERIVATION

In this appendix, values for the full range of parameter values used in this study are discussed and/or derived. This study consists of two broad analytical thrusts. First, chapters 3 to 6 analyse a number of general issues in the area of user economies of scale. Second, chapter 7 presents a detailed case study of the Adelaide bus network. The data and parameter value requirements for each of these differ. The general analysis requires only indicative parameter values, whereas the Adelaide analysis requires more detailed data at a disaggregated level. These differences are highlighted in the discussion below, with a summary of the parameter values used for the two types of analysis presented in section B.4. Where possible, parameter values are based on Adelaide conditions, and/or derived from Adelaide data, including those for the general analysis. All values are in 1993 units and values.

B.1 User Parameters

B.1.1 v_V, Value of In-Vehicle Travel Time Savings (cents/min)

v_V is a key parameter because v_w and f, the two major value of time parameters in this study, are often expressed in terms of v_V. All in-vehicle time spent on public transport is assumed to be valued as non-working time. This is a commonly used assumption (e.g. Dodgson, 1985) and is based on the fact that a very low proportion of public transport users travel in working time. Dodgson reports that, for Melbourne in 1979 and Sydney in 1981 respectively, 4.8% and 4.7% of all public transport journeys were made in working time (i.e. on employers business) (Dodgson, 1985, p.24).

The field of estimation of value of time savings, which is based on the economic theory of time allocation (see discussion in section 3.2), has seen major empirical developments over a number of decades (Beesley and Kemp, 1987; Small, 1992; Truong and Hensher, 1985; Hensher, Barnard and Truong, 1988). A recent survey, covering both theoretical and empirical aspects, is contained in MVA Consultancy (1987). A wide spectrum of values of travel time saving time have been reported in the literature (BTE, 1982; Starrs, 1984) without any overall agreement about universal values.

In this study, three previous Australian studies have guided the development of an estimate of v_V. The first is survey work undertaken in Perth using the preference evaluation technique (Director-General of Transport, Western Australia, 1976). Starrs (1984), looking at Adelaide public transport, and Tisato (1990; 1991; 1992) base their analyses on this source. A value for v_V of 1.66 cents/minute was reported for Perth for 1976. The common practice for updating values of time savings into units of a later year is to adjust them in line with movements in earnings (Starrs, 1984; Hensher, 1989b),with the appropriate series for this purpose being average hourly earnings (*AHE*). To update the Perth 1976 figure, however, average weekly earnings (*AWE*) was used as a proxy since *AHE* figures for 1976 were unavailable. From 1976 to 1993, *AWE* grew by a factor of 3.41 (ABS, 1993a; 1993b), thus yielding a Perth v_V value in 1993 dollars of $v_V = 5.7$ cents/min, or $3.40 /hour.

The second study was that by Dodgson (1985). He valued non-working time savings at 25% of hourly earnings. The 1993 *AWE* for South Australia was $604.20 for 37.7 hours of work, with *AHE* thus being $16.0, and thus $v_V = \$4.00$ /hour, or 6.7 cents/min, in 1993 dollars.

The third study was the recent update by Hensher (1989b) for Australia based on surveys undertaken in Sydney. Hensher reports, for 1982, v_V for urban bus commuters as being 32% of average gross wage rate. Applying the *AHE* figures from the last paragraph, this equates to $v_V = \$5.12$ /hour, or 8.5 cents/min.

These three sources yield reasonably consistent values. Dodgson does not specify why v_V was valued at 25% of hourly earnings, and thus is somewhat arbitrary. The figure probably reflects his perception of a commonly used percentage. Hensher's Sydney values are the most recent. On the other hand, with Adelaide being more like Perth in respect to the type of city, transport system and population, the Perth figures may be more appropriate for use in Adelaide. On balance, a figure of $v_V =$ **7 cents/min** was adopted for application in this study.

B.1.2 v_w, Value of Waiting Time Savings (= Unit Waiting Time Cost) (cents/min)

The marginal value of saving a unit of time previously spent in an intermediate activity *j*, $MVTS_j$ (i.e. transferring that time unit from activity *j* to pure leisure)[1] is the difference between the marginal monetary value of the utility gained from pure leisure time, MVT_{PL}, and the marginal monetary value of the utility gained from spending time in activity *j*, MVT_j, (MVA Consultancy, 1987)[2] i.e.

$$MVTS_j = MVT_{PL} - MVT_j \tag{B.1}$$

Time spent in travel activities (e.g. in-vehicle, waiting at a bus stop, etc) usually yield disutility (Bates, 1987)[3], i.e. $MVT_j < 0$, and thus with $MVT_{PL} > 0$, from (B.1), the value of time savings ($MVTS_j$) is expected to exceed the value of pure leisure. In comparison to in-vehicle time, waiting time is considered by passengers to be more "distressing" (BTE, 1982), i.e. it yields greater disutility. As a result, the value of waiting time savings, v_w, will exceed the value of in-vehicle time savings, v_V. A factor of 2 or more is regularly used as an approximate rule of thumb (BTE, 1982; Truong and Hensher, 1985). A factor of 2 is adopted here, i.e. $v_w = 2\ v_V$, yielding $\boldsymbol{v_w} =$ **14 cents/min.**

B.1.3 *f*, Unit Planned Frequency Delay Cost (cents/min)

Several approaches have been proposed in the literature for determining a value for unit frequency delay cost (*f*) for planned user behaviour (see section 3.6.1). The first and simplest approach, is to argue that v_V acts as an upper limit for the purpose of estimating *f* (Douglas and Miller, 1974)[4]. This argument rests on the observation that a greater range of (useful) activities can be undertaken when experiencing frequency delay than when in-vehicle (Mohring, 1976; Douglas

[1] The terms intermediate activity and pure leisure were defined in section 3.2.

[2] The reader should clearly note the distinction between the two concepts : value of time savings, and the value of time. The latter is a measure of the satisfaction gained from spending time undertaking an activity. On the other hand, the former is the difference between two values of time. This distinction has sometimes been confused in the literature with the term value of time being used to refer to the value of time savings.

[3] The exception is recreational travel which generally yields utility.

[4] This will generally be the case, although as Forsyth and Hocking (1978) and Tisato (1990) argue there will be cases where this upper limit rule breaks down.

and Miller, 1974), i.e. MVT_j in (B.1) is greater for frequency delay than for in-vehicle time. As a result, the loss of utility (and thus the value of time savings (*MVTS*)) from the former is therefore smaller than from the latter. The second approach entails empirical estimation of demand functions (De Vany, 1975) although it is difficult to obtain statistically significant results (Forsyth, 1983). The third approach is to impute a value by observing the choice of plant size on the part of the service provider, a method used by Forsyth (1983) in the case of air transport.

Lack of data, and the above mentioned difficulties, suggests that the first approach provides the best guide to estimating f for this study. However, the upper limit approach provides little guidance as to where f will lie in the range 0 to v_V. For the purpose of this study, low and high estimates of f are taken to be 25% and 75% of v_V respectively. This yields, rounding to the nearest half cent, corresponding values for f of : $\boldsymbol{f_{low}}$ = **2 cents/min** and $\boldsymbol{f_{high}}$ = **5 cents/min**. For the Adelaide case study in Chapter 7, a value of $\boldsymbol{f}$ **= 3 cents/min** is used.

B.1.4 *I*, Information Cost (cents/trip)

I, the information cost per trip, is interpreted in the model as the cost incurred in acquiring bus departure times details by consulting a timetable, which would then allow him/her to access a bus in a planned fashion.

Lack of data prevented *I* from being estimated in any formal way. A nominal value of ***I*** **= 5 cents/trip** is used, based on Tisato (1990) which found that use of this figure produced H_c values which were broadly consistent with empirical evidence on the switching between random and planned behaviour.

Reflecting on the discussion of f in section B.1.3 also suggests that this value is not unreasonable. *I*, is just another example of a time cost incurred in the process of trip-making, as is waiting time, frequency delay, etc. As a result, the time spent consulting a timetable can be valued in the same way as other time costs. Since the act of consulting a timetable can be undertaken in a location of the users choosing (e.g. at home), but is unlikely to yield utility, it seems reasonable to suggest that this activity can be costed at between f (2 to 5 cents/min) and v_V (7 cents/min). I = 5 cents would then be consistent with about 1 minute spent consulting a timetable, possibly an upper, but not unreasonable estimate.

As Tisato (1990) notes, one could model *I* more rigorously to account for a number of other factors, including, the decision to obtain a timetable as an investment decision, the potential inconvenience cost experienced by the user from having to carry the timetable with them, and the role of other mechanisms for obtaining information such as using a timetable telephone information service.

B.1.5 L_t, Trip Length (km)

For the disaggregated analysis in chapter 7, the only disaggregated data available on trip lengths was average trip length for aggregations of ROSIS corridors (defined in chapter 7, section 7.2.2) : Inner Suburban (corridors 2,5,7,10,12); Middle Suburban (corridors 4,11,13); Busway (corridor 3); Outer Suburban (corridors 1,6,8,9) and Cross Suburban (corridor 14) (STA, 1992a). The respective average km trip lengths in 1992 were (STA, 1992a) 5.7, 7.2, 11.4, 13.0 and 8.0. L_t for the representative trip in a given corridor was derived by weighting these average aggregated trip lengths by the ratio of average route length in the corridor to average route length for the corresponding aggregated corridor grouping within which the corridor lies.

For general analysis in the earlier chapters, the aggregated values listed above yield an average trip length throughout the bus network of approximately L_t = 8 km, a figure also used elsewhere (Commonwealth Grants Commission, 1988).

B.1.6 AC_o, "Other" (Non-Frequency Related) User Costs (cents/trip)

AC_o comprises the sum of two time costs : walk time, and in-vehicle time. As was the case with waiting time, the value of walking time savings is conventionally valued at twice in-vehicle time, i.e. $v_{wk} = 2v_V$. Therefore :

$$AC_o \;=\; wk.2.v_V + t_V.v_V \;=\; v_V\,(2wk \;+\; t_V) \qquad \text{(B.2)}$$

where wk is walk time (mins)

t_V is in-vehicle time (mins)

$$= \frac{L_t}{speed}$$

and *speed* is bus speed

A walk time of 5 mins is indicative of the average in the Adelaide bus network (Wilson, 1994). In corridors where there are major interchanges with park and ride facilities, users avoid a walk cost, but incur the car travel costs of reaching the interchange. For simplicity it is assumed that the 5 min walk time cost is also indicative of the access costs of these park and ride users. This was considered a reasonable assumption since this cost accounts for only 15%-20% of total generalised cost. A 5 min walk time cost is therefore used to represent the access cost for all users.

For general analysis, a 22 km/hr average bus speed (STA, 1993a) and an 8 km average trip length (see section B.1.5) were adopted, resulting in t_V = 22 mins. Combining this with wk = 5 mins, (B.2) reduces to $AC_o = 32v_V$. Substituting for v_V from section B.1.1, yields $\boldsymbol{AC_o}$ **= 224 cents/trip**.

For disaggregated analysis in chapter 7, a unique AC_o value was derived for each bus corridor based on unique *speed* (STA, 1993a) and L_t (see discussion in section B.1.5) values for each corridor.

B.1.7 β, Constant in Demand Function

The value of β depends on the own price elasticity of demand, ε, where $\varepsilon = -\beta P$ (Evans, 1987). Use of a value ε = -0.3 has found common use around the world (Transport and Road Research Laboratory, 1980). Philipson and Willis (1990) argue, however, that experience has shown elasticities in Australia to be lower. Further, it is also generally accepted that peak (*PK*) elasticity is lower than off-peak (*OP*) elasticity, with Nash (1982) indicating by a factor of around 2. Combining these observations, this study adopts (following Chalmers, 1990) peak own price elasticity, ε_{PK} = -0.2, and off-peak own price elasticity, ε_{OP} = -0.35. The 1993 peak and off-peak average fares in Adelaide of 71 cents and 57 cents (STA, 1993a) then yield β values of $\boldsymbol{\beta_{PK}}$ **= 0.0028**, and $\boldsymbol{\beta_{OP}}$ **= 0.0061** respectively. These values are used in the disaggregated analysis of Adelaide buses in chapter 7.

The more general analysis in earlier chapters adopts a single β value, **β = 0.0036**, which is the weighted average of the peak and off-peak values, weighted by the proportion of users in each period.

B.1.8 b_1, b_2 and μ

These parameter values are discussed in the main text. Section 3.5.3 reports $b_1 = 1.25$, and $b_2 = 1.65$. Section 3.8.3(c) suggests a range of μ values ranging from 0.03 to 0.22 depending on how rapidly users switch between random and planned behaviour.

B.2 Adelaide Bus Network Parameters

B.2.1 Corridor Route Representation and the Effective Number of Routes (N_{re})

The ROSIS system of bus corridors was adopted as the framework for disaggregated analysis of the Adelaide bus network in chapter 7 (see section 7.2.2). As explained in chapter 7, each ROSIS corridor (see Figure 7.1 for a listing of corridors) is analysed separately. Each corridor consists of N_r routes of different length, of varying running frequency, and with some routes overlapping. The analytical approach adopted was to analyse the "representative" route for the corridor, and then factor up to yield corridor aggregates like corridor subsidy.

The existence of overlapping routes, or route duplications, i.e. routes that run along the same road, and thus service the same catchment of potential users, poses a complication for representative route definition. If the existence of route duplications are ignored, users along these route segments would be being modelled as simply facing the service frequency of the route they actually travel on. In reality, however, they also receive the benefit of being able to catch the other route(s) forming the duplication. As a result, users actually face a higher service frequency, and thus incur lower user costs. Ignoring route duplications therefore overestimates user costs.

The parts of routes which are duplicated are, in effect, acting as a single route with the combined frequency of the duplicated routes. Potential user cost distortions can, therefore, be avoided if those portions of the corridor where duplications occur are modelled as such, i.e. as a single route with the combined frequency of the duplicated routes. The net effect at the corridor level is a reduction of the number of routes to what is defined as the "effective" number of routes in the corridor, N_{re} ($< N_r$), and an increase in route frequency (i.e. $F(N_{re}) > F(N_r)$).

Care is required in determining N_{re}. The aim of adjusting N_r is to avoid distortions to user cost from route duplications. With this in mind, the adjustment adopted should only be reflecting the impact of duplication within built-up "service areas" of the corridor, i.e. where the users are.[5] The following expression for N_{re} achieves this for all corridors except Outer NorthEast (to be discussed shortly) :

$$N_{re} = 1 + (N_r - 1)\,(1 - d/100) \qquad \text{(B.3)}$$

where d is the degree (%) of route duplication in the corridor's service areas

$$= L_d/L_2 \qquad \text{(B.4)}$$

where L_d is duplicated route-kms in the corridor's service areas

L_2 is total route-kms in the corridor's service areas

$$= L - L_1 \qquad \text{(B.5)}$$

L is total route-kms in the corridor

and L_1 is route-kms in the corridor's non-service areas (i.e. those parts of routes passing through open areas where there are no users to serve, or those parts of routes which run as express or limited stop buses).

From expression (B.3), $N_{re} \rightarrow 1$ as $d \rightarrow 100$ (i.e. when there is 100% duplication, with all routes running along the same road for their entire length), $N_{re} \rightarrow N_r$ as $d \rightarrow 0$ (i.e. when there is no route duplication at all), and N_{re} varies in a linear fashion between 1 and N_r as d varies between 0 and 100. L was determined from route length data, whilst L_1 and L_d were determined by studying network and route maps, all of which were obtained from TransAdelaide.

In the case of the Outer NorthEast corridor, the above adjustment is also adequate for those parts of the corridor outside of the North East Busway, where the routes ply through suburbs serving dispersed users. Along the busway, however, users board at three interchanges (Modbury, Paradise and Klemzig). For these users, with all the corridor's bus routes funnelling down the Busway, the corridor should be modelled as a single overall route with the combined frequency of all the scheduled routes which run along the busway. Denoting b as the proportion of total corridor

[5] Both Dodgson (1985) and Chalmers (1990) also recognise the importance of removing the impact of route duplications. Unfortunately, the Chalmers study does not outline how this was done. Dodgson, working at the aggregated entire network level, removes the impact of route duplications by using unduplicated (rather than duplicated) network length in his analysis. With network length being the product of the number of routes and route length, the approach adopted here is implicitly broadly consistent with Dodgson's. The analysis here benefits, however, from the distinction between duplications in service areas vs non-service areas, something which Dodgson does not differentiate.

boardings which occur along the busway, N_{re} for the corridor is then a weighted average of $N_{re} = 1$ for the *b* per cent of users, and expression (B.3) for the other (1 - *b*) percent of users, i.e.

$$N_{re} = (b/100).1 + (1 - b/100)\,[1 + (Nr - 1)\,(1 - d/100)] \qquad \text{(B.6)}$$

B.2.2 Representative Route Data

Representative route parameter values in each corridor were derived by averaging data across all routes within the corridor. Data for some of these route parameter values, or the corridor data used in their derivation, were directly available on a corridor/time period basis from TransAdelaide. This was the case for L_r (route length)[6], B_y (corridor boardings per year), *%conc* (% concession users), VK_y (corridor veh-kms per year), *speed* (average bus speed), P_1 (average fare paid by full fare paying users), *%DR* (the % of dead running kms in the corridor), *%tr* (the % of corridor boardings which are transfers), and *SJPD* (social justice price discount = reimbursement revenue/concession boardings).

Other parameter values required derivation, the details of which are discussed below.

B.2.2.1 OH_y, *Hours of Operation per Year*

As discussed in chapter 7 (section 7.2.2(i)), buses in Adelaide operate under two networks : Day and Night networks. Table B.1 summarises the derivation of OH_y for a number of time periods split, into Day and Night network hours. Thus $OH_{y,PK}$ = 1512 hrs; $OH_{y,OP}$ = 4832 hrs. As discussed in section 7.2.2(i), derivation of number of routes and route length for *OP* requires

Table B.1 : Hours of Operation, Adelaide Buses

	Day Netw Hrs/Day	*Night Netw Hrs/Day*	*Days/ Year*	*Day Netw Hrs/Year*	*Night Netw Hrs/Year*
PK	6	-	252	1512	-
OP :					
. Interpeak	6	-	252	1512	-
. M to Th Ev	1	5	200	200	1000
. Fr Ev	4	2	52	208	104
. Sat	11	5	52	572	260
. Sun/PH	0	16	61	-	976
OP Total				2492	2340

Note : Abbreviations : Netw = Network; Ev = Evening; PH = Public Holiday.
Source : STA (1994).

[6] For general analysis, the average route length is L_r = 16 kms (Kerin, 1990).

weighting of Day and Night network data by their respective hours of operation. Based on Table B.1, these weighting factors are 0.516 and 0.484 respectively.

B.2.2.2 *%dir, % Directional Split*

Limited data was available on bus boarding directional splits by corridors or routes. For the off-peak (*OP*) a 50:50 directional split was assumed. For the peak (*PK*), the time when the directional split would be most pronounced due mainly to journey to/from work travel, the only information source was network travel demand modelling undertaken by the SA Department of Transport. Based on an investigation of a sample of routes throughout the network, an 80:20 directional split was adopted for the *PK* period.

B.2.3 Graphical Summary of Key Statistics

Figures B.1 to B.7 provide an overview of the current situation in the Adelaide bus system by presenting the variation across the system (- see Figure 7.1 in chapter 7 for a corridor listing by name -) of a number of key parameter values and variables which result from the derivations outlined above : N_r, N_{re}, L_r, q, F, H, $\%R$ and LF.

Several observations can be made from these plots. Figure B.1 compares the actual number of routes (N_r) and the effective number of routes (N_{re}). This shows the difference between N_r and N_e as being greater in *PK* (in which only the Day network operates) than *OP* (in which both Day and Night networks operate), implying that route duplications are more frequent in the Day network.

The key feature of Figure B.3, which reports average patronage per route, is the extent to which Outer NorthEast routes are patronised in the *PK* compared to other corridors, an outcome largely due to the positive impact of the NorthEast Busway which greatly reduces in-vehicle time for travel to the *CBD*. The Eastern corridor is the next best patronised, partly due to travel by private school children to and from the Eastern suburbs (Wilson, 1995). Patronage figures reported for the NorthWest, Outer North and Outer South are likely to have some negative bias due to the high proportion of feeder routes in those corridors.

Figures B.4 and B.5 report service frequency and headway experienced by the user after the effect of route duplications has been removed. The differences between *PK* and *OP* are as one would expect. It is worth noting, however, that the values reported are influenced by the way the

Figure B.1 : Adelaide Bus System - Number of Routes

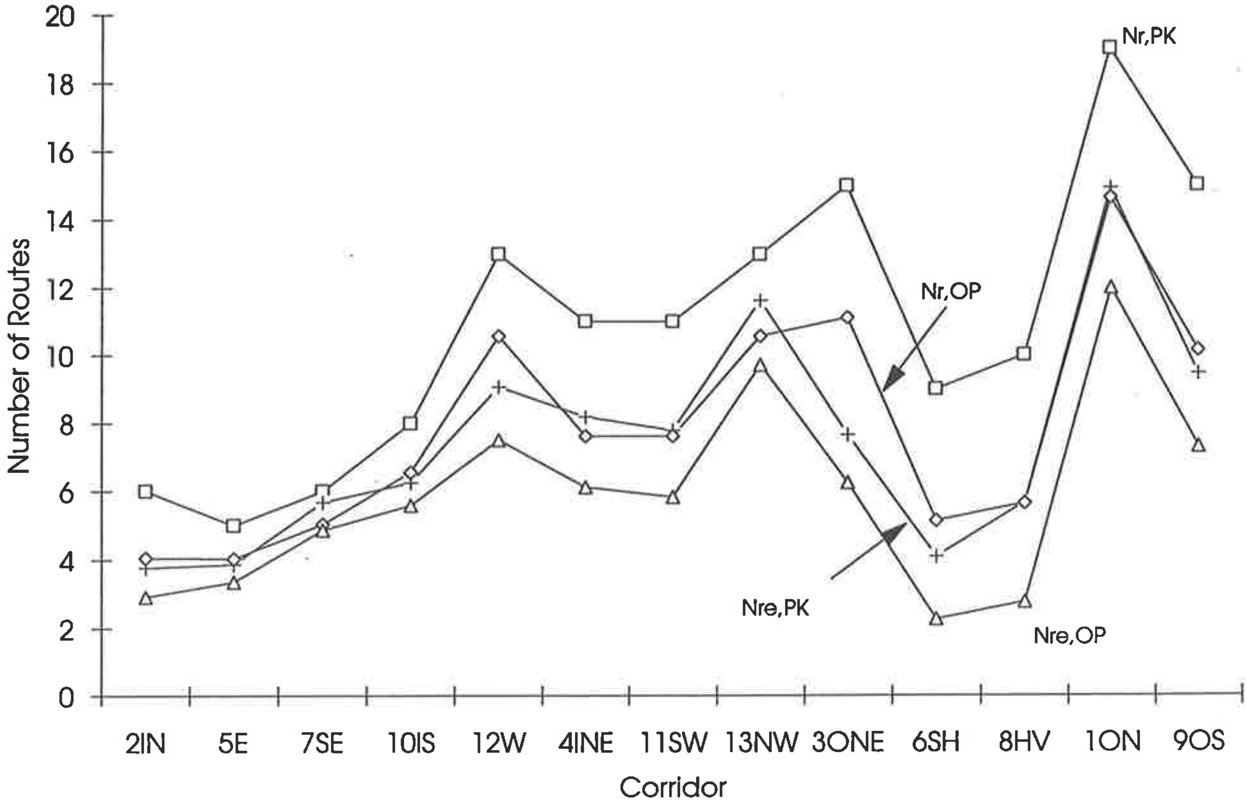


Figure B.2 : Adelaide Bus System - Average Route Length

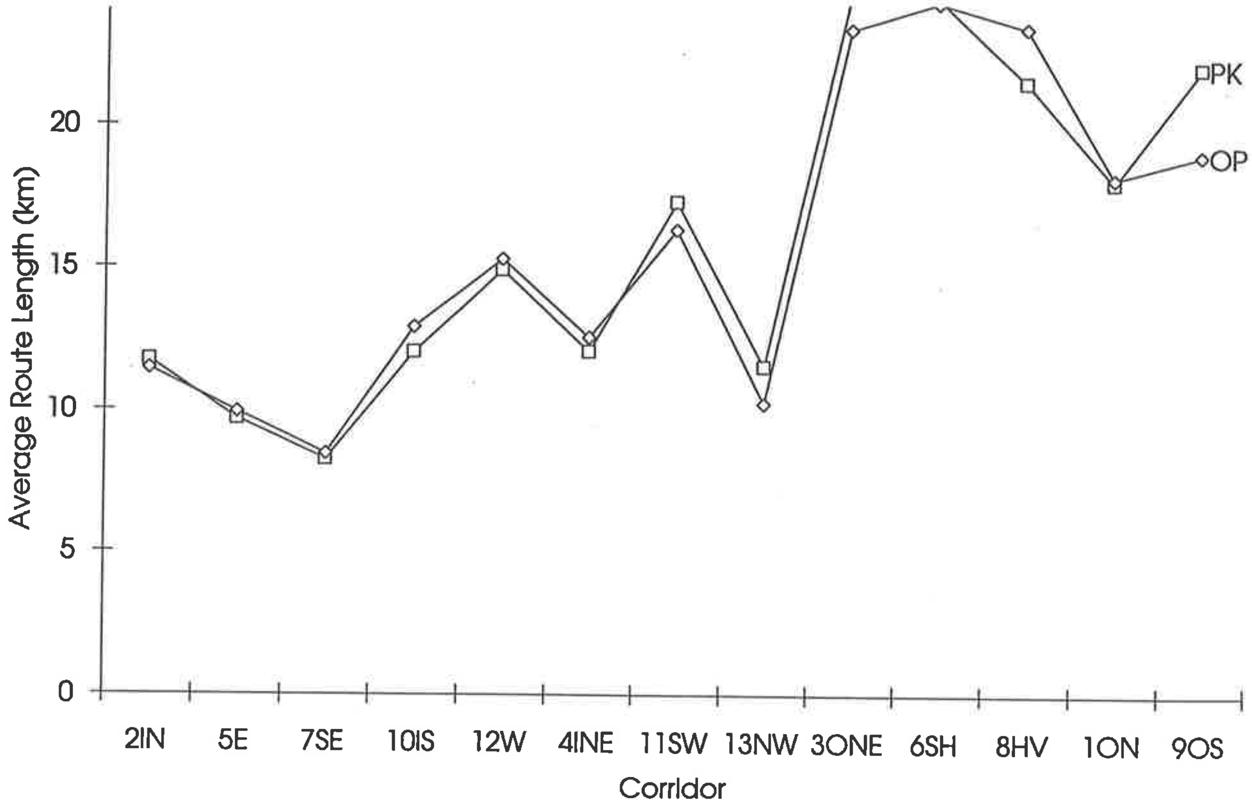


Figure B.3 : Adelaide Bus System - Current Boardings/Route/Hour

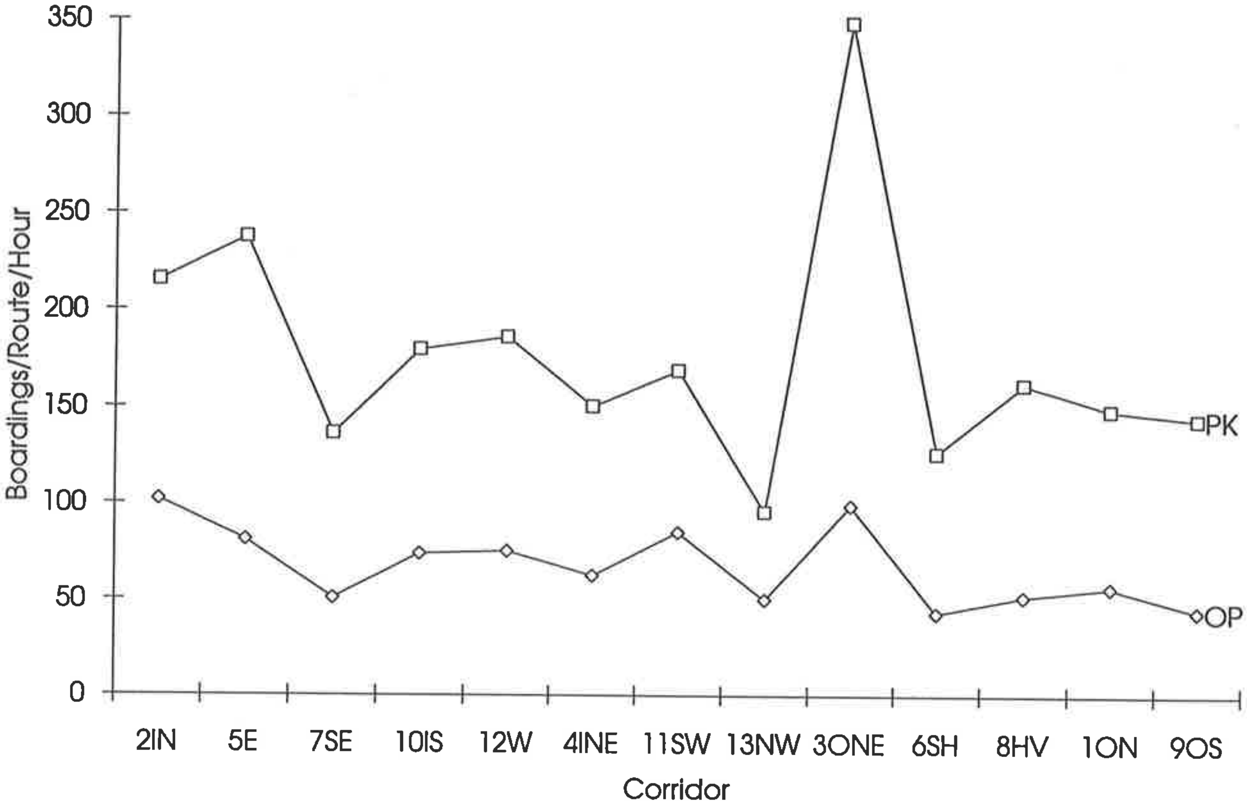


Figure B.4 : Adelaide Bus System - Current Average Frequency

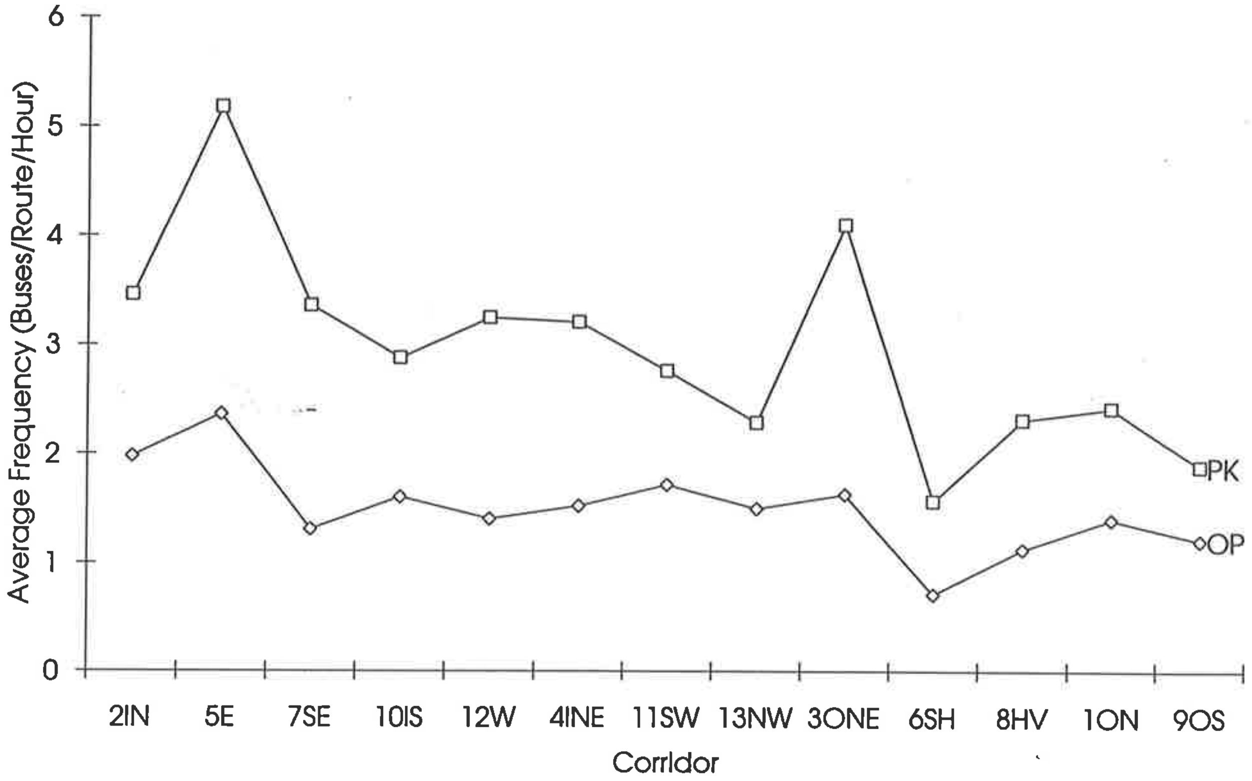


Figure B.5 : Adelaide Bus System - Current Headway

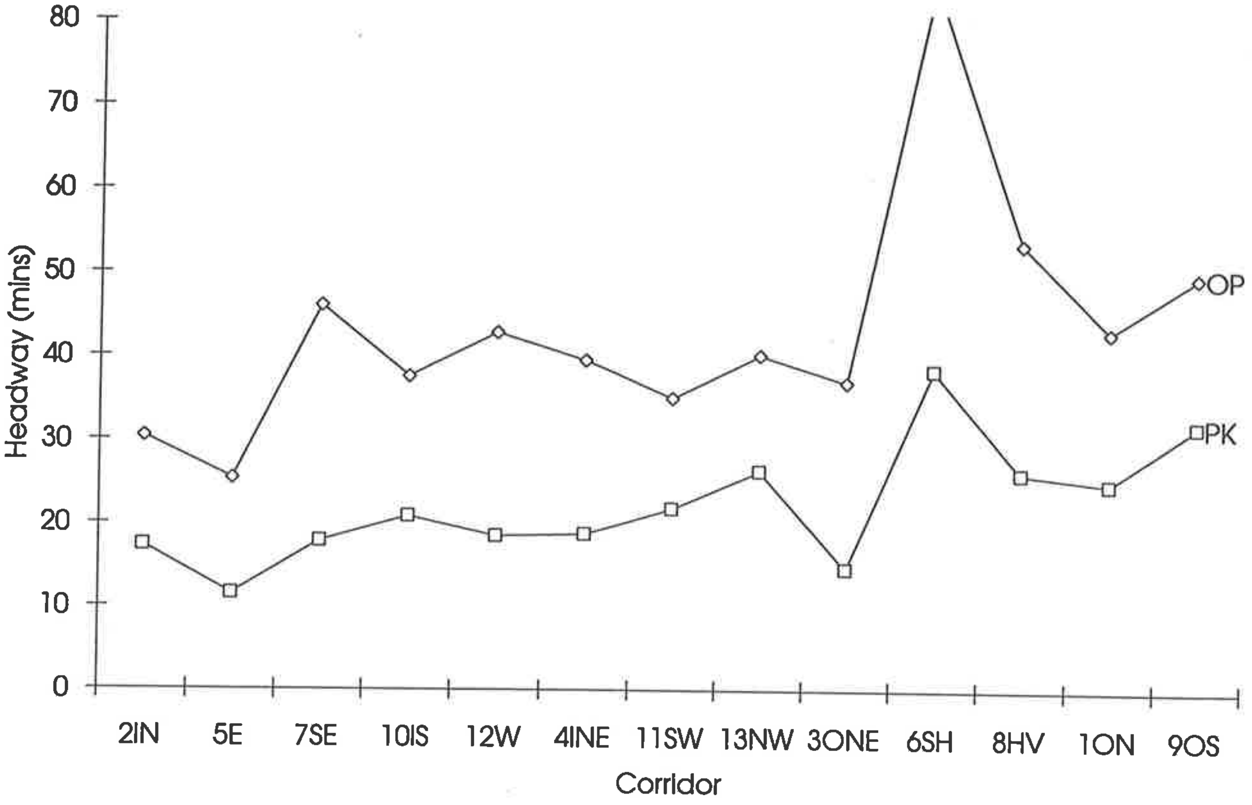


Figure B.6 : Adelaide Bus System - Current Proportion of Random Users

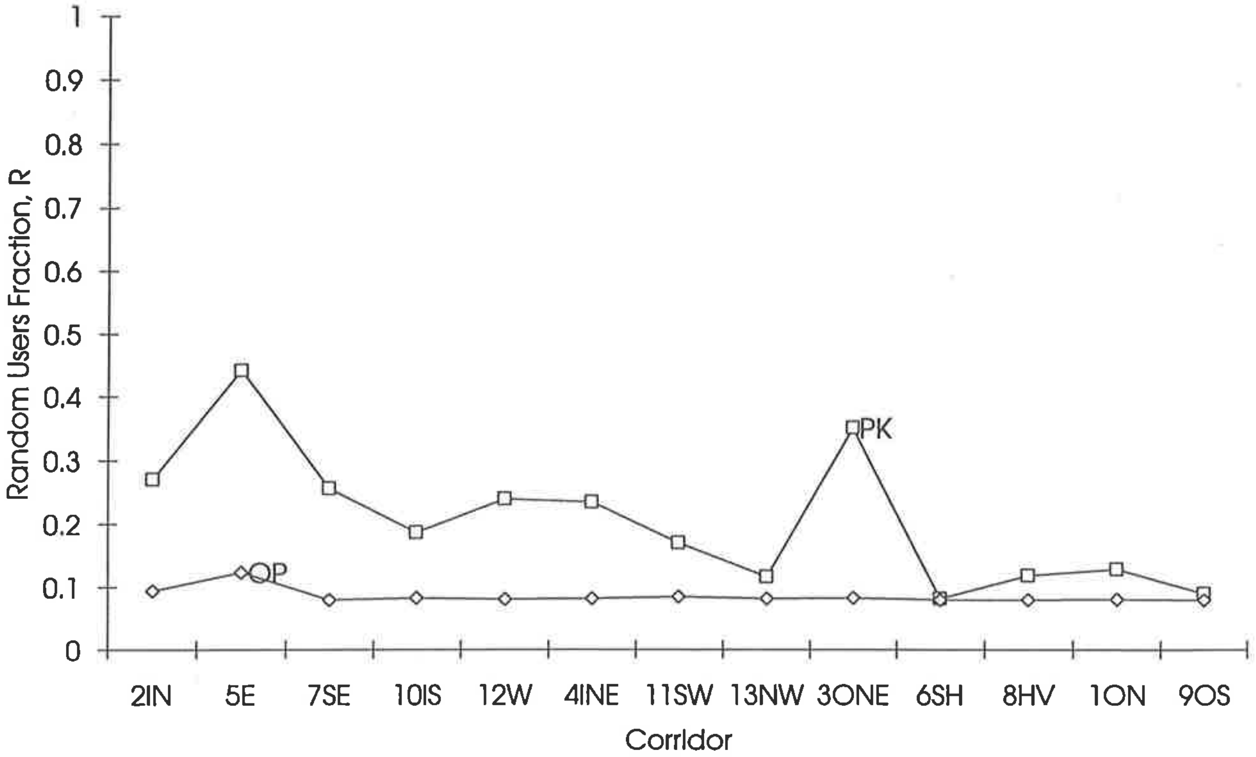
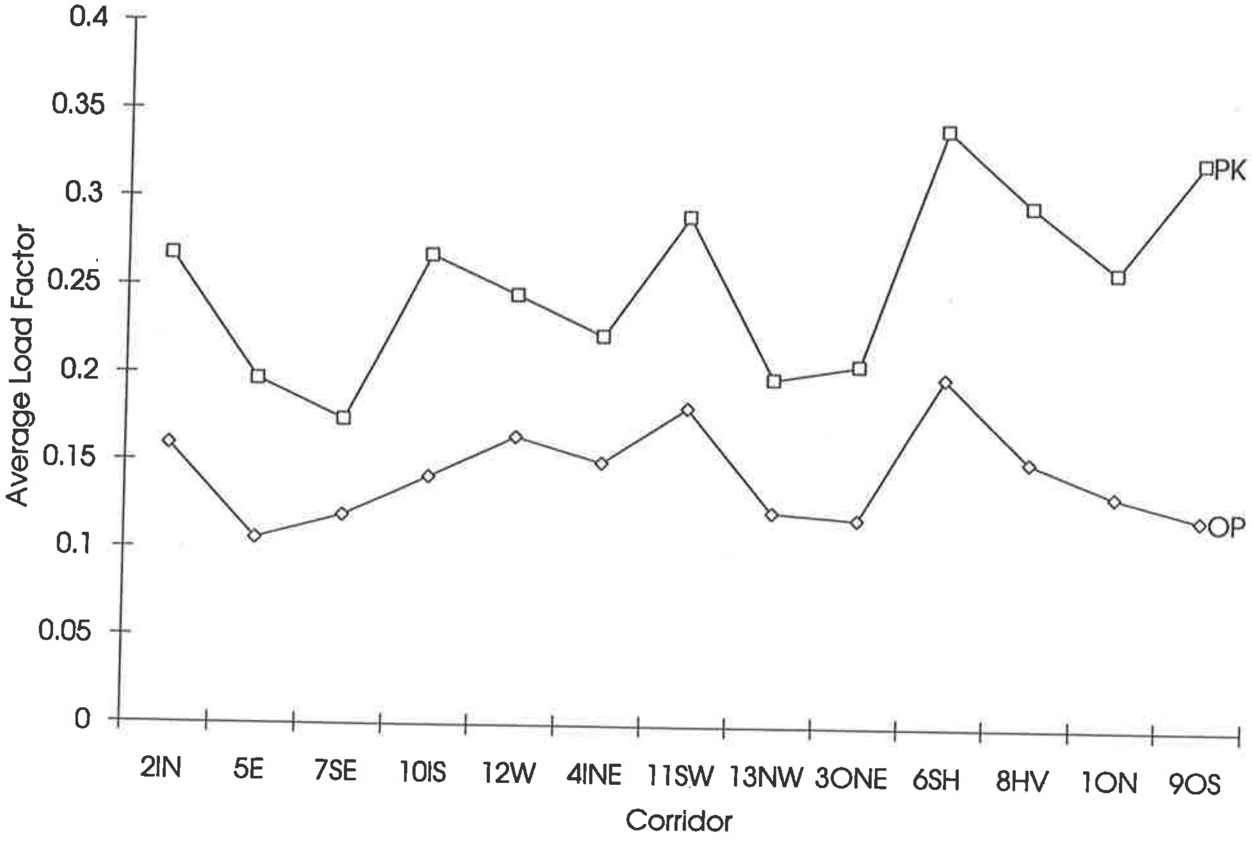


Figure B.7 : Adelaide Bus System - Current Average Load Factor



time periods have been defined. For example, the *F* values reported for the *PK*, a three hour period, are lower than the values which occur in the peak of the peak, say the hour with the busiest demand. Further, with *OP* being constructed by combining the Day and Night networks, the *F* values reported for *OP* are higher than those which occur at the periods of lowest demand (i.e. during Night network operation).

It should also be noted that the *F* values for corridors 9 and 10 (Outer South and Outer North are lower than what was expected a priori. This is largely due to the significant proportion of (lower *F*) feeder routes in these corridors. Corridor 13, NorthWest, where over 50% of routes are feeders, also has a markedly lower frequency. Corridor 6, Stirling Hills serves some low density areas with corresponding lower frequencies.

Figure B.6 reports the proportion of random users across the corridors. With the high *H* values in *OP*, almost all users act in a planned manner. In *PK*, with nearly all existing *H* values being above 10 mins (= H_c, see appendix to chapter 7), planned behaviour still dominates. Random behaviour is most frequent in the corridors with significantly higher *F* values, namely corridors 3 and 5 (Outer NorthEast and Eastern).

B.3 Bus Parameters

B.3.1 σ, Service Unreliability

A comprehensive survey of unreliability levels for a wide range of cities around the world was beyond the scope of this study. Instead, data on service unreliability was found for two US studies, and, in addition, the Adelaide situation was assessed.

The first US study which reports unreliability levels is Bowman and Turnquist (1981) for Chicago and Evanston, Illinois. In modelling waiting times (see discussion in section 3.6.3), the study collected data on unreliability, which it treated as an independent variable. Data obtained from a collection of bus stops yielded values for σ, the standard deviation of bus departure times from scheduled times (an indicator of the level of unreliability), in the range 0.5 to 1 .

The second US study is the recent analysis of Portland, Oregon by Strathman and Hopper (1993) which empirically investigates the factors which contribute to service unreliability using comprehensive network-wide data. Strathman and Hopper report a histogram summarising the frequency of occurrence of time deviations of actual departure times from scheduled service departure times. σ was calculated as the weighted average deviation in the histogram. The standard expression for sample variance (σ^2, the square of the standard deviation, σ) is (Kreysig, 1972) :

$$\sigma^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

$$\text{where } \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

is the mean of the n x_j values

In the case of the Strathman and Hopper histogram, this reduces to :

$$\sigma^2 = \frac{1}{n-1}\sum_{j=E\max}^{L\max} x_j (j - \bar{j})^2 \qquad \text{(B.7)}$$

$$\text{where } \bar{j} = \frac{1}{n}\sum_{j=E\max}^{L\max} jx_j \qquad \text{(B.8)}$$

$$\text{where } n = 1\Big/\sum_{E\max}^{L\max} x_j \qquad \text{(B.9)}$$

where j indicates the size of actual departure time deviation from scheduled time

E_{max} is the biggest time deviation (mins) for early buses

L_{max} is the biggest time deviation (mins) for late buses

x_j is the number of occurrences of time deviation j

and $\bar{j}$ is the mean time deviation (mins).

Strathman and Hopper reported values of E_{max} and L_{max} of -7 and +27 (with - and + designating early and late buses). The resulting σ value lies in the range 2.5 to 5 minutes. A range was calculated, rather than a single value, due to the fact that the heights of some of the histogram bars in the tails of the histogram were difficult to discern from the published article. Accordingly, lower and upper estimates of these bars were postulated, thus yielding a range estimate.

Finally, expressions B.7 to B.9 were also used to determine σ for the Adelaide bus network. Data on time deviations between actual and scheduled bus departure times was obtained from the STA (1993b) on a comprehensive network-wide basis (aggregated across all bus depots), similar to

Strathman and Hopper's Portland data. Figure B.8 plots the frequency distribution of these time deviations (in minutes). Once again, a negative deviation refers to buses arriving early, whilst a positive deviation infers late buses. The biggest deviations were 12 minutes for both early and late buses (i.e. E_{max} = -12, and L_{max} = +12). The distribution fairly closely approximates a normal distribution but has a slight amount of skewness favouring late arrivals. The resulting σ value was 4.2 minutes.

Unfortunately, however, the Adelaide data is likely to be deficient in the following sense. The data is collected by the bus driver pressing a button when the bus departs each stop. Seaman (1994) advised that some drivers regularly forget to press the button whilst at the stop, but then do so when the bus is between stops, distorting the data. Alternative STA manual counts suggest the spread of the distribution, and thus σ, is likely to be smaller, with a figure of σ = 2.5 mins providing an indicative estimate for Adelaide buses.

On the basis of the above discussion, a σ range of **σ = 1 to 4 mins** was adopted for general analysis, especially for chapter 6 which considers the link between service unreliability and subsidy. For the analysis of Adelaide buses in chapter 7, a value of **σ = 2.5 mins** was adopted.

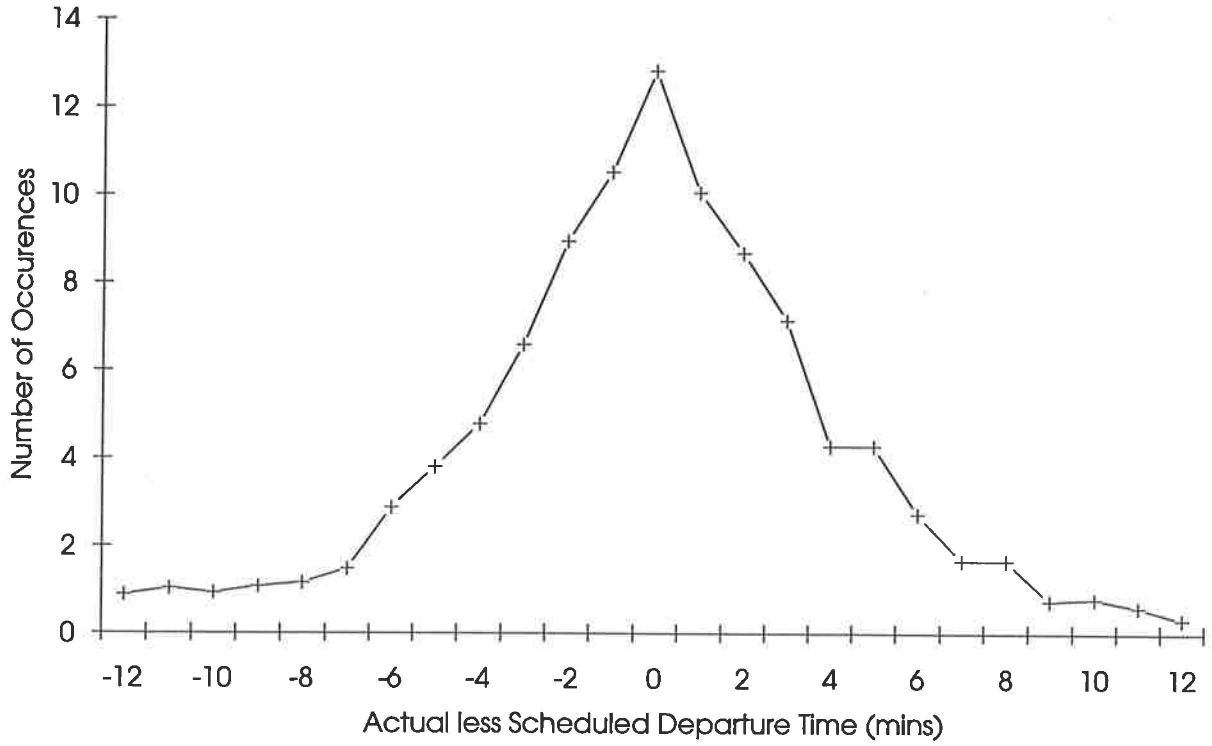
B.3.2 <u>*N*, Bus Size</u>

The term bus size (*N*) denotes the passenger carrying capacity of a bus, or the Maximum Allowable Load (*MAL)*. This consists of the number of seats plus the maximum number of standing passengers allowed, with the latter determined by the STA in negotiation with the relevant union. *MAL*, which is marginally below the physical carrying capacity of the bus, is reported in Table B.2 for four bus sizes considered by Kerin (1990) in his study : Mini, Midi, Rigid, Artic.

Table B.2 : Bus Size (*N*), i.e. Maximum Allowable Load

	Mini	*Midi*	*Rigid*	*Artic*
. seating	15	28	46	68
. standing	5	10	26	35
Total	20	38	72	103

Figure B.8 : Adelaide Bus System - Bus Departure Time Distribution



Note : - denotes an early bus, + denotes a late bus

The fleet of buses operated in Adelaide by the STA consists of a mix of Rigid and Artic buses[7], the majority being Rigids. The sizes vary to a limited extent even within a bus size category, with the sizes reported in Table B.2 being average sizes for each class. Based on fleet size, fleet mix and seating numbers reported in STA (1992b), there are, on an average weighted basis, 49.6 seats/bus. From Table B.2, the corresponding maximum number of standing passengers allowed per average bus is approximately 28 passengers, yielding a **current average bus size of 78 passengers per Adelaide bus.**

B.3.3 $C_p/VK\ (= \partial C_p/\partial VK)$[8]

For the Adelaide analysis in chapter 7, disaggregated costs are required by ROSIS corridor and by time period. STA (1993a) reports C_p/VK by corridor and time period, with appropriate costing (STA, 1994) so that all bus costs are assigned to the peak only (making them suitable for use in the peak/off-peak analysis of chapter 7). Inspection of this data by corridor revealed several distinct features. First, the C_p/VK value for corridor 3, Outer NorthEast, was substantially higher in the *PK* period than all other corridors. This is due to the fact that the capital cost of the NorthEast Busway, a guided busway with exclusive right of way, had been included in the determination of C_p/VK for this corridor. To enable sensible comparisons between corridor 3 and the remaining corridors, the cost of the busway has subsequently been removed from C_p/VK determination. Second, other than the busway capital cost discrepancy, there was reasonable uniformity in C_p/VK across corridors within any given time period. Consequently, within each time period, a common C_p/VK value was used across all corridors. The resulting disaggregated C_p/VK values used in chapter 7 are given in Table B.3.

Table B.3 : Producer Cost per Vehicle-Kilometre, C_p/VK (\$/veh-km)

Time Period	C_p/VK *: operating cost*	C_p/VK *: bus cost*	C_p/VK *: total*
PK	3.01	1.25	4.26
OP	2.71	-	2.71

[7] There is also a very small number of midis currently operating.

[8] As noted in section 4.2.2 of chapter 4, with constant returns with respect to veh-kms, $C_p/VK = \partial C_p/\partial VK$.

Table B.3 reveals a marked difference in C_p/VK (total) values between time periods. Two factors contribute to this. First, as discussed earlier, bus capital costs are treated exclusively as peak costs. Second, operating costs are greater in the *PK*. Operating costs reflect both unit labour costs and operating conditions. Penalty rates (for example an early start penalty), plus slower average bus speed, act to make *PK* operating costs per veh-km higher than in *OP*.

A 20% reduction in *operating* costs was applied to the figures in Table B.3 to simulate the introduction of competitive tendering (see discussion in section 7.7.1).

In the general analysis in the earlier chapters prior to chapter 7, only an indicative figure of C_p/VK is required, with the daily average for Adelaide buses being used (STA, 1993a). In addition, a relationship is also required between C_p/VK and bus size, *N*. A relationship in 1993 dollars was derived from the work of Kerin (1990). Kerin investigated the bus costs for the four bus sizes discussed in section B.3.2, with cost data for Rigid and Artic buses coming from STA internal data sources, whilst costs for Mini and Midi buses were based on experience of private operators in Perth and experience overseas. Regressing the C_p/VK values for the four bus sizes (R^2 = 0.98), yielded the following relationship (in 1993 dollars) :

$$\frac{C_p}{VK}\left(=\frac{\partial C_p}{\partial VK}\right) = c_1 + c_2 N \qquad \text{(B.10)}$$

where $\boldsymbol{c_1}$ **= $ 2.54** and $\boldsymbol{c_2}$**= $ 0.0157**

B.4 Summary of Parameter Values

Tables B.4 and B.5 provide a summary of parameter values. Table B.4 gives the parameter values used in general analysis prior to the Adelaide case study. Many of these are also used in the Adelaide case study, with Table B.5 summarising the additional parameter used in the case study. Table B.6 provides three parameter value sets of parameters which are used in chapters 3 and 5 for sensitivity testing.

Table B.4 : Summary of Parameter Values for General Analysis

Parameter	*Value*	*Units*	*Section Derived*
v_V	7	cents/min	B.1.1
v_w	14	cents/min	B.1.2
f	2 - 5	cents/min	B.1.3
I	5	cents/trip	B.1.4
L_t	8	kms	B.1.5
AC_o	224	cents	B.1.6
β	0.0036	boardings/hour	B.1.7
b_1	1.25	unitless	B.1.8
b_2	1.65	unitless	B.1.8
μ	0.03-0.22	unitless	B.1.8
σ	1 - 4	mins	B.3.1
N	see Table B.2	passengers	B.3.2
c_1	2.54	$	B.3.3
c_2	0.0157	$	B.3.3
L_r	16	kms	B.2.2

Table B.5 : Additional Parameters Used in Adelaide Case Study

Parameter	*Value*	*Units*	*Section Derived*
f	3	cents/min	B.1.3
β_{PK}	0.0028	boardings/hour	B.1.7
β_{OP}	0.0061	boardings/hour	B.1.7
$OH_{y,PK}$	1512	hours per annum	B.2.2.1
$OH_{y,OP}$	4832	hours per annum	B.2.2.1
$\%dir_{PK}$	80 : 20	% : %	B.2.2.2
$\%dir_{OP}$	50 : 50	% : %	B.2.2.2
σ	2.5	mins	B.3.1
N	78	passengers	B.3.2
C_p/VK	see Table B.3	$/veh-km	B.3.3

Table B.6 : Parameter Value Sets

Parameter	*PV Set 1*	*PV Set 2*	*PV Set 3*
v_w	14.0	14.0	14.0
I	5	5	5
f	5.0	3.0	2.0
σ	3.0	2.0	1.5
H_C	20.4	14.5	11.8

REFERENCES

ABS, 1993a, *Average Weekly Earnings : States and Australia*, Catalogue no. 6302.0.

ABS, 1993b, *Distribution and Composition of Employee Earnings and Hours - Australia*, Catalogue no. 6305.0

Adebisi, O., 1986, A mathematical model for headway variance of fixed-route buses, *Transportation Research B,* 20B(1), 59-70.

Advertiser, 1994, November 11, March date set for public bus tenders.

Akcelik, R., 1978, A new look at the Davidson's travel time function, *Traffic Engineering and Control*, 19 (October), 459-463.

Akerlof, G.A. and Yellen, J.L., 1985, Can small deviations from rationality make significant differences to economic equilibria?, *American Economic Review*, 75(4), 708-720.

Allen, R.R., 1987, The future of urban transport subsidies, *Proceedings of the Australian Transport Research Forum*, 457-472.

Amos, P. and Starrs, M., 1984, Public transport subsidies in Adelaide, *Proceedings of Australian Transport Research Forum*, 595-611.

Australian Urban and Regional Development Review, 1994, *Urban public transport futures*, Workshop Papers No. 4, prepared by the National Capital Planning Authority.

Australian Urban and Regional Development Review, 1995, *Timetabling for tomorrow : An agenda for public transport in Australia*, Strategy Paper No. 2, Department of Housing and Regional Development.

Bates, J.J., 1987, Measuring travel time values with a discrete choice model : A note, *Economic Journal*, 97, June, 493-498.

Baumol, W.J. and Vinod, H.D., 1970, An inventory theoretical model of freight transport demand, *Management Science*, 16, March, 413-421.

Becker, G.S., 1965, A theory of the allocation of time, *Economic Journal*, 75, 493-517.

Beesley, M.E. and Glaister, S., 1985a, Deregulating the bus industry in Britain : A response, *Transport Reviews*, 5(2), 133.

Beesley, M.E. and Glaister, S., 1985b, Deregulating the bus industry in Britain : A reply, *Transport Reviews*, 5(3), 223-224.

Beesley, M.E. and Kemp, M.A, 1987, Urban transportation, in Mills, E.S. (ed), *Handbook of Regional and Urban Economics, Volume II : Urban Economics*, Amsterdam : North Holland, 1023-1052.

Ben-Akiva, M. and Lerman, S.R., 1985, *Discrete choice analysis,* MIT Press, Cambridge.

Bly, P. and Oldfield, R., 1987, An analytic assessment of subsidies to bus services, in Glaister, S. (ed), *Transport Subsidy*, Policy Journals, 40-51.

Bly, P., Webster, F. and Pounds, S., 1980, *Subsidisation of urban public transport*, Great Britain Department of Transport, Supplementary Report S41.

Bowman, L.A. and M.A. Turnquist, 1981, Service frequency, schedule reliability and passenger wait times at transit stops, *Transportation Research*, 15A(6), 465-71.

Bray, D., 1995, *Transport strategy for Adelaide : The past and present*, Report to the Minister for Transport SA.

Brown, J.B. and Sibley, D.S., 1986, *The theory of public utility pricing*, Cambridge University Press.

Browning, E.K., 1976, The marginal cost of public funds, *Journal of Political Economy*, 84, 283-298.

Bruzelius, N., 1979, *The value of travel time*, Croom Helm, London.

BTE, 1982, The value of travel time savings in public sector evaluations, *Occasional Paper 51*.

Cervero, R. and Wachs, M., 1982, An answer to the transit crisis : The case for distance-based fares, *Journal of Contemporary Studies*, 5(2), Spring, 59-70.

Chalmers, M., 1990, *Efficiency in Public Transport Pricing, Service Levels and Subsidy Levels*, unpublished Honours Economics Thesis, University of Adelaide.

Commission of Audit SA, 1994, *Charting the way forward : Improving public sector performance*.

Commonwealth Grants Commission, 1988, *Report on general revenue grant relativities*, vol. II.

Della-Torre, K. 1994, *Community benefit study*, State Transport Authority.

DeNeufville, R. and Mira, L.J., 1974, Optimal pricing policies for air transport networks, *Transportation Research*, 8, August, 181-192.

DeSerpa, A.C., 1971, A theory of the economics of time, *Economic Journal*, 81, 828-846.

DeVany, A., 1975, The effect of price and entry regulation on airline output, capacity and efficiency, *Bell Journal of Economics*, 6, Spring, 327-345.

Director-General of Transport, Western Australia, 1976, *Transport policies for central Perth : Review and formulation.*

Dodgson, J.S, 1985, Benefits of urban public transport subsidies in Australia, *Bureau of Transport Economics, Occasional Paper 71.*

Dodgson, J.S., 1986, Benefits of changes in urban public transport subsidies in the major Australian cities, *Economic Record*, 62, 224-235.

Dodgson, J.S. and Topham, N., 1987, The shadow price of public funds : A survey", in Glaister, S. (ed), *Transport Subsidy*, Policy Journals, 114-119.

Domberger, S., 1993, Privatisation : What does the British experience reveal ?, *Economic Papers*, 12(2), 58-68.

Douglas, G.W. and Miller III, J.C., 1974, *Economic regulation of domestic air transport : Theory and policy*, The Brookings Institution, Washington DC.

Duldig, P. and Gaudry, B., 1993, The equity incidence of the State Transport Authority subsidy in South Australia: An update, *Papers of the Australasian Transport Research Forum,* 18(2), 895-914.

Else, P.K., 1985, Optimal pricing and subsidies for scheduled transport services, *Journal of Transport Economics and Policy*, 19(3), September, 263-79.

Evans, A.W., 1987, A theoretical comparison of competition with other economic regimes for bus services, *Journal of Transport Economics and Policy*, 21, January, 7-36.

Evans, A.W., 1990, Are urban bus services natural monopolies, *Flinders University, Research Paper No. 90-19.*

Fielding, G.J., 1988, *Public transport in metropolitan Adelaide in the 1990's : The Fielding Report*, for the Minister of Transport, South Australia.

Findlay, C.C., 1983, Optimal air fares and flight frequency and market results, *Journal of Transport Economics and Policy*, 17(1), January, 49-66.

Findlay, C.C. and Jones, R.L., 1982, The marginal cost of Australian income taxation, *Economic Record* 58, 253-262.

Forsyth, P.J., 1983, The cost of convenience in transport : The case of airlines, *mimeo*.

Forsyth, P.J. (ed), 1992, *Microeconomic Reform in Australia, Sydney*, Allen and Unwin.

Forsyth, P.J. and R.D. Hocking, 1978, Optimal air fares, service quality and congestion, supplement to *Forum Papers of the 4th Annual General Meeting of the Australian Transport Research Forum*.

Freebairn, J., 1995, Reconsidering the marginal welfare cost of taxation, *Economic Record*, forthcoming.

Gargett, A., 1994, Passenger Transport Board, personal communication.

Glaister, S., 1981, *Fundamentals of transport economics*, Basil Blackwell : Oxford.

Glaister, S., 1982, *Urban public transport subsidies : An economic assessment of value for money*, Technical Report, for the UK Dept of Transport.

Glaister, S., 1984, The allocation of urban public transport subsidy, in J. LeGrand and R. Robinson (eds), *Privatisation and the Welfare State*, Allen & Unwin.

Glaister, S., 1986, Bus deregulation, competition and vehicle size, *Journal of Transport Economics and Policy*, 20(2), May, 217-244.

Glaister, S., 1987, Allocation of urban public transport subsidy, in Glaister, S. (ed.), *Transport Subsidy*, Policy Journals, 27-39.

Glaister, S. and D. Lewis, 1978, An integrated fares policy for transport in London, *Journal of Public Economics*, 9, 341-55.

Gwilliam, K.M., Nash, C.A. and Mackie, P.J., 1985, Deregulating the bus industry in Britain - (B) The case against, *Transport Reviews*, 5(2), 105-132.

Henderson, J.V., 1981, The economics of staggered work hours, *Journal of Urban Economics*, 9, 349-364.

Hensher, D.A., 1986, Simultaneous estimation of hierarchical logit mode choice models, *Transport Research Group, Macquarie University, Working Paper No 24*.

Hensher, D.A., 1989a, The externality benefits (road congestion) of public transport subsidies, Draft final report, *Transport Research Centre, Macquarie University*.

Hensher, D.A., 1989b, Behavioural and resource values of travel time savings : A bicentennial update, *Australian Road Research*, 19(3), 223-229.

Hensher, D.A., 1993, Local urban bus services : Natural monopoly and benchmark contestability, *Institute of Transport Studies, Working Paper ITS-WP-93-13,* University of Sydney.

Hensher, D.A., Barnard, P.O. and Truong, T.P., 1988, The role of stated preference methods in studies of travel choice, *Journal of Transport Economics and Policy*, 22(1), 45-58.

Hensher, D.A. and Bullock, R.G., 1979, Price elasticity of commuter mode choice : Effect of a 20% fare reduction, *Transportation Research A*, 13A, 193-202.

Hensher, D.A., and Daniels, R., 1994, Performance measurement in the urban bus sector, Appendix E in IC, 1994, *Urban Transport, Volume 2.*

Holroyd, E.M. and D.A. Scraggs, 1966, Waiting times for buses in Central London, *Traffic Engineering and Control*, July, 158-160.

IC, 1994, *Urban Transport.*

Jackson, R., 1975, Optimal subsidies for public transit, *Journal of Transport Economics and Policy*, January, 3-15.

Jansson, J.O., 1979, Marginal cost pricing of scheduled transport services, *Journal of Transport Economics and Policy*, 13(3), 268-94.

Jansson, J.O., 1980, A simple bus line model for optimisation of service frequency and bus size, *Journal of Transport Economics and Policy*, 14(1), 53-80.

Jansson, K., 1993, Optimal public transport price and service frequency, *Journal of Transport Economics and Policy*, 27, January, 33-50.

Jolliffe, J.K., and Hutchinson, T.P., 1975, A behavioural explanation of the association between bus and passenger arrivals at a bus stop, *Transportation Science*, 9, 248-282.

Kain, P., 1981, Urban transport crisis : A study of Adelaide bus operations in transition 1967-81, BEc (Hons) thesis, University of Adelaide.

Kerin, P.D., 1987, Why subsidise state transport authorities?, *The Australian Quarterly*, Autumn, 60-72.

Kerin, P.D., 1989, On the marginal capital costs of peak and off-peak transit services, *Journal of the Transportation Research Forum*, 24(2), 349-355.

Kerin, P.D., 1990, *Efficient transit management strategies and public policies : Radial commuter arteries*, unpublished PhD dissertation, Harvard University, Cambridge, Massachusetts.

Kerin, P.D., 1992, Efficient bus fares, *Transport Reviews*, 12(1), 33-47.

Kreyszig, E., 1972, *Advanced engineering mathematics*, Wiley.

Larsen, O.I., 1983, Marginal cost pricing of scheduled transport services, *Journal of Transport Economics and Policy*, 17(3), Sept, 315-17.

Lesley, L.J.S., 1975, The role of timetables in maintaining bus service reliability, *Proceedings Symposium on Operating Public Transport,* University of Newcastle-Upon-Tyne.

Lewis, D., 1978, Estimating the influence of public policy on road traffic levels in Greater London, *Journal of Transport Economics and Policy*, 12, 99-102.

Liberal Party of South Australia, 1993, *Passenger transport strategy*.

Mohring, H., 1972, Optimisation and scale economies in urban bus transportation, *American Economic Review*, 62, 591-604.

Mohring, H., 1976, *Transportation Economics*, Ballinger, USA.

Mohring, H., 1979, The benefits of reserved bus lanes, mass transit subsidies, and marginal cost pricing in alleviating traffic congestion, in *Current Issues in Urban Economics*, edited by P. Mieszkowski and M. Straszheim, Baltimore : John Hopkins University Press.

Mohring, H., 1983, Minibuses in urban transportation, *Journal of Urban Economics*, 14, 293-317.

MVA Consultancy, 1987, *The value of travel time savings,* A report for the UK Dept of Transport, Policy Journals.

Nash, C.A., 1982, *Economics of public transport*, Longman.

Nash, C.A., 1988, Integration of public transport : An economic assessment, in Dodgson, J.S. and Topham, N. (eds), *Bus Deregulation and Privatisation*, Aldershot, UK : Gower, 97-118.

Obeng, K., 1985, Bus transit cost, productivity and factor substitution, *Journal of Transport Economics and Policy*, 19(2), 183-203.

Oldfield, R.H., and Bly, P.H., 1988, An analytic investigation of optimal bus size, *Transportation Research B*, 22B(5), 319-337.

Panzar, J.C., 1979, Equilibrium and welfare in unregulated airline markets, *American Economic Review Papers and Proceedings*, 61, May, 92-95.

Passenger Transport Board, 1994, *Proposed public transport service contracting : Information paper for prospective tenderers.*

Philipson, M., and Willis, D., 1990, Free public transport for all ?, *Proceedings of the Australian Transport Research Forum*, 623-640.

Pucher, J., Markstedt, A. and Hirschman, I., 1983, Impact of subsidies on the cost of urban public transport, *Journal of Transport Economics and Policy*, 17(2), 155-176.

Ramsey, F.P., 1927, A contribution to the theory of taxation, *Economic Journal*, 37, 47-61.

SA Government, 1993, *Submission to Industry Commission Inquiry on Urban Transport.*

Scrafton, D., 1985, *Transport policy and strategic planning: A mid 1980's status report*, Director-General of Transport, South Australia.

Seaman, R., 1994, STA, personal communication.

Seddon, P.A. and M.P. Day, 1974, Bus passenger waiting times in Greater Manchester, *Traffic Engineering and Control*, 442-45.

Sherman, R., 1971, Congestion interdependence and urban transit fares, *Econometrica*, 39(3), 565-576.

Sherman, R., 1972, Subsidies to relieve urban traffic congestion, *Journal of Transport Economics and Policy*, 6(1), 22-31.

Small, K., 1982, The scheduling of consumer activities : Work trips, *American Economic Review*, 72, 467-479.

Small, K., 1992, *Urban transportation economics*, Harwood, Switzerland.

Smith, D., and Street, J., 1992, Estimating the net welfare gains from Australian domestic aviation reforms, Paper presented at the Conference of Industry Economics, Australian National University.

STA, 1992a, *Performance indicators report : 1984/85 - 1991/92.*

STA, 1992b, *Annual Report 1992.*

STA, 1993a, *Performance indicators report, 1992/93 financial year .*

STA, 1993b, Unreliability data obtained from Roger Seaman, STA.

STA, 1994, Routes and services information system (1992 version), Main report.

Stanford, L., 1992, Competitive tendering bus services in Adelaide, Report to the Director-General of Transport, South Australia.

Starrs, M.M., 1984, *Urban transport funding and pricing*, unpublished Master of Transport Economics thesis, University of Tasmania.

Strathman, J.G. and Hopper, J.R., 1993, Empirical analysis of bus transit on-time performance, *Transportation Research A,* 27A(2), 93-100.

Steiner, P.O., 1957, Peak loads and efficient pricing, *Quarterly Journal of Economics*, 71, November, 585-610.

Stuart, C., 1984, Welfare costs per dollar of additional tax revenue in the United States, *American Economic Review*, 74, 352-362.

Tisato, P., 1990, *An improved bus user cost model and its impact on subsidy*, unpublished Master of Economics thesis, University of Adelaide.

Tisato, P., 1991, User costs in public transport : A cost minimisation approach, *International Journal of Transport Economics*, 18(2), 71-97.

Tisato, P., 1992, User cost minimisation and transport subsidy, *Economics Letters*, 39, 241-47.

TM, 1978, *Adelaide bus costing study*, Report to the Director-General of Transport SA.

TM, 1984, *The incidence of public transport subsidies in Adelaide*, for Director-General of Transport, South Australia.

TM, 1988, The subsidisation of urban transport services : Stage 2, Report to the *Urban Transport Council, New Zealand.*

TM, 1991a, *Efficient pricing policy study*, for Queensland Department of Transport.

TM, 1991b, *Optimum financial support level for metropolitan public transport services*, for Department of Transport, Western Australia.

TM (NZ) Ltd, 1994, Urban bus operations : Productive efficiency and regulartory reform - International experience, Appendix F in IC, 1994, *Urban Transport, Volume 2.*

Transport and Road Research Laboratory, 1980, *The demand for public transport : Report on the international collaborative study of the factors affecting public transport patronage.*

Truong, T.P. and Hensher, D.A., 1985, Measurement of travel time values and opportunity cost from a discrete choice model, *Economic Journal,* 95, June, 438-451.

Turk, C. and Sullivan, P., 1987, Effects of subsidy on bus operating costs, in *Transport Subsidy*, in Glaister, S. (ed), Policy Journals, 109-113.

Turnquist, M.A., 1978, A model for investigating the effects of service frequency and reliability on bus passenger waiting times, *Transportation Research Record,* 663, 70-73.

Turnquist, M.A., 1982, Notes on derivation and use of the passenger-choice wait time model, *Cornell University*, unpublished paper.

Turvey, R. and Mohring, H., 1975, Optimal bus fares, *Journal of Transport Economics and Policy*, 9(3), 280-286.

Vickrey, W., 1980, Optimal transit subsidy policy, *Transportation*, 9(4), 389-409.

Wallis, I., 1991, Competitive tendering in New Zealand : Evolving policies and experience, *International Conference on Privatisation and Deregulation in Passenger Transportation*, Tampere, Finland.

Walters, A.A., 1961, The theory and measurement of private and social cost of highway congestion, *Econometrica*, 29(4), 676-697.

Walters, A.A., 1968, The economics of road user charges, *World Bank Staff Occasional Paper No. 5,* John Hopkins Press, Baltimore.

Walters, A.A., 1982, Externalities in urban buses, *Journal of Urban Economics*, 11, 60-72.

Waters, W.G. II, 1982a, Waiting time and increasing returns in scheduled transport services: A geometric treatment, *International Journal of Transport Economics*, 9(1), 45-52.

Waters, W.G. II, 1982b, Waiting time and increasing returns in scheduled transport services: A geometric treatment - A correction, *International Journal of Transport Economics*, 9(2), 223.

White Paper, 1984, *Buses*, Great Britain Department of Transport, Cmnd 9300 (HMSO London).

Williamson, O.E., 1966, Peak load pricing and optimal capacity under indivisibility constraints, *American Economic Review*, 56, September, 810-827.

Willis, D., 1995, TransAdelaide, personal communication.

Wills, A., 1995, TransAdelaide, personal communication.

Wilson, T., 1994, Passenger Transport Board, personal communication.

Wilson, T., 1995, Passenger Transport Board, personal communication.

Windle, R.J., 1988, Transit policy and the cost of urban bus transportation, in Dodgson, J.S. and Topham, N. (eds), *Bus deregulation and privatisation*, Aldershot, UK : Gower, 119-140.