TRANSPORT PLANNING AND INVESTMENT

only study guide for **TRL3702**

author: Mr JW Barendrecht
curriculum developer: Dr BV Nduna

department of transport economics, logistics and tourism
university of south africa, pretoria

Printed and published by the
University of South Africa
Muckleneuk, Pretoria

TRL3702/1/2013–2015

98832263

InDesign

New_PR_Tour_Style

CONTENTS

FOREWORD

You are most welcome to this applied discipline in which you plan transport facilities and infrastructure for all the transport modes. You have to know how to invest in these and how to make sound transport policy judgements.

You will have to strive to achieve the outcomes below and we will assess you on the assessment criteria that follow:

Specific outcome 1:

Learners can demonstrate an understanding of the relationship between transport and development on a local, regional, national and international scale.

Assessment criteria

- Explain the link between transport and economic development
- Indicate the influence transport has in less developed countries
- Indicate the influence of transport costs on international trade
- Distinguish between the developmental roles that transport can play in a regional versus an urban area

Specific outcome 2:

Learners can give reasons for government involvement in transport and explain the transport planning process.

Assessment criteria

- Explain the necessity for government involvement in transport planning
- Indicate the cyclical nature of the urban transport planning process graphically
- Plan transport infrastructure and facilities for an urban or metropolitan area by following the seven steps of urban transport planning
- Explain the different models that can be used in simulating the transport system

Specific outcome 3:

Learners can demonstrate an understanding of the methodology used in determining the suitability of transport infrastructure and the techniques used to determine the suitability of and economic justification for transport infrastructure such as cost/benefit analysis and multicriteria analysis.

Assessment criteria

- Indicate what is meant by the technical and economic suitability of infrastructure
- Indicate when the optimal period for investment will theoretically be
- Explain how the suitability theory is applied

- Define user surplus and explain what is meant by the optimal allocation of factors of production
- Distinguish between the various criteria for project evaluation
- Indicate what role the time value of money plays in project evaluation
- Explain the various techniques that can be used in project evaluation
- Do project evaluation using the various techniques
- Explain what the technique of multicriteria analysis entails
- Explain how elementary consequences, dimensions and criteria are identified
- Explain the process of multicriteria analysis

Specific outcome 4:

Learners can demonstrate an understanding of how effective transport infrastructure in each of the four modes of road, sea, air and rail transport can be provided so that the investment in these fixed structures will be operationally and economically efficient.

Assessment criteria

- Identify the main characteristics of modal infrastructure such as roads, ports, airports, stations and rail infrastructure
- Describe how the provision of the infrastructure is planned, including the role played by the government, international trade and ownership of the infrastructure
- Identify the issues that play a role in investment decisions about the infrastructure provision for these modes
- Distinguish between the different types of financing and cost recovery methods available to the providers of the infrastructure
- Indicate that planning of these facilities cannot take place in isolation and should be seen as an interrelated system
- Indicate the role played by ownership of the infrastructure in decisions on planning and investing in the infrastructure

Specific outcome 5:

Learners can demonstrate an understanding of the need for a national transport policy, government involvement in transport, legislation and transport regulation.

Assessment criteria

- Give reasons for government involvement in transport through the formulation of a transport policy
- Explain the development of a transport policy with reference to its objectives and elements, the factors to be considered, the instruments used and the levels of such a policy.
- Indicate the existing South African transport policy
- Explain what is meant by safety regulation and indicate the elements of safety regulation

Syllabus:

The syllabus for this course is as follows:

- Transport and development
- Transport planning
- Suitability of transport infrastructure
- Cost/benefit analysis
- Multicriteria analysis
- Investment in road infrastructure
- Planning and investing in seaports
- Planning and investing in airports
- Rail transport investment
- Transport policy and regulation

The purpose of this study guide is to enable you to do transport planning in the correct chronological order. We explain the investment decisionmaking which underlies transport planning according to different approaches and show you how to interpret these approaches correctly so that you can choose the correct options in a practical situation. To provide further clarity, we consider the uniqueness of each mode, namely road, water, air and rail transport, with a view to planning and investment. Finally, we explain transport policy in South Africa, which should be taken into consideration at all times in respect of transport and planning and investment.

The following is an overview of the study guide:

For a long time now, the effect of transport patterns on the economic development in the environment in which transport operates has been a topical issue among transport economists. It is generally accepted that transport plays a crucial role in economic development, but transport economists are also beginning to realise that we should periodically re-examine the impact of the role of transport. The first thing to do is to consider the underlying problem, namely how an economy develops.

Transport (which can also be referred to as spatial interaction) reflects the socioeconomic, spatial and political dynamics of a community. During the 1960s in Europe, a period characterised by little economic growth, transport policy was aimed at network and capacity expansion. However, from the 1970s onwards, the emphasis shifted to the effective use of existing infrastructure as opposed to its physical expansion. The 1980s were characterised by environmental awareness and consequently questions about the negative side-effects of transport for general quality of life. From the 1990s onwards, interest in the potential of modern technology to improve networks waned.

Planning, and hence transport planning, is under increasing scrutiny. This is because the traditional characteristics of planning, as a discipline with a strong normative character, are changing rapidly. These changes stem from profound changes in the context and environment of planning. In the past few years transport planning has also changed in respect of its context and the environment in which it operates.

Before we make investment decisions, it stands to reason that we should consider the existing situation, also known as the zero alternative. It is not enough to have a transport infrastructure – it must also be suitable for the purpose for which it is used. Improved technology plays a vital role in this regard: at the time of building the transport infrastructure

may have been suitable for the purpose for which it was planned, but improvements in technology might have made it obsolete.

The necessity of planning transport infrastructure according to economic selection criteria stems from the fact that the transport infrastructure uses productive resources (factors of production) for long periods, and during that time these resources cannot be used for other purposes. Planning errors cannot be rectified in the short term, while medium-term adjustments can only be made at a high cost. This aspect of investment is especially important when the factors of production in infrastructure construction are scarce. The method used to control investment is to allocate the available factors of production to the best alternative – in other words, the optimal allocation of scarce factors of production.

Analytical methods such as social cost/benefit analyses, economic impact studies, environmental impact evaluation and traditional multicriteria analyses are sometimes regarded as exclusively value assessment methods. A social cost/benefit analysis measures the economic (monetary) welfare effect of a project, while an economic impact study measures economic development in terms of value added, the creation of job opportunities, economic growth and so on. An environmental impact study, in turn, measures the influence of the project on the specific environment, while a traditional multicriteria study determines the most acceptable or optimal solution according to a number of criteria.

When considering planning and investment for specific modes, it is important that we take the various operating and economic characteristics of the specific mode into consideration. In the case of road transport, investment in and upgrading of infrastructure such as roads is important, while harbours, airports and runways, and railway lines and stations are of decisive importance in sea, air and rail transport respectively. Planning for and investment in the operating vehicles of each type of mode are approached in a similar way. Finally, it is imperative to take cognisance of the "limitations" of transport planning in respect of planning and investment.

In study unit 1 we explain transport and development. Transport planning is the topic of study unit 2, while study unit 3 deals with the suitability of infrastructure, study unit 4 with cost/benefit analysis and study unit 5 with multicriteria analysis. The planning of and investment in the various modes (road, water, air and rail transport) are dealt with in study units 6, 7, 8 and 9. Finally, in study unit 10 we discuss the general principles of transport policy.

Recommended Reading

This study unit is based mainly on Button (1993:223–240).

STUDY UNIT 1

Transport and development

UNIT OUTCOMES

After working through this study unit you should be able to:

- explain the link between transport and economic development
- discuss the influence of transport in less developed countries
- explain regional development and the role of transport in this regard
- analyse transport and improvement of the urban core
- elaborate on the link between transport and development
- indicate how transport can lead to development and how transport is needed when development takes place

KEY CONCEPTS

- Transport planning
- Economic development
- Regional investment
- Urban core

1.1 Introduction

How changes in transport patterns influence economic development in the environment in which transport is operating has long been a topical issue among transport economists. It is generally accepted that transport plays a significant role in economic development while the impact of this role of transport needs to be periodically reviewed. In this study unit we will also examine the underlying issue of what underlies economic development.

It is traditionally held that transport has a strong positive influence on economic development and that an increase in production is directly related to improved transport. Anderson and Stromquist (1988), for example, argue that all major periods of transition in the European economy were related to large-scale changes in transport and telecommunications infrastructure. Four main categories of radical changes in transport and logistics can be distinguished:

- the period from the 13th century onwards, during which water transport, through the linking of cities and coastal areas, originated as a logistical system (the Hansa economy).
- the period from the 16th century onwards ("the Golden Age"), which was characterised by a dramatic improvement in navigation and the introduction of a new banking system which stimulated trade with the East and West Indies.
- the period from the middle of the 19th century (the period of the Industrial Revolution) during which the steam engine was invented (The steam engine generated new transport modes which, in turn, created new market opportunities in North America.)
- the period from 1970, which was characterised by increased information and greater flexibility such as just-in-time (JIT) procedures.

1.2 The link between transport and economic development

The link between transport and economic development is the result of direct and indirect transport inputs. The direct link relates to the following:

The shipping costs for the transportation of goods are generally low, which means that markets can be dispersed over a wide area and that large-scale production involving a variety of activities can take place. Thus, in such circumstances, the transportation of goods has a direct effect on the establishment of markets and the production process. The Industrial Revolution, for example, was successful because it was preceded by the revolution in transport technology. In a similar vein, Owen (1964) argues that the expansion of markets as a result of improved transport services is a prerequisite for economic development. Furthermore, for a variety of geographic, economic and historical reasons, undeveloped countries are reliant on international trade which, in turn, is a requirement for growth. In such circumstances, efficient port facilities are of vital importance.

The indirect effect of transport on the link between transport and economic development relates to the job opportunities generated by the development of infrastructure and the operation of transport services. Moreover, the use of, say, steel, wood and stone, which are required for the construction of transport infrastructure, has a multiplier effect.

However, the causal approach to transport and economic development has lost popularity. Through his econometric studies, Fogel (1964) shows that the growth in the American economy during the 19th century would in fact have been possible without the development of railways, because waterways can establish an intensive transport system at a comparable cost.

Nowadays, economic development is regarded as a complex process, and transport as the means to utilise and exploit natural resources. Transport is no longer necessarily seen as a driver of effective development. Transport can change what is operating capital in one area to fixed capital in another area, on condition that appropriate production opportunities exist in the potential market. However, in this regard, public infrastructure must be constructed in proportion to the availability of private capital.

Improved transport can prevent a bottleneck in production processes and therefore promote economic expansion, while inadequate transport (from a socioeconomic point of view) can prevent development and national integration. For example, inadequate transport hampers the establishment of infrastructure for, say, educational and medical services, which has a social impact.

1.3 Transport economics in less developed countries

1.3.1 Introduction

Transport investment is a principal component of the capital formation of less developed countries and transport expenditure is usually the single largest item of the national budget. Contributions to transport expenditure are received from international organisations, such as the World Bank, or in the form of direct support from individual countries. However, it is important to determine whether this kind of support is the most practical and efficient, and what effect it has on the development of individual transport systems. These considerations will now be discussed.

1.3.2 Social overhead capital as opposed to the total cost of productive activities

The contribution of transport to the economic development of a country can be identified on the basis of the following functions:

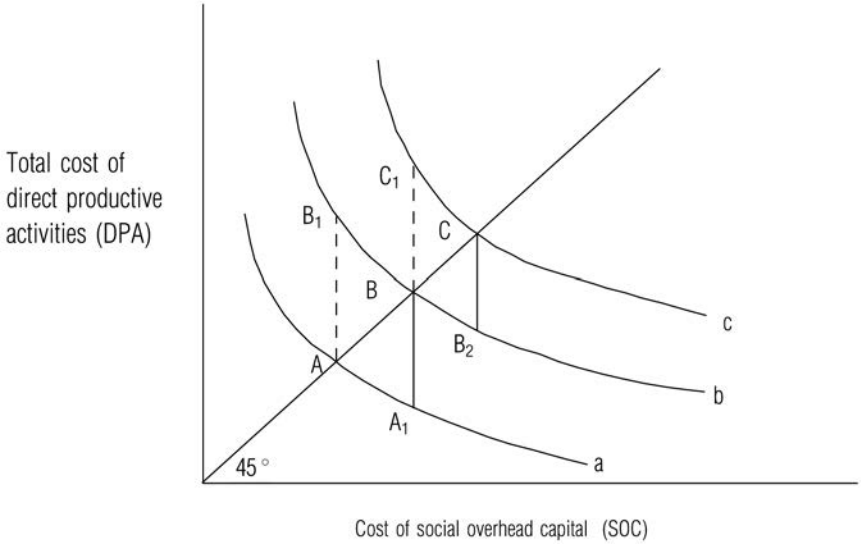
- First, transport provides input in the total production process by placing people and goods in the production process and consumption centres. Because some of these movements are between rural and urban areas, the economic process spills over into the agricultural sector.
- Secondly, improvements in transport can influence the production function because transport can indirectly reduce factor costs. For example, efficient transport can reduce high inventory levels tied up in the production process.
- Thirdly, mobility is increased and factors of production (especially labour) are transferred to places where they can be applied most effectively.
- Fourthly, transport improves the welfare of people by giving them access to various social facilities and providing national defence.

Transport economists have made important contributions by determining in detail the support role that transport can play in economic development. At microeconomic level, techniques have been developed that determine the cost and benefits of individual transport projects according to a scientific approach (cost-benefit techniques are discussed in detail in study unit 4). These techniques for appraising investment possibilities are applied in both developed and developing countries. In developing countries, the local situation sometimes requires that adjustments be made in the application of a technique. In certain countries, for example, canoes are still used to transport goods, while this technique uses mechanical transport modes. Reliable information is also not always available.

At macroeconomic level, the focus is on the contribution that transport in general can make to economic development. While one can argue in general that transport can be extended to balance development in other sectors of the economy, this is not always the case. The balanced approach is based on the assumption that if transport is inadequate, it will cause a bottleneck in the production process, but that if transport capacity is excessive, then scarce resources will be wasted, in the sense that these resources could earn a better return in another sector of the economy. However, Hirschman (1958) argues that the relationship between economic development and the provision of social overhead capital is less flexible than the proponents of the balanced approach believe.

Figure 1.1

Balanced and unbalanced growth of social overhead capital and direct productive activity



Source: Adapted from Button (1993:228)

Figure 1.1 shows the provision and cost of social overhead capital (SOC) on the horizontal axis. This is generally provided by the social sector and includes transport as the principal component. The vertical axis represents the total cost of direct productive activities (DPA), and is based on purely commercial criteria.

The balanced approach accepts that DPA output and SOC activities grow together (ie according to the 45-degree line from the origin), and pass through the various curves from A to C which show the successive increases in DPA/SOC ratios. However, according to Hirschman (1958), it is not practically possible for developing countries to follow this growth path, because social overhead capital schemes, especially in developing countries, are inherently indivisible. Hence the growth path in developing countries is inevitably unbalanced, and may following one of the following paths:

- One path is based on excessive social overhead capital. Here the path from A => A_1 => B => B_2 => C is followed.
- The second path is based on a shortage of social overhead capital – that is, A => B_1 => B => C_1 => C.

If a strategy of excessive overhead social capital is followed, direct productive activities should be less costly, which should encourage investment. Alternatively, if there is a shortage of social overhead capital, direct productive activities should first increase, together with the costs involved. Thus the construction of more intensive social overhead capital products (A => A_1 => B => B_2 => C) should result in considerable savings. However, the actual effectiveness of the alternatives will be determined by the strength of the profit motive in respect of DPA, and in the case of SOC, the reaction of the government to public demand.

1.3.3 Provision of transport

The type of transport provided which will be suitable for a developing economy is generally not as important as the total provision of transport. Some developing countries tend to use their limited development funds for prestigious transport projects, such as expensive

international airports, so that in the eyes of the world, they appear just as important as more developed countries. In other words, X efficiency is sacrificed for a superficial image. However, it is more important to spend money on internal transport provision, in an effort to ensure that benefits can be generated through the application of limited capital resources for road and rail transport. (In the case of developing countries, air and sea transport are defined as external sources of transport because they are related to international trade.)

The suitability of specific transport modes in a country is mainly determined by the geographic and demographic nature of the country. As a rule, less developed countries can be characterised as follows (Fromm 1965):

- densely populated tropical countries
- tropical countries with a low population density
- mountainous temperate countries with a low total population density concentrated in the coastal area
- desert areas in which the low-density population is concentrated along irrigated channels

The suitability of various transport modes changes according to the type of country. Poorly populated tropical countries, for example, have different problems from countries with densely populated urban areas.

1.3.3.1 Internal transport

While rail development was vital for economic and colonial development in the 19th century, in the past few years the emphasis has shifted to the development of an appropriate road transport infrastructure. The advantages of this are as follows:

- In countries with an existing basic road network, it is possible to upgrade and extend this network.
- Remote agricultural areas can be linked.
- New development, which depends on transport, is generated.

Despite the positive economic influence of road development, Wilson (1966) argues that the indirect influence of road development on social aspects such as education is greater than in the case of other modes. However, a disadvantage of road development is that urbanisation could polarise the spatial economy, and this could have social and economic disadvantages.

1.3.3.2 External transport

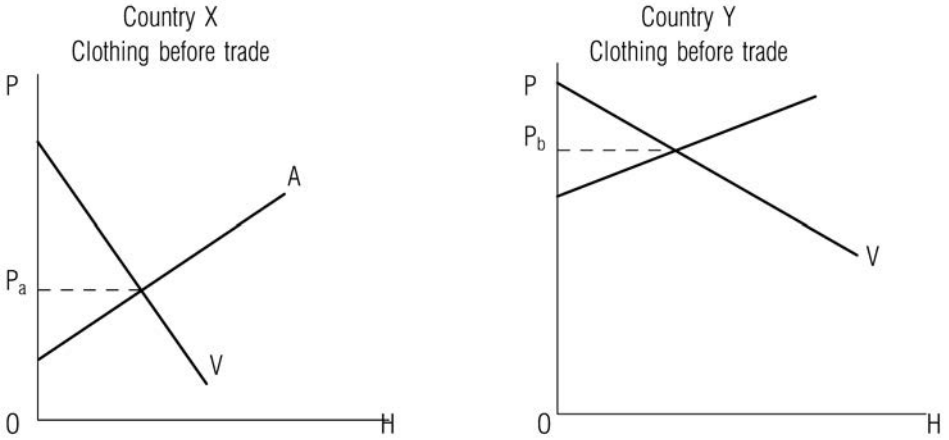
For the purposes of our discussion, external transport in developing countries refers primarily to sea transport because air transport is more geared to passenger transport and is less important for the development of the economy than for the transportation of products.

Improved port and shipping facilities can result in developing countries exporting their products to a greater variety of markets. However, the costs involved in improving or constructing a port should be measured in terms of international exchange, because the production costs of products (of which transport costs constitute a major portion) differ from one country to the next, as do the monetary units of different countries.

The influence of transport costs on international trade which should be taken into consideration in the development or upgrading of sea transport infrastructure will now be discussed. The demand for and supply of a single product (such as clothing) are used to indicate the influence of transport costs. We will first examine the situation between two countries, say country X and country Y before trade (Hogendorn & Brown 1979:225–226).

Figure 1.2

Demand for and supply of clothing

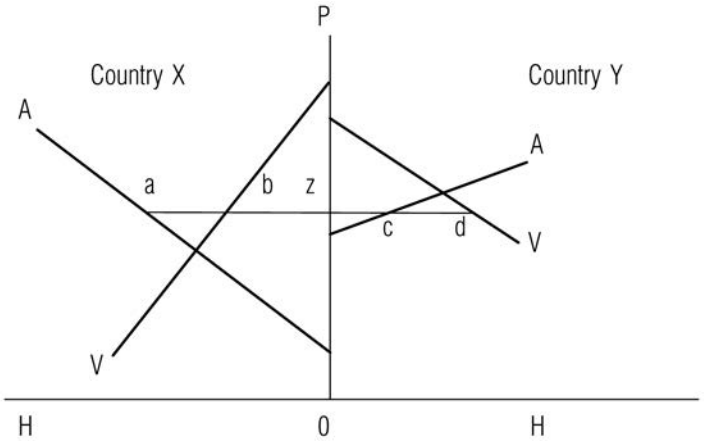


Source: Adapted from Hogendorn & Brown (1979:225)

In figure 1.2 we see the local demand for and supply of clothing in the two countries. In country X, the price of clothing (P_a) is lower than that in country Y (P_b). If the two countries start trading with each other (transport costs excluded), country X will export clothing to country Y which will result in an increase in the price of clothing in country X, while in country Y, the price will decline until it reaches a break-even point. Figure 1.3 depicts this situation.

Figure 1.3

Trade without transport costs



Source: Adapted from Hogendorn & Brown (1979:226)

Figure 1.3 is a back-to-back diagram. The demand and supply in country Y is usually indicated on the right-hand side. The demand and supply in country X has been rearranged and appears on the left-hand side. The values of the horizontal axis are from left to right, which (as usual) indicate that the demand for quantities decreases if the price of clothing increases. The slope of the curves, however, is the opposite of the normal situation – the supply curve increases to the left and the demand curve declines to the right.

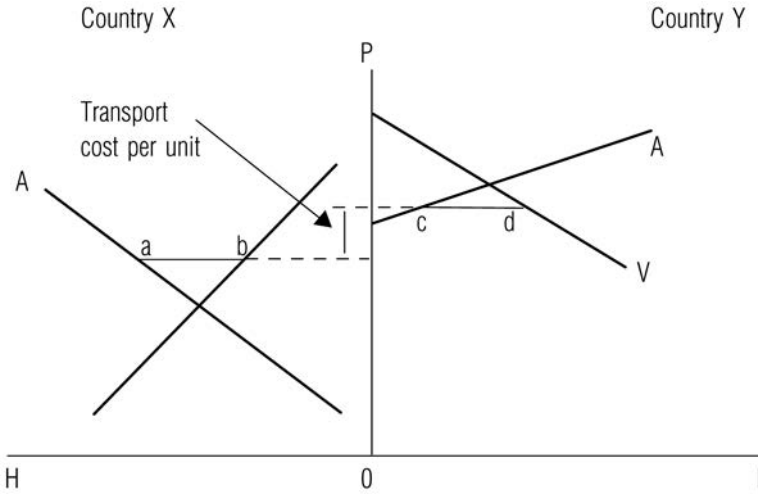
The equality in the price of clothing is indicated by *0z*. Export *ab* from country X is the surplus of its own local supply against local demand, while import *cd* from country Y is the surplus of its own local demand against its local supply. Because the exports from the one country must equal the imports of the other, *ab* = *cd*. Country X manufactures *az*, consumes *bz* and exports *ab*, while country Y manufactures *zc*, imports *cd* and consumers *zd*. In terms of the trade between the two countries, *0z* is the equilibrium. Thus the price of clothing increased in country X and decreased in country Y.

The cost of transport is now taken into consideration. Transport costs are defined as not only the cost of transport itself, but also the indirect transport costs such as insurance and handling. The result of these transport costs is that the price of clothing in the two countries is different. Trade between the two countries will increase the price of clothing in the export country (although not so much if there were no transport costs) and reduce the price in the import country (although not so much if there were transport costs).

In figure 1.4, *ab* once again represents the exportation of clothing from country X and *cd* the import of clothing from country Y. The transport cost per unit is measured along the vertical axis. The result of these transport costs is a considerable reduction of trade between these two countries because *ab*, which is equal to *cd*, has declined. (Bear in mind that the relationship between country X and country Y remains the same in respect of demand and supply. The variables which are influenced by transport costs are *ab* and *cd*.)

Figure 1.4

Trade with transport costs

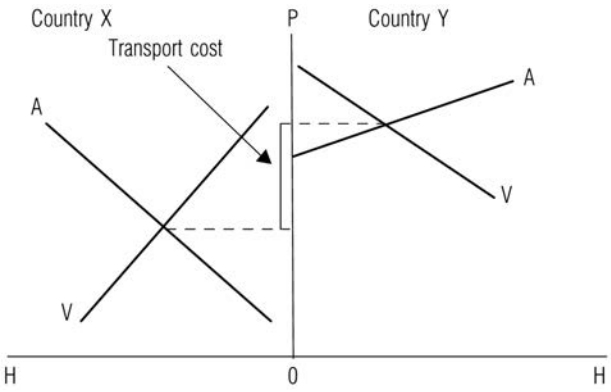


Source: Adapted from Hogendorn & Brown (1979:227)

If transport costs are excessively high, as shown in figure 1.5, this may cancel out the difference in the price of clothing before trade between the countries, and there is thus no economic sense in the two countries trading.

Figure 1.5

Trade restriction by transport costs



Source: Adapted from Hogendorn & Brown (1979:227)

This analysis of transport costs generally applies to international trade. Where no such transport costs are present, there will be trade in various commodities between countries because they will enjoy comparative advantages (except in cases where there are identical tastes, identical production factor ratios, identical technology and therefore identical prices ratios between countries). However, when transport costs are high, certain commodities may have *natural* protection, which makes international trade in them difficult or impossible.

As indicated earlier, investment in sea and air transport facilities is complex. It is clear that different variables, which are sometimes uncontrollable, need to be taken into consideration. The advantages of economic development by means of sea and air transport facilities therefore need to be considered in terms of international circumstances.

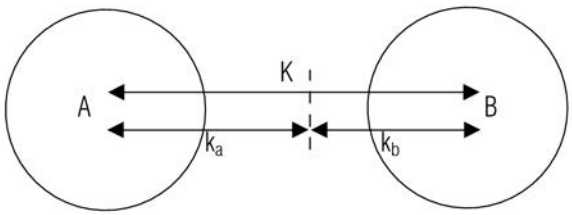
1.4 Transport and regional development

1.4.1 Introduction

The distribution of economic activities between the various regions of a country is an important issue for any government. Geographical differences relating to employment, income, migration and industrial structures are significant because they can handicap welfare and in many cases the total operation of the national economy. Governments therefore endeavour to follow a policy that stimulates economic activities in underdeveloped areas, while excessive economic growth in a developed area, which may have detrimental results in the long term, is curtailed. Hence direct financial support is made available in an underdeveloped area, and mobility of labour improved. The economic structure of developed areas is supported by improving transport facilities.

Figure 1.6

Market areas served by regions A and B



Source: Adapted from Button (1993:235)

1.4.2 Regional investment

The effectiveness of a policy favouring transport investment in underdeveloped areas has been questioned, especially in developed countries. In developed countries where an infrastructure has already been developed fairly intensively, transport is seldom a factor that is used to explain inequalities in the economic performance of regions. A transport policy supported by regional policy objectives should therefore be approached cautiously because improved transport may be counterproductive. The following simplified hypothetical example illustrates this problem:

Take two regions, namely A and B, which manufacture a single homogeneous commodity. The centres of the regions (see fig 1.6) are K kilometres apart, and the commodity that is manufactured can be transported at a fixed cost per kilometre of t per ton. The markets serving the area differ because the cost per ton to produce the product in region A is C_a, and in region B, C_b. Thus a distribution boundary can be drawn between the two regions (the assumption being that there are no production centres between the two regions). This boundary, which is indicated by the dotted line in figure 1.6, is k_a kilometres from the centre of region A, and k_b kilometres from the centre of region B ($k_a + k_b = K$). The regional boundary is determined by the relative production in the specific region and the cost of transport – that is, $C_a + tk_a = C_b + tk_b$. Mathematical manipulation of this equation (you need not study it) gives the following equation:

$$k_a = k_b + (C_b - C_a)/t$$

This equation shows that if the production costs in region A are relatively cheaper, the distribution boundary (k_a) will increase if improved infrastructure reduces the cost of transport. Hence if region A is underdeveloped, an improvement in the transport infrastructure can help to expand the region's potential market and generate an increase in revenue and job provision. However, if region B is underdeveloped, obviously investment in transport infrastructure will exacerbate the regional problem because the market area will be curtailed. An extreme case would be if region B were to be forced out of the market as a result of the expansion of the low-cost region.

Bear in mind that the above example is rather simplistic. For example, as a rule, regions do not specialise in a single product, but manufacture a variety of products. The improvement in the transport infrastructure may be detrimental to some industries, while increasing competition in others. The ultimate influence of transport investment will depend on the relative production costs between regions and on the significance of transport compared with production costs in the total cost function of the products.

1.5 Transport and urban development

1.5.1 Introduction

Over the years, changes in transport technology have had a profound influence on the shape and patterns of urban areas. The development of the steam engine in the 19th century considerably improved interurban transport and stimulated urban growth. However, local distributional services developed at a slower pace because activities tended to develop in concentric patterns around rail/sea terminals. Affluent people lived on the outer limits of the area because they could afford to use the appropriate transport, while industries and the lower-income groups tended to be concentrated around the urban core where interurban transport was available. In South Africa we have had the opposite situation where many of the less affluent lived far from the CBD.

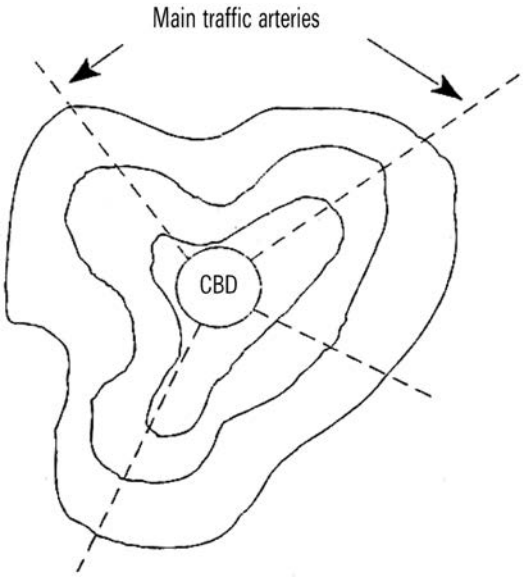
The introduction of motorised local public transport, initially in the form of trams and later buses and motorcars, resulted in the axial development of urban areas with the origin in the central business district (CBD). The axial patterns of development extended the earlier

concentric circles of housing into ribbon-like developments along the main arteries (see fig 1.7).

Finally, the universal use of motor vehicles, improvements in road systems and therefore more efficient road freight transport has led to the development of numerous urban cores and their extensions.

Figure 1.7

Transport and urban development



Source: Adapted from Button (1993:237)

1.5.2 Modern trends

Nowadays, problems in respect of transport and urban development are not so much concerned with the development and control of transport modes, but rather the degeneration of urban areas. A case in point is Johannesburg's CBD. The Carlton Centre which has office facilities on 50 storeys has become obsolete because businesses are moving to the suburbs. This trend is also common in Europe and the USA. Since the beginning of the 1970s, the focus on transport problems in urban areas has shifted to the redevelopment of urban areas.

The depopulation of urban cores which goes hand in hand with the rapid exodus of industries to the outskirts of cities, has led to the degeneration of the economies of city cores. There is escalating unemployment and as a result, the tax revenue required for services and development has to be obtained from the older and less affluent residents of these areas. The reasons for the degeneration of urban cores are complex and can be ascribed to the following, *inter alia:*

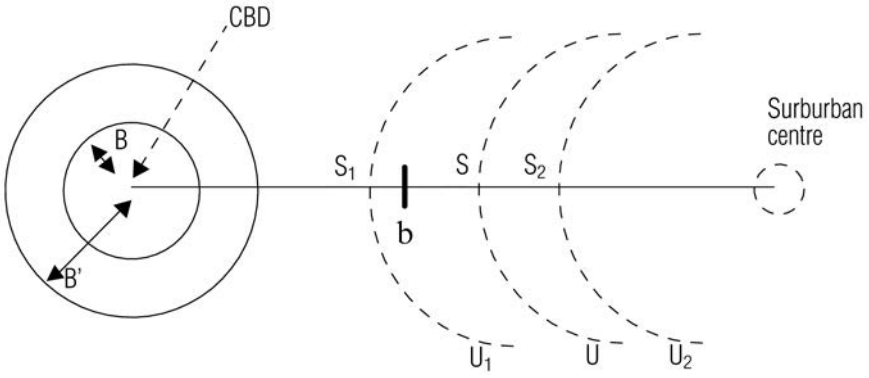
- The urban lifestyle becomes less attractive as incomes increase.
- Industries have to handle the increased land-output ratio as one of the production functions.
- Improvements in personal transport, particularly the increase in ownership of private motor vehicles, has promoted commuting to and from the suburbs.

1.5.3 Improving the urban core

Figure 1.8 depicts a typical urban situation. The central business district (CBD) is the focal point of employment and dominates the suburban centre which is linked to it by road.

Figure 1.8

The influence of transport on the CBD



Source: Adapted from Button (1993:239)

The urban core is served by efficient public transport and people up to point *B* use the bus service. Equilibrium is maintained – in other words, each household has the same utility level. Workers can choose one of three main employment/residential locations:

- live within the urban core (radius *B*) and use the bus to commute between the workplace (inside the CBD) and home.
- live outside the immediate bus transport area and travel by car to the workplace inside the CBD.
- travel by car to the suburban centre.

In the third situation (excluding city residents who may choose a fourth option, namely working from home) the boundary *U* will separate individuals working in the urban cores from those working in the suburban centres. This means that no household can improve its utility by changing its workplace, place of residence or transport mode.

We shall now examine the effect of two possible transport strategies to improve the economy of the urban core.

(1) The first strategy is to generate more income for the urban core by means of transport. Here parking tariffs can be increased or a road usage tariff imposed. However, there will be a rapid decline in activities in the urban core because increased motoring costs will cause the immediate commuting boundary belt to increase (say, to *B'* in fig 1.8). The real income of people living *B' – B* kilometres from the CBD will decline because a bus transport service is not available, and as mentioned earlier, higher motoring costs. This strategy will encourage more people to work from home or to withdraw from the labour market, with the result that the supply of labour in the CBD will decrease. Similarly, motorists who have to contend with increased transport costs will start working in the suburban centre, which will shift the employment boundary to, say, U_1. Increasing competition to work in the suburban centre will place real income under pressure, which in turn will lead to a general decrease in overall supply of labour in the urban core.

It should be clear from the above that transport restraints in the CBD tend to make labour conditions less favourable, while (in the long term) making suburban centres more attractive for the establishment of industries. The empirical fact that skilled labour is more mobile than unskilled between home and work intensifies this effect on industrial location.

(2) A second strategy is to use a subsidised express bus service in bus lanes. The service is provided between depot *b* and the urban core without any stops in between to take on or offload passengers. This strategy should not have a major impact on the transport patterns of existing bus commuters in the CBD, although it may result in fewer car trips (the cost of bus transport may be lower than that of motor transport) and therefore less congestion, which should make the labour market in the CBD more accessible. Previous car travellers from as far as S_2 will find that travelling by car to depot *b* and then taking the express bus service to the CBD will decrease the total transport cost to the CBD. Thus the boundary demarcating labour for the CBD will shift to U_2. People to the left of U_2 will notice an increase in their real income as a result of the lower transport cost. The supply of labour in the CBD will also increase, which makes the establishment of industries in the CBD more attractive. People who previously worked from home or stopped working may now find it attractive to work in the CBD. The supply of labour has therefore decreased in the suburban centre and salaries in this area will have to be increased to attract workers to the area. In the long run, the suburban centres will become less attractive for employees.

This theoretical analysis shows that although one common result of both the traffic restraint policy and the public transport improvement policy is to increase use of the bus transport service, the long-term effect on the population and economic activities will probably be different. Thus, although transport cannot solve the deterioration of the urban core, it can at least delay the process.

1.6 Conclusion

In this study unit we discussed the effect that changes in transport patterns can have on economic development in the environment in which transport operates. There is no doubt that transport plays a vital role in economic development. However, the impact of this role on transport should be periodically reviewed. It is important to understand how economic development occurs, because it is closely intertwined with the impact of transport.

1.7 Self-evaluation questions

(1) Explain in detail the relationship between transport and economic development.

(2) Discuss the relationship between social overhead capital and the total cost of productive activities. How does this influence transport economics in less developed countries?

(3) Explain by means of a graph the influence of transport costs on international trade.

(4) Explain why a transport policy supported by regional objectives should be approached with caution.

(5) Discuss two possible transport strategies that can be used to improve the economy of the urban core.

STUDY UNIT 2

Transport planning

UNIT OUTCOMES

After working through this study unit you should be able to:

- discuss the reasons for government involvement in transport planning
- discuss the steps in transport planning
- explain the models that can be used to simulate a transport system
- explain the cycle of the transport planning process using a diagram
- relate transport planning to the urban transport environment if you live in a city and to the rural town if you live in a rural area

KEY CONCEPTS

- Transport planning
- Cyclical nature
- Phases of planning

2.1 Introduction

Transport, and consequently spatial interaction, reflects the socioeconomic, spatial and political dynamics of a society. During the sixties in Europe, a period of unprecedented economic growth in several western countries, transport policy was geared to network and capacity expansion. From the seventies onwards the emphasis shifted to the efficient use of existing infrastructure, rather than its physical expansion. The eighties were characterised by environmental awareness, and consequently questions about the negative side-effects of transport for people's general quality of life. Interest in the potential of modern technology (including telecommunications) for network improvements increased from the nineties onwards (Fokkema & Nijkamp 1994:141).

Planning, and consequently transport planning as well, is receiving more and more attention. The reason is that the inherent characteristics of planning as a discipline with a strongly normative character are changing rapidly. This change is the consequence of the radical change in the planning context and environment. Bolan (1991:7) expresses this as follows:

> Planning today faces a challenging new puzzle. On the one hand, the experiments of Communist central planning failed. The block of Central and Eastern European countries are rejecting central command planning and seeking to decentralize governance and move to free market economies. On the surface at least, there appears to be a distinctive failure of collective planning. On the other hand, the past two decades have seen Japan, other Pacific rim nations, and Western Europe become major industrial states with the help of strong public and private planning mechanisms. This presentation explores the thesis that this contradiction evolves from the character of "institutional settings" in which planning takes place. The need to contingently anticipate, shape and control the future is a fundamental condition of human existence – for both individuals and collectivities. At the collective level, however, planning is embedded in a series of existential dilemmas, including: the desire for autonomy and freedom of action versus the need for connection and community; the desire for spontaneity and novelty versus the need for predictability; the need for individual expression versus the need for social control. These dilemmas are faced, negotiated, resolved, and re-negotiated in different societies through an infrastructure of historical agreements, norms, customs, rules, laws, rights, and obligations which we loosely conceive of as institutions.

Transport planning has changed accordingly over the past few years, both in context and in nature. Transport planning is currently not merely the planning of a "fixed route" but increasingly requires a flexible and illuminating policy structure in an uncertain environment. The requirements of a democratic society, such as South African society, and consequently the external trends and the internal systems of planning are client-oriented.

2.2 The role of government

2.2.1 Introduction

Investment in transport infrastructure requires capital investment on a scale which can normally only be met by the government. Transport infrastructure is regarded as part of the total infrastructure of a country and is provided for the benefit of all the inhabitants of the country. It is therefore obvious that the government will play a leading part in any investment in transport infrastructure. We now take a brief look at the principles of government involvement in the market situation.

2.2.2 Government involvement

According to traditional welfare theory, manufacturers and consumers enjoy the greatest possible degree of welfare when transactions take place on a free barter basis in markets which are perfectly operated. A prerequisite for this situation is the government's adoption of a *laissez-faire* approach. Public involvement will influence not only the Pareto optimality of the market but also related markets.

Governments do however become involved in countries with a market system. Here two questions arise: Why does the government try to become involved in the independent mechanism of a market and will this involvement improve the efficiency of the market system?

Traditionally there are two reasons why the government becomes involved in market systems, namely:

- to prevent market failure
- ethics or equity considerations

These two reasons will be discussed in the following section.

2.2.3 Market failure

In practice a market economy does not always result in a Pareto-optimal apportionment. Pareto optimality is subject to strict requirements, which are not always attainable in practice and consequently give rise to market failure – the market mechanism fails and the price system cannot guarantee the Pareto-efficient allocation of resources. Consequently the goal of remedying market failures may be given as a reason for government involvement. The aim of government measures should therefore be to correct deviations from the optimal allocation of factors of production.

Well-known causes of market failure are imperfect competition, inadequate information and a shortage of markets. Imperfect competition occurs when resources/facilities are indivisible and public monopolies are created as a result in an attempt to meet the consumer's requirements. Where there is a lack of (reliable) information the government can issue regulations concerning the quality of products in order to protect less well-informed market participants. Examples of causes of a shortage of markets are externalities, such as damage to the environment and expenditure on public goods such as national defence.

2.2.4 Is government involvement ethical or reasonable?

A second reason for government involvement in the economy is when the public is of the opinion that the economy is functioning in an unethical or unfair manner, from an ethical or political perspective. In such a case the government could redistribute income and wealth through instruments such as taxation, an income policy (regulation of minimum income) or an interest policy. An example would be reduced transport tariffs for children or older people who use public transport.

The government may also be of the opinion that consumers are underestimating the value of certain goods and services in certain cases, in which case the use of such goods and services may be made compulsory (such as compulsory insurance) or the goods may be made available free of charge or at reduced prices. It is also possible that the use of certain goods will decrease because this is in the interests of the public, for example when a high tax is placed on fuel or when consumption is prohibited (eg speeding).

It is clear from traditional welfare theory that there may be several reasons for public involvement which could possibly restore the Pareto-efficient allocation of resources. The danger is, however, that the government's good intentions could have the opposite effect (Fokkema & Nijkamp 1994:130–132).

2.3 The theory of transport planning

2.3.1 Introduction

The movement from physical transport planning to structural planning during the sixties increased the economic input into the transport planning process. One problem regarding the use of urban territory and consequently transport planning is the large number of options available. In a country such as Britain this problem is less important because land use patterns have already been established. In this case the transport system should be optimised for the existing urban system. A systematic approach to urban transport planning should therefore be followed, with four different levels of planning:

- an overall land use plan
- a strategic transport plan

- a detailed land use plan
- a detailed transport plan

Transport planning is inherently a complex process. Generally speaking it can be divided into several subsections, each with a particular economic input. You should always bear in mind, however, that there is no rigid planning guideline. Specific circumstances play an important part and specific planners will always exercise their preferences. The steps in a typical transport planning process are:

- Determine the aims and objectives to be achieved.
- Make a survey of the existing situation.
- Simulate the transport system.
- Make a forecast of the physical influence of alternative plans.
- Evaluate the alternative plans in economic terms.
- Implement the chosen plan.
- Monitor the way it is functioning.
- Revise goals and targets if necessary.

These steps are shown in figure 2.1 and may be explained as follows:

2.3.2 Aims and objectives

It is clear that the aims of (urban) transport planning have changed over time. Periods during which social and environmental considerations are emphasised are usually followed by periods when the efficiency of the system is given priority. The general aim should, however, be to make provision for weighing up the utility of specific projects and taking appropriate action. Opportunity costs should be taken into account if resources are to be efficiently utilised. Since aims and objectives are formulated early on in the planning process, when information is sketchy, it is possible that they will have to be reformulated at a later stage of the planning process.

Targets must be either directly or indirectly measurable. The “units of measurement” used to measure targets are known as efficiency criteria. Since targets are achieved via the performance of components of transport systems, it is obvious that these (efficiency) criteria also measure the performance of the components of the transport system.

For example, traffic flow is measured in the number of vehicles per hour, traffic safety in the number of accidents per million vehicle kilometres, and deaths and injuries per million passenger kilometres. Similarly, your performance as a student is measured by the percentage points you achieve.

Efficiency criteria should comply with three basic requirements (Wohl & Martin 1967:8):

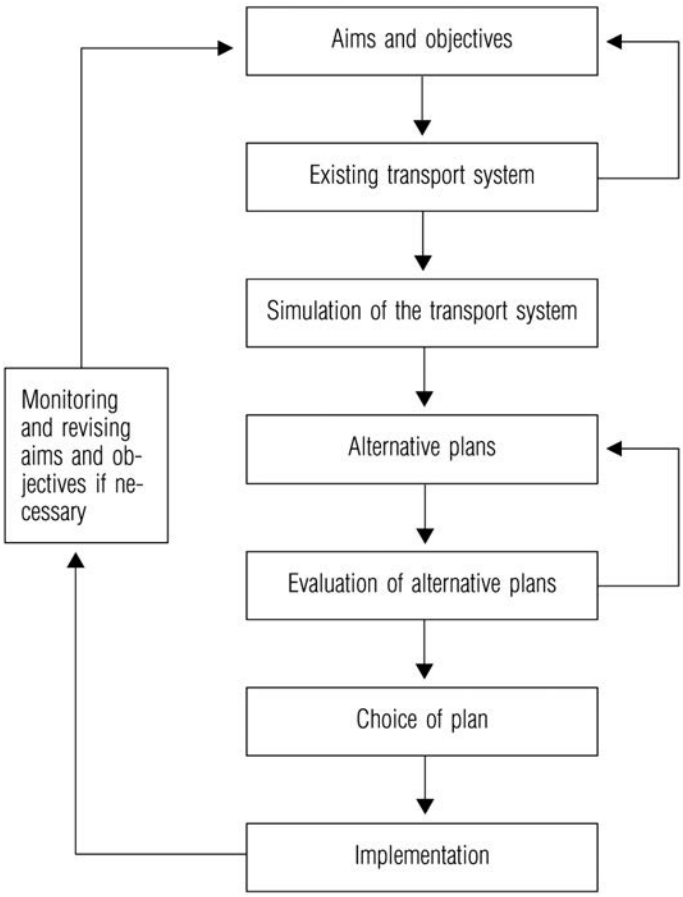
(1) They should measure the efficiency of the whole system.

(2) They should be quantifiable, that is they should be expressed in numerical terms.

(3) They should be statistically reliable, and it should be possible to collect the relevant data within a short period, at a reasonable cost.

Figure 2.1

Urban transport planning



Efficiency criteria must be appropriate and aimed at the actual problem. For example, if we are analysing an intersection that is controlled by a traffic light and is functioning unsatisfactorily, it is important to choose the right efficiency criteria. So, what should be investigated: the number of times vehicles have to stop before they can move through the crossing, the number of seconds' delay each vehicle experiences or the average length of the queue at the intersection? Criteria should also be applied to the whole transport system component. It would be an inadequate solution to favour the main street in the intersection for an unnecessarily long period so that vehicles in the smaller side street experience long delays. Similarly, the performance of a single transport component should be seen in the context of the rest of the system. It would be of little use, for example, if the criteria showed that the problem at an intersection had been solved but the batches of cars let through were to arrive at the next traffic light at the beginning of a red phase. A thorough analysis would also reveal whether the traffic light could be coordinated with other traffic lights.

There are certain factors that limit the available options when it comes to solving problems. The commonest restrictions are economic or financial. Statutory and bureaucratic restrictions are frequently present and can result in the introduction of a system that appeared more or less ideal to the planner being delayed for long periods or even brought to a standstill – think for example of the time-consuming process involved in town planning and establishment. Then, certain solutions may be in either political favour or disfavour at a particular time. An example here is the commissions of inquiry which have a prime opportunity to follow a true systems analysis approach but whose results risk being rejected

by the government of the day. It should also be remembered that the existing system may impose restrictions on an alternative solution. An improved or new systems component has to fit in and be compatible with the subsystem or system in which it is going to operate. An example of the problems that could occur in this area is the Franco-British Concorde supersonic passenger aircraft which could never be properly accommodated in the larger or international air transport system, partly because the advantage of reduced flying times was virtually cancelled out at flight terminals by slow procedures. Many excellent proposals have come to nothing because the restrictions were ignored. (The Concorde was of course retired after the fatal crash in Paris in 2000.)

Once targets and goals have been set and the scope of existing problems has been estimated, one can proceed to carry out a survey of the existing situation within the framework of existing restrictions.

2.3.3 The existing transport system

Information on the transport system and travel patterns is obtained by means of sampling or a physical count of the people who make use of the transport system. Information on travelling behaviour is obtained from households and also from organisations, which make their transport requirements known. Additional information can be obtained from official sources, such as the Central Statistical Service.

The kinds of surveys and the questions asked have changed over time. Emphasis is now placed on the reasons why people embark on journeys, rather than on the development of transport prediction models. Consequently more detailed information is required from households. Furthermore, modern measuring techniques, which are more efficient, have led to a decrease in sample sizes.

The existing transport system can, however, be efficiently evaluated by identifying existing problems. Problem detection and identification can be done in various ways. Within the transport sector a problem is typically *first* observed and experienced by the users of a transport system who become dissatisfied with the performance and service delivery level of a systems component. A problem therefore arises when the inefficiency present in a system makes inroads on users' standards of mobility. When a system no longer comes up to users' expectations, the resulting dissatisfaction can lead to changes in consumer behaviour, which can lead in turn to a new set of problems. For example, a decline in the efficiency of a transit system's service could cause passengers to use their own transport, which could lead to traffic jams and parking problems on the one hand and land the transit system in financial problems on the other hand, leading to a further decline in the standard of service. Users of a transit system often base their criticism on an implicit assumption regarding the standards that existing systems components and services should meet. Because of technological advances and increased prosperity, users may also view the lack of sophisticated new alternatives (such as the absence of a throughway of a particular geometric standard or a high-speed rail system) as a problem.

A *second* important method of identifying problems is to look at the number of potential systems users who are not using the system because they find it inaccessible. Additional infrastructure is usually required to make it possible to utilise economic and development opportunities, and the problem of making infrastructure available is a typical source of transport problems in developing areas. It may be impossible, for example, to exploit raw materials and minerals that have been discovered because of inadequate accessibility.

A *third* method of identifying problems, and one that is more formally geared to diagnosis, is to obtain information from investigating officers who are responsible for the functioning and coordination of individual systems components, and monitor and investigate this information regularly. Some examples of such people are road and railway track inspectors, vehicle inspectors, quality controllers and helicopter traffic patrols.

A *fourth* method of problem detection and diagnosis uses financial and economic control. Accountants, transport economists and financial experts usually draw on budget variance control, economic monitoring processes and financial techniques in an attempt to apply raw materials and other transport inputs effectively and in optimal proportion to one another, so that efficient service delivery can take place and acceptable levels of income can be maintained.

As soon as a problem has been recognised and properly diagnosed, its extent must be assessed. This is important for various reasons:

(1) The urgency of a problem determines its priority on the list of problems to be solved.

(2) The extent of the problem will determine the attention and the resources that are devoted to solving it.

(3) In view of the durability of transport facilities it is necessary to establish whether the measures adopted to solve a problem should make provision for initial overcapacity.

(4) The scale of a problem serves as a guideline in the search for alternative solutions.

Any attempt to solve transport problems should always be geared to the future. For example, where a road is operating at service levels E and F at peak hours, and service level D would be acceptable, it may be desirable for economic reasons to plan the road improvements in such a way that overcapacity exists initially and service level C is initially experienced at peak hours.

2.3.4 Simulation of the transport system

2.3.4.1 Introduction

Because of the complexity of transport markets, the representative data required to reflect the situation accurately are difficult to come by. Transport data can also be too cumbersome for future forecasts if statistical techniques are used that set certain limitations. The following should be the requirements for a functional transport model:

- It should explain the transport situation/behaviour in simple terms.
- It should make a contribution to policy formulation.
- It should predict the transport situation/behaviour meaningfully.

Transport demand and travel models are used to predict the transport facilities that will be required. It is therefore important that the variables in question should be accurately predicted. The models can also be used to assess various planning options, which implies that the models should be simple and that it should be fairly easy to evaluate the influence of various alternative strategies.

Transport models have their limitations, however, because each ratio and the related variables should be determined unambiguously. The ratios should be determined strictly according to mathematical methods, and all the related variables should be quantifiable. For this reason, and because a large number of variables can influence the entire system, a large number of data are required. The acquisition, storage and handling of these data cause various practical problems which have to be solved before the data can be placed in a computer data bank. The cost of this process and of programming the computers could be an obstacle to the development of models and should not be lost sight of.

2.3.4.2 *Types of analytical models*

We shall now discuss *three* kinds of analytical transport models, namely descriptive models, predictive models and planning models.

(a) Descriptive models

Descriptive models are used to describe the behaviour of systems. Their aim is to describe the typical characteristics of the functioning of a system in concise and simple terms and replicate this in mathematical terms. By replicating the functioning or performance of an observed system in other terms (either words or symbols), the descriptive model is able to describe and predict performance in similar systems.

Descriptive models do not rely very much on logical and causative aspects. To build a descriptive model it is adequate to discern sufficient regularity. For example, if historical data point to a significant correlation between two variables, a descriptive model may be used to determine the relationship between these variables without taking into account what the causative or logical basis of the relationship may be.

Descriptive models are usually built on the basis of empirical observation and analysis. In other words, they are based on the results of the observation of the system in question. Statistical analysis is useful for the building of descriptive models, and there are certain analytical techniques which are widely used in transport systems analysis. These techniques include regression analysis, correlation analysis and factor analysis. In transport studies regression analysis is frequently used to calibrate models of trip generation. These are models which describe how the number of trips generated by a group of people are related to their socioeconomic characteristics. A model of this kind is based on observations, usually at one specific period, of various groups of people with different socioeconomic characteristics and different trip generation patterns. Correlation analysis determines which variables are included in the models and regression analysis is used to estimate parameter values and to test the importance of various forms of comparison. A typical descriptive generation model might look like this:

$R_w = a + b\,HH,$

where

R_w = the number of daily commuter trips generated in an area

HH = the number of households in the area

a, b = the coefficients determined by regression

A descriptive model is therefore a faithful representation of how a system works, based on observations of the system.

(b) Prediction models

The purpose of prediction models is to predict the performance of a system at any period in the future or under certain hypothetical conditions. They are sometimes referred to as predetermination models, but this is probably an exaggeration. A prediction is a statement about the possible occurrence of phenomena in the future, and is generally accompanied by a degree of uncertainty. A predetermination, on the other hand, is a more definite or deterministic statement about future events. There are a number of differences of opinion about whether predetermination is actually possible. In most instances it appears that predetermination cannot be done scientifically. Prediction, on the other hand, can be done in various ways, and may vary from simple extrapolation from past trends (eg time series analysis) to complex causative models (econometric models).

The most important requirement in prediction models is that there should be a logical or causative connection between the variables. This is the only means of guaranteeing that relationships that have been observed at some point will continue to be present in future. The analyst must conduct an empirical investigation into the stability of observed relationships between variables. In prediction models there is a strong correlation between stability and a logical basis, which implies that they must have a logical content.

For example, if we observe that the ratio between the number of households and the number of commuter trips generated in an area is constant, and if we accept that there is a logical or a cause-effect relationship between these two magnitudes, we can use the model that relates to them as a prediction model. It is also important to note, however, that prediction can only be done in one direction, whereas description can be done in two directions. In other words, in the model of household trips, it is possible to predict the number of trips if we know the number of households but we could not predict the number of households if we knew the number of trips. A prediction of this kind would be senseless, since it would have no logical basis. Households cause trips to be generated, and we can therefore predict the number of trips, since we know how many households there are. But trips do not give rise to the presence of households, so we cannot make a prediction in that direction. This distinction between descriptive and predictive models is important, because in the case of descriptive models it does not matter in which direction the comparison works. Because they are simply based on observed correlation, we can replace or exchange the dependent and independent variables in a descriptive model, which is not the case in a prediction model.

In forecasting, an understanding of the relationship between form and process is of decisive importance. In a descriptive model it may be sufficient to note that X and Y are covariable (eg that the variable Y consistently has the value 5X, or similarly that X = 0,2Y), but when the aim is to calculate the value of Y at some time in the future, the model must specify a causative sequence (eg that a one-unit change in the value of X will cause the value of Y to change by five units). If we can postulate the direction of causation, knowledge of the future value of the "cause" allows us to determine the value of the "result" in advance.

The first task of anyone building a prediction model is therefore to provide a logical framework within which the variables of interest are placed at the end rather than at the beginning of a causative series of variables. (Variables on the right hand of the equal sign are often described as exogenous.) The second task is to make certain that those variables that are placed at the beginning (known as endogenous variables) can be estimated in an acceptable manner as far into the future as may be necessary.

The second requirement is partially relaxed in the case of provisional forecasts, which are more important to planners than the unconditional kind in any case. The forecaster is usually interested in the condition of the world after he has taken planned action of one kind or another, or after some possible but uncertain event outside his control has taken place. The model can then be allowed to react in the form "if X occurs, Y will follow", without any explicit declaration that X will probably occur. But an explicit prediction still has to be made for other exogenous events, since these could reinforce or counteract the effect of the hypothetical change in X.

A special type of conditional prediction is known as impact analysis. The focus here is the result a specific exogenous impact (change in X) is expected to have if the environment is not disturbed in any other way. An example of this kind of forecast is when the macroeconomic consequences of having a new road in a region are estimated.

Prediction models may be used either to extrapolate trends or to provide predictions of reaction variables on the basis of exogenous predictions of the causative variables. In the case of trend extrapolation, a prediction model may take the form of a continuous

time-dependent function, a differential equation, or a discrete differential function that relates the variables to time, measured in stages.

(c) Planning models

These models are used to arrive at strategies for systems planning. In comparison with descriptive and predictive models, planning models are not used to give the analyst an indication of how the system works, or of what is likely to happen to the system; rather they indicate how the system should work, or what should happen to the system. Planning models are useful for the development of alternatives in the systems planning process. They are also useful for the analysis of alternative systems. Planning models often provide a forecast of the consequences of a course of action, together with an evaluation of these consequences in terms of the planning goals and performance criteria. They are normative and fall into one of two large groups: optimisation models and equilibrium models.

Optimisation models are used to derive operational or design strategies for systems that will produce a restrictive optimum on a goal function. In other words, a function is built which describes the goal of the system, measured by performance criteria. This function is known as the goal function, and it is usually a function of systems variables. The goal of an optimisation model is to obtain the values of the operating variables that will optimise the goal function (minimise or maximise it). There are a number of techniques that are used in building optimisation models, including integral and differential calculus, simulation and mathematical programming.

A typical example of an optimisation model is a model of traffic light phase regulation. A simple optimisation model could look like this:

> $G_1/G_2 = V_1/V_2$, for minimum average delay, where G is the length of green time allocated to the approach to a crossing, and V is the traffic volume of each of the approaches.

The second kind of planning model is the *equilibrium model*. In contrast with the optimisation models, which rely heavily on goal setting, equilibrium models rely on good descriptive models of the system. These models are used to work out operational strategies for behavioural systems. Equilibrium models are used for systems with a reaction feature where the reaction of the operating conditions depends on the system. An example of this kind of model is the supply-and-demand equilibrium analysis in transport demand modelling. When transport is allocated to networks, the routes one chooses depend on the travelling times, and the travelling times depend on the amount of traffic. Equilibrium models are used to describe the allocation of capacity to traffic links.

2.3.5 Alternative plans

Different alternatives are usually investigated (apart from the existing situation or null alternative, which is the norm according to which alternatives are measured). These alternatives might be a transport package oriented towards public transport, a transport package oriented towards private transport, or a combination of the two. Once the forecasts or evaluations have been carried out, they may provide new information on the detailed effect of the alternative plans and a new mix of the packages may emerge. Without this feedback the option that might turn out to be the most advantageous may well be overlooked.

The problem-solving process is essentially dynamic, and this leads to the postulation of numerous possible alternatives; however, there are two sets of limiting factors that will result in a provisional screening process that will weed out the non-starters. Firstly, proposals must promise to be economically viable and capable of achieving the aims and objectives. The second set of factors takes the form of regulations which, especially in the transport industry, impose physical restrictions on a solution, such as design standards and regulations, safety regulations, health requirements and environmental standards.

Within this somewhat limited framework, the search for solutions is guided largely by two economic guidelines. Firstly, the restrictions imposed by raw material availability and

budget ceilings will force analysts to be very selective in submitting alternative solutions. Secondly, analytical investigations should be directed at options that are likely to prove the most productive.

Economic considerations would normally also indicate when a search for a solution should be abandoned. First, once the funds budgeted for the investigation have been exhausted, the search would be abandoned. Secondly, the process might be suspended when the marginal costs attached to the search reach break-even point with the net additional benefits that might result from the search. All alternative problem solutions which appear to be compatible with the systems restrictions should then be evaluated.

Any change to an urban transport system is extremely complex and far-reaching. It is therefore difficult to formulate more than a limited number of detailed plans. A detailed plan would involve the development of projects and schemes, and for a large city the possible combinations can be enormous. It may in fact only be possible to draw up one alternative plan in detail after it has been selected from among other possibilities that were investigated at the beginning of the planning process.

2.3.6 Evaluation of alternative plans

The funds required to meet the continual demand for transport improvement generally exceed available funding. There are usually a number of capital-intensive projects competing for the available funds. It is therefore essential that alternative transport projects should be scientifically assessed to determine whether they are viable so that the maximum advantage can be gained for the community without expenditure exceeding certain limits.

Road user costs which are influenced by road improvement projects include vehicle running costs, value of travellers' time and accident costs. To enable the analyst to evaluate the savings potential of a planned road transport facility objectively, it is essential that road user costs should be accurately and realistically calculated.

Although benefit/cost analyses of transport projects are generally seen in the literature as being synonymous with economic evaluation, what is actually meant is evaluation from a transport economics point of view. Such an evaluation merely amounts to a micro-economic evaluation of those aspects that are directly related to the physical transport aspects of the project, the aim being to assess the viability of community savings.

All predicted benefits and costs which are directly related to the provision, use and maintenance of a facility are related to one another and evaluated in respect of cost efficiency. Facility costs (which jointly represent community costs) can be divided into three components, namely:

- start-up costs (all capital costs related to the establishment of the facility).
- maintenance costs (eg all maintenance costs to keep a road negotiable and also the cost of providing for the traffic flow on the road).
- end-value or residual value (The former is the reuse or salvage value of any components of the road and the value of the land reserve at the end of the service life of the road; residual value is the remaining value of the road and the value of the land reserve at the end of the period of analysis if the service life of the road has not yet expired).

Taking a road facility as an example, the components of user costs are as follows:

- vehicle running costs (fuel consumption, tyre wear, engine oil consumption, vehicle capital costs and maintenance costs)
- accident costs
- value of travellers' time

Whereas a microeconomic evaluation or evaluation from a transport economics point of view usually concentrates on a project itself and evaluates its effectiveness in terms of predicted savings in total transport costs, a macroeconomic evaluation concentrates on the economic benefits which are generated outside the project. In the latter case an evaluation would typically be geared to predicting the potential economic development and growth in an area that would be stimulated by the building of a road. Among other things, the multiplier and accelerator effects that could be expected in the regional economy would be noted. In addition, an estimate can be made on the basis of input/output analyses of the forward and backward linkage effects in the flow of goods and services. A macroeconomic analysis of a proposed road is therefore an investigation into the effect of economic plus factors that could benefit non road users in many ways. Individual evaluation techniques used in transport economics usually determine the microeconomic viability of proposed projects in one of the following three ways:

- absolute advantage – which is determined by the net current value technique
- relative advantage – which is usually determined either by the benefit/cost ratio technique or by the yield rate technique
- minimum total social costs – which are determined by the technique of current value of costs

2.3.7 Choice of plan

Because resources are scarce, budgets are usually limited – at government and all other decision-making levels. However, society's need for transport is virtually unlimited, and therefore potential transport projects should not only be evaluated economically but also very carefully chosen. Two criteria usually apply when the economic aspect is taken into account:

- Project expenditure should be within the budgeted amount.
- The economic principle should be strictly pursued.

The projects that best meet these two requirements make the grade in economic terms. The individual techniques of transport economic evaluation usually determine viability according to one of three criteria, namely: absolute advantage, relative advantage and total community costs.

The economic choice of a specific project for implementation takes two forms, namely project formulation and project prioritisation.

Project formulation is the selection of the best option from the point of view of transport economics, the options being mutually exclusive. Mutually exclusive projects are projects which are pursuing the same goal, for example to link points A and B. If there are three routes by means of which two places could be linked, the choice of one route would exclude the choice of the other two routes. It can therefore be said that project formulation means selecting the most beneficial method of solving a specific transport problem.

Project prioritising means arranging all functionally independent projects in order of micro-economic viability. According to this method projects are ranked in order of attractiveness in terms of transport economics, starting from the top and working downwards, until the point is reached where the capital budget has been exhausted. Every independent project that is in competition for selection on microeconomic grounds is already naturally the “winner” out of a small group of mutually exclusive projects. Functionally independent projects are alternatives which are aiming to achieve different goals or objectives. The selection of one independent project can at most delay the selection of another, but not exclude it.

In this discussion of project selection our point of departure is that project developers should always take account of the preferences of a local community, such as that the minimum amount of damage should be done to the environment, that the interests of landowners should not be prejudiced and that any system limits should not be exceeded. If the impact of candidate projects is reconcilable with the sentiments of the people they affect, the ultimate criterion should be economic considerations.

One of the characteristics of the functioning of the multidisciplinary team of systems analysts is that there should be continual contact with the decision makers or their representatives so that the project does not founder unexpectedly because of political undercurrents.

2.3.8 Implementation

In the transport industry (as in any other field) projects and plans cannot be implemented before certain tasks have been scheduled. There must be clarity on matters such as which tasks will be funded at which times. The question of task programming, methods of determining critical routes, project management and the design of time frameworks for transport capital investments affect implementation very closely. If capital dries up before the completion of a project, or funds are temporarily inadequate, the result may be interruptions in the implementation phase, with the following disadvantages:

- The longer implementation is delayed, the bigger the original problem becomes, and the more acute the effect it has on existing transport facilities.
- Because of inflation, after a budget discontinuity even more serious budgetary problems are experienced.
- Interruptions in construction programmes generally bring about additional implementation costs in the sense that work teams and equipment cannot readily be allocated to other tasks and factors of production may be unutilised.
- Since the decisionmaking body is committed to transport improvement, opportunity costs are negatively influenced in any case by the assets which had already been tied up before the interruption, but are now not being utilised.

Implementation should take place in such a way that it is timely in terms of the needs of the users of the system, at the lowest cost by the presenter, and with the least possible disruption from the point of view of the community as a whole.

2.3.9 Monitoring of performance and revision of aims and objectives

Twenty or 30 years can elapse from the time when a transport problem is first observed, or a transport need emerges, and the implementation of a project. The service life of road facilities, for example, spans about 20 to 30 years. It is self-evident that a community's goals may change over such a long period, and the operation of any new system should therefore be monitored on a continuous basis from the time when it is implemented.

It is easier to make adjustments to a project immediately before, rather than immediately after, the final detailed planning stage. Once the detailed planning has been completed and the implementation stage has begun, there is less opportunity to make adjustments. Because technical obsolescence begins to operate right from the beginning of a project (which can interfere with performance) and economic obsolescence can arise as a result of innovation, technological progress, and an increase in social and consumer standards, the functioning of systems components should be continuously monitored and revised.

If the projections show that a permanent gap is developing between system functioning and system goals, this is an indication of a serious problem in the system which requires more than simply short-term action and the cycle of analysis has to start from the beginning

again. As the flow charts on the left side of figure 2.1 indicate, the process of transport systems analysis (TSA) does not have a definite cut-off point, and the monitoring and revision stage will take the whole process back to the initial stages of the analytical cycle as soon as a new problem emerges.

A fairly long period elapses from the initial experience of a problem, through all the development stages, up to the time when a solution is implemented, and problems begin to develop with the solution after a while, so the team that lives with the project and gets to know it in intimate detail will naturally change over this period. One of the chief reasons, as we have said, is that such a project runs for a very long time, but secondly it should be remembered that the full spectrum of analytical techniques is too comprehensive and diversified for a single team of analysts.

2.4 Summary

In this study unit we emphasised the fact that transport planning has changed radically in the past few years. Modern transport planning is no longer the planning of a "fixed route" but increasingly requires a flexible and informative policy strategy in an uncertain environment. The demands of a democratic society, such as that of South Africa, play an important part. Consequently the external trends and the internal systems of planning are client-oriented.

The improvement of transport infrastructure requires a level of capital investment which can usually be provided only by the government. Transport infrastructure is also regarded as part of the total infrastructure of a country which is provided for the benefit of all the inhabitants. It is therefore obvious that the government plays a leading part in investment in transport infrastructure.

The movement from physical transport planning to structural planning has raised the economic input into the transport planning process. One problem attached to the use of urban land and transport planning is the large number of options available. In this case the transport system should be optimised, given the limitations of the existing system. A systematic approach to urban transport planning should therefore be implemented, with various levels of planning. We have now learned how infrastructure for transport is planned; next we look how we determine the suitability of such infrastructure in study unit 3.

2.5 Self-evaluation questions

(1) Explain the necessity for government involvement in transport planning.

(2) Discuss the transport planning process and demonstrate its cyclical nature by means of a figure.

(3) Explain the models that can be used to simulate the transport system.

STUDY UNIT 3

Suitability of transport infrastructure

UNIT OUTCOMES

After working through this study unit you should be able to:

- discuss the suitability of infrastructure
- explain the technical and economic suitability of infrastructure
- illustrate the advantages of capacity expansion schematically
- explain the determination of the optimal period of investment
- discuss the application of the suitability theories

KEY CONCEPTS

- Technical suitability
- Economic suitability
- Optimal period of investment
- Suitability theory

3.1 Introduction

Before transport investment decisions can be made, the existing situation, which is also known as the null alternative in benefit/cost analyses, should be considered. The mere existence of transport infrastructure is not sufficient; it is important that it should be suitable for the purpose for which it is used. In this regard improved technology plays an important part: the transport infrastructure may have been suitable for the purpose for which it was created, but it becomes obsolete as technology improves. An example of technological renewal of transport infrastructure is the continuous improvement of the OR Tambo International Airport.

In this study unit we explain the methodology used to evaluate the suitability of transport infrastructure. This study unit is based on research by the Australian Bureau of Transport and Communications Economics which mainly takes the form of an assessment of the transport infrastructure for the next 20 years (Harvey 1995).

3.2 Suitability of transport infrastructure

3.2.1 Introduction

There is a link between the suitability of transport infrastructure and the issue of whether or not investment in additional transport infrastructure is required. Investment in transport infrastructure may be required if the existing quality of infrastructure service delivery is inadequate for the following reasons:

- high operating costs
- long periods in service
- unreliable provision of service

The above problems may arise when capacity is inadequate and the infrastructure deteriorates physically and becomes obsolete over time. This obsolescence is the result of:

- technological change
- a change in demand requirements
- increases in input prices
- a change in safety requirements

It is not always easy to define what would be regarded as a "low level of service delivery". If the efficient use of resources is the criterion used to determine whether the level of service delivery is low and the infrastructure consequently needs upgrading, this is an economic consideration. The cost of investment could then be weighed up against the benefits of service delivery, using a cost/benefit analysis. (Cost/benefit techniques are explained in detail in study unit 4.) Cost/benefit techniques are, however, complex, data-intensive and time-consuming. Simpler and faster methods can be used to take decisions on smaller investments, where the use of cost/benefit analysis techniques is not justified. The usual approach is to apply a practical rule, namely that upgrading should be considered in cases where the service quality of infrastructure has dropped below a certain acceptable level.

The evaluation of infrastructure by measuring it against a technical definition can only serve as a broad guideline in determining whether or not the upgrading is economically justified. It is possible that infrastructure which is not technically suitable (and therefore requires upgrading) may be economically suitable if the cost of upgrading is higher than the benefits that it would bring about. In this case a cost/benefit analysis would show that the upgrading was uneconomic. In other words, if the advantages of upgrading exceed the cost, the upgrading would be economically justified, even if the infrastructure was technically suitable.

3.2.2 Technical suitability

The above-mentioned practical rule, which serves as a broad guideline on whether or not to invest, is applied in order to determine technical suitability. Transport infrastructure is technically suitable if the physical or operating results are above a certain minimum acceptable level. Technical suitability can therefore be said to be determined on the basis of physical or operating characteristics.

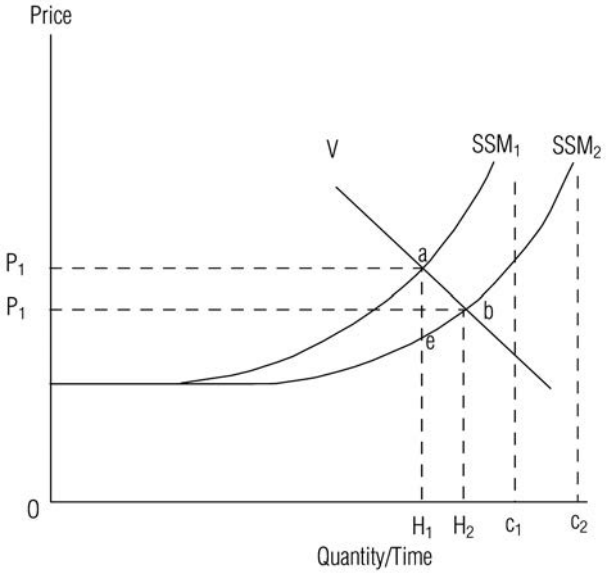
An example of physical characteristics in road transport is the number of vehicles per day per lane. The minimum acceptable operating characteristics may be specified in technical

terms, such as average or peak hour speed, or in terms of costs, such as user costs per kilometre.

In view of the fact that the suitability of transport infrastructure is essentially an economic question, the setting of technical standards is subject to economic criteria. One approach would be to accept a certain average for the network being assessed, because the standard for a given type of infrastructure at any given time is usually more or less correct in economic terms. The physical or operating characteristics of a large section of the existing infrastructure can be compared accordingly and the infrastructure that falls below a certain minimum standard in terms of the practical rule may be regarded as technically unsuitable. In the final analysis, the dividing line between suitability and unsuitability is a matter of opinion.

Figure 3.1

Benefits of capacity expansion



Source: Adapted from Harvey (1995:61)

3.2.3 Economic suitability

The economic suitability of infrastructure is determined on the basis of a social cost/benefit analysis. Where investment in order to improve the quality of the service is not economically justified, transport infrastructure is regarded as economically suitable.

Any investment is justified at any particular time if:

- the current value of benefits exceeds the current value of costs
- no net advantage in welfare would be produced by delaying the investment

The first condition ensures that the resources used for investment will produce at least the same returns as they would produce elsewhere in the economy, and the second condition ensures that the investment will be made at the best possible time.

The economic suitability of investment is explained with the aid of figure 3.1. The specific investment is represented by a demand curve, V, and two short-term social marginal cost

curves, SSM_1 and SSM_2. The horizontal axis represents the quantity provided or requested per time period and the vertical axis represents the "generalised" social cost of using the infrastructure.

Generalised costs consist of all the costs related to the use of the infrastructure, irrespective of who is involved. In the case of roads, generalised costs include the following:

- the cost of providing and maintaining roads
- the cost of operating vehicles
- passenger time and the time linked to freight transport
- external costs, such as those attached to accidents, air pollution and noise

The marginal costs of using infrastructure are brought about by an "additional" user, and the expression "short-term" indicates the period during which it is not possible to invest in order to change the infrastructure. The capital cost and fixed operating cost of the infrastructure are excluded because they are not influenced by short-term use.

According to figure 3.1 the short-term social marginal costs (SSM_1) rise in proportion to the use of the infrastructure to a maximum capacity (c_1), and operating costs, delays and unreliability also increase. If the maximum capacity increases to, say, c_2, the short-term social marginal cost will move to the right – to SSM_2.

The demand curve (D) shows the amount of infrastructure demanded at each level of generalised cost by the user. User costs represent the costs of users themselves, taxes and the cost of using the infrastructure. The composition of costs is simplified by assuming that the taxes and levies charged are representative of the short-term social marginal cost of the resources used by the users. This is the economic optimal price.

As a result of capacity expansion, users will experience a decrease from P_1 to P_2 generalised costs and consequently their use will increase from H_1 to H_2. The net advantage of capacity expansion for the community is equal to the shaded area abf in figure 3.1. This is the difference between the advantage to the consumers, which is represented by the height of the demand curve, and the social costs required for the additional demand represented by the height of the SSM_2 curve. It is clear from this that the greater the shaded area – and obviously the benefits of capacity expansion for the community – the higher the demand in relation to capacity will be.

A social cost/benefit analysis would compare the capital costs of the capacity expansions with the discounted current value of the benefit per time period. The first condition of economic suitability requires that the latter should exceed the former.

If it were possible to expand infrastructure in equal sections, capacity could be added progressively, on condition that the current value of the benefit exceeds the additional value of the expense. In practice, capacity expansion usually takes place in bulk, chiefly because of the economies of scale that can be obtained with the construction of capacity. In other cases technical factors play a part – a road has to have a certain number of lanes.

Activity 3.1

Briefly distinguish between the technical and economic suitability of infrastructure. Give examples of such infrastructure investments in South Africa.

3.2.4 Optimum period for investment

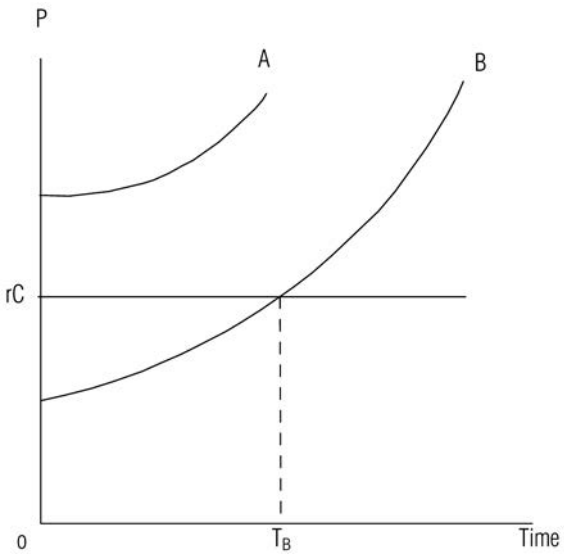
(Please note that it is not compulsory to study the formulas given below. They are merely given to explain the relationships between the different variables. A study of the formulas would naturally be to your advantage, however.)

The timing of infrastructure investment is important in economic terms. Even if an analysis shows that the current value of benefits exceeds their cost, it may be more advantageous to delay the investment.

Suppose the upgrading of infrastructure is of a permanent nature. If the upgrading is delayed, the user or community is denied the benefits of the upgrading. However, the capital that would have been used for the infrastructure can be invested and the interest earned can benefit the users or community.

Figure 3.2

Optimal period for investment



Source: Adapted from Harvey (1995:62)

The optimum period for investment is shown in figure 3.2. The horizontal axis represents time and the vertical axis the annual benefits and costs. The annual benefit is represented by curves A and B. The horizontal curve rC represents the discounted rate and the capital cost of investment.

The assumption of "perfect" capital markets is made. According to this the interest rate is equal to the discounted rate and consequently equal to the opportunity cost of the cost of capital. In these circumstances the community will enjoy the benefit rK – in other words, the interest return on the capital earmarked for investment. The proposed investment should therefore be delayed while the benefit in time [B(t)] is less than the interest yield [B(t) < rC]. Briefly, the project should be delayed if the interest yield for the first year is lower than the discounted rate. If it is necessary to know at any specific time whether the infrastructure is suitable, information on the benefits in the first year of the project and the cost of capital is required. Forecasts of future benefits will have to be done to determine exactly when capital will be required in future.

As the demand for the infrastructure increases in time, the annual benefits will also increase, with the result that the investment will be justified at a certain time. This is illustrated in

figure 3.2. The annual benefits curves (A and B) show a rising trend to the right because, as the demand curve in figure 3.1 moves to the right with time, the distance between the SSM_1 curve and the SSM_2 curve increases.

If there is an annual benefit as represented by curve A, the circumstances warrant immediate investment. In this case the optimal period for investment lay in the past. An annual benefit, as represented by curve B, implies that the period of investment should be delayed until curve B intersects $rC - T_B$ in figure 3.2.

You must make certain that you understand and are able to apply the above ratios. The following explanation should be of assistance:

We assume that that the annual growth over time takes place in accordance with the following function:

$$b(1 + g)^t$$

where b is the benefit in the year nil of the investment and g is the annual growth rate in benefits. If the demand curve shifts to the right at a constant growth rate, as is generally accepted in cost/benefit studies, the growth in benefits will usually be higher than the growth in the demand for the benefits. The benefits of infrastructure expansion are expected to increase at a faster rate than a shift in the demand curve because the gap between marginal costs increases with and without investment – as figure 3.1 shows. Remember, however, that the benefits of a project do not necessarily follow this simple functional pattern over time – the above example is merely intended to serve as an illustration.

By using the formula for annual benefits instead of optimal period conditions, we determine the optimal period for investment as follows:

$$[\ln (rC/b] / [\ln(1 + g)]$$

It is clear from this formula that if the discounted rate and capital costs are high, the optimal period of investment is delayed, whereas higher benefits and an increase in the growth rate of benefits will put forward the optimal period for investment.

The benefit/cost ratio (BCR), which is the current value of benefits divided by the current value of capital costs, is as follows:

$$BCR = b(1 + g)^T/C[r - \ln(1 + g)]$$

where T is the period of implementation. Consequently the BCR increases over time with an increase in annual benefits.

If the investment is made at the optimal period, the BCR formula changes as follows:

$$1/[1 - \ln(1 + g)/r]$$

The benefit in the nil year, b, and C, the capital cost of investment, fall away. It is clear from this formula that with a positive growth rate and an optimal period, the benefit/cost ratio (BCR) cannot be lower than one. Consequently a BCR with a value of less than one would be held back until such time as this value was above one.

The value of BCR above one at the optimal period will be determined by the growth rate of the benefits relative to the discounted rate. If the optimal period of the project was in the past, as illustrated by curve A in figure 3.2, the BCR will still be high, with a value determined by exactly when in the past the optimal period occurred. The application of the optimal period criterion to identify investment projects and determine the period implies that the BCR will be above the value of one. The BCR value will be significantly higher than the value of one where the growth rate in benefits is high relative to the discounted rate, and where there is considerable underinvestment.

The focus on the optimal period of investment simplifies studies which determine the sensitivity of change in demand, because the change in the period of the project can reasonably be estimated.

Activity 3.2

Explain the relationship between the benefits of a project and the optimal period of implementation.

3.2.5 Investments other than capacity expansion

The short-term social marginal costs curve (SSM) in figure 3.1 has been drawn in such a way that investment shifts this curve to the right. The short-term marginal costs remain the same at low inputs and improvement in service quality takes place because there is sufficient capacity to deal with any volume of demand.

Certain investments will shift the SSG curve downwards, or both to the right and downwards. An example here is investment which is aimed at a saving of variable maintenance costs. The principles for determining whether or not the investment is justified and the principles for estimating the optimal period are the same. In terms of figure 3.1 the demand curve in this case would move through the flat areas of the SSG curves and the annual benefit would still be determined by the area between the SSG curves and the demand curve.

3.2.6 Nonoptimal pricing

In order to simplify the discussion in figure 3.1, it is assumed that taxes and levies paid by users would include short-term social costs throughout. This is the optimum pricing rule for achieving economic efficiency because, in addition to their own private costs, marginal users pay the full cost which devolves on the community as a result of their decision to use a road and cause noise and air pollution, for instance.

In practice prices will never reflect identical social marginal costs. When the prices of social marginal costs differ, the benefits derived from upgrading the infrastructure will be more complicated than is evident from the shaded area in figure 3.1. Benefits that arise from the increasing willingness of users to pay for the facilities that are being used will be measured according to the demand curve and the private generalised costs which occur, including taxes and levies. Benefits, in the form of net cost savings, will be measured by the social marginal costs curve.

If the prices are higher than the social marginal costs, the infrastructure will be underutilised by efficiency standards, which could lead to a decrease in investment. If, however, the use of the infrastructure is underpriced, congestion will increase according to the efficiency standards, which could result in an increase in investment.

3.2.7 The influence of investment on networks

Investment in transport capacity could, as a result of the increase or decrease in the traffic making use of the facility, cause an increase or decrease in the traffic in other parts of the transport network, which would bring about additional costs or benefits through changes in the congestion levels. This could occur in the same type of mode of transport or in another type if the modes of transport are competitive or complementary. The influence of investment on networks should be taken into account when cost/benefit analyses

are carried out, even though it is difficult to estimate the influence of investment. It could be omitted from strategic assessments, provided that it does not play a significant part.

The influence of investment in one project could result in the period of investment in another project being accelerated or retarded. This is an important aspect which should be taken into account when evaluating a transport network where simultaneous investment in several projects is planned. If this aspect is ignored, any estimate of future expenditure may be skewed. However, in order to evaluate the interaction between various investment projects we would have to use complex mathematical programming techniques that require detailed data on the demand relations and traffic flow at the starting and finishing points of the route. However, if detailed data are not available, sensitivity tests can be carried out which are able to determine what the significance of the influence of the various projects would be in general terms. In an analysis of intercity highways in Australia, Harvey (1995:64) for example, postulated a more price-sensitive demand curve that would be applicable to a single project in isolation, in order to simulate the influence of the implementation of a number of projects.

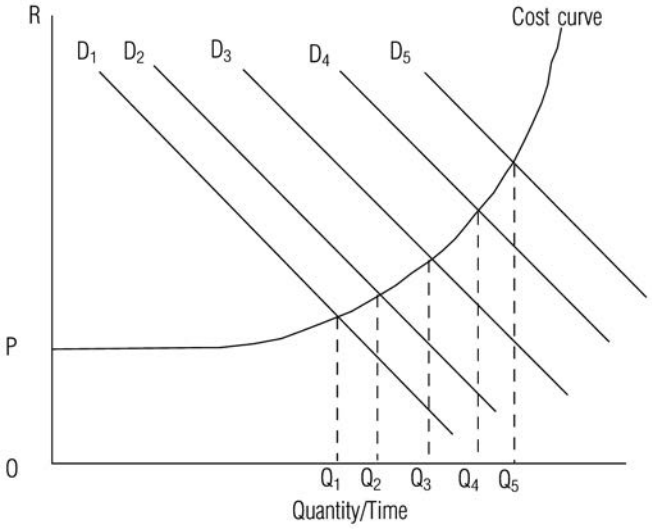
3.3 Application of suitability theory

3.3.1 Demand projections

Before the suitability of the transport infrastructure is assessed, it is essential to carry out demand projections. As demand increases to the level of fully utilised capacity, the quality of the service will decrease and consequently depress demand. Investment in new or improved capacity could, however, stimulate the demand for the use of the facility. In order to study the correct demand trends, it is necessary to distinguish between the influence of growth in demand and the influence congestion has on demand. When the demand for the facility increases, the demand curve shifts to the right and the influence of congestion is associated with movements along the demand curve.

Figure 3.3

The demand for travel quantities



Source: Adapted from Harvey (1995:65)

Demand projections are illustrated in figure 3.3. The demand projections can be simplified by making the assumption that the quality of the service is constant. In line with this, figure 3.3 shows a demand curve which moves to the right over five periods. The price level P represents the generalised cost at time 1 when the demand curve is at V_1 time. As the demand increases and with generalised costs at the level of P, the quantity demanded will follow the projections of the demand curves along the horizontal axis. These are the quantities the demand projections aim to estimate. If the quality of the service is taken into account, however, the quantities will be obtained where the demand and cost curves intersect.

3.3.2 Information requirements

3.3.2.1 Introduction

An important requirement for efficient evaluation is to collect relevant information on the infrastructure that is to be studied. Information on recent utilisation is important in order to make demand projections, and time series information is required for certain prediction techniques.

3.3.2.2 Technical assessment

The basic information required for technical evaluation is the following:

- the physical properties of each part of the infrastructure to be evaluated
- the level of utilisation

Technical evaluation is carried out by drawing comparisons between the physical properties of each part of the infrastructure and predetermined standards. Comparisons are also made between existing parts of the infrastructure in order to identify the least efficient part of the infrastructure. Utilisation information is required when the physical properties are expressed in relation to output, such as vehicles per day per lane. Note that it can be useful to consider the information on technical evaluation together with utilisation information, because part of the infrastructure which is of a low standard but also has a low utilisation could be suitable in economic terms.

A sophisticated form of technical evaluation is based on performance characteristics such as delays, reliability or operating costs. The information required is the current level of service delivery, or a model which is capable of estimating it. A model requires more detailed information on the physical characteristics and utilisation than would be necessary for the technical evaluation of physical characteristics. A projection of future service levels, using the estimated demand for the existing facilities, also requires modelling.

If a project is identified by means of a technical evaluation and the project is affordable, the costs of possible future projects can be deduced from it. The project investment identified will be that which improves the level of service delivery.

3.3.2.3 Economic evaluation

The evaluation of economic suitability requires that the upgrading of the infrastructure should be specified so that the benefits and costs of the upgrading can be determined accordingly. If alternative plans/methods are available which could produce the same improvement in the infrastructure, these alternatives should be compared with one another. As we have already indicated, technical evaluation can be useful in identifying these alternatives.

In the case of a large-scale strategic study with a limited time frame or limited resources, it is necessary to keep the economic evaluation at an elementary level and simply to provide a broad guideline as to whether the investment is justified. If the information and models are available to estimate the service level that would be provided by the infrastructure, as required by the technical evaluation of the performance characteristics, a basic economic analysis is possible but the proviso would be that additional information requirements are met. This additional information would include the capital cost of investment projects and information on operating costs. The value of time and reliability, which are the chief benefits of project investment, should also be included.

3.4 Summary

The approach followed for evaluating the suitability of infrastructure offers a high degree of flexibility, which is essential given the variations in the availability of information and the relaxation of modelling between modes.

A technical overview of the physical characteristics of the infrastructure is carried out at the lowest level of evaluation of infrastructure. The next level of evaluation is a technical evaluation of suitability, based on the current and projected performance of the infrastructure, in terms of the level of service delivery. This evaluation has the additional advantage that it can be included in demand projections. By following the technical evaluation approach to identify potential projects, and then estimating the costs of these projects, it is possible to determine future investment needs.

Lastly, if it is possible to specify investment projects and to estimate the costs and benefits, economic evaluation will be applied. This technique can be applied at various strategic levels.

In general, little attention is given to the optimal period of investment. It has been shown, however, that a benefit/cost ratio with a value greater than one is not necessarily a sufficient reason in economic terms for considering a project. In the case of marginal projects, one could consider delaying the project.

3.5 Self-evaluation questions

(1) Explain what is meant by the “suitability of infrastructure”.

(2) When we measure the suitability of infrastructure, we measure it first according to technical requirements and then according to economic requirements. Discuss the investigative process in detail.

(3) Explain the advantages of capacity expansion schematically.

(4) Why is the optimal period of investment important? Discuss the reasons in full.

(5) Explain the interaction between the respective variables when the optimal period of investment is determined.

(6) How is suitability theory applied? Discuss in full.

STUDY UNIT 4

Cost/benefit analysis

UNIT OUTCOMES

After working through this study unit you should be able to:

- explain the optimal allocation of factors of production in terms of user surplus
- distinguish between and discuss the various criteria for project evaluation
- define the time value of money
- explain the influence of the time value of money on project evaluation
- discuss the various techniques of project evaluation
- carry out project evaluation according to the various techniques

KEY CONCEPTS

- Cost/benefit analysis
- Economic costs
- Financial costs
- Time value of money
- Project evaluation

4.1 Introduction

A cost/benefit analysis may be defined as a practical way of evaluating the desirability of a project, an exercise which takes the form of adding up and evaluating all the relevant costs and benefits (Prest & Turvey 1965:685). This method of analysis aims to evaluate the desirability of the project, and specifically make a selection of infrastructure, as well as arrive at a broad assessment of public expenditure. In addition a cost/benefit analysis, or its results, can be used to develop or modify transport policy, because the focus here is is long-term transport, which involves far-reaching positive and negative consequences.

We plan transport infrastructure using economic selection criteria because transport infrastructure uses economically productive resources (factors of production) for long periods,

during which those factors of production cannot be used for other purposes. This means that errors in planning cannot be rectified in the short term, and even a correction in the medium term would be very expensive. This aspect of investment is especially important when the factors of production of infrastructure construction are scarce. This method of controlling investment consequently involves allocating the available factors of production to the best alternative. What is at issue here is the optimal allocation of scarce factors of production, which we will discuss next.

4.2 The optimal allocation of factors of production

4.2.1 Introduction

Economists work on the assumption that the factors of production are scarce in relation to demand. As a result, factors of production have a particular value. When factors of production are allocated, the aim is to maximise value in respect of:

- distribution among the applicants for the factors of production
- distribution of the factors of production over time

The distribution of the factors of production among the applicants is usually determined by supply and demand in the market. The price arrived at reflects the value of the factor of production for the person who wants it, rather than its actual cost. The efficient application of factors of production over time means that the time value of money should also be taken into account in the cost/benefit evaluation of investment possibilities.

The goals to be achieved should also be borne in mind. Where the optimum division of scarce factors of production is aimed at benefiting the community as a whole, the allocation is judged on the basis of the social benefits obtained versus the cost of the factors of production used. When organisational goals have to be achieved, the allocation should be judged on the basis of the profit/loss position of the organisation after the factors of production have been used. The difference between these two approaches lies in the fact that a change in the value of a new or technically better product such as an improved road network is not dependent only on the price of the product or service. The user surplus should also be taken into account when evaluating investment possibilities.

4.2.2 User surplus

The reasons why user surplus should be taken into account are explained with reference to figure 4.1.

DD represents the demand curve of the new improved product or service (road network) and is a function of the quantity and price (all other factors remain constant). At a price P_0 the demand is Q_0. If the price drops to P_1 (as a result of savings in travelling costs), the volume requested would rise to Q_1. The quantity by which it rises has two components:

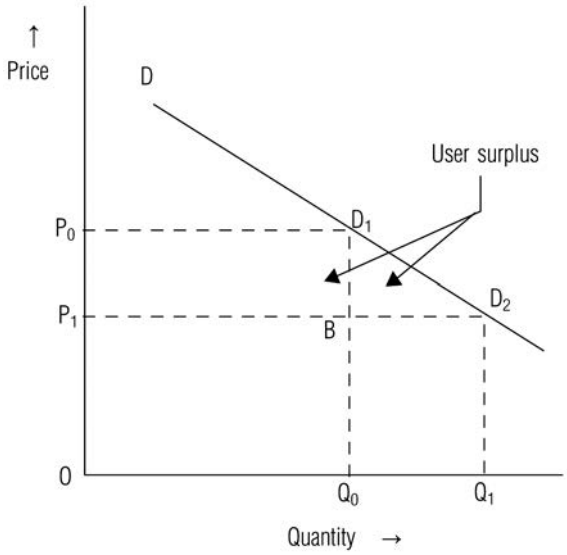
(1) Benefits to existing users $= P_0Q_0 - P_1Q_0$

$= Q_0(P_0 - P_1)$

(2) Benefits to new users $= ½(Q_1 - Q_0)(P_0 - P_1)$

Figure 4.1

Illustration of user surplus



Source: Homburger & Kell (1982:418)

The first component corresponds to the usual financial analysis of increased income. In this case the organisational goals are put first and the second component is ignored, because it does not produce any surplus income. We therefore merely compare the areas of the rectangles $P_0D_1Q_0O$ and $P_1D_2Q_1O$ (ie income) and leave the area D_1BD_2 out of account.

We should however always take the second component into account when doing an economic analysis because the total benefit obtained consists of the benefit that existing users obtain from the reduction in costs (first component) plus the saving in costs to new users (second component). Where the benefits derived (profit) are used as a criterion for decision making in financial analysis, the savings consumers would derive from the acceptance of a project are used as decision-making criterion in economic analysis.

The difference between the financial and economic analysis lies in the fact that the values of the costs and benefits of the actual monetary or financial flow may vary, and furthermore the discounted rate may not necessarily be the same. In the first place these differences are the result of the different ways in which the problem occurs — financial viability is related to financial planning and economic viability is related to the planning of investment. In this regard calculations based on total economic costs and benefits differ from those of the decision-making models of private investors. In general the viability of infrastructure investment is measured according to overall economic targets, whereas investment in the private sector is aimed at a maximum rate of return.

The aim of a financial or economic evaluation, on the other hand, is to prioritise alternative possible solutions to a problem by allocating a monetary value to each of them. This is done by quantifying both the benefits that could be obtained and the costs that would be incurred, discounting these values to current values and comparing the current values of each project in order to draw up a list of priorities.

4.3 Criteria for project evaluation

(Georgi 1973:10–15)

4.3.1 Introduction

As a result of the change in income and costs over time caused by infrastructure investment, it is necessary to compare the various alternatives in order to determine their economic

viability. This suggests that we require a criterion that will give the decisionmaker an indication of how necessary the investment is. It is important, however, to use a uniform standard of evaluation to make certain that the comparisons are uniform. In other words, the different investment possibilities in question should be evaluated according to the same criteria. On the basis of the results, a priority list of projects can be drawn up.

Such comparisons can best be carried out if the following assumptions are made:

- There are no budgetary restrictions.
- The products being compared are not interdependent in any way – direct comparisons are being drawn.
- The change in costs and benefits is assessed according to the objectives in view.

These assumptions do not alter the fundamental theoretical principles of cost/benefit analyses in any way. If these assumptions are not made the problem simply takes a different form; it does not become a new problem based on different principles.

Criteria which are used to determine the desirability of investment in particular sectors of the overall economy should be based on a "with or without" principle, which necessitates a comparison between discounted future values of revenues and costs so that different time spans can be taken into account. This is the only way to judge whether the economy can continue with or without the planned investment. Appropriate investment criteria are those that not only eliminate non-variable investment possibilities but also make it possible to make a choice between variable investments. This choice is especially important when two or more projects which are being assessed have the same goal. Another function of the assessment criteria is to rank the projects in order of importance, a matter which is naturally of economic importance when funds are limited.

Various criteria which comply with the requirements of the above investment criteria may be used. These criteria are discussed below with the following objects in view: to eliminate uneconomic projects and select the most economic product.

4.3.2 A self-sustaining economy in comparison with real costs

A standard for the evaluation of infrastructure investment in respect of the total economy centres on the question of whether or not a project is likely to raise the productivity of the total economy. According to this criterion productivity increases if a total benefit can be produced for the total economy at a lower actual cost after investment than would have been the case before investment. Owing to the influence of infrastructure investment over time, a project is economically justified if the capital costs of the new project (eg the building of a road) are lower for the economy than the cost of building competitive infrastructure that would fulfil the same function in the economy (eg a railway line).

This comparison of actual costs does not take all the socio-economic consequences generated by the investment into account. Only the marketable costs, internal to the specific sector, are taken into account. Since these evaluation criteria do not take all the consequences which the investment holds for the economy into account, this method does not qualify as a standard for assessment.

4.3.3 The increase in the national product test

This criterion for economic viability attempts to calculate the contribution of the specific investment to the present and future national product (the total value of the production by factors of production during a specific period). This criterion is supported by input/output models which are divided according to regions and sectors, so that not only the direct influence but also the indirect influence can be identified.

If infrastructure investment is aimed at growth, the infrastructure is justified if it leads to an increase in real national income during the period of analysis. This is the "exclusion criterion" for investment, which in a certain sense is the minimum condition. If a number of projects pass the test the one that shows the highest productivity is implemented.

4.3.4 Linear programming

In contrast to the "increase in the national product test" which identifies the investment with the highest productivity, the linear programming decision-making model attempts to identify the optimum project by means of a simultaneous solution. An optimum project of this kind would be based on the following specifications:

- the operational formulation of goals (goal function)
- information on real structural relations
- information on the nature of the influence of infrastructure investment
- conditions of a factual and normative nature which limit the scope of permissible solutions

Linear programming makes the following assumptions:

- the linearity of ratios with resulting constant economies of scale
- the linearity of the target function
- the perfect divisibility of projects
- the independence of processes

It is clear from the above that the linear programming technique is merely an aid to decision-makers because the empirical ratios between the individual infrastructure investments and the rest of the economy can at best only be represented approximately. The representation of these ratios is further complicated by the fact that in most cases infrastructure investment shows a declining cost pattern and, in addition, externalities are present. Consequently the interdependence of various factors is not properly taken into account.

4.3.5 Project-related investment criteria

Unlike the previous criteria, project-related investment criteria aim to apply a uniform standard to test the influence of all infrastructure investment costs and revenue on the economy as a whole. Such investment costs and revenue are determined, classified and evaluated by means of economic evaluation.

For the purposes of economic project evaluation, projects can be divided into two groups:

(1) mutually exclusive projects

(2) independent projects

Mutually exclusive proposals are alternative methods used to perform the same task. If one piece of equipment is chosen for a task, another will not be required. For example, if three different alignments for the road between A and B are being considered, it is clear that only one can be chosen. The economic evaluation of mutually exclusive projects is aimed at identifying the most efficient alternative in economic terms.

Independent items, by contrast, are pieces of capital equipment which are used to execute various types of projects or tasks. Examples of independent projects are a proposed urban throughway in Johannesburg and another throughway in Cape Town. More than

one independent project can be selected for implementation. All the independent projects under consideration could in fact be undertaken if they were all economically justified and sufficient funds were available. The economic evaluation of independent projects therefore involves the ranking of economically justified projects in order of economic merit.

The sequence in which mutually exclusive and independent projects should be evaluated is important. This sequence results from the organic nature of the planning process, which begins with finding the best solution to a single "undesirable" situation. The mutually exclusive alternatives which are proposed to rectify the situation must be evaluated in order to find the best alternative in economic terms. The various independent alternatives (ie the best solutions in economic terms for the various "undesirable" situations) should be compared in order to rank them on the basis of economic merit.

4.4 Economic evaluation of projects

4.4.1 Introduction

The economic evaluation of projects is based on the above project-related investment criteria. Several aspects that influence the evaluation process and the techniques associated with them will now be discussed.

4.4.2 Aspects related to project evaluation

4.4.2.1 The analytical period

The analytical period refers to the period in which alternatives are evaluated. It should be the same in all cases in order to ensure that the different analyses are compatible. The following factors influence the duration of the analysis period:

(1) The physical properties of the facility usually indicate a long physical life and consequently a long analysis period.

(2) Forecasting for long periods is always a problem, however. Future land use and technological change are only two of the possible factors that complicate forecasting.

(3) Discounting for periods of longer than 20 to 25 years remains a problem. The series of current value factors of 10 percent per annum, for example for 25 and 100 years, varies between just 9,077040 and 9,999274.

For these reasons, an analytical period of 20 years is recommended. If the life of the infrastructure is expected to be more than the analytical period, the residual value should be included in the analysis.

4.4.2.2 Discount rate

The discount, which is also known as the cut-off rate or the minimum acceptable rate of return, is related to the time value of money, which will be discussed in detail in the next section. With the PVOC (present value of costs technique) and PNV (net present value) techniques, as well as the B/C (benefit/cost) ratio technique, the discount rate can be interpreted as the "interest rate" or the value of i that should be used in calculating the interest factors. With the IRR (internal rate of return) technique the discounted rate can be interpreted as that rate of return beneath which a particular alternative is not economically acceptable. The expected IRR is therefore always compared with the discount rate.

When evaluating public projects, such as road construction projects, the Treasury usually recommends a rate that is applicable to all projects.

4.4.2.3 Inflation

If all the cost components for the duration of the analysis period are subject to the same inflation rate, inflation should not be taken into account during the analysis. Constant base year prices (as at the beginning of the analysis period) should be used.

If, however, the fuel price is expected to rise more rapidly than the inflation rate applicable to other cost components, this should be taken into account in the discounting of future fuel prices by using the "general inflation rate", in order to calculate the relative change in prices.

Although it is theoretically correct to include inflation when the cost components are subject to differential inflation, many transport economists use constant prices in the economic evaluation of road-building projects, owing to the problem of predicting inflation rates for a 20 year period.

4.4.2.4 Null alternative

The null alternative represents the existing situation; it is also known as the "do-the-minimum" alternative, and serves as the standard by which the other mutually exclusive alternatives are measured. It is extremely important that the null alternative should be correctly identified – even in cases where there is no "obvious" null alternative – in order to ensure valid results. The "construction costs" of the null alternative should, however, not be included in the analysis, since these represent a historical or sunk cost.

4.4.2.5 Sensitivity analysis

Since economic evaluation involves the forecasting of benefits and costs over a long period, it is important to examine the relative effect of the various assumptions on the results of the evaluation. It is therefore desirable to repeat the analysis by assigning different values to parameters such as the following:

(1) *Discount rate.* The analysis should be carried out with different discount rates. For example, if the Treasury recommends a rate of 6 percent per annum, the analysis should also be carried out with rates of 4 and 8 percent per annum.

(2) *Traffic growth rate.* It is advisable to do the analysis for a spectrum of traffic growth rates, such as 2, 4 and 6 percent per annum.

(3) *Time costs.* Since the expected savings in travel time constitute an intangible benefit, it is recommended that the analyses should be carried out with and without the inclusion of time savings.

4.4.2.6 Residual value

The residual value (remaining value) of a capital asset may be defined as the economic value of the asset at the end of the analysis period. Most capital assets, such as roads, have an economic life of over 20 years, and consequently this residual economic value must be included in the analysis, as indicated in the example in section 4.5. It is especially important that the present value of the residual value should be deducted from the present value of the construction costs of the road, and not added to the benefits.

4.4.2.7 *Shadow prices*

Shadow prices may be defined as the real, economic or intrinsic values that should be assigned to benefits and costs. Shadow prices are used when there is reason to believe that market prices do not reflect the economic value or cost of resources. There are several factors which could contribute to this discrepancy between market and shadow prices, such as:

(1) imperfect markets or the absence of markets

(2) government intervention in the economy in the form of regulation of the prices of certain commodities or changing of price levels via taxes and subsidies

Taxes and subsidies should therefore be omitted from the analysis. They do not represent economic resources but are merely a vehicle for transferring funds between the public and the private sectors. If taxes and subsidies cannot be excluded from the analysis, benefits or costs or both may be either understated or overstated. This means that factor costs must be used, which can be defined as market prices minus taxes, plus subsidies.

4.4.3 The time value of money

The "time value of money" means that a sum of money has a higher value at the present time than the same sum will have at some time in the future. This assertion is correct if the value of money remains unchanged, since money which is immediately available can be invested at a certain interest rate in order to earn further income in the interim.

The following example serves as an illustration:

Suppose R100 is given to a person who is 20 years old. The recipient has the following options:

(1) The sum can be paid out immediately.

(2) The sum of R100 is available one year later when the person attains his or her majority.

Choice (1) is more beneficial to the recipient than choice (2) because the R100 is immediately available for use or can be invested immediately so that it can earn interest. If it is invested for a year at an interest rate of six percent per annum, it will be worth R106, whereas with choice (2) the value would still be R100 after a year. In other words, the value of choice (2) is not R100 at present, since it will take a year before the investment is worth R100. The value of choice (2) is currently equivalent to an amount that would have to be invested today at an interest rate of six percent per annum in order to yield an amount of R100 in a year's time.

The position at this stage can be set out as follows:

Choice	**Sum available in cash**	**Time when sum will be available**	**Value of cash currently available**
(1)	R100	Today	R100
(2)	R100	In 1 year's time	R94,33

Compound interest can be used to calculate the future value, using the following mathematical formula (the present or current value is the converse of the future value):

Future value = Current value$(1 + i)^n$

Present or current value = Future value$/(1 + i)^n$

= R100$/(1 + ,06)^1$

= 100/1,06

= R94,33

(where i = interest rate per annum and n = number of years)

This can also be calculated as follows if we assume that X = the amount immediately available:

106% of X = R100

106/100(X) = R100

X = R100(100/106)

= R94,33

The conclusion which can be drawn is that choice (1) is the more favourable choice because it immediately represents an amount of R5,67 more than choice (2).

It is clear from the discussion so far that the present value of a sum of money which falls due in the future cannot be calculated unless a discount rate has been established. The discount in the existing example is R5,67, and the discount rate is six percent per annum.

Just as interest is used to calculate future values, a system known as discounting can be used to calculate the present values of amounts that are receivable in the future.

The following illustration is self-explanatory:

PRESENT VALUE + INTEREST = FUTURE VALUE

FUTURE VALUE − DISCOUNT = PRESENT VALUE

When we speak of the time value of money, we are referring to the following types of money:

(1) Single amounts

(a) the present value of a single amount which will be received/paid after a certain period

(b) the future value of a single amount that will be received/paid today

(2) Annuities

(a) the present value of an amount that will be received/paid per period for a certain number of periods

(b) the future value of an amount that will be received/paid per period for a certain number of periods

Basically what the above amounts to is that, because of the return that can be earned on money, a sum invested today for a certain term, at a certain interest rate, will grow into a larger sum than the initial sum invested today, and that a sum that will be received in future would be traded today at less than the future Rand value.

4.4.4 The timewise comparability of costs

To compare alternative projects in terms of transport economics, it is necessary for costs to be assessed on a collective time basis, because a continuous time value is attached to money or capital.

The greater importance attached to the current power of disposal over funds, as opposed to the later power of disposal over the same amounts, is called *time preference inclination*. This time value of money has nothing to do with inflation. Even during periods when there is no inflation, a time preference is linked to money and it is related to the average earnings that can be obtained within a community on savings and investments. Therefore if one withdraws a sum of money from all economic activities for a period and keeps it inactive, there is no opportunity for that money to grow in an alternative manner. The average time preference attached to funds can therefore be equated to this opportunity or alternative cost as reflected by the average capital yield over a period. It is only when all future values

have been expressed in equivalent or equal terms, that is when they have been reduced to a common period by means of a representative discounted rate, that one can take a decision on transport economics.

4.4.5 Techniques for economic evaluation

4.4.5.1 Introduction

Various techniques, all based on the principle of discounted cash flow, can be used in economic evaluation. In this section we shall explain four of the techniques which are most widely used:

(1) the present value of costs technique (PVOC)

(2) the net present value technique (NPV)

(3) the benefit/cost ratio technique (B/C)

(4) the internal rate of return (IRR)

These techniques can be classified into two groups on the basis of their underlying principles. For the *first* group, all that is calculated is the cost of each alternative, the argument being that the alternative with the lowest cost is the best option. The PVOC technique falls into this group. For the *second* group of techniques, both the benefits and the cost of alternatives are calculated. Benefits are defined as savings on recurring costs in relation to the null alternative. The techniques in this group depend on the assumption that an alternative is economically justified if the benefit exceeds the cost. The method used to identify the best alternative will depend on the specific technique. Three techniques fall into this group, namely the NPV, the B/C and the IRR techniques.

When mutually exclusive alternatives are compared, these techniques, if correctly applied, will all give the same answer. However, when independent projects are compared, only the B/C and the IRR techniques can be used, since the PVOC and the NPV techniques do not make provision for possible differences in scale that may exist between independent projects.

The various techniques are explained below with reference to typical situations that require economic evaluation, namely (1) the comparison of mutually exclusive alternatives, and (2) the ranking of independent projects.

(The application of these techniques is explained in section 4.5.)

In order to simplify the explanation of the techniques, we have made the following assumptions:

(1) The construction of the infrastructure takes place over a short period; the construction costs are incurred at the null point in time and the infrastructure can be used immediately.

(2) The recurring costs, indicated by the symbol T (where T = KI + KG; see sect 10.4), grows exponentially during the analytical period and at the same rate as the expected traffic growth rate.

(3) The infrastructure will have a residual value at the end of the analysis period (indicated by Res). (The concept “residual value” is discussed in sect 4.5.)

4.4.5.2 Comparison of mutually exclusive alternatives

(a) The PVOC technique

The PVOC technique expresses all cost items (CC, CM and CU) associated with a particular alternative in terms of present value. The time value of money is taken into account by discounting future costs to their present, using a discount rate. From an economic point of view the alternative with the lowest PVOC is the best option.

For any particular alternative the PVOC is calculated by means of the following formula:

$$\begin{aligned} PVOC &= \sum_{n=0}^{20} c_n/(1+i)^n \\ &= CC - PV(Res) + PV(T) \\ &= CC - Res(PV, i\%, n)] + T(E_tPV, i\%, n) \end{aligned}$$

where

C_n = the cost in year n

PV = the present value

and the other symbols have the same meaning as in the previous section.

Remember that we are dealing with an assumption that both elements of T (recurring costs) increase exponentially. Where road user costs (CU) increase exponentially, but road maintenance costs (CM) remain constant over that period, the recurring cost item should be divided into its two components because different discount rates are applicable to them. The formula could look like this:

$$PVOC = CC - Res(PV, i\%, n) + CU(E_tPV, i\%, n) + CM(SPV, i\%, n)$$

(b) The NPV technique

This technique measures the difference between the present value of benefits resulting from an investment in road infrastructure and the present value of the construction costs of the road. These benefits are represented by the saving on recurring costs which is made possible by the investment. The term "present value" implies that the time value of money is taken into account through the use of a discount rate. The answer is given in absolute monetary terms and should be equal to or greater than zero in order to be acceptable. This means that an alternative is only justified if the value of the benefits exceeds the costs. (If the value of the benefits was less than the costs, the difference between the benefits and the costs would be negative.) The alternative with the highest NPV is preferred, the argument being that this alternative would make the biggest contribution to the economic welfare of the community. Note that the NPV of an alternative can only be calculated by comparing the alternative with the null alternative – therefore the NPV of the null alternative cannot be calculated.

The NPV is calculated by means of the following formula:

$$\begin{aligned} NPV &= \sum_{n=0}^{20} b_n/(1+i)^n - \sum_{n=0}^{20} c_n/(1+i)^n \\ &= PV(T_0 - T_A) - (CC - PV(Res)) \\ &= (T_0 - T_A)(E_tPV, i\%, n) - (CC - Res(PV, i\%, n)) \end{aligned}$$

where

b_n = the benefits in year n

T_0 = the recurring costs for the null alternative

T_A = the recurring cost for the particular alternative

and the other symbols have the same meaning as in the previous section.

As with the NPV technique, the current value of CM and CU must be separately calculated when the one increases exponentially and the other does not. The formula can then be adjusted as follows:

$$NPV = (CU_0 - CU_A)(E_tPV, i\%, n) + (CM_0 - CM_A)(SPV, i\%, n) - (CC - Res(PV, i\%, n))$$

It should be self-evident that a similar adjustment will be required with all four techniques.

(c) B/C ratio

This technique measures the ratio of the present value of benefits to the present value of the construction costs of the infrastructure. The time value of money is again taken into account by using a discount rate. A project is economically justified if this ratio is equal to or greater than one. Note that, as in the case of the NPV technique, the B/C ratio of an alternative can only be calculated in relation to another alternative.

In simpler terms: an alternative is said to be economically justified if its B/C ratio is equal to or greater than one, since this implies that the benefits exceed the costs or are at least equal to the costs. If the benefits are less than the costs, the ratio will be less than one.

The B/C ratio is calculated by means of the following formula:

$$B/C = \sum_{n=0}^{20} b_n/(1+i)^n \Big/ \sum_{n=0}^{20} c_n/(1+i)^n$$

$$= PV(T_0 - T_A)/(CC - PV(Res))$$

$$= (T_0 - T_A)(E_tPV, i\%, n)/(CC - Res(PV, i\%, n))$$

where the symbols have the same meaning as in the previous section.

When the B/C ratios for the comparison of mutually exclusive projects are used, it is necessary to do an incremental analysis in order to identify the best alternative. The comparison of individual projects with the null alternative merely indicates whether a project is economically justified in comparison with the null alternative. For example, if there are two alternatives apart from the null alternative and alternative 1 shows a B/C ratio of 1,6:1, whereas alternative 2 shows a B/C ratio of 1,8:1, alternative 2 is not necessarily the best. It must first be compared with alternative 1.

To conduct an incremental analysis (ie to compare the various alternatives), the alternatives first have to be ranked on the basis of construction costs, starting with the alternative with the lowest construction costs. The alternative with the lowest construction costs then serves as the base with which the next alternatives in the ranking are compared.

For example, when the incremental (additional) investment in alternative 2 (the difference between the construction costs of alternatives 1 and 2) is justified by the incremental benefits (ie the ratio of alternative 2 to alternative 1 is greater than 1), alternative 2 becomes the base for justifying the incremental investment required by alternative 3. However, if alternative 2 cannot be justified (ie the ratio is less than one), alternative 1 remains the base on which the incremental investment required by alternative 3 must be justified. The process is repeated until all the alternatives have been compared. The remaining alternative (1) is therefore the best option.

This process is illustrated by the example in section 4.6.

(d) The IRR technique

This technique calculates the expected internal rate of return of an alternative, that is the rate of return that will cause the present value of the stream of benefits to be equal to the present value of the construction costs. The IRR of an alternative can also be defined as that rate of return at which the NPV of an alternative would be equal to zero.

In order to be economically viable, a project must show a rate of return which is equal to or higher than the discount rate. Again, mutually exclusive alternatives cannot be compared in terms of their expected rate of return relative to the null alternative, and incremental analysis should be used to identify the best alternative. As in the case of the NPV and B/C techniques, the IRR of an alternative can only be calculated relative to another alternative. This means that the IRR of the null alternative cannot be calculated.

The IRR is calculated by finding a rate (i) at which:

$$\sum_{n=0}^{20} c_n(1+i)^n = \sum_{n=0}^{20} b_n/(1+i)^n$$

therefore

$$CC - PV(Res) = PV(T_0 - T_A) \text{ and}$$

$$CC - Res(PV, i\%, n) = (T_0 - T_A)(E_tPV, i\%, n)$$

where the symbols have the same meaning as in the previous section.

Independent projects can be ranked in respect of either their B/C ratio in relation to the null alternative or their IRR in relation to the null alternative.

Activity 4.1

Towns A and B are currently linked by a gravel road which is 23 kilometres long. Every rainy season the road gets into a very poor condition, which causes a lot of dissatisfaction among road users and places an excessive burden on the local authority responsible for maintenance.

After representations had been received, funds were voted for the building of a new road, subject to the condition that an economic evaluation should be undertaken and that the best option in terms of transport economics should be chosen.

The following three alternatives are being investigated:

Alternative 0: The existing situation is retained and no new road is built.

Alternative 1: The gravel road is tarred without altering the existing road geometry in any way.

Alternative 2: A new tarred road 20 kilometres in length is built.

Either alternative 1 or alternative 2 would take two years to carry out. All the initial costs (planning, expropriation and building costs) are paid on the day on which the road is opened. The day on which the road is opened is regarded as “year nil”. The analysis period is 20 years (that is, up to the end of year 20). On the basis of a communication from the Treasury, a real discount rate of 6 percent is used.

The road authority has requested you to evaluate the project in economic terms and has given you the following information:

Alternative	**0**	**1**	**2**
Initial costs	0	R 4 000 000	R6 000 000
Road maintenance costs	R220 000	R55 000	R47 800
Road user costs	R1 300 000	R1 040 000	R900 000
Residual value of road	0	R1 000 000	R1 600 000

Notes

(1) Annual road maintenance costs remain constant throughout the analysis period in respect of all three alternatives.

(2) As a result of traffic growth, road user costs are expected to increase geometrically (ie exponentially) at a rate of 4 percent per annum in respect of all three alternatives.

(3) No differential inflation is expected.

Required

(1) Evaluate the available alternatives with reference to the following techniques:
   (a) present value of costs
   (b) net present value of benefits
   (c) benefit/costs ratio
   (d) incremental benefit/cost ratio between alternatives 2 and 1

(2) Explain what the aim(s) of each technique are.

(3) Recommend an alternative on the basis of your analysis.

Hints

Discount all future amounts to year 0 values. This makes the calculations easier, especially because road maintenance costs are a simple progression and road user costs an exponential progression.

(A similar evaluation is carried out in section 4.4.)

4.5 The application of economic evaluation: an example

As we mentioned previously, mutually exclusive alternatives must first be analysed before independent projects can be ranked in order of merit. This is illustrated in the example given below.

4.5.1 The comparison of mutually exclusive alternatives

4.5.1.1 Problem statement

A section of road between Tzaneen and Polokwane cuts across a very mountainous area. The section, which links point A and point B, is five kilometres in length and has a maximum gradient of six percent. It is in a poor condition and does not comply with modern design standards. Engineers have proposed the following schemes for improving the road:

Alternative 1: The same alignment as for the null alternative is followed, but the maximum gradient is reduced to 3 percent.

Alternative 2: A new horizontal and vertical alignment with a maximum gradient of 6 percent is used, which reduces the distance to 3,5 kilometres.

Alternative 3: The same alignment as in alternative 2 is used, but the maximum gradient is reduced to 5 percent.

Data relating to these alternatives are summed up in the table below.

Table 4.1

Data relating to alternatives in the example

	Mutually exclusive alternative			
	0 (R'000)	1 (R'000)	2 (R'000)	3 (R'000)
CC	0	727,8	4 693,0	4 937,2
T (in base year)	1 707,7	1 596,0	1 145,3	1 129,1
Residual	0	145,6	938,6	987,4

The roads authority requests that these alternatives should be compared over an analysis period of 20 years, on the assumption that the traffic growth rate is five percent per annum and the discount rate six percent.

Both road maintenance costs and road user costs are increasing exponentially relative to the traffic growth rate.

4.5.1.2 *Comparison of alternatives*

A simple method of comparing the alternatives is first to discount all future amounts to year-zero values and then simply to apply the formulas. These discounted amounts are reflected in table 4.2 and were determined as follows:

(1) *Construction costs*. These remain the same as the original amount, because they were incurred in year zero.

(2) *Recurring costs*. Since this is an exponential series which must be discounted to current value, the factor is obtained from appendix 3 at a growth rate of 5 percent per annum, a discount rate of 6 percent per annum and a period of 20 years. This factor is 18,1324. The recurring cost of each alternative is then multiplied by this factor.

(3) *Residual value*. The residual value is a single amount supplied for year 20. The factor for the present value of R1 after 20 years at a discounted rate of 6 percent is given in the PV column of appendix 1 as 0,3118. The residual value of each alternative should therefore be multiplied by this factor to calculate the discounted residual value.

Table 4.2

Discounted amounts

	Factor	Alternative			
		0 (R'000)	1 (R'000)	2 (R'000)	3 (R'000)
CC	—	0	727,8	4 693,0	4 937,2
T	18 1324	1 707,7	1 596,0	1 145,3	1 129,1
Residual	0,3118	0	145,6	938,6	987,4

Since the amounts have already been discounted, simplified formulas can be used for the various techniques.

(a) Present value of costs

NPV = CC − Res + T

Alt 0: NPV = 0 + 30 964,7

Alt 1: NPV = 727,8 − 45,4 + 28 939,3

= 29 621,7

Alt 2: NPV = 4 693,0 − 292,7 + 20 767,0

= 25 167,3

Alt 3: NPV = 4 937,27 − 307,9 + 20 473,3

= 25 102,6

Note that the formula given in section 4.4.5.2 for this technique is exactly the same as here, except that the present value has already been calculated and substituted into the formula. For example, take alternative 2:

NPV = CC − Res(PV, i%, n) + T (E_tPV, i%, n)

= 4 963,0 − 938,6 (0,3118) + 1 145,3 (18,1324)

= 25 167,3

The above remarks apply to all the techniques.

(b) Nett present value

NPV = $(T_0 - T_A) - (KK_A - Res_A)$

Alt 0: = not applicable

Alt 1: NPV = (30 964,7 − 28 939,3) − (727,8 − 45,4)

= 1 343,0

Alt 2: NPV = (30 964,7 − 20 767,0) − (4 693,0 − 292,7)

= 5 797,4

Alt 3: NPV = (30 964,7 − 20 473,3) − (4 937,2 − 307,9)

= 5 862,1

(c) Benefit/cost ratio

B/C = $(T_0 - T_A) / (CC_A - Res_A)$

Alt 1: B/C = (30 964,7 − 28 939,3) / (727,8 − 45,4)

= 2,97

Alt 2: B/C = 2,32

Alt 3: B/C = 2,27

Note that the benefit/cost ratio is expressed with reference to the null alternative. As we have already mentioned, an incremental analysis should be done to find the most economic alternative.

(d) Incremental B/C

First arrange the alternatives according to their construction costs.

	Construction costs (R1 000)
Alternative 0	0
Alternative 1	727,8
Alternative 2	4 693,0
Alternative 3	4 937,2

The calculation B/C 1,0 has already been done and it has produced a B/C ratio that is greater than one. Alternative 2 therefore now has to be compared with alternative 1. This is done by dividing the saving on recurring costs which is produced by implementing alternative 2 by the additional or incremental initial costs attached to it.

Therefore

$$B/C_{2,1} = (T_1 - T_2) / [(CC_2 - Res_2) - (CC_1 - Res_1)]$$

$$= (28\,939,3 - 20\,767,0) / [(4\,693,0 - 292,7) - (724,4 - 45,4)]$$

$$= 8\,172,3 / 3\,717,9$$

$$= 2,20$$

Although the B/C ratio of alternative 1 is relatively greater than that of alternative 2 when they are both compared with the null alternative, alternative 2 is more economic because it still produces a ratio greater than one when compared with alternative 1.

In accordance with the explanation of the B/C ratio alternative 2 should now be used as the base for justifying the additional investment which alternative 3 requires.

Therefore

$$B/C_{3,2} = (T_2 - T_3)/[(CC_3 - Res_3) - (CC_2 - Res_2)]$$

$$= (20\,767,0 - 20\,473,3)/(4\,937,2 - 307,9) - (4\,693,0 - 292,7)$$

$$= 293,7/229,0$$

$$= 1,28$$

(e) Internal rate of return

A trial and error method is used to determine the internal rate of return. A project is economically justified only if it produces a rate of return which is equal to or higher than the discount rate.

We therefore need to find a rate (i) where:

$$CC_A - Res_A(PV, i\%, n) = (T_0 - T_A)(E_tPV, i\%, n)$$

For alternative 1: Begin with a rate of 15 percent:

$$727,8 - 145,6(0,0611) < (1\,707,7 - 1\,596,0)(8,7978)$$

$$718,9 < 982,71$$

Because $T_0 - T_1 > CC_1 - Res_1$ a higher rate should be used. Therefore, use 20 percent:

$$727,8 - 145,6\ (0,0261 < (111,7)(6,5155)$$

$$724,0 < 727,78$$

$T_0 - T_1 > CC_1 - Res_1$. It is therefore necessary to use an even higher rate. Try 25 percent:

727,8 − 145,6 (0,0115) > 111,7 (5,0894)

726,1 > 568,49

It would therefore appear that the IRR of alternative 1 is about 21 percent. (Since our tables are not so complete we shall do an estimate.)

For alternative 2: IRR = ±15%

For alternative 3: IRR = ±15%

As in the case of the B/C ratio, an incremental analysis also has to be conducted. The results are as follows:

$IRR_{2,1}$ = ± 16%

$IRR_{3,2}$ = ± 8,5%

4.5.1.3 Conclusion

The comparison of the alternatives shows that alternative 3 is the most economic, for the following reasons:

(1) It has the lowest present value of costs.

(2) It has the highest net present value.

(3) Although it produces the lowest B/C ratio when all three alternatives are compared with the null alternative, the incremental B/C analysis shows that this is the most advantageous alternative.

(4) The incremental IRR analysis also shows that the rate of return would be about 8 percent when alternative 3 is compared with alternative 2, which is therefore higher than the discount rate.

It is true, however, that the “best” alternative from the point of view of the community is not necessarily the best from the point of view of the authority concerned. If the authority's point of departure is that economy is necessary or that expenditure should be cut, the null alternative would be chosen, since it involves no construction costs. However, if the public insists that “something” should be done about the road, the government might choose alternative 1, since after the null alterative it is the one that requires the least expenditure. It has been proved, however, that alternative 3 is the best option and the authority in question, which is acting on behalf of the community, should be prepared to opt for this alternative. Although the above example is an oversimplification of the problem, it does illustrate an important principle. A discussion of the line that should be taken in cases which justify a deviation from this principle does not fall within the scope of this study guide.

4.5.2 The ranking of independent projects

The ranking of independent projects presupposes that the mutually exclusive alternatives for every “undesirable situation” have already been compared and that the best alternatives have been chosen from each of these “sets of mutually exclusive alternatives”. The following step involves the ranking of these “best” alternatives, which are now treated as independent projects, according to economic merit, for example in terms of their IRR relative to the null alternative. Projects can then be selected for implementation, starting at the top of the list, until all the funds have been exhausted. This is illustrated in table 4.3

Table 4.3

The ranking of independent projects		
Project No	IRR	R
C	27	1.3
A	24	2.7
D	23	4.5
G	22	20.3
S	18	0.7
T	17	5.5
-	-	<= 35.0
B	16	17.2
Z	15	3.3
-	-	-
-	-	-
-	-	-
C	9	1.7
E	7	7.9
-	-	110.0

If sufficient funds are available (ie R110 million), all the above projects can be undertaken since they are all economically justified. (Note that only economically justified projects will be put on this list.) If only R35 million is available, then only projects C, A, D, G, S, and T should be selected, since this combination of projects would best satisfy the objective of maximising the net benefits to the community.

Lastly, it is important to note that the “optimum” choice of projects mentioned above has been obtained by applying a “limited” criterion of economic efficiency, since only certain costs and benefits were included in the analysis. Other important factors, such as benefits of a macroeconomic nature and strategic and political considerations, were not taken into account. The inclusion of these considerations could result in a different “optimum” combination of projects. A discussion of the inclusion of considerations of this kind falls outside the scope of this course.

4.6 Summary

The roads authority, which is regarded as the community's agent, should always act in the best interests of the community, not only when faced with choices between mutually exclusive alternatives, but also when independent projects have to be selected for implementation. The nature and purpose of the economic evaluation of roads are described in this chapter. We also explained how this evaluation can make a useful contribution to the official decisionmaking process, and so help to promote the economically efficient allocation of scarce resources, while every effort would still be made to provide adequate roads.

4.7 Self-evaluation questions

(1) Explain in detail what is meant by “the optimal allocation of factors of production”.

(2) How is the user surplus determined by graphical means?

(3) Give a full discussion of the different criteria used in project evaluation.

(4) Explain the role of the time value of money when project evaluation is carried out.

(5) Define the various techniques that are used in the economic evaluation of projects.

STUDY UNIT 5

Multicriteria analysis

UNIT OUTCOMES

After working through this study unit you should be able to:

- define consequences, dimensions and criteria as used in this study unit
- explain how elementary consequences, dimensions and criteria are identified
- explain the process of pre-multicriteria analysis
- represent multicriteria analysis schematically

KEY CONCEPTS

- Consequences
- Dimensions
- Criteria
- Multicriteria
- "Unambiguously monetisable effects synthetic criterion" (UMESC)

5.1 Introduction

(This study unit is based on the article by De Brucker, De Winne, Peeters, Verbeke & Winkelmans 1995.)

Analytical methods such as social cost-benefit analyses, economic impact studies, environmental impact assessment and traditional multicriteria analysis are sometimes seen as mutually exclusive appraisal methods. A social cost-benefit analysis measures the economic (monetary) welfare effect of a project, while an economic impact study measures economic development in terms of, among other things, value added, the creation of job opportunities and economic growth. An environmental impact study in turn measures the influence that a project will have on a specific environment, while a traditional multicriteria study determines the most acceptable or optimal solution according to a number of criteria.

However, the above methods can be used to complement one another. The results of the methods can be integrated and used to conduct an exclusive multicriteria analysis. This exclusive analysis differs from the traditional multicriteria analysis in two respects. First,

the output of the other analytical methods is used as the input and, secondly, the exclusive multicriteria analysis helps to answer the fundamental question about demand, namely how desirable the specific project is. In addition, the exclusive multicriteria analysis is characterised by a high degree of transparency because decisionmakers can understand it easily.

5.2 Choice of criteria and dimensions: the theory

5.2.1 Definitions of consequences, dimensions and criteria

Any evaluation model should be based on the results of specific actions and on a value judgment of these results by the parties concerned. Hence the establishment of a set of criteria starts with a study of the results of actions that are regarded as relevant.

In practice, however, only those consequences that can be clearly identified and defined should be taken into account. These are referred to as the elementary consequences. Any elementary consequence that is evaluated on the basis of a preference scale is referred to as a preference dimension, or simply a dimension.

The evaluation of one action according to the preference scale is known as a score or an indicator. A score may be complemented by a dispersion indicator which measures the probability that the corresponding score will be obtained (say, by means of a probability distribution).

The corresponding dimensions provide the basis for determining the final criteria. A criterion implies the evaluation of a certain action on the basis of one or more dimensions, by using a function. Obviously this function corresponds with a specific point of departure.

Any consequence and dimension has an objective meaning, while a criterion is characterised by a subjective meaning since the value function of one or more individuals (or a community) associated with the specific consequences or dimensions is indicated.

There is no properly defined method that can be used to link criteria and dimensions. One possibility is to use the elementary consequences and dimensions as one's point of departure, and then to establish criteria on the basis of these dimensions. In practice, criteria (on the basis of dimensions) can be established in the following three ways:

(1) A criterion can be related to a corresponding dimension. If the result is expressed in terms of a quantitative preference scale, it has to be encoded in terms of utilisation. In cases where it is expressed according to a qualitative scale, it must also be expressed according to a quantitative scale (ie 1–20), or, if possible, according to a quantitative scale that can be converted to a monetary scale, as in the case of social cost-benefit analyses.

(2) In certain cases, where the elementary results correspond with the dimensions, there is too much information to form a single criterion. A solution to this problem would be to use the excess information to determine two or more criteria, a procedure that is referred to as splitting the dimension. Thus provision is made for all the available information. One disadvantage, however, is that the splitting process increases the complexity of decisionmaking.

(3) A single criterion can be linked to more than one dimension. Here the number of criteria decreases and decisionmaking becomes less complex and more transparent. However, the following two conditions need to be met in order to establish criteria in this format:

- The dimensions that are synthesised must correspond to elementary consequences which complement each other.
- Each participant in the decisionmaking process must accept the value implications of the synthesis.

5.2.2 Establishing a set of criteria

When a set of criteria is established on the basis of dimensions, it is imperative to evaluate the quality of the set. A set of criteria should meet three requirements, namely exhaustivity, coherence and independence. As far as the last requirement is concerned, there are two types of interdependence. Firstly, interdependence may occur in a structural or statistical relationship between two criteria, and secondly, it may be preferential. The first type of interdependence does not cause problems if the relevant criteria (between which a structural or statistical relation of this kind exists) correspond to different opinions or influence the different participants in the decisionmaking process. The second kind of interdependence may be problematic and therefore requires careful investigation.

5.3 A multicriteria evaluation of investment in transport infrastructure: the analytical phase

5.3.1 Identifying the elementary consequences

Any effect or characteristic that influences the goal of at least one participant in the decisionmaking process can be regarded as a consequence. A survey of the sum total of the elementary consequences of transport infrastructure can be conducted in the following four ways:

(1) Use your common sense or identify the transport policy objectives.

(2) Re-examine any completed projects.

(3) Consider the legal requirements (if there are any).

(4) Look at similar consequences in other countries.

The consequences of investment in transport infrastructure can be subdivided into three main categories, namely monetary effects, environmental and safety effects and socio-economic effects.

5.3.1.1 Monetary effects

The following monetary effects can be identified:

- construction, maintenance and repair costs
- vehicle operating cost
- the gain of travel time enjoyed by users of the facility
- a change in the demand for traffic (traffic generation)
- the consequences for other transport modes
- the effect on the value of property (including possible losses experienced by traders during the construction phase)
- economic activities resulting from the project (transparent value added)
- revenue such as tolls, if applicable

5.3.1.2 *Socioeconomic effects*

The following socioeconomic effects can be identified:

- the redistribution of income between regions
- the redistribution of income among different categories of road users
- the redistribution of income among socioeconomic groups
- contributions to the creation of job opportunities

5.3.1.3 *Environmental and safety effects*

The following environmental and safety effects should be taken into consideration:

- noise pollution
- air pollution
- safety
- the effect on material assets and cultural heritage
- the effect on the quality of the environment such as visual intrusion and the breaking up of communities

Some of the consequences identified in the third category can be transferred to the first category if the decision-makers agree that monetary values can be assigned to these consequences and if there is consensus about the value appraisal procedures.

5.3.2 Identifying the dimensions

A dimension implies linking an elementary consequence to a preference scale. The following dimensions serve merely as examples of dimensions which are either included or excluded on the basis of the type of investment project being considered.

The monetisable effects, namely construction costs (CC), maintenance costs (MC), exploitation costs (EC), repair costs (RC), vehicle operating costs (VOC), cost implications for other transport modes (IOTM), changes in the value of property (VP) and financial receipts (FR) are measured according to a monetary scale. All these effects represent different dimensions.

Gain in travel time (GT) for users does not in itself represent a dimension because the scale used, namely the number of seconds, minutes, and so on, does not represent a preference scale. The value that different users attach to this differs according to specific categories of users. If gain in travel time can be classified according to the different categories, this classification can be used as a dimension.

The project's implications for economic development (ED) are evaluated in terms of value added (contribution to GDP) which can obviously be expressed in monetary terms. Because of the measurability of this implication, it is represented by a dimension. This information is obtained from an economic impact study.

Regarding the socioeconomic effect of a project, it is argued that the effect of, say, income distribution cannot be used as an elementary consequence. This is because it is difficult to define such an effect and it cannot therefore be tabulated according to a preference scale. Such a preference scale is possible only if an empirically measurable function of

the marginal utility of income can be developed. In addition, the redistribution of income in industrial countries is regarded as irrelevant (De Brucker et al 1995:266).

The effect of the project that contributes to the creation of job opportunities does in fact represent a dimension because it can be measured according to the number of jobs created directly or indirectly. The information can be obtained from an economic impact study.

Lastly, the environmental and safety effects of the project can be subdivided into noise pollution (NP), air pollution (AP), visual intrusion and community severance (VI + CS) and damage to cultural heritage (CH). Each of these effects can be represented in terms of a dimension. Visual intrusion, community severance and damage to cultural heritage can only be evaluated quantitatively or defined in physical terms by environmental impact experts. Noise and air pollution are measured in decibels (dB) and CO_2 respectively.

Safety implications can be subdivided in a similar fashion – in other words, as casualties where there is material damage only, slight injury casualties, serious injury casualties and fatal casualties. In the case of casualties where there is material damage (MAC) only, this can be expressed in monetary terms and therefore placed in the category of unambiguously monetisable effects.

This classification also applies to all material damage that occurs in the other categories of casualties. Thus only the following safety-related dimensions are taken into account:

- slight injury (SLIC)
- serious injury casualties (SEIC)
- fatal injury casualties (FIC)

To the extent that there is consensus among decisionmakers, monetary values can be attached to these categories.

5.3.3 Identifying the criteria

As far as the unambiguously monetary effects are concerned, the dimensions of constructions costs (CC), maintenance costs (MC) and repair costs (RC) can be placed in a single (synthetic) criterion. The following conditions apply:

- The dimensions have to be synthesised with the elementary results which complement one another.
- The value implications of synthesis should be acceptable to each participant in the decisionmaking process, provided that the conditions set in section 5.2.1.3 are adhered to.

The following dimensions in this category are also assessed in monetary terms:

- vehicle operating costs (VOC)
- the implications for other transport modes (IOTM)
- changes in the value of properties (VP)
- financial receipts (FR)
- the material aspects of casualties (MAC)

These dimensions, however, differ from the previous three, because different participants can be affected, depending on each dimension. If the values of these dimensions are determined on the strength of their monetary values, and the resulting compensation among the dimensions is acceptable, these dimensions can be classified under the synthetic criterion as outlined above.

The same applies to dimensions relating to increasing the gain in travelling time (GT). However, one condition is that the monetary value for the increase in the gain in travelling time should be correctly determined for each category. Another important consideration is that any possibility of a double score should be avoided. For example, the financial receipts of a transport infrastructure operator may constitute a significant part of the gain in travel time which is to the advantage of users.

The performance of a project, according to the synthetic criterion, can be obtained by adding up the monetary value for each dimension. The effects which are scattered over time can be discounted to the same period, as explained in study unit 4. The criterion can be referred to as the "unambiguously monetisable effects synthetic criterion" (UMESC). In the case of the budgetary restraint, the UMESC values of projects tie in with the total capital cost (C). Use of the UMESC/C permits the maximisation of the monetisable effects for each investment unit. Except for two differences, these criteria are similar to those of the net present value technique, as discussed in study unit 4. The first difference is that the environmental and safety effects are not included in the criteria of the UMESC except when there is absolute consensus among the decision-makers about the monetary effect of each value. Secondly, no adjustments are made for market imperfections except in respect of the first difference, namely that there should be absolute consensus among decisionmakers. In other words, when the UMESC/C is used, the actual preferences of the decisionmakers are taken into account. External effects and other market imperfections are not given a monetary value on the strength of the "perceptions" of welfare economists, except when there is total consensus about exactly what the monetary value should be. If there is any doubt about this evaluation, the monetary value should be assessed according to other (nonmonetary) criteria.

The dimension of economic development can be expressed in terms of value added (VA) generated by the project. This value added should also be linked to invested capital. As far as high budget deficits are concerned, public decision-makers are not only interested in the value added resulting from the project, but also in the portion of value added that may flow back to the government (BFG). Two criteria are therefore used. The first [(VA–BFG)/C] represents the value added generated by the project minus the portion that flows back to the government, divided by the total investment cost (also known as the total cost of capital). The second criterion (BFG/C), includes the value added that flows back to the government, divided by the total cost of investment.

In the case of the socioeconomic effects of the project, and more specifically job opportunities, the dimension can be represented by the number of job opportunities (NJ) generated by the project. Only new job opportunities are relevant here. Job opportunities in previous projects which are simply transferred to the new project, are not taken into consideration because they are regarded as existing job opportunities. As in the case of economic development, the number of jobs created should be linked to capital investment.

Hence the criterion (NJ/C) reflects the number of jobs created, divided by the total investment costs.

At this stage it is not possible to formulate criteria for the environmental and safety effects. Experts in environmental impact studies can provide qualitative and quantitative

observations. However, these observations do not provide information on the utility associated with environmental and safety effects that can be used in the decisionmaking process.

5.4 A multicriteria evaluation of investment in transport infrastructure: the synthetic phase

5.4.1 Methodological foundations

It is imperative to follow a suitable aggregation procedure when conducting a multicriteria evaluation. Three such procedures can be distinguished, namely complete aggregation, partial aggregation and local (iterative) aggregation. Because projects that relate to transport infrastructure are generally discrete, local aggregation is usually not appropriate. Hence a choice needs to be made between complete and partial aggregation.

If one opts for a complete aggregation procedure, one assumes that the criteria will be completely intercomparable, so that trade-offs can be make. In the case of a partial aggregation procedure, the requirements for transivity, complete ranking and transparency are relaxed. Partial aggregation seldom involves complete ranking. As a rule, an a-classification is obtained. (The purpose of an a-classification is to identify the best action out of a group of actions.) Some of these methods do in fact result in a g-classification, but they are extremely complex methods. (In the case of a g-classification, the various actions are ranked according to preference.) These methods sometimes lack transparency and transitivity, and therefore have to be artificially imposed.

Complete aggregation multicriteria methods do in fact give rise to complete ranking. The final preference relation is transitive. However, the dimensions and criteria that are taken into consideration in the evaluation of the transport infrastructure are not so heterogeneous as to render the partial aggregation procedure more appropriate. In fact many dimensions tend to be homogeneous because they are expressed in monetary values and are synthesised in a single synthetic criterion. Furthermore, the dimensions can be synthesised hierarchically so that they focus on a common goal, such as the economic welfare of the community. Of course trade-offs between different sub-elements of a common goal are also possible.

The belief that multicriteria analysis with complete aggregation is the most appropriate method of evaluating investment in a transport infrastructure is further reinforced by ongoing research in Europe where various member countries of the European Union are using multicriteria analysis (De Brucker et al 1995:270).

Multicriteria analyses have the following disadvantages:

- Excessive trade-offs can occur between the high scores of certain criteria and the low scores of others – in other words, the possibility of trading off the disadvantage of a certain characteristic against a major benefit of another characteristic.
- From a welfare point of view, the desirability or suitability of a project (as discussed in study unit 2) cannot be explicitly addressed. A possible solution to this problem would be to include “desirability criteria” in the multicriteria analysis. Such criteria would ensure that a minimum value is obtained in order to evaluate the specific criterion as desirable. The multicriteria method developed in the next section does in fact include such desirability criteria.

5.4.2 Applying a multicriteria analysis to transport infrastructure investments

5.4.2.1 A pre-multicriteria analysis

A pre-multicriteria analysis is important for the following three reasons:

(1) A pre-analysis prevents excessive trade-offs between the high scores on certain criteria and low scores on others. In practice, however, limited trade-offs between good and poor scores are in fact acceptable, and only excessive trade-offs should be avoided (De Brucker et al 1995:272). Hence a pre-multicriteria analysis can be used to eliminate projects with a number of unacceptable negative scores that cannot be replaced by positive scores on other criteria.

(2) The aim of the pre-multicriteria analysis as discussed below is to ascertain whether or not the project is desirable. As a rule multicriteria analyses do not pursue this goal because they only determine *where* a particular airport should be built, *what* railway lines and stations should be constructed and so on. The desirability of the project is thus regarded as obvious.

In the case of a pre-multicriteria analysis, the project can be regarded as desirable if the UMESC is strictly positive and if the nonmonetary external effects are within acceptable bounds.

(3) The pre-multicriteria analysis makes provision for the effective implementation of nonlinear weights. If a project does not pass the pre-multicriteria test, this implies that the weight(s) that is (are) allocated to criteria are infinite.

A pre-multicriteria analysis is based on a ß-classification (selecting all the actions that appear “good”), which divides the group of actions (projects) into three categories:

(1) K_1, a category with alternatives whose desirability is incontestably established.

(2) K_3, a category with alternatives whose undesirability is incontestably established.

(3) K_2, a category with alternatives whose desirability requires further study.

Once these studies have been conducted, the various K_2 projects can be added to the desirable K_1 projects or undesirable K_3 projects.

The desirability of a project is confirmed if the UMESC is strictly positive and if, at the same time, all the negative external effects are below the lowest ceiling. This implies that K_1 represents all those projects that simultaneously include the following:

- unambiguously monetisable effects synthetic criterion (UMESC) > 0 and
- slight injury casualties (SLIC) $\leq x_2$ and
- serious injury casualties (SEIC) $\leq x_3$ and
- fatal injury casualties (FIC) $\leq x_4$ and
- noise pollution (NP) $\leq x_5$ and
- air pollution (AP) $\leq x_6$ and
- visual intrusion & community severance (VT + CS) $\leq x_7$ and
- cultural heritage (CH) $\leq x_8$

The undesirability of a project is confirmed if the unambiguously monetisable effects synthetic criterion (UMESC) is negative and one or more external effects exceed the upper

ceiling. This implies that K_3 represents all those projects that simultaneously include the following:

- unambiguously monetisable effect synthetic criterion (UMESC) ≤ 0 or
- slight injury casualties (SLIC) > y_2 or
- serious injury casualties (SEIC) > y_3 or
- fatal injury casualties (FIC) > y_4 or
- noise pollution (NP) > y_5 or
- air pollution (AP) > y_6 or
- visual intrusion and community severance (VI + CS) > y_7 or
- cultural heritage (CH) > y_8 or

K_2 represents all the other projects – in other words, all those projects whose negative external effects are within the upper ceiling (y_2 ... y_8), but of which at least one negative external effect does not respect the lower ceiling (x_2 ... x_8). In mathematical terms this means that all the projects simultaneously meet the following requirements:

- unambiguously monetisable effects synthetic criterion (UMESC) > 0 and
- slight injury casualties (SLIC) ≤ y_2 and
- serious injury casualties (SEIC) ≤ y_3 and
- fatal injury casualties (FIC) ≤ y_4 and
- noise pollution (NP) ≤ y_5 and
- air pollution (AP) ≤ y_6 and
- visual intrusion and community severance (VI + CS) ≤ y_7 and
- cultural heritage (CH) ≤ y_8

and

x_2 < slight injury casualties (SLIC), or

x_3 < serious injury casualties (SEIC), or

x_4 < fatal injury casualties (FIC), or

x_5 < noise pollution (NP), or

x_6 < air pollution (AP), or

x_7 < visual intrusion and community severance (VI + CS), or

x_8 < cultural heritage (CH)

where: $K_1 \cup K_2 \cup K_3 = A$

$$\forall i = 1...8: x_i < y_i$$

Environmental impact analysts should determine the values x_2 ... x_8 and y_2 ... y_8, which represent the upper and lower ceilings respectively. Decisionmakers' preferences should be taken into consideration in the decisionmaking process. The ceilings of these values may vary over time, depending on the project or characteristics of the specific sector. In this case, a solution to the desirability problem would be to determine a *threshold*

value (y_1) for the unambiguously monetisable effects synthetic criterion (UMESC). If the UMESC reaches this *threshold value*, one could argue that the unambiguously monetisable effect outweighs the negative external effects. Thus the project will shift from category K_2 to category K_1, which now becomes the new category K_1^+. If this does not happen, the project will move to category K_3, which will become the new category K_3^+. Category K_2 now contains no project.

The value y_1, which plays a significant role in the final decision about whether a project is desirable and varies from one project to the next, needs to be determined by the decision-makers themselves. However, consultations with specific analysts in this regard are essential. In other words, decision-makers themselves must themselves decide whether the value of the UMESC exceeds the value of the negative effects in order to make a final decision about the desirability of the project.

Projects that are placed in category K_3^+ should no longer be taken into account, and a ranking for this category is therefore no longer necessary. If there are no budgetary constraints, all the projects in category K_1^+ should be implemented because they are all desirable and will make a positive contribution to the economic welfare of the community. However, if there are budgetary constraints, the projects in K_1^+ should be ranked according to the g-classification.

5.4.2.2 Definition of the hierarchy

The dimensions and criteria that should be considered in the multicriteria analysis were discussed in sections 5.3.2 and 5.3.3 respectively and are indicated in figure 5.1, which is based on Saaty's (1980) “Analytic Hierarchy Process” (AHP).[1] All the criteria are included in figure 5.1. The results of the various alternatives can be included in an evaluation matrix, which can be used as an instrument for the pairwise comparison of alternatives according to Saaty's scale.

The overarching objective of transport infrastructure projects is to generate economic benefits for a community, which is the focus of the multicriteria analysis. The second-level elements represent the subobjectives that represent the various aspects of economic benefits for the community. The dimensions are at the third level of the hierarchy as discussed in section 5.3.2.

Elements at the fourth level such as (VA-BFG)/C, BFG/C and NJ/C, which we discussed in section 5.3.3, are “real” in the sense that they use the utility function to summarise the contribution of each alternative to one or more dimensions. The relative priority allocated to each alternative in terms of each of these criteria can be determined automatically, by normalising the value of each instead of constructing a matrix of pairwise comparisons and determining an eigenvector. In terms of these criteria, the pairwise comparisons using Saaty's scale is obsolete.

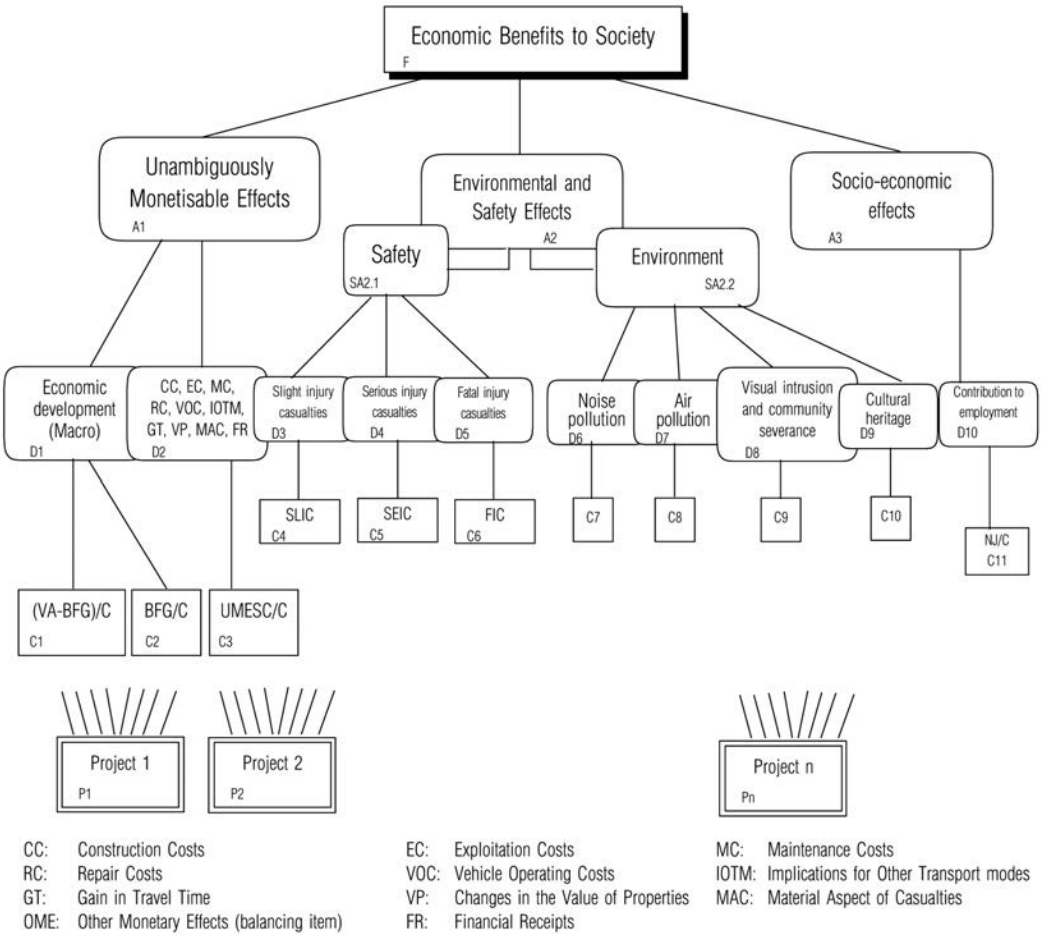
As explained in section 5.3.3, no “real” criteria that coincide with environmental and safety dimensions are constructed. The relative priority of each alternative, according to these dimensions, should be determined by means of pairwise comparisons, using Saaty's scale. A similar pairwise comparison should also be undertaken to evaluate the relative contribution of each dimension to the subobjectives of the environmental and safety impacts. Lastly, a pairwise comparison is necessary to determine the weight of the contribution of each subobjective to the focus, namely the economic benefit of the community. These pairwise comparisons can be based on the quantitative and qualitative information that is available. The core of these elements shows that they do not represent a criterion. A pairwise comparison of the alternatives (using the Saaty scale), according to each element,

1 The AHP method involves complex mathematical computations. The theoretical foundation of this is important and is set out at the end of this study unit (Saaty 1986:841–844).

shows implicitly the utility function that can be related to the relevant dimension. According to the conventional Analytic Hierarchy Process, all relative priorities are determined according to such a pairwise comparison.

Figure 5.1

The economic benefits of transport infrastructure projects



Source: De Brucker et al (1995:276)

5.4.2.3 *The uniformity of the process*

The uniformity of the results derived from the pairwise comparisons can be controlled according to the eigenvector developed by Saaty (1980:20–25, 49–51 & 83–84; Saaty 1986:850). The uniformity test for criteria such as (VA–BFG)/C, BFG/C and UMESC/C is redundant if the relative priorities of the alternatives are determined by means of normalisation techniques.

In terms of the uniformity of the total hierarchy as depicted in figure 5.1, only the uniformity of the results determined by pairwise comparisons should be taken into account.

5.5 Practical application of the multicriteria evaluation

The multicriteria evaluation methods discussed thus far are extremely flexible. The dimensions and criteria taken into account can be adapted according to the types of project and

the decisionmakers' objectives. A number of criteria such as the UMESC/C which represent the influence of the project on the environment should always be taken into consideration. In the case of smaller projects, the criteria affecting economic development and employment can be excluded, while in the case of larger projects, in sea and rail transport in particular, these criteria should always be considered.

The multicriteria method in this study unit can be applied for two purposes, namely:

- to determine the weights of the criteria
- to rank the desired project

Here the number of projects should be limited to no more than 15. The pairwise comparison of about 15 projects, in respect of the contribution of each one to each criterion, could become extremely complex. However, if one is involved in a large number of projects, a weight can be determined for each criterion. To determine the priority of desirability, it is necessary to use multicriteria of total aggregation. Further research should also be conducted to ascertain the most appropriate method to use.

5.6 Conclusion

The multicriteria method discussed in this study unit can be used to evaluate complex transport infrastructure projects. The method comprises two main steps, namely determining the relevant criteria and then aggregating them. First, a b-classification is used, which entails eliminating undesirable projects from the perspective of “economic benefits for the community”. Secondly, all desirable projects (the g-classification) are ranked according to Saaty's *Analytic Hierarchy Process*.

The multicriteria method used in this study unit differs from the more conventional one because this method emphasises the importance of the desirability of a specific project – in other words, whether or not the project does in fact make a net contribution to the economic benefits for society. Furthermore, excessive compensation becomes impossible because of the use of the minimum scores obtained for specific criteria in the pre-multicriteria analysis. However, the actual multicriteria analysis is characterised by a total aggregation procedure.

5.7 Self-evaluation questions

(1) Define the terms “consequences”, “dimensions” and “criteria”.

(2) The consequences of transport infrastructure investments can be subdivided into three categories. Explain the identification of elementary consequences in terms of these categories.

(3) Explain the identification of dimensions and consequences.

(4) Discuss fully the synthesisation phase of the multicriteria evaluation of transport infrastructure.

(5) What are the application possibilities of multicriteria analysis in practice?

MANAGEMENT SCIENCE
Vol. 32, No. 7, July 1986
Printed in U.S.A.

AXIOMATIC FOUNDATION OF THE ANALYTIC HIERARCHY PROCESS*

THOMAS L. SAATY
Graduate School of Business, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

This paper contains an axiomatic treatment of the Analytic Hierarchy Process (AHP). The set of axioms corresponding to hierarchic structures are a special case of axioms for priority setting in systems with feedback which allow for a wide class of dependencies. The axioms highlight: (1) the reciprocal property that is basic in making paired comparisons; (2) homogeneity that is characteristic of people's ability for making comparisons among things that are not too dissimilar with respect to a common property and, hence, the need for arranging them within an order preserving hierarchy; (3) dependence of a lower level on the adjacent higher level; (4) the idea that an outcome can only reflect expectations when the latter are well represented in the hierarchy. The AHP neither assumes transitivity (or the stronger condition of consistency) nor does it include strong assumptions of the usual notions of rationality. A number of facts are derived from these axioms providing an operational basis for the AHP. (CHOICE MODELS)

1. Introduction

The basic problem of decision making is to choose a best one in a set of competing alternatives that are evaluated under conflicting criteria. The Analytic Hierarchy Process (AHP) provides us with a comprehensive framework for solving such problems. It enables us to cope with the intuitive, the rational, and the irrational, all at the same time, when we make multicriteria and multiactor decisions. We can use the AHP to integrate our perceptions and purposes into an overall synthesis. The AHP does not require that judgments be consistent or even transitive. The degree of consistency (or inconsistency) of the judgments is revealed at the end of the AHP process.

Most of us have difficulty examining even a few ideas at a time. We need instead to organize our problems in complex structures which allow us to think about them one or two at a time. We need *simplicity* and *complexity*. We need an approach that is conceptually simple so that we can use it easily. And at the same time, we need an approach that is robust enough to handle real world decisions and complexities.

The Analytic Hierarchy Process is such a problem-solving framework. It is a systematic procedure for representing the elements of any problem. It organizes the basic rationality by breaking down a problem into its smaller constituent parts and then calls for only simple pairwise comparison judgments to develop priorities in each hierarchy.

There are three principles which one can recognize in problem solving. They are the principles of decomposition, comparative judgments, and synthesis of priorities.

The decomposition principle calls for structuring the hierarchy to capture the basic elements of the problem. An effective way to do this is first to work downward from the focus in the top level to criteria bearing on the focus in the second level, followed by subcriteria in the third level, and so on, from the more general (and sometimes uncertain) to the more particular and definite. One can then start at the bottom, identifying alternatives for that level and attributes under which they should be

*Accepted by Ambar G. Rao; received October 1, 1984. This paper has been with the author 6 months for 1 revision.

0025-1909/86/3207/0841$01.25

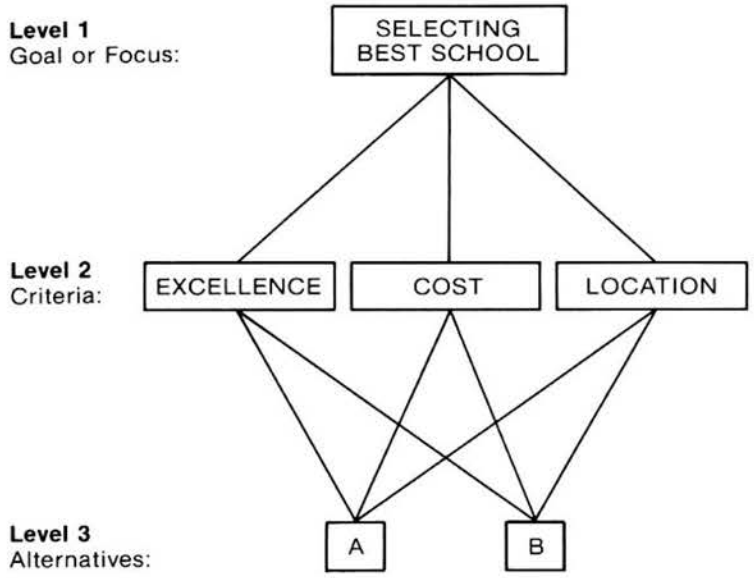


FIGURE 1

compared which fall in the next level up. Then one finds an intermediate set of higher criteria that can both be decomposed into these attributes and are themselves decompositions of the higher level criteria or subcriteria identified in the downward process. In this way, one can link the focus of the hierarchy to its bottom level in a sequence of appropriate intermediate levels. The levels of a decomposition are an essential part of measurement, and, hence, adjacent ones should generally not be too disparate, that is they do not differ by more than a "qualitative" order of magnitude. In general, the bottom level of the hierarchy contains the resources to be allocated, or the alternatives from which the choice is to be made. (See Figure 1.)

The principle of comparative judgments calls for setting up a matrix to carry out pairwise comparisons of the relative importance of the elements in the second level with respect to the overall objective (or focus) of the first level. In the case where no scale of measurement exists, this is a judgment made by the individual or group solving the problem. The scale for entering judgments is given in Table 1. Additional comparison matrices are used to compare the elements of the third level with respect to the appropriate parents in the second, and so on down the hierarchy. The process could be started at the bottom level and move upward. An entry of each matrix belongs to a fundamental scale employed in the comparisons. These entries are used to generate a derived ratio scale. The next step deals with the composition of the derived ratio scales.

The synthesis of priorities principle is now applied. Priorities are synthesized from the second level down by multiplying local priorities by the priority of their corresponding criterion in the level above, and adding them for each element in a level according to the criteria it affects. (The second level elements are each multiplied by unity, the weight of the single top level goal.) This gives the composite or global priority of that element which is then used to weight the local priorities of elements in the level below compared by it as criterion, and so on to the bottom level.

The AHP contains an intrinsic measure of inconsistency for each matrix and for the whole hierarchy. Knowledge of inconsistency enables one to determine those judgments which need reassessment.

When a group uses the AHP, their judgments can be combined after discussion by applying the geometric mean to the judgments which derives from the requirement

TABLE 1

Scale of Relative Importance

Intensity of Relative Importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective.
3	Moderate importance of one over another	Experience and judgment slightly favor one activity over another.
5	Essential or strong importance	Experience and judgment strongly favor one activity over another.
7	Demonstrated importance	An activity is strongly favored and its dominance is demonstrated in practice.
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation.
2, 4, 6, 8	Intermediate values between the two adjacent judgments	When compromise is needed.
Reciprocals of above non-zero numbers	If an activity has one of the above numbers assigned to it when compared with a second activity, then the second activity has the reciprocal value when compared to the first.	
Rationals	Ratios arising from the scale	If consistency were to be forced by obtaining n numerical values to span the matrix.

REMARK. When only two objects are compared it may be desirable to expand the interval 1, 2 (from equal to slight importance) by inserting the values, 1.1, 1.2, . . . , 1.9, starting with 1.1 as very slight, 1.2 as slight, 1.3 as moderate, etc.

that the collective judgment itself must satisfy the reciprocal property (Aczel and Saaty 1983).

The AHP can be applied to set priorities on the criteria and subcriteria of the hierarchy. The alternatives may be evaluated by paired comparisons (relative measurement). When there are many alternatives, and neither their number nor their kind affect the importance of the criteria, they can be absolutely measured or scored on each criterion according to merit or degree to which they meet the standards (see §4).

Many decision problems involve dependence of criteria on alternatives and of

higher order criteria on lower order ones; also alternatives may depend on other alternatives. A particularly useful generalization of the theory to deal with such dependence situations has been formalized within a network system with feedback of which a hierarchy is a special case.

The purpose of this paper is to state the axioms on which the AHP is based and to show how the theory of the AHP is derived from these axioms. For a more basic introduction to the AHP and its many applications, the reader is referred to Saaty (1980).

2. Axioms for Deriving a Scale from Fundamental Measurement and for Hierarchic Composition

Let $\mathfrak{A}$ be a finite set of n elements called alternatives. Let $\mathfrak{C}$ be a set of properties or attributes with respect to which elements in $\mathfrak{A}$ are compared. Philosophers distinguish between properties and attributes. A property is a feature that an object or individual possesses even if we are ignorant of this fact. On the other hand an attribute is a feature we assign to some object: it is a concept. Here we assume that properties and attributes are interchangeable and generally refer to them as criteria. A *criterion* is a primitive.

When two objects or elements in $\mathfrak{A}$ are compared according to a criterion in $\mathfrak{C}$, we say that we are performing binary comparisons. Let $>_C$ be a binary relation on $\mathfrak{A}$ representing "more preferred than" with respect to a criterion in $\mathfrak{C}$. Let $\sim_C$ be the binary relation "indifferent to" with respect to a criterion C in $\mathfrak{C}$. Hence, given two elements, $A_i, A_j \in \mathfrak{A}$, either $A_i >_C A_j$ or $A_j >_C A_i$ or $A_i \sim_C A_j$ for all $C \in \mathfrak{C}$. We use $A_i \gtrsim_C A_j$ to indicate more preferred or indifferent. A given *family of binary relations* $>_C$ with respect to a criterion C in $\mathfrak{C}$ is a primitive.

Let $\mathfrak{P}$ be the set of mappings from $\mathfrak{A} \times \mathfrak{A}$ to $\mathbb{R}^+$ (the set of positive reals). Let $f: \mathfrak{C} \to \mathfrak{P}$. Let $P_C \in f(C)$ for $C \in \mathfrak{C}$. P_C assigns a positive real number to every pair $(A_i, A_j) \in \mathfrak{A} \times \mathfrak{A}$. Let $P_C(A_i, A_j) \equiv a_{ij} \in \mathbb{R}^+$, $A_i, A_j \in \mathfrak{A}$. For each $C \in \mathfrak{C}$, the triple $(\mathfrak{A} \times \mathfrak{A}, \mathbb{R}^+, P_C)$ is a *fundamental* or *primitive scale*. A fundamental scale is a mapping of objects to a numerical system.

DEFINITION. For all $A_i, A_j \in \mathfrak{A}$ and $C \in \mathfrak{C}$

$$A_i >_C A_j \quad \text{if and only if} \quad P_C(A_i, A_j) > 1,$$

$$A_i \sim_C A_j \quad \text{if and only if} \quad P_C(A_i, A_j) = 1.$$

If $A_i >_C A_j$ we say that A_i dominates A_j with respect to $C \in \mathfrak{C}$. Thus P_C represents the intensity or strength of preference for one alternative over another.

Axiom 1 (Reciprocal). For all $A_i, A_j \in \mathfrak{A}$ and $C \in \mathfrak{C}$

$$P_C(A_i, A_j) = 1/P_C(A_j, A_i).$$

Whenever we make paired comparisons we need to consider both members of the pair to judge the relative value. If one stone is judged to be five times heavier than another, then the other is automatically one fifth as heavy as the first because it participated in making the first judgment. The comparison matrices that we consider are formed by making paired reciprocal comparisons. It is this simple, but powerful means of resolving multicriteria problems that is the basis of the AHP.

Let $A = (a_{ij}) \equiv (P_C(A_i, A_j))$ be the set of paired comparisons of the alternatives with respect to a criterion $C \in \mathfrak{C}$. By Axiom 1, A is a positive reciprocal matrix. The object is to obtain a *scale of relative dominance* (or *rank order*) of the alternatives from the paired comparisons given in A.

There is a natural way to derive the relative dominance of a set of alternatives from

a pairwise comparison matrix A. Let $R_{M(n)}$ be the set of $(n \times n)$ positive reciprocal matrices $A = (a_{ij}) \equiv (P_C(A_i, A_j))$ for all $C \in \mathfrak{C}$. Let $[0,1]^n$ be the n-fold cartesian product of $[0,1]$ and let $\psi : R_{M(n)} \to [0,1]^n$ for $A \in R_{M(n)}$, $\psi(A)$ is an n-dimensional vector whose components lie in the interval $[0,1]$. The triple $(R_{M(n)}, [0,1]^n, \psi)$ is a *derived scale*. A derived scale is a mapping between two numerical relational systems.

It is important to point out that the rank order implied by the derived scale ψ may not coincide with the order represented by the pairwise comparisons. Let $\psi_i(A)$ be the ith component of $\psi(A)$. It denotes the relative dominance of the ith alternative. By definition, for $A_i, A_j \in \mathfrak{A}$, $A_i >_C A_j$ implies $P_C(A_i, A_j) > 1$. However, if $P_C(A_i, A_j) > 1$, the derived scale could imply that $\psi_j(A) > \psi_i(A)$. This occurs if row dominance does not hold, i.e., for $A_i, A_j \in \mathfrak{A}$ and $C \in \mathfrak{C}$, $P_C(A_i, A_k) \geqslant P_C(A_j, A_k)$ does not hold for all $A_k \in \mathfrak{A}$. In other words, it may happen that $P_C(A_i, A_j) > 1$, and for some $A_k \in \mathfrak{A}$ we have

$$P_C(A_i, A_k) < P_C(A_j, A_k).$$

A more restrictive condition is the following:

DEFINITION. The mapping P_C is said to be *consistent* if and only if

$$P_C(A_i, A_j) P_C(A_j, A_k) = P_C(A_i, A_k) \qquad \text{for all} \quad i, j, \text{and } k. \tag{1}$$

Similarly the matrix A is consistent if and only if $a_{ij} a_{jk} = a_{ik}$ for all i, j and k.

If P_C is consistent, then Axiom 1 automatically follows and the rank order induced by ψ coincides with pairwise comparisons.

Hierarchic Axioms

DEFINITION. A *partially ordered set* is a set S with a binary relation $\leqslant$ which satisfies the following conditions:

(a) Reflexive: For all $x \in S$, $x \leqslant x$,

(b) Transitive: For all $x, y, z \in S$, if $x \leqslant y$ and $y \leqslant z$ then $x \leqslant z$,

(c) Antisymmetric: For all $x, y \in S$, if $x \leqslant y$ and $y \leqslant x$ then $x = y$ (x and y coincide).

DEFINITION. For any relation $x \leqslant y$ (read, y includes x) we define $x < y$ to mean that $x \leqslant y$ and $x \neq y$. y is said to *cover* (*dominate*) x if $x < y$ and if $x < t < y$ is possible for no t.

Partially ordered sets with a finite number of elements can be conveniently represented by a directed graph. Each element of the set is represented by a vertex so that an arc is directed from y to x if $x < y$.

DEFINITION. A subset E of a partially ordered set S is said to be *bounded* from above (below) if there is an element $s \in S$ such that $x \leqslant s$ ($\geqslant s$) for every $x \in E$. The element s is called an upper (lower) bound of E. We say that E has a supremum (infimum) if it has upper (lower) bounds and if the set of upper (lower) bounds U (L) has an element u_1 (l_1) such that $u_1 \leqslant u$ for all $u \in U$ ($l_1 \geqslant l$ for all $l \in L$).

DEFINITION. Let $\mathfrak{H}$ be a finite partially ordered set with largest element b. $\mathfrak{H}$ is a *hierarchy* if it satisfies the conditions:

(1) There is a partition of $\mathfrak{H}$ into sets called levels $\{L_k,\ k = 1, 2, \ldots, h\}$, where $L_1 = \{b\}$.

(2) $x \in L_k$ implies $x^- \subseteq L_{k+1}$, where $x^- = \{y \mid x \text{ covers } y\}$, $k = 1, 2, \ldots, h-1$.

(3) $x \in L_k$ implies $x^+ \subseteq L_{k-1}$, where $x^+ = \{y \mid y \text{ covers } x\}$, $k = 2, 3, \ldots, h$.

DEFINITION. Given a positive real number $\rho \geqslant 1$ a nonempty set $x^- \subseteq L_{k+1}$ is said to be ρ-homogeneous with respect to $x \in L_k$ if for every pair of elements $y_1, y_2 \in x^-$, $1/\rho \leqslant P_C(y_1, y_2) \leqslant \rho$. In particular the reciprocal axiom implies that $P_C(y_i, y_i) = 1$.

Axiom 2. Given a hierarchy $\mathfrak{H}$, $x \in \mathfrak{H}$ and $x \in L_k$, $x^- \subseteq L_{k+1}$ is ρ-homogeneous for $k = 1, \ldots, h-1$.

Homogeneity is essential for comparing similar things, as the mind tends to make large errors in comparing widely disparate elements. For example we cannot compare a grain of sand with an orange according to size. When the disparity is great, the elements are placed in separate clusters of comparable size giving rise to the idea of levels and their decomposition. This axiom is closely related to the well-known Archimedean property.

The notions of fundamental and derived scales can be extended to $x \in L_k$, $x^- \subseteq L_{k+1}$ replacing C and $\mathfrak{A}$ respectively. The derived scale resulting from comparing the elements in x^- with respect to x is called a *local derived scale* or *local priorities.* Here no irrelevant alternative is included in the comparisons and such alternatives are assumed to receive the value of zero in the derived scale.

Given L_k, $L_{k+1} \subseteq \mathfrak{H}$, let us denote the local derived scale for $y \in x^-$ and $x \in L_k$ by $\psi_{k+1}(y/x)$, $k = 2, 3, \ldots, h-1$. Without loss of generality we may assume that $\sum_{y \in x^-} \psi_{k+1}(y/x) = 1$. Consider the matrix $\psi_k(L_k/L_{k-1})$ whose columns are local derived scales of elements in L_k with respect to elements in L_{k-1}.

DEFINITION. A set $\mathfrak{A}$ is said to be *outer dependent* on a set $\mathfrak{C}$ if a fundamental scale can be defined on $\mathfrak{A}$ with respect to every $c \in \mathfrak{C}$.

Decomposition implies containment of the small elements by the large clusters or levels. In turn, this means that the smaller elements depend on the outer parent elements to which they belong, which themselves fall in a large cluster of the hierarchy. The process of relating elements (e.g., alternatives) in one level of the hierarchy according to the elements of the next higher level (e.g., criteria) expresses the dependence of the lower elements on the higher so that comparisons can be made between them. The steps are repeated upward in the hierarchy through each pair of adjacent levels to the top element, the focus or goal.

The elements in a level may depend on one another with respect to a property in another level. Input-output dependence of industries is an example of the idea of inner dependence. This may be formalized as follows:

DEFINITION. Let $\mathfrak{A}$ be outer dependent on $\mathfrak{C}$. The elements in $\mathfrak{A}$ are said to be *inner dependent* with respect to $C \in \mathfrak{C}$ if for some $A \in \mathfrak{A}$, $\mathfrak{A}$ is outer dependent on A.

Axiom 3. Let $\mathfrak{H}$ be a hierarchy with levels $L_1, L_2, \ldots, L_h$. For each L_k, $k = 1, 2, \ldots, h-1$,

(1) L_{k+1} is outer dependent on L_k,

(2) L_{k+1} is not inner dependent with respect to all $x \in L_k$,

(3) L_k is not outer dependent on L_{k+1}.

Principle of Hierarchic Composition. If Axiom 3 holds, the global derived scale (rank order) of any element in $\mathfrak{H}$ is obtained from its component in the corresponding vector of the following:

$$\begin{aligned}
&\psi_1(b) = 1,\\
&\psi_2(L_2) = \psi_2(b^-/b),\\
&\quad\vdots \qquad\qquad \vdots\\
&\psi_k(L_k) = \psi_k(L_k/L_{k-1})\psi_{k-1}(L_{k-1}), \qquad k = 3, \ldots, h.
\end{aligned}$$

Were one to omit Axiom 3, the Principle of Hierarchic Composition would no longer apply because of outer and inner dependence among levels or components which need not form a hierarchy. The appropriate composition principle is derived from the supermatrix approach of which the Principle of Hierarchic Composition is a special case (Saaty 1980).

A hierarchy is a special case of a system, the definition of which is given by:

DEFINITION. Let $\mathfrak{S}$ be a family of nonempty sets $\mathfrak{C}_1, \mathfrak{C}_2, \ldots, \mathfrak{C}_n$, where $\mathfrak{C}_i$ consists of the elements $\{e_{ij}, j = 1, \ldots, m_i\}$, $i = 1, 2, \ldots, n$. $\mathfrak{S}$ is a system if

(i) It is a directed graph whose vertices are $\mathfrak{C}_i$ and whose arcs are defined through the concept of outer dependence; thus

(ii) Given two components $\mathfrak{C}_i$ and $\mathfrak{C}_j \in \mathfrak{S}$ there is an arc from $\mathfrak{C}_i$ to $\mathfrak{C}_j$ if $\mathfrak{C}_j$ is outer dependent on $\mathfrak{C}_i$.

Therefore, many of the concepts derived for hierarchies also relate to general systems with feedback. Here one needs to characterize dependence among the elements. We now give a criterion for this purpose.

Let $D_A \subseteq \mathfrak{A}$ be the set of elements of $\mathfrak{A}$ outer dependent on $A \in \mathfrak{A}$. Let $\psi_{A_i,C}(A_j)$, $A_j \in \mathfrak{A}$ be the derived scale of the elements of $\mathfrak{A}$ with respect to $A_i \in \mathfrak{A}$ for a criterion $C \in \mathfrak{C}$. Let $\psi_C(A_j)$, $A_j \in \mathfrak{A}$ be the derived scale of the elements of $\mathfrak{A}$ with respect to a criterion $C \in \mathfrak{C}$. We define the dependence weight

$$\phi_C(A_j) = \sum_{A_i \in D_{A_j}} \psi_{A_i,C}(A_j)\psi_C(A_i).$$

If the elements of $\mathfrak{A}$ are inner dependent with respect to $C \in \mathfrak{C}$, then $\phi_C(A_j) \neq \psi_C(A_j)$ for some $A_j \in \mathfrak{A}$.

Expectations are beliefs about the rank of alternatives derived from prior knowledge. Assume that a decision maker has a ranking, arrived at intuitively, of a finite set of alternatives $\mathfrak{A}$ with respect to prior knowledge of criteria $\mathfrak{C}$. He may have expectations about rank order.

Axiom 4 (Expectations).

$$\mathfrak{C} \subset \mathfrak{H} - L_h, \qquad \mathfrak{A} = L_h.$$

This axiom simply says that those thoughtful individuals who have reasons for their beliefs should make sure that their ideas are adequately represented for the outcome to match these expectations; i.e., all alternatives are represented in the hierarchy, as well as all criteria. It neither assumes rationality of the process nor that it can only accommodate a rational outlook. People have many expectations that are irrational.

3. Results from the Axioms

Note that if P_C is consistent, then Axiom 1 follows, i.e., consistency implies the reciprocal property. The first few theorems are based on this more restrictive property of consistency.

The theorems show that paired comparisons and the principal eigenvector are useful in estimating ratios. We use perturbation arguments to demonstrate that the principal eigenvector solution is the appropriate one to surface rank order from inconsistent data and that the eigenvector is stable to small perturbations in the data. These results are also obtained by means of graph theoretic arguments.

Let $R_{C(n)} \subset R_{M(n)}$ be the set of all $(n \times n)$ consistent matrices.

THEOREM 1. *Let* $A \in R_{M(n)}$. $A \in R_{C(n)}$ *if and only if* $\operatorname{rank}(A) = 1$.

PROOF. If $A \in R_{C(n)}$, then $a_{ij}a_{jk} = a_{ik}$ for all i, j and k. Hence, given a row of A, $a_{i1}, a_{i2}, \ldots, a_{in}$, all other rows can be obtained from it by means of the relation $a_{jk} = a_{ik}/a_{ij}$ and $\operatorname{rank}(A) = 1$.

Let us now assume that $\operatorname{rank}(A) = 1$. Given a row a_{jh} $(j \neq i,\ h = 1, 2, \ldots, n)$, $a_{jh} = Ma_{ih}$ $(h = 1, 2, \ldots, n)$ where M is a positive constant. Also, for any reciprocal matrix, $a_{ii} = 1$ $(i = 1, 2, \ldots, n)$. Thus, for $i = h$ we have $a_{ji} = Ma_{ii} = M$ and $a_{jh} = a_{ji}a_{ih}$ for all i, j and k, and A is consistent.

THEOREM 2. *Let* $A \in R_{M(n)}$. $A \in R_{C(n)}$ *if and only if its principal eigenvalue* $\lambda_{\max}$ *is equal to* n.

PROOF. By Theorem 1 we have $\operatorname{rank}(A) = 1$.

Also, all eigenvalues of A but one vanish. Since $\text{Trace}(A) = \sum_{i=1}^{n} a_{ii} = n$ and $\text{Trace}(A) = \sum_k \lambda_k = n$, then $\lambda_{\max} \equiv \lambda_1 = n$.

If $\lambda_{\max} = n$,

$$n\lambda_{\max} = \sum_{i,j=1}^{n} a_{ij} w_j w_i^{-1} = n + \sum_{1 \leqslant i < j \leqslant n} \left(a_{ij} w_j w_i^{-1} + a_{ji} w_i w_j^{-1} \right)$$

$$\equiv n + \sum_{1 \leqslant i < j \leqslant n} (y_{ij} + 1/y_{ij}).$$

Since $y_{ij} + y_{ij}^{-1} \geqslant 2$, and $n\lambda_{\max} = n^2$, equality is uniquely obtained on putting $y_{ij} = 1$, i.e., $a_{ij} = w_i/w_j$. The condition $a_{ij}a_{jk} = a_{ik}$ holds for all i, j and k, and the result follows.

THEOREM 3. *Let $A = (a_{ij}) \in R_{C(n)}$. There exists a function $\psi = (\psi_1, \psi_2, \ldots, \psi_n)$, $\psi : R_{C(n)} \to [0, 1]^n$ such that*

(i) $a_{ij} = \psi_i(A)/\psi_j(A)$,

(ii) *The relative dominance of the ith alternative, $\psi_i(A)$, is the ith component of the principal right eigenvector of A,*

(iii) *Given two alternatives $A_i, A_j \in \mathfrak{A}$, $A_i \gtrsim_C A_j$ if and only if $\psi_i(A) \geqslant \psi_j(A)$.*

PROOF. $A \in R_{C(n)}$ implies that $a_{ij} = a_{ik}a_{jk}^{-1}$ for all k, and each i and j. Also by Theorem 1, we have $\text{rank}(A) = 1$ and we can write $a_{ij} = x_i/x_j$, where $x_i, x_j > 0$ $(i, j = 1, 2, \ldots, n)$. Multiplying A by the vector $x^T = (x_1, x_2, \ldots, x_n)$ we have $Ax = nx$. Dividing both sides of this expression by $\sum_{i=1}^{n} x_i$ and writing $w = x/\sum_{i=1}^{n} x_i$ we have $Aw = nw$, and $\sum_{i=1}^{n} w_i = 1$. By Theorem 2 we have n as the largest positive real eigenvalue of A and w as its corresponding right eigenvector. Since $a_{ij} = x_i/x_j = w_i/w_j$ for all i and j, we have $\psi_i(A) = w_i$, $i = 1, 2, \ldots, n$ and (i) and (ii) follow.

By Axiom 1, for $A \in R_C(n)$, $A_i \gtrsim_C A_j$ if and only if $a_{ij} \geqslant 1$ for all i and j, and hence we have $\psi_i(A) \geqslant \psi_j(A)$ for all i and j.

It is unnecessary to invoke the Perron–Frobenius Theory to ensure the existence and uniqueness of a largest positive real eigenvalue and its eigenvector. We have already proved the existence of an essentially unique solution in the consistent case. A similar result follows using the perturbation argument given below.

THEOREM 4. *Let $A \in R_{C(n)}$, and let $\lambda_1 = n$ and $\lambda_2 = 0$ be the eigenvalues of A with multiplicity 1 and $(n - 1)$, respectively. Given $\epsilon > 0$, there is a $\delta = \delta(\epsilon) > 0$ such that if*

$$|a_{ij} + \tau_{ij} - a_{ij}| = |\tau_{ij}| \leqslant \delta \qquad \text{for} \quad i, j = 1, 2, \ldots, n,$$

the matrix $B = (a_{ij} + \tau_{ij})$ has exactly 1 and $(n - 1)$ eigenvalues in the circles $|\mu - n| < \epsilon$ and $|\mu - 0| < \epsilon$, respectively.

PROOF. Let $\epsilon_0 = \frac{1}{2}(n)$, and let $\epsilon < n/2$. The circles C_1: $|\mu - n| = \epsilon$ and C_2: $|\mu - 0| = \epsilon$ are disjoint. Let $f(\mu, A)$ be the characteristic polynomial of A. Let $r_j = \min|f(\mu, A)|$ for μ on C_j. Note that $\min|f(\mu, A)|$ is defined because f is a continuous function of μ, and $r_j > 0$ since the roots of $f(\mu, A) = 0$ are the centers of the circles.

$f(\mu, B)$ is a continuous function of the $1 + n^2$ variables μ and $a_{ij} + \tau_{ij}$, $i, j = 1, 2, \ldots, n$, and for some $\delta > 0$, $f(\mu, B) \neq 0$ for μ on any C_j, $j = 1, 2$, if $|\tau_{ij}| \leqslant \delta$, $i, j = 1, 2, \ldots, n$.

From the theory of functions of a complex variable, the number of roots μ of $f(\mu, B) = 0$ which lie inside C_j, $j = 1, 2$, is given by

$$n_j(B) = \frac{1}{2\pi i} \int_{C_j} \frac{f'(\mu, B)}{f(\mu, B)}\, d\mu, \qquad j = 1, 2,$$

which is also a continuous function of the n^2 variables $a_{ij} + \tau_{ij}$ with $|\tau_{ij}| \leqslant \delta$.

For $B = A$, we have $n_1(A) = 1$ and $n_2(A) = n - 1$. Since $n_j(B)$, $j = 1, 2$, is continuous, it cannot jump from $n_j(A)$ to $n_j(B)$ and the two must be equal and have the value $n_1(B) = 1$ and $n_2(B) = n - 1$, for all B with $|a_{ij} + \tau_{ij} - a_{ij}| \leq \delta$, $i, j = 1, 2, \ldots, n$.

THEOREM 5. *Let $A \in R_{C(n)}$ and let w be its principal right eigenvector. Let $\Delta A = (\delta_{ij})$ be a matrix of perturbations of the entries of A such that $A' = A + \Delta A \in R_{M(n)}$, and let w' be its principal right eigenvector. Given $\epsilon > 0$, there exists a $\delta > 0$ such that $|\delta_{ij}| \leq \delta$ for all i and j, then $|w_i' - w_i| \leq \epsilon$ for all $i = 1, 2, \ldots, n$.*

PROOF. By Theorem 4, given $\epsilon > 0$, there exists a $\delta > 0$ such that if $|\delta_{ij}| \leq \delta$ for all i and j, the principal eigenvalue of A' satisfies $|\lambda_{\max} - n| \leq \epsilon$. Let $\Delta A = \tau B$. Wilkinson (1965) has shown that for a sufficiently small τ, $\lambda_{\max}$ can be given by a convergent power series $\lambda_{\max} = n + k_1\tau + k_2\tau^2 + \cdots$. Now, $\lambda_{\max} \to n$ as $\tau \to 0$, and $|\lambda_{\max} - n| = o(\tau) \leq \epsilon$.

Let w be the right eigenvector corresponding to the simple eigenvalue n of A. Since n is a simple eigenvalue, $(A - nI)$ has at least one nonvanishing minor of order $(n - 1)$. Suppose, without loss of generality, that this lies in the first $(n - 1)$ rows of $(A - nI)$. Then from the theory of linear equations, the components of w may be taken to be $(A_{n1}, A_{n2}, \ldots, A_{nn})$ where A_{ni} denotes the cofactor of the (n, i) element of $(A - nI)$, and is a polynomial in n of degree not greater than $(n - 1)$.

The components of w' are polynomials in $\lambda_{\max}$ and τ, and since the power series expansion of $\lambda_{\max}$ is convergent for all sufficiently small τ, each component of w' is represented by a convergent power series in τ. We have

$$w' = w + \tau z_1 + \tau^2 z_2 + \cdots \quad \text{and} \quad |w' - w| = o(\tau) \leq \epsilon.$$

By Theorems 4 and 5, it follows that a small perturbation A' of A transforms the eigenvalue problem $(A - nI)w = 0$ to $(A' - \lambda_{\max} I)w' = 0$.

THEOREM 6 (Ratio Estimation). *Let $A \in R_{M(n)}$, and let w be its principal right eigenvector. Let $\epsilon_{ij} = a_{ij} w_j w_i^{-1}$, for all i and j, and let $1 - \tau < \epsilon_{ij} < 1 + \tau$, $\tau > 0$, for all i and j. Given $\epsilon > 0$ and $\tau < \epsilon$, there exists a $\delta > 0$ such that for all $(x_1, x_2, \ldots, x_n)$, $x_i > 0$, $i = 1, 2, \ldots, n$, if*

$$1 - \delta < \frac{a_{ij}}{x_i / x_j} < 1 + \delta \qquad \text{for all } i \text{ and } j, \tag{2}$$

then

$$1 - \epsilon < \frac{w_i / w_j}{x_i / x_j} < 1 + \epsilon \qquad \text{for all } i \text{ and } j. \tag{3}$$

PROOF. Substituting $a_{ij}\epsilon_{ij}^{-1}$ for w_i / w_j in (3) we have

$$\left| \frac{w_i / w_j}{x_i / x_j} - 1 \right| = \left| \frac{1}{\epsilon_{ij}} \frac{a_{ij}}{x_i / x_j} - 1 \right| \leq \frac{1}{\epsilon_{ij}} \left| \frac{a_{ij}}{x_i / x_j} - 1 \right| + \left| \frac{1}{\epsilon_{ij}} - 1 \right|.$$

By definition $\epsilon_{ij} = 1 / \epsilon_{ji}$ for all i and j, and we have

$$\left| \frac{w_i / w_j}{x_i / x_j} - 1 \right| = \epsilon_{ji} \left| \frac{a_{ij}}{x_i / x_j} - 1 \right| + |\epsilon_{ji} - 1| < (1 + \tau)\delta + \tau.$$

Given $\epsilon > 0$ and $0 < \tau < \epsilon$, there exists a $\delta = (\epsilon - \tau)/(1 + \tau) > 0$ such that (2) implies (3).

This theorem says that if the paired comparison coefficient a_{ij} is close to an underlying ratio x_i / x_j then so is w_i / w_j and may be used as an approximation for it.

THEOREM 7. *Let $A = (a_{ij}) \in R_{M(n)}$. Let $\lambda_{\max}$ be its principal eigenvalue and let w be its corresponding right eigenvector with $\sum_{i=1}^{n} w_i = 1$, then $\lambda_{\max} \geqslant n$.*

PROOF. Let $a_{ij} = w_j w_i^{-1} \epsilon_{ij}$, $i, j = 1, 2, \ldots, n$. Since $Aw = \lambda_{\max} w$, and $\sum_{i,j=1}^{n} a_{ij} w_j = \lambda_{\max}$, we have

$$\lambda_{\max} - n = \sum_{i,j=1}^{n} a_{ij} w_j - n = \sum_{i,j} \epsilon_{ij} - n.$$

By definition, the matrix $(\epsilon_{ij}) \in R_{M(n)}$. We have $\epsilon_{ii} = 1$ for all i, and $\epsilon_{ij} > 0$ for all i and j. Hence, we have $\sum_{i,j=1}^{n} \epsilon_{ij} - n = \sum_{i \neq j} \epsilon_{ij} > 0$ and the result follows.

THEOREM 8. *Let $A \in R_{M(n)}$. Let $\lambda_{\max}$ be the principal eigenvector of A, and let w be its corresponding right eigenvector with $\sum_{i=1}^{n} w_i = 1$. $\mu \equiv (\lambda_{\max} - n)/(n-1)$ is a measure of the average departure from consistency.*

PROOF. For $A \in R_{C(n)} \subset R_{M(n)}$, *by Theorem 2 we have $\lambda_{max} = n$, and hence, we have* $\mu = 0$.

For $A \in R_{M(n)} - R_{C(n)}$, let $a_{ij} = w_i \epsilon_{ij} / w_j$ for all i and j. We have

$$\lambda_{\max} = \sum_{j=1}^{n} a_{ij} \frac{w_j}{w_i} = \sum_{j=1}^{n} \epsilon_{ij},$$

$$n\lambda_{\max} = \sum_{i,j=1}^{n} \epsilon_{ij} = n + \sum_{1 < i < j < n} \left(\epsilon_{ij} + \frac{1}{\epsilon_{ij}} \right),$$

$$\frac{\lambda_{\max} - n}{n-1} = -1 + \frac{1}{n(n-1)} \sum_{1 < i < j < n} \left(\epsilon_{ij} + \frac{1}{\epsilon_{ij}} \right).$$

As $\epsilon_{ij} \to 1$, i.e., consistency is approached, $\mu \to 0$. Also, μ is convex in ϵ_{ij}, since $(\epsilon_{ij} + 1/\epsilon_{ij})$ is convex, and has its minimum at $\epsilon_{ij} = 1$, $i, j = 1, 2, \ldots, n$. Thus, μ is small or large depending on ϵ_{ij} being near to or far from unity, respectively, i.e., near to or far from consistency, and the result follows.

Note that $\sum_{i,j=1}^{n} a_{ij} w_j w_i^{-1} - n^2 = n(n-1)\mu$ is also a measure of the departure from consistency.

It is also possible to show that $(A - nI)w = 0$ is transformed into $(A' - \lambda_{\max} I)w' = 0$ by means of graph theoretic concepts.

DEFINITION. The intensity of judgments associated with a path from i to j called the *path intensity* is equal to the products of the intensities associated with the arcs of that path.

DEFINITION. A *cycle* is a path of pairwise comparisons which terminates at its starting point.

THEOREM 9. *If $A \in R_{C(n)}$, the intensities of all cycles are equal to a_{ii}, $i = 1, 2, \ldots, n$.*

PROOF. $A \in R_{C(n)}$, implies $a_{ij} a_{jk} = a_{ik}$ for all i, j and k. Hence, we have $a_{ii} = a_{ij} a_{jk} a_{ki} = 1$ for all $i = 1, 2, \ldots, n$. By induction, if $a_{ii_1} \ldots a_{i_{n-1} i} = 1$ for all $i_1 \ldots i_{n-1}$, then $a_{ii_1} \ldots a_{i_{n-1} i_n} a_{i_n i} = a_{ii_n} a_{i_n i} = 1$ and the result follows.

THEOREM 10. *If $A \in R_{C(n)}$, the intensities of all paths from i to j are equal to a_{ij}.*

PROOF. Follows from $a_{ij} = a_{ik} a_{kj}$ for all i, j and k.

COROLLARY 1. *If $A \in R_{C(n)}$, the entry in the (i, j) position can be represented as the intensity of paths of any length starting with i and terminating with j.*

PROOF. Follows from the proof of Theorem 10.

COROLLARY 2. *If $A \in R_{C(n)}$, the entry in the (i, j) position is the average intensity of paths of length k from i to j, and $A^k = n^{k-1}A$ $(k \geqslant 1)$.*

PROOF. From Theorem 10, the intensity of a path of any length from i to j is equal to a_{ij}.

An arbitrary entry of A^k is given by

$$a_{ij}^{(k)} = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_{k-1}=1}^{n} a_{ii_1} a_{i_1 i_2} \cdots a_{i_{k-1} j}.$$

Since $a_{ij}a_{jk} = a_{ik}$ for all i, j and k we have

$$a_{ij}^{(k)} = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_{k-1}=1}^{n} a_{ij} = n^{k-1} a_{ij}.$$

By induction, if $a_{ij}^{(k)} = n^{k-1}a_{ij}$ for $k = 1, 2, \ldots, m-1$, for $k = m$ we have

$$a_{ij}^{(m)} = \sum_{i_1=1}^{n} \cdots \sum_{i_{m-1}=1}^{n} a_{ii_1} \cdots a_{i_{m-1} j}$$

$$= n^{m-2} \sum_{i_{m-1}=1}^{n} a_{ii_{m-1}} a_{i_{m-1} j} = n^{m-1} a_{ij}.$$

Hence, we have

$$a_{ij} = \frac{1}{n^{m-1}} a_{ij}^{(m)} \qquad \text{for all} \quad m \geqslant 1,$$

and the result follows.

THEOREM 11. *If $A \in R_{C(n)}$ the entry in the (i, j) position is given by the average of all path intensities starting with i and terminating with j.*

PROOF. By Corollary 2 of Theorem 10, we have

$$a_{ij} = \frac{1}{n^{m-1}} \sum_{i_1=1}^{n} \cdots \sum_{i_{m-1}=1}^{n} a_{ii_1} \cdots a_{i_{m-1} j}.$$

Hence, we have

$$a_{ij} = \lim_{m \to \infty} \frac{1}{n^{m-1}} a_{ij}^{(m)},$$

and the result follows.

THEOREM 12. *If $A \in R_{C(n)}$ the scale of relative dominance is given by any of its normalized columns, and coincides with the principal right eigenvector of A.*

PROOF. Let a^j be the jth column of A.

$$A \cdot a^j = \left(\sum_{k=1}^{n} a_{ik} a_{kj} \right) \qquad (i, j = 1, 2, \ldots, n),$$

$$= \left(\sum_{k=1}^{n} a_{ij} \right) = (na_{ij}) \qquad (i, j = 1, 2, \ldots, n),$$

and any column of A (whether or not it is normalized to unity) is a solution of the eigenvalue problem $Ax = nx$. By Corollary 2 of Theorem 10 we have $A^k = n^{k-1}A$. We have

$$\psi(A) = \lim_{m \to \infty} \frac{1}{m} \sum_{k=1}^{n} \frac{A^k e}{e^T A^k e} = \lim_{m \to \infty} \frac{1}{m} \sum_{k=1}^{n} \frac{Ae}{e^T Ae} = \frac{Ae}{e^T Ae}.$$

Hence, we have

$$\psi_i(A) = \sum_{j=1}^{n} a_{ij} \Big/ \sum_{i,j=1}^{n} a_{ij} = a_{ih}\left(\sum_{j=1}^{n} a_{hj}\right) \Big/ \left(\sum_{i=1}^{n} a_{ih}\right)\left(\sum_{j=1}^{n} a_{hj}\right) = \frac{a_{ih}}{\sum_{i=1}^{n} a_{ih}}$$

for all i and h, and the result follows.

COROLLARY. *The principal eigenvector is unique to within a multiplicative constant.*

PROOF. Follows from the proof of Theorem 12.

THEOREM 13. *If $A \in R_{M(n)}$ the intensity of all paths of length k from i to j is given by*

$$\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_{k-1}=1}^{n} a_{ii_1} a_{i_1 i_2} \cdots a_{i_{k-1} j}.$$

PROOF. It is known that the number of arc progressions of length n between any two vertices of a directed graph whose incidence matrix is V is given by V^n. If in addition each arc has associated a number ($\neq 1$) representing the intensity (or capacity) of the arc, then V^n represents the intensity of all arc progressions of length n between two vertices.

Let $V = A$. The entries of A^k give the intensity of all paths of length k between two vertices. Let $A^k = (a_{ij}^{(k)})$. By construction we have

$$a_{ij}^{(k)} = \sum_{i_1=1}^{n} \cdots \sum_{i_{k-1}=1}^{n} a_{ii_1} \cdots a_{i_{k-1} j}$$

and the result follows.

THEOREM 14. *Let $A \in R_{M(n)}$, $A \notin R_{C(n)}$. The principal right eigenvector of A is given by the limit of the normalized intensity of paths of length k,*

$$w_i = \lim_{k\to\infty} \frac{a_{ih}^{(k)}}{\sum_{i=1}^{n} a_{ih}^{(k)}}, \qquad i = 1, 2, \ldots, n,$$

for all $h = 1, 2, \ldots, n$.

PROOF. It can be shown that

$$\lim_{k\to\infty} \frac{a_{ih}^{(k)}}{\sum_{i=1}^{n} a_{ih}^{(k)}} = \lim_{k\to\infty} \frac{a_{is}^{(k)}}{\sum_{i=1}^{n} a_{is}^{(k)}}, \qquad h, s = 1, 2, \ldots, n. \tag{4}$$

The proof of this statement is given in Saaty and Vargas (1984b). Also we know that the principal right eigenvector of A is given by

$$w_i' = \lim_{k\to\infty} \frac{\sum_{h=1}^{n} a_{ih}^{(k)}}{\sum_{i=1}^{n}\sum_{h=1}^{n} a_{ih}^{(k)}}, \qquad i = 1, 2, \ldots, n. \tag{5}$$

Multiplying and dividing the right side of (5) inside the limit by $\sum_{i=1}^{n} a_{ih}^{(k)}$ and rearranging the terms we have

$$w_i' = \lim_{k\to\infty} \left[\sum_{h=1}^{n} \frac{a_{ih}^{(k)}}{\sum_{i=1}^{n} a_{ih}^{(k)}} \cdot \frac{\sum_{i=1}^{n} a_{ih}^{(k)}}{\sum_{i=1}^{n}\sum_{h=1}^{n} a_{ih}^{(k)}} \right]$$

$$= \sum_{h=1}^{n} \left[\lim_{k\to\infty} \frac{a_{ih}^{(k)}}{\sum_{i=1}^{n} a_{ih}^{(k)}} \right] \left[\lim_{k\to\infty} \frac{\sum_{i=1}^{n} a_{ih}^{(k)}}{\sum_{i,h=1}^{n} a_{ih}^{(k)}} \right]$$

From (5) we have

$$w_i' = \left[\lim_{k\to\infty} \frac{a_{is}^{(k)}}{\sum_{i=1}^n a_{is}} \right] \sum_{h=1}^n \frac{\sum_{i=1}^n a_{ih}^{(k)}}{\sum_{i,h=1}^n a_{ih}^{(k)}}$$

and the result follows.

COROLLARY. *Let $A \in R_{M(n)}$, $A \notin R_{C(n)}$. The principal right eigenvector of A is unique to within a multiplicative constant.*

PROOF. Follows from the proof of Theorem 14, and Theorem 5 in Saaty (1980).

THEOREM 15. *Let $\mathfrak{A}$ be a finite set of n elements $A_1, A_2, \ldots, A_n$, and let $C \in \mathfrak{C}$ be a criterion which all the elements in $\mathfrak{A}$ have in common. Let A be the resulting matrix of pairwise comparisons. The ith component of the principal right eigenvector of the reciprocal pairwise comparison matrix A gives the relative dominance of A_i, $i = 1, 2, \ldots, n$.*

PROOF. By Theorem 14, the principal right eigenvector of A is given by

$$w_i = \lim_{m\to\infty} \frac{a_{ih}^{(m)}}{\sum_{j=1}^n a_{jh}^{(m)}}, \qquad i = 1, 2, \ldots, n,$$

for any $h = 1, 2, \ldots, n$. By Theorem 7.13 in Saaty (1980) we have

$$w_i = \lim_{m\to\infty} \frac{\sum_{j=1}^n a_{ij}^{(m)}}{\sum_{i,j=1}^n a_{ij}^{(m)}}, \qquad i = 1, 2, \ldots, n.$$

Thus, the relative dominance of an alternative along all paths of length $k \leqslant m$ is given by

$$\frac{1}{m} \sum_{k=1}^m \frac{a_{ih}^{(k)}}{\sum_{i=1}^n a_{ih}^{(k)}}.$$

Let

$$s_k = \frac{a_{ih}^{(k)}}{\sum_{i=1}^n a_{ih}^{(k)}} \quad \text{and} \quad t_m = \frac{1}{m} \sum_{k=1}^m s_k.$$

It can be shown that if $\lim_{k\to\infty} s_k$ exists then $\lim_{m\to\infty} t_m$ also exists and the two limits coincide. By Theorem 14, we have $s_k \to w$ as $k \to \infty$, where w is the principal right eigenvector of A. Thus $t_m \to w$ as $m \to \infty$ and $\psi_i(A) = w_i$, $i = 1, 2, \ldots, n$.

This theorem highlights the fact that the right eigenvector gives the relative dominance (rank order) of each alternative over the other alternatives along paths of arbitrary length. It holds for a reciprocal matrix A which need not be consistent.

4. Relative and Absolute Measurement-Rank Preservation

The AHP can be used to make relative measurement through paired comparisons (scaling) of criteria and of alternatives, or to make absolute measurement (scoring) of the alternatives with respect to the criteria. The former is now familiar. The latter has been used when the number of alternatives is large and the decision is standard such as admitting students to a college based on well-established criteria whose weights are not affected by the number of students and their scores.

When the AHP uses paired comparisons it assumes structural dependence of the criteria on the number of alternatives and on their priorities. As a result, when alternatives are scaled through paired comparisons, adding a new alternative can

change the relative ranking of the old ones when the judgments are inconsistent or when several criteria are used. Under a single criterion rank never changes with the addition of a new alternative when the judgments are consistent (Saaty and Vargas 1984a). Note that if structural criteria are an integral part of a decision theory, the weights of these criteria would change with the introduction or deletion of alternatives and hence both the priorities and the ranks of the old alternatives can change. Thus structure is an important aspect of all systems and needs to be considered for better understanding of decisions. How to interpret such structural criteria has been covered in other works by this author now in process of publication.

If, in spite of structural dependence, for some practical reason one insists that the old rank remain in place and a new alternative be added, the new alternative can be measured by comparing it with one of the original ones and assigning it the appropriate value under each criterion without renormalizing. Normalization is then applied to the composite result. The priorities will change, but the ranking will be the same.

With absolute measurement there can be no rank reversal under a single or under multiple criteria. One compares the criteria, with respect to the goal, subcriteria with respect to the criteria and then the intensities of the subcriteria such as: excellent, very good, good, average, below average, poor, and very poor, with respect to each subcriterion. This yields a set of priorities for the intensities of the subcriteria. Each alternative is then scored with respect to each subcriterion by selecting the appropriate intensity. Once the weights of the intensities have been established the question of consistency in scoring the alternatives does not occur. Finally one adds all the priorities of the intensities to obtain a score for the alternative. In the end, these priorities may be normalized for all the alternatives.

5. Conclusion

We conclude with general remarks about the use of the AHP.

Because the AHP does not separate intangible factors from tangible ones and conducts its measurement by making pairwise comparisons, it is a useful way for analysis and decision making in complex social and political problems. In general, other methods such as multiattribute utility theory would first quantify individual intangible factors before calculating utility functions.

The AHP is also useful when many interests are involved and a number of people participate in the judgment process. Here debate may be to no avail and several answers must be developed. The results would then be weighted by the priority of the corresponding individuals according to that individual's relevance to the problem. These priorities are derived by extending the hierarchy upwards to include the individuals and criteria for evaluating them, with their assistance or participation when possible.

Judgments from different people on a single comparison must satisfy the reciprocal property for the group. This implies that these judgments must be synthesized into a single judgment according to the geometric mean (Aczel and Saaty 1983).

The AHP deals with problem decomposition in a systematic way. It requires that elements in each level be homogeneous, decreasing in size from the top to the bottom level of the hierarchy. While there is flexibility in structuring a problem, it is clear from the start that one proceeds by arranging the issues in decending (or ascending) order. It is also possible through the AHP to structure a problem which has dependencies and feedback to set priorities and make a choice.

Most of the difficulties encountered in using the AHP relate to the need for judgments. If a problem is complex and requires careful analysis, then time would be needed to elicit judgments. However, people can become tired and need to return to

the process after some rest. The more complex problems have needed nearly two days for this kind of participation. Furthermore, the AHP calls for occasional repetition of the process to make sure that the participants have not changed their minds dramatically. Patrick Harker of Wharton has recently developed a procedure for shortening the judgmental process.

It should now be clear that designing the analytic hierarchy, like the structuring of a problem by any other method, necessitates a substantial knowledge of the system in question. A strong aspect of the AHP is that the knowledgeable individuals who supply judgments for the pairwise comparisons usually also play a prominent role in specifying the hierarchy. Another key aspect in structuring a hierarchy is that any element in a level can be compared with respect to some elements in the level immediately above. The hierarchy need not be complete; that is, an element at an upper level need not function as a criterion for all the elements in the lower level. It can be partitioned into nearly disjoint subhierarchies sharing only a common topmost element. Thus for instance, the activities of separate divisions of an organization can be structured separately. The analyst can insert and delete levels and elements as necessary to clarify the task or to sharpen the focus on one or more areas of the system.

The AHP has already been successfully applied in a variety of fields. These include: a plan to allocate energy to industries; designing a transport system for the Sudan; planning the future of a corporation and measuring the impact of environmental factors on its development; design of future scenarios for higher education in the United States; the candidacy and election processes; setting priorities for the top scientific institute in a developing country and the faculty promotion and tenure problem (Saaty 1982, Wind and Saaty 1980, Toné 1986). The use of the AHP has been facilitated greatly by the availability of the microcomputer software package Expert Choice (1985).

References

ACZEL, J. AND T. L. SAATY, "Procedures for Synthesizing Ratio Judgments," *J. Math. Psych.*, 27, 1 (March 1983), 93–102.

Expert Choice, Decision Support Software, Inc., McLean, VA, 1985.

GASS, SAUL I., *Decision Making, Models and Algorithms: A First Course*, Wiley Interscience, New York, 1985.

HARDY, G. H., *Divergent Series*, Oxford University Press, London/New York, 1949.

HARKER, P. T. AND L. G. VARGAS, "The Theory of Ratio Scale Estimation: Saaty's Analytic Hierarchy Process," 1984 (submitted for publication).

HILDEBRAND, F. B., *Methods of Applied Mathematics*, Prentice-Hall, Englewood Cliffs, N.J., 1952.

SAATY, T. L., *The Analytic Hierarchy Process*, McGraw-Hill, New York, 1980.

———, "Structural Dependence and Rank Preservation in the Analytic Hierarchy Process," 1985 (submitted for publication).

——— AND L. G. VARGAS, "The Legitimacy of Rank Reversal," *OMEGA*, 12, 5 (1984a), 513–516.

——— AND ———, "Inconsistency and Rank Preservation," *J. Math. Psych.*, 28, 2 (June 1984b).

——— AND ———, "Comparison of Eigenvalue, Logarithmic Least Squares and Least Squares Methods in Estimating Ratios," *Math. Modelling*, 5, 5 (1984c), 309–324.

TONÉ, KAORU, *The Analytic Hierarchy Process*, in Japanese, Jan. 1986.

WILKINSON, J. H., *The Algebraic Eigenvalue Problem*, Claredon Press, Oxford, 1965.

WIND, Y. AND T. L. SAATY, "Marketing Applications of the Analytic Hierarchy Process," *Management Sci.*, 26, 7 (1980), 641–658.

STUDY UNIT 6

Investment in road infrastructure

The aim of this study unit is to explain how an effective road transport infrastructure can be provided against the background of policies on the provision of infrastructure and the principles of user payment so that scarce economic resources can be efficiently applied.

UNIT OUTCOMES

After working through this study unit, you should be able to:

- give an indication of the demand for roads in a community
- identify the total cost of road transport in a community
- meaningfully discuss the policy on road infrastructure in South Africa
- make suggestions about the implementation of South African policy on road infrastructure
- identify the benefits of improvements in infrastructure accruing to road users and nonusers
- allocate road costs to users
- critically evaluate different practical methods of recovering road costs

KEY CONCEPTS

- Road infrastructure
- Policy on provision of road infrastructure
- Road user benefits
- Types of traffic eg derived traffic
- Nonroad user benefits
- Cost allocation

6.1 Introduction

Road networks have evolved as a result of developments and the increase in the number of motor vehicles of all classes. The advent of motor vehicles created a demand for the

construction of better roads, which in turn served as a stimulus for technical improvements to and increased production of motor vehicles.

Roads obviously play an important role in road transport – without them it would be impossible to transport passengers and goods by road. Hence existing road networks cover wide geographical areas and extend over long distances.

This study unit is concerned with the provision (supply) of road infrastructure and responsibility for the costs involved. We therefore start the study unit with a concise discussion of the demand for roads so as to emphasise the interaction between supply and demand. Since road infrastructure is provided mainly by governments (at all levels), we shall also be examining the general role of government and the South African policy on the provision of infrastructure. The establishment of a sustainable or economically viable infrastructure requires an analysis of the costs and benefits involved as well as a study of possible methods of user payment. Since evaluation techniques were discussed in a previous study unit, a brief overview of the possible benefits of a new road and theoretical principles for the quantification of these benefits will suffice. This will be followed by a discussion of cost allocation methods and the main methods of recovering road costs.

6.2 Demand for roads

Demand for a road is derived from the demand for trips. In this context, a “trip” is defined as the transportation of goods or passengers in a vehicle on a road between two places separated geographically.

According to conventional demand theory and the market mechanism, transport demand should depend on the price paid for it. This price is largely determined by total transport costs, which comprise the following:

(1) infrastructure costs

(2) community costs

(3) haulage costs

(The total cost of transport will be discussed in greater detail in the next section.)

Since road users are not directly responsible for the costs of road provision but merely make an indirect contribution (eg in the form of fuel tax and vehicle licences), the demand for trips cannot always be effectively determined by the market mechanism (supply and demand). However, the market mechanism is used in decision-making on road construction programmes.

Transport demand is a prerequisite for the construction of a road network. This demand (between two places) depends mainly on the structure (size, composition and density) of the population, kinds of industries and job opportunities in both places, while the effectiveness of existing transport facilities between the two places also plays a decisive role.

The demand for infrastructure is influenced by the following features of the demand for transport:

(1) It varies according to the time of day (peak hours).

(2) It increases during the holiday seasons.

(3) It increases with economic growth.

(4) Poor road conditions sometimes prevent trips from being made.

(5) Traffic conditions discourage motorists from undertaking trips – since traffic congestion compels road users to travel at lower speeds, trips take longer.

Using transport demand as a point of departure, the authorities now have to make provision for the "production" of trips by providing adequate roads. Any addition to or improvement of the existing road network is aimed at greater transport efficiency, which is why the economics of road provision focuses mainly on an effective road investment programme.

6.3 Total road transport costs

In a universal sense, road transport costs are the total costs associated with road transport, infrastructure costs included. The term "costs" is used here in its widest possible sense and includes the direct costs and side effects or externalities of road transport, regardless of whether or not it is possible to measure them in money terms.

Activity 6.1

Make a list of the costs, which in your opinion, you incurred or generated on your last trip to (1) work, and (2) the shops. Study the section below and add to your list if necessary.

Total road transport costs have three components, namely infrastructure costs, community costs and road haulage costs, which can be subdivided as follows:

(1) Infrastructure costs include all costs relating to the construction, maintenance and administration of roads and law enforcement and traffic control on roads.

(2) Community costs comprise all social and other external costs arising from road transportation that cannot be recovered directly from road transport users through the market mechanism. These costs are also referred to as externalities. Examples of these are the consequences of road accidents, traffic jams (congestion) and environmental damage (all kinds of pollution).

(3) Road haulage costs (generally referred to as transport costs) are the costs associated with the ownership and operation of road transport services catering for passengers or goods. These services are rendered either by professional carriers (transport for remuneration) or by own transport (ancillary transport if an enterprise undertakes its own transport, or private transport in the case of private households).

The above interpretation of road transport costs is represented schematically in figure 6.1.

It should now be clear that the costs involved in your trip to work or the shops involves a lot more than your vehicle's operating costs or your bus or taxi fare!

This brings us to a critical question that will be dealt with in greater detail in the sections to follow:

- Who pays for these costs?

Figure 6.1

(1) INFRASTRUCTURE COSTS	Capital costs	Initial costs: preliminary investigations, planning, design, land expropriation, construction with a view to providing roads by establishing permanent facilities/structures Reconstruction and structural changes to existing roads
	Maintenance costs	Routine maintenance Repairs
	Administration costs	Routine administration and management Control: vehicle licensing and registration legislation Regulation: transport permits, inspections
	Law enforcement costs	Law enforcement, protection and lifesaving services
	Traffic-control costs	Control of traffic-flow to promote the road's effectiveness
	Research costs	Developing and improving methods of supplying infrastructure
(2) COMMUNITY COSTS	Contingency costs	That portion of resource wastage, attributable to accidents, that cannot be recovered from those causing accidents and is therefore borne by the community: loss of output, the subsidised portion of lifesaving services, hospitalisation and medical services, police investigations, free (pro deo) legal services and nonchargeable court costs. Nonquantifiable costs: pain, suffering and inconvenience
	Congestion costs	Waste of time Additional vehicle running costs Nonquantifiable costs: frustration, stress, et cetera
	Damage to the environment	Pollution Noise Vibration Unsightliness and visual intrusion Community costs
(3) HAULAGE COSTS	Vehicle running costs	Running or movement costs of vehicles and standing costs (time costs) associated with vehicles
	Overhead costs	Fixed costs not directly related to vehicles

6.4 The authorities' role in the provision of road infrastructure

As a rule, central and regional authorities are responsible for the provision of public roads and recover the costs from road users through indirect taxes and levies which form an integral part of the price of transport inputs.

These charges are levied in the form of value-added tax, customs and excise duty, import duty, ad valorem tax payable on vehicles, spare parts, fuel and so on, which are generally regarded as *general government revenue*. In addition, a levy can be included in the price of each litre of fuel sold for road use and earmarked for road provision. Annual revenue obtained from vehicle licences is also supposed to be allocated to the provision of roads. Thus government revenue received from road users can be subdivided into two categories – *allocated* (earmarked revenue) and *general* revenue, both of which can be used to provide road infrastructure.

With the exception of certain community costs that are not recoverable from road users, suppliers of transport services (carriers) must at some stage or other bear all the other transport costs. They, in turn, recover the costs from the users of their services by charging a tariff (fares in respect of passengers and freightage in the case of goods). However, for various reasons, passenger transport services are often subsidised.

In their efforts to economise, road authorities invariably face a complex problem. First, there is the *scarcity phenomenon*: unlimited transport needs have to be satisfied with limited resources. Secondly, there is a *conflict of choice*: a choice has to be made between different modes of transport in order to attain maximum need satisfaction. Thirdly, maximum need satisfaction depends on the *efficiency* of transport operations and an adequate road network.

Unfortunately, road infrastructure which is necessary for efficient road transport cannot always be provided at an acceptable profit. This obviously deters the suppliers of credit, and the consequent inadequate provision of road infrastructure has a negative effect on the economy. This situation is indicative of a suboptimal allocation of resources, and compels the government to assist the free-market mechanism in its efforts to optimise resource allocation. It does so by satisfying the community's needs for those collective goods and services that are supplied only partially by the free-market system, such as hospitals, health services and education. Government is also responsible for law enforcement, defence and the provision of road infrastructure, albeit often for strategic and political rather than economic reasons.

As mentioned earlier, the state's sources of revenue are limited, and it must therefore provide goods and services in such a way that society will benefit the most. At the same time, it has to keep private enterprise tax within reasonable limits to ensure that private enterprises are left with sufficient available capital for reinvestment and the generation of profits. In other words, government has to maintain a balance between the amount of money available to pursue the profit objective and the service objective. These funds should be applied simultaneously, with due consideration of the economic principle. According to this principle, economic efficiency requires that for each consumer, the marginal utility of a product or service supplied by the private sector should be equal to its marginal cost; and, at the same time, the total marginal utility for all consumers of a product or service supplied by the public sector must be equal to its marginal cost.

Thus far we have assumed that government has a fixed roads budget. In practice, however, the budgets of most authorities vary considerably, depending on the availability of loan funds. Thus, given a variable roads budget and relatively unlimited funds in the hands of the authority, the next step is to prioritise proposed road construction projects. Purely on the basis of economic considerations, the authority should undertake all projects with a cost-benefit ratio greater than one (ie with a rate of return exceeding the community's current opportunity cost of capital). Under perfect competition, it is assumed that projects with benefits greater than cost are normally undertaken by the private sector, since private enterprises will continue to incur costs up to the point where marginal costs equal marginal utility. However, as long as the benefits of a public project exceed its costs, a transfer of capital from the private to the public sector is justified since it would result in a gain for the community (Pienaar 1981:13).

To make a meaningful study of the scope and nature of road costs and benefits, they first have to be identified, quantified and related to each other. Transport or road economists must be able to appraise these costs and benefits in economic terms. One of the following three evaluation techniques can be used to determine the microeconomic viability of a project:

(1) absolute advantage, which may be determined by the net present value technique

(2) relative advantage, which is usually determined by means of the cost-benefit ratio technique or the internal rate of return technique

(3) minimum total community cost, which may be determined by the present value of cost technique

These techniques were discussed in a previous study unit. The identification and calculation of the advantages of improvement in a road network will be discussed in section 6.6.

6.5 South African policy on road infrastructure

6.5.1 Core principles of the policy

Against the background of the above costs of road transport and the general role of government in road provision, we shall now briefly examine the present policy on the provision of road infrastructure in South Africa, as set out in *Moving South Africa: the action agenda (a 20-year strategic framework for transport in South Africa)* (South Africa 1999). In this study unit we shall refer throughout to this work as the MSA document.

The current policy follows a holistic approach in which investment in transport focuses on prioritised customer groups. In this regard the policy is aimed at the following actions:

- Focus the scope of the transport system. This will be achieved by concentrating assets and investment to consolidate high volume routes and nodes to make up a national and various urban and rural strategic networks. The strategic networks would form the backbone of the transport system, underpinned by supporting networks.
- Deploy transport modes in the strategic and supporting networks (and their component routes) in order to capture the best economies of scale possible according to the ability of modes to meet customer needs.
- Create an environment in which customers are empowered and transport service providers are enabled to improve efficiency, productivity and competitiveness (South Africa 1999:12).

The strategy is therefore geared to a viable or *sustainable* customer-oriented transport sector in which road transport and the provision of road infrastructure play a vital role.

Strategic principles to implement the strategy were formulated. The important principles relating to road transport infrastructure are as follows:

- Recover full costs from users – charge users the full cost for operations, infrastructure and externalities.
- Optimise modal economies – pricing in corridors should facilitate the optimal use of modes based on demand density and distance (South Africa 1999:22).

The development of strategic and supporting networks is a core element of the policy:

- For both urban and freight customers, the strategy is to consolidate core transport assets into high volume corridor networks and dense development nodes.
- These corridors and nodes will concentrate demand for services into a focused area that will enable the low cost, high quality and affordable backbone of the total transport system. This is the *Strategic Network*. The dense demand and the

simpler corridor network will lead to higher vehicle utilisation, larger volumes per vehicle and a resulting lower cost per passenger and per unit of freight.

- In support of this strategic network, feeding into it, distributing from it and serving the needs of customers for transport between points outside of the core is the *Supporting Network*. The supporting network must itself be sustainable as part of the total system, but because of lower demand density patterns, its operations will be characterised by lower levels of fixed costs and higher levels of variable costs (South Africa 1999:21).

6.5.2 Infrastructure development for different customer groups

Earlier we mentioned particular customer groups and a customer-oriented transport system. The following four broad categories of customers are identified:

- urban passengers
- rural passengers
- tourists and long-distance passengers
- freight transport users

The policy for each group in respect of infrastructure development is summarised below.

6.5.2.1 Urban passengers

In the case of urban passengers, the policy is aimed at creating high-volume corridors over parts of the existing network where demand is highest. Investment in infrastructure will also focus primarily on public transport and the more effective utilisation of existing infrastructure. Instead of building new roads, investment will focus on the core public transport network with the development of public transport facilities such as transfer facilities, multimodal transfer facilities, bus and train stations and *densification*.

Corridor-supporting infrastructure investment such as extensions to a railway system and allocated public transport road infrastructure such as bus lanes is encouraged.

The policy on investment in road transport infrastructure for urban passengers is therefore aimed at promoting public transport and focuses on public transport infrastructure instead of the provision of roads that benefit private cars.

The aim of the strategy is to manage road space by discouraging the use of private motor vehicle transport in certain urban areas with huge congestion problems. A combination of control measures, pricing and improvement in public transport systems are indicated as possible methods to improve the utilisation of urban roads.

A general principle in the MSA document (South Africa 1999:30) is to make the user (the private motorist too) responsible for the full cost of externalities.

To summarise, as far as urban passenger transport is concerned, the policy focuses on corridor development, public transport along corridors and better utilisation of road space by means of control measures and the pricing of externalities.

6.5.2.2 Rural passenger transport

In rural areas, the policy focuses on the provision of a suitable infrastructure such as the upgrading of links to primary road networks. The MSA document identifies a general shortage of proper roads in rural areas. The challenge involves providing sustainable roads

in communities that need them most. This requires the establishment of a framework for prioritising communities and the roads they require on the basis of development needs, development potential and other social criteria (South Africa 1999:35).

The MSA document proposes the development of a coordinated national framework for infrastructure investment in rural areas. This entails the coordination of investments in transport, water, education, health care, electricity and other infrastructure in rural areas to achieve the maximum benefits for the community as a whole. Investment decisions should be prioritised on the basis of the sustainability of communities. Since most rural communities are not economically self-sufficient, the first step should be to formulate criteria and develop a framework for measuring *sustainability*.

Since only certain roads in rural areas will yield an economic return (and recover costs through user levies), the following broad actions are recommended:

- Identify which roads can be self-sustaining and fund their upgrading by means of user charging, including externality costs.
- Identify the roads that are mainly required to stimulate development (and therefore unable to create sufficient return to be self-sustaining) and prioritise them according to criteria such as development potential, the size and density of the population and accessibility.

6.5.2.3 Tourists and long-distance passengers

An integrated approach is followed in that the Department of Transport is intensely involved in formulating tourist strategies. The following actions are proposed to support the tourism industry:

- Focus investment around locations that result from the tourism strategy's targets.
- Limit investment in assets (like some roads) with little tourism or long-distance passenger potential.
- Coordinate large infrastructure decisions on an intermodal basis, especially within a tourist corridor. This is required to meet the objective of a seamless global tourism service. An example is that of co-ordinating road infrastructure with airport expansion decisions. This will require a co-ordinating mechanism which enables large tourism infrastructure providers (eg the National Roads Agency and the Airports Company of South Africa) to collaborate in directing big investments (South Africa 1999:42).

Strategic actions relating to road infrastructure for tourism and long-distance passengers have been formulated as follows:

- *Define the strategic tourism road network*. Roads and road networks that are critical to the growth of tourism should be identified on the basis of the tourism strategy for different customer segments.
- *Manage road infrastructure investment*. Investment patterns should be based on the importance of a particular road to the tourism strategy. Roads that are identified as priorities should be adequately funded to handle increasing traffic, for example roads between airports and city centres or between airports and strategic tourist destinations.
- *Payment for road usage and associated externalities*. In the long term, road pricing for tourism should be linked to the overarching road user charging system. The technology for direct road user charging of tourists and long-distance users should, however, be studied so that tourists can pay the full cost of using the transport system.

6.5.2.4 *Freight transport*

The provision of infrastructure for road freight transport falls within the framework of an integrated transport system in which all freight modes and networks are integrated. Infrastructure development focuses on upgrading the transport system in an effort to promote the export of value-added manufactured goods.

As in the case of passenger transport, the focus is on a corridor development with a strategic network of dense corridors supplemented with a supporting network.

Four strategic actions are formulated in the MSA document:

- *Define the freight transport network*. A strategic road network should be defined by means of consultation between the Department of Transport, the National Road Agency, provinces and other suppliers of road infrastructure. A major consideration in the definition of a strategic network is the choice of freight corridors which can serve as a high-level, high fixed cost[2], world-class strategic backbone for promoting the export of high value-added products. These strategic networks should be supported by a diversified, high variable cost-supporting network that is well maintained. This network will mainly serve the domestic flow of freight and also feed goods into and from the strategic backbone.
- *Manage road infrastructure investment*. This strategic action involves the funding of roads and emphasises:
  (1) the allocation of more funds to dense routes constituting the strategic network
  (2) increased funding for the general road network
  (3) the formulation of criteria for funding the supporting network
- *Charge road carriers for road use and externalities*. Recovery of the total cost of infrastructure provision and maintenance as well as the externality costs generated by road users is necessary for the sustainability of road infrastructure and to restore the current imbalances between land freight transport modes.
- *Gross vehicle mass limits must be strictly enforced*.

6.5.3 Integration of the strategic framework

All segments of transport infrastructure should be regarded as part of an integrated whole. Any efforts to develop density corridors for land transport (passengers and freight) should be supplemented with adequate infrastructure at nodes or terminals. A high-density corridor for freight exports should, for example, be supplemented with an adequate port infrastructure. Similarly, intermodal networks in the case of passengers require coordinated infrastructure to render a continuous and complete service to customers.

Investment in roads should contribute to an increase in the density of the system and facilitate the use of appropriate modes. The road strategy should also increase the flexibility of the system. Roads can contribute to flexibility by creating lower fixed cost infrastructure and make provision for vehicles with different capacities.

The road strategy covers four main actions:

- *Align roads strategy with customer strategies* (industrial, tourism, urban and rural development). Investment in roads for freight and urban passengers should focus on corridor development. Roads for tourism should support the tourism strategy

2 It should be clear from previous studies that high fixed costs mainly emphasise rail transport and high variable costs, road transport.

and investments in rural areas should be in accordance with the prioritisation framework for funding roads in sustainable communities.

- *Investment in roads should be focused*. The network should comprise different but appropriate levels of quality in order to focus on scarce resources in core networks. This means that roads should be prioritised on the basis of different customer segments.
- *Recovery of total costs from users*. Road users should pay the total cost (provision and maintenance of roads and externality costs) of road use.
- *Create an institutional framework to support the strategy*. Sources for road funding should be related to the relevant institution which will coordinate all the priorities of road investment and integrate them with the various corridor initiatives.

The policy emphasises the recovery of externality costs from road users. Efforts will be made to internalise the externalities by charging road users for the costs involved in externalities. By paying the full costs of road usage, users will make more economically rational decisions about where and when to use the road system. This will reduce road congestion and pollution and encourage motor vehicle users to use public transport.

6.5.4 Funding

The MSA document proposes the following for the funding of road transport infrastructure:

> *The MSA option is therefore a system based on dedicated transport funds in all three spheres (national, provincial and local). All three levels of transport funding would obtain their revenue from a range of sources. These could include fiscal allocations for developmental programmes, fuel levies (as a proxy general road-user charge), licence fees, area access charges and other funding sources. At the local level this could include revenues from developmental charges where developers are required to pay into the funds according to the level of transport infrastructure and service demand that a new development generates. It is envisaged that the transport funds would have to ensure the appropriate levels of externality charges, and to make decisions to invest such funds in programmes to reduce externalities, or to disburse such funds to the health system or to the environmental authorities* (South Africa 1999:77).

Activity 6.2

Summarise the South African policy on the provision of road transport infrastructure. What is your opinion of the policy? Do you think it will help to improve the mobility of the country's inhabitants and establish an efficient transport system with the minimum use of scarce resources?

In my opinion, for the first time, South Africa has an effective and workable transport policy that is based on sound economic principles and endeavours to utilise resources optimally and encourage economic development. The principle of user payment is especially attractive. In this regard, the government should guard against imposing excessive fuel levies. The present fuel levy is not an earmarked road fund and serves as a general source of income. I would like to see this general revenue (which is currently coming from the fuel levy) being obtained from another source and the fuel levy being earmarked for transport infrastructure (and services which, for some or other reason, warrant financial support). Cost recovery methods for road infrastructure will be dealt with in more detail in section 6.9.

6.6 Road-user benefits (savings)

6.6.1 General

In the discussion of the role of government in the provision of infrastructure and the South African policy, it would seem that improvements in infrastructure should either be economically viable or sustainable from a community perspective. Viability and sustainability require that the proposed infrastructure development is subject to an evaluation of the costs and benefits involved. In this section we shall briefly examine the identification and theoretical principles used to quantify the benefits of a new road for users. Non road-user benefits will be discussed in section 6.7.

Activity 6.3

Write down a few benefits that you think a motorist would enjoy if a new road were to be constructed. Study the section below and add to your list if necessary.

Roads are primarily built to provide access to places. New roads are therefore often built as an *instrument of development* to stimulate investment in economically dormant areas or regions. The forecasting and evaluation of the benefits accruing to non-road users (or plus factors) usually require a macro-economic analysis, which strictly speaking falls outside the scope of transport economics. However, when roads are built mainly to improve existing mobility, and only secondarily to stimulate latent travel demand (the usual motive for road construction in developing regions), the aim is to improve the quality and quantity of traffic flow, to reduce vehicle running costs and travel time and enhance traffic safety. The improvement of mobility requires a micro-economic analysis which falls within the scope of transport economics and focuses on the potential saving in the community's total transport costs as a result of the new road.

6.6.2 Identifying road-user benefits

Road-user benefits usually comprise savings for users resulting from the new road. These savings can be divided into three categories:

(1) reduction in vehicle running costs

(2) time savings

(3) fewer accidents and a reduction in associated costs

6.6.2.1 Reduction in vehicle running costs

Running costs usually comprise the direct variable cost of the road user and are incurred when a vehicle is in motion. The principal running costs elements are fuel, tyres, maintenance and oil.

Savings in running costs which can be attributed to the establishment of new transport facilities can normally be measured fairly accurately by calculating the difference between vehicle running costs with and without the new or improved facility.

6.6.2.2 *Time savings*

A time saving means that trip times on the new road are shorter compared with trips on the existing road between the same origin and destination. When evaluating the effects of road improvements, time savings should be assessed in money terms. We shall now discuss a number of principles that are used to distinguish between time savings for individuals as opposed to vehicles.

(a) Times savings for individuals

When evaluating times savings for individuals one should distinguish between working time and leisure time. Working time can be measured in terms of the average per capita wage rate per time unit.

Regardless of whether time savings are being evaluated in terms of working time or leisure time, the question arises whether negligible time savings have a significant value. Do 60 savings of one minute each have the same value as one saving of an hour? The answer will depend on circumstances. A commuter who manages to arrive one minute earlier at a bus stop, enabling him to catch a bus departing seven minutes earlier than the next one, actually gains seven minutes. Another person may not be able to do anything constructive in 60 savings of one minute each, but may be able to complete an entire economic operation during 60 consecutive minutes. Savings in working time will be of economic value only if the time saved is constructively utilised. The value of road users' working time is normally measured on the basis of average monthly income per capita.

(b) Time savings for vehicles

Time savings in relation to vehicles can be measured on the basis of two criteria:

- Can the same trip be covered in less time?
- Can a longer distance or more journeys be made in the same time?

Both measures indicate the potential of generating greater income or saving money, or both because of better utilisation. If time savings are not sufficient to utilise vehicles more intensively, an alternative benefit can be gained by utilising idle time for maintenance work on the vehicle. Time savings constitute a net benefit if existing fleet capacity is being fully utilised and if time is already a factor. Time savings resulting in reduced vehicle running costs are assessed in the same way as the running-cost saving accruing to existing traffic.

6.6.2.3 *Reduction in accident risk and cost of damage*

One of the main reasons for constructing a freeway may be to reduce the number and severity of accidents. This particular benefit may even be the deciding consideration in planning a facility such as a grade-separated railway crossing.

To quantify the benefits of accident prevention, it is necessary to predict the accident rate (usually on the basis of similar road standards and traffic conditions). The cost of the anticipated benefits is then subtracted from the cost of current accidents under existing conditions. This poses two problems, namely determining

(1) to what extent accidents are in fact attributable to poor road conditions and quality

(2) what money value to place on loss of human life and personal injury

6.6.3 Estimating road-user savings

6.6.3.1 Identifying the components of traffic

To determine the savings accruing to road users, it is necessary to classify the anticipated users of the new or improved road according to the components indicated below. The savings of each component are then calculated separately.

(1) *Existing traffic*. This refers to current traffic on a road due to be replaced or improved. When a new route is to be introduced without effecting any changes to existing routes in the area, the existing traffic on the new route equals zero. However, when a new road replaces an existing one, the existing traffic on the old route is regarded as the existing traffic on the new road.

(2) *Diverted traffic*. This refers to traffic that is diverted from other roads because of the opening of a new road. When a new route is introduced without any changes to existing routes, all the traffic transferred from the existing routes to the new one is regarded as diverted traffic.

(3) *Converted traffic*. This kind of traffic is created when travellers who previously made use of nonroad transport modes decide to use a new or improved road.

(4) *Normal-growth traffic*. This is traffic that would have normally developed in spite of the opening of the new facility. Such growth can be attributed to the following factors:

(a) general population growth

(b) an increase in the per capita ownership of vehicles

(c) an increase in average use per vehicle

(5) *Development traffic*. The construction of new roads in a region improves accessibility and thus stimulates economic development and the establishment of industries. Improved access generally results in changed and more intensive land use, which in turn attracts (develops) more traffic to the corridor through which the road passes. The volume of development traffic can be estimated by means of appropriate trip-attraction and trip-development indices for the new land uses.

(6) *Generated traffic*. This is entirely new traffic that has been generated solely by the improved or new road. Generated traffic consists of potential road users who have been encouraged to join the traffic because the new or improved road, by reducing the economic distance, has put new destinations within their reach. While development traffic can be attributed to improved accessibility, generated traffic can be attributed to increased mobility.

6.6.3.2 Estimating savings for each traffic component

The potential saving in road-user costs made possible by a new facility can be estimated in two ways:

(1) If the proposed alternative will completely *replace* the existing facility, a projection is made of road-user costs, assuming (a) a continuation of the existing situation and (b) the introduction of a new facility. Savings in road-user costs are estimated by subtracting the proposed facility's projected road-user costs from the existing facility's projected road-user costs.

(2) If the *addition* of a new facility to an existing transport network is being considered (ie existing routes and services by other modes are retained), road-user costs are determined by projecting these costs for a network that includes the additional facility and subtracting the result from the projected cost for the existing network with the additional facility.

The savings in road-user costs accruing to each individual traffic component are calculated as follows:

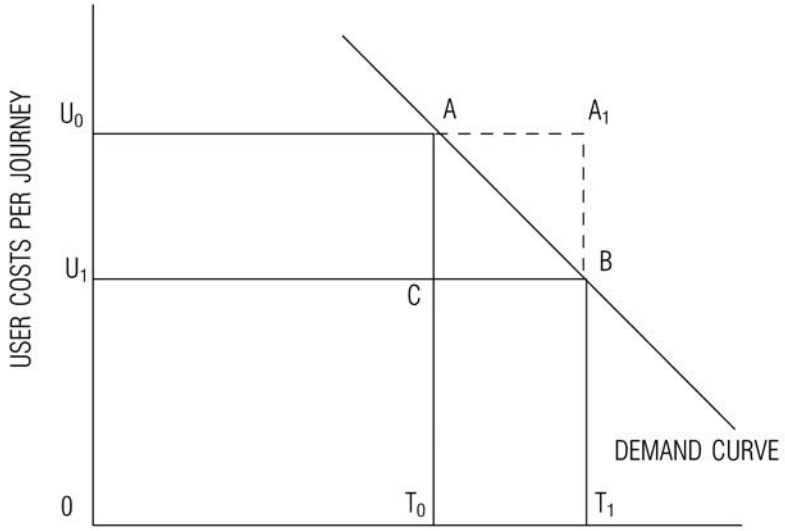
(1) The savings accruing to existing, diverted, development and normal growth traffic trips are computed by projecting each component's road-user costs between origin and destination, first with and then without the facility, and subtracting the first result from the second.

(2) The (usually negative) saving accruing to converted traffic is calculated by subtracting the cost of converted journeys from any cost saving that may have been effected by those transport modes from which traffic has switched.

(3) The saving accruing to generated trips is equal to half the saving per trip accruing to existing traffic. This arbitrary convention is based on the assumption that the demand for trips is represented by a straight line as depicted in figure 6.2.

Figure 6.2

Potential user cost saving over a specific period in respect of a new facility



Suppose that user costs for existing traffic on a new facility are reduced from U_0 to U_1 and the existing traffic volume is T_0. The saving for existing users is represented by the rectangle U_0ACU_1 in figure 6.2. If $T_1 - T_0$ represents the generated traffic volume, the "saving" achieved by generated traffic is not represented by the rectangle AA_1BC; instead the demand curve AB halves the "saving" to the area of triangle ABC.

6.6.4 Summary

Road-user benefits can be summarised as follows:

(1) reducing vehicle running costs in respect of:

(a) existing traffic.

(b) additional traffic:

- normal-growth traffic
- diverted traffic

- converted traffic
- development traffic
- generated traffic

(2) time savings in respect of:

(a) individuals

(b) vehicles

(3) lowering the accident rate and cost of damage:

(a) suffered by people

(b) to vehicles

(c) to goods and other property

6.7 Nonroad-user benefits (plus factors for the community)

6.7.1 A macroeconomic analysis

Road-user benefits (savings accruing to users) should not be regarded as a road's only benefits, since in fact they merely constitute the microeconomic advantages.

Nonroad-user benefits are indirect gains derived by the community from new or improved roads. Unlike road-user benefits, they are not real savings but represent a group of plus factors that emanate partially from incentives and investments in other economic sectors. Nonroad-user benefits thus constitute macroeconomic benefits, such as economic activity induced by generated traffic, the role of new roads in the establishment of industries or the stimulation of regional development, increases in land values when a region becomes more accessible, or increased strategic value for national and civil defence purposes.

While a transport economics or microeconomic evaluation usually focuses on the internal components of a project and assesses its economic efficiency in terms of expected savings in total transport costs, a macroeconomic evaluation is concerned with the project's external economic advantages. The purpose of the macroeconomic evaluation is to predict economic development and growth in an area as a result of the opening of a new road. Factors to consider include the anticipated multiplier and accelerator effects in the economy of a region or country. Forward and backward linkage effects in the flow of goods and services should also be estimated by means of input and output analyses. A macroeconomic analysis of a proposed road therefore entails a study of the impact of economic benefits on nonroad users in other sectors (Pienaar 1985:10).

Broadly speaking, nonroad-user benefits can be divided into economic and social benefits.

6.7.2 Economic benefits

Investments in roads may stimulate production, firstly by attracting new investors to an area and permitting the utilisation of idle resources, and secondly by permitting the release of resources for application elsewhere.

These effects are usually closely related to the economic environment, development potential, availability of factors of production and the state of the economy.

The economic benefits arising from investment in road infrastructure can be summarised as follows:

(1) Expenditure on road projects injects funds into the private sector and promotes production. This, together with increased demand for transport, stimulates the economy. If, during an economic upswing, the road network is able to cope with the increased traffic volumes and if excessive or frequent congestion can be prevented, the road network will fully serve its purpose as an economic activator.

(2) The stimulation of economic activity is associated with a rise in company profits and personal income, and the resultant increase in tax revenue boosts government revenue. Ideally, a portion of this government revenue should be used to cover some of the cost of the roads that originally induced the economic activity.

(3) New and improved road infrastructure facilitates access to property and increases mobility within and between suburbs and industrial areas, thereby giving rise to new economic land-use patterns. This in turn boosts land values, and increases the benefits for land and property owners. The revenue of local authorities from property tax also increases. Theoretically, local authorities should therefore have more funds available for improving the urban road infrastructure.

(4) New and improved roads not only boost economic development indirectly, but also have a direct impact on the establishment of manufacturing industries, distributors and utility (service) industries, especially in urban areas. The proximity of a major road is also important to retail organisations such as service stations, food retailers, hotels and businesses catering for tourists. Effective link roads also allow for the decentralisation of industries which may detrimentally affect the environment.

There are six main classes of nonroad-user beneficiaries:

- the general public
- land owners
- roadside enterprises
- roadside advertisers
- utility enterprises
- goods consigners and consignees

6.7.2.1 The general public

It can be argued that all the inhabitants of a country benefit from the existence of a road system because accessibility enables society to function efficiently. Without roads people would need to live close to their place of employment and would be denied a wide range of goods and services. Roads and streets provide access to fixed property (not just for occupants, but also for the suppliers of emergency services and facilities) and they facilitate personal and commercial transportation and the administration of law and order. For example, roads constructed in remote areas primarily for defence purposes could be said to serve the interests of society at large, rather than only the small group of users living in those areas.

6.7.2.2 Landowners

The value of land or fixed property is inseparably linked to its accessibility. In urban areas, roads provide access to virtually all properties, while in rural areas new or improved roads can considerably increase the value of property and land by bringing amenities and markets closer to the inhabitants. New or improved arterial roads in urban areas have a similar, but less marked, effect on land values. However, amenities are brought closer and commuter

distances reduced. A new freeway in an urban area often increases the value of properties in outlying suburbs by reducing commuting time to the central business district. On the other hand, the value of properties adjoining a freeway may drop because of excessive noise, vibration, pollution and visual intrusion.

6.7.2.3 Roadside enterprises and roadside advertisers

The survival of many enterprises depends on the traffic on a road nearby. Examples are service stations, fuel and food outlets, drive-in theatres and refreshment vendors. They benefit from road improvement, but suffer financially if traffic is diverted to a new route.

6.7.2.4 Utility companies

Utility companies may enjoy a right of way beneath or below a road, especially in urban areas, for example for the construction of water and gas pipelines and installation of electric and telephone cables. In Europe and North America, rail transport enterprises enjoy air rights above certain roads for the operation of elevated rail services.

6.7.2.5 Goods consignors and consignees

Goods consignors may also enjoy time savings in their capacity as nonroad users. These time savings have a greater value in developing than in developed regions, mainly because of a more limited supply of capital. Time savings in the consignment of goods can be beneficial in two ways:

(1) Faster deliveries mean lower storage costs.

(2) Highly perishable products can be distributed over a wider area.

Since speed can be an important factor in total production costs, and, as explained above, can put new outlets within a distributor's reach, the value of the time savings can be measured in terms of how much consumers and retailers are willing to pay for faster deliveries. However, one should keep in mind that the value of time varies with the time of day. When, for example, a delivery is made outside of business hours, a time saving has no value whatsoever.

6.7.3 Social benefits

The community enjoys a greater sense of security when a new road is built or an existing one is improved, for the following reasons:

(1) greater accessibility

(2) increased mobility

(3) more efficient civil defence services

(4) potentially more efficient protection services

(5) improved accessibility from a military point of view

Although nonroad-user benefits are extremely difficult to quantify, they are an important consideration in many road construction or improvement decisions, especially when a road has the potential of stimulating economic activity and development in new areas.

6.8 Road cost allocation

6.8.1 The principle of user charging

It is an accepted economic principle that in order to ensure the equitable allocation of scarce resources, users or consumers of these resources should, if possible, bear the full and actual cost of their use or consumption. This will ensure among other things that the various transport modes compete on an equal footing and that road users and other beneficiaries bear their fair share of the costs, so that they do not have to be subsidised from other sources.

This principle is strictly pursued in the South African policy as discussed in 6.5 above. It is therefore clear that road cost allocation should be based on a theoretically sound and equitable road costing procedure. Two methods outlined below exist for costing such a system.

6.8.2 Historical cost method

According to this method, the *historical* costs (which are sunk costs) of constructing, expanding or improving existing roads are spread over time between successive generations of users. The method involves two steps: first an estimate is made of the value of the capital tied up in the physical facility, and then an appropriate discount rate is selected whereby the value of this capital amount can be spread uniformly over the service life of the road network.

6.8.3 Development cost method (current expense method)

This method ignores the historical costs of existing roads and concentrates on recovering current or future costs associated with the construction, maintenance, expansion or improvement of the road system by one of two methods:

(1) the development cost or long-term marginal cost method

(2) the incremental method

These two methods are briefly explained below.

6.8.3.1 The long-term marginal cost method

The cost of providing an additional road or future road space is based on the value of future services made possible by the road. Current users are thus charged in advance for future costs.

6.8.3.2 Incremental method

The incremental method is similar to the long-term marginal cost method, except that it also incorporates marginal cost per period – in other words, “pay as you go”. Investment costs are regarded as current costs in the year of expenditure. In effect, this method regards road expenditure per period as the road cost for that period.

6.9 Practical road cost recovery methods

The principle of user charging requires that mechanisms should be put in place to recover the costs of road use from users. This principle is emphasised in the South African policy in this regard, and in our discussion of this policy in 6.5 above we referred to some of these methods or mechanisms. The discussion below investigates these and other methods in more detail.

Activity 6.4

Before continuing, go back to section 6.5, and see whether you can identify a few road cost recovery methods which comply with the principle of user charging. Try to add to the list, and decide whether the methods you came up with are justified and are actually associated with road use. Study the discussion below and then add to your list.

The revenue sources for financing roads can be divided into five groups, ranging from specific taxes on vehicle use or ownership to general taxes levied on society as a whole. Each revenue source is discussed under a separate heading. This discussion is not exhaustive and is confined to the methods commonly applied, both locally and abroad.

6.9.1 Tax relating to vehicle use

6.9.1.1 Fuel tax

Fuel tax, the most common levy on road users, is found in almost every country in the world. Its relative popularity is due not only to its simplicity, the ease with which it can be levied, and its general cost effectiveness, but also because of its basic characteristics:

(1) It is paid for every kilometre travelled.

(2) It varies according to the nature of the vehicle, for example, with mass and power.

(3) It varies according to the speed at which vehicles travel.

(4) It varies according to the manner in which vehicles are driven.

Because of the comparatively direct relationship between fuel tax and road use, it provides an attractive basis for road cost recovery. Fuel tax avoidance is virtually impossible and administration costs are low in proportion to total revenue generated. It is moreover acceptable to the public because it is paid in small quantities at frequent intervals to satisfy immediate needs. Even so, it is not the ideal solution because it cannot recover the full costs associated with heavy vehicles. One possible solution would be to impose differential taxes on petrol and diesel fuel, with diesel fuel users (mostly heavy vehicles) having to pay a larger levy than petrol users.

The disadvantages of this method are firstly that users are not realistically charged according to the geographic distribution of their trips. The majority of trips in South Africa take place in urban areas, while a relatively small percentage of road lane kilometres are found within these areas. Secondly, fuel consumption on a good road, which costs more to build, is less per vehicle than on a low-quality road at the same speed.

6.9.1.2 *Tyre tax*

Tyre tax is based on another kind of variable cost. Tyre tread wear (and hence tyre tax) is progressive and related to use. Tyres wear more quickly in urban traffic conditions, where frequent stops, speed change cycles and corners increase friction and cause serious tread abrasion.

Tyre levies have serious disadvantages. Although the demand for tyres is not invariably price elastic, an increase in the price of tyres could induce vehicle users to prolong the use of existing tyres which detrimentally affects safety. At present prices of tyres in South Africa do not include a dedicated road levy. Apart from VAT, the only form of government levy on tyres is a duty on certain imported additives used during the manufacture of tyres and the imported material used to reinforce radial tyres.

6.9.1.3 *Levies on vehicle spare parts*

Duties charged on vehicle spare parts are relatively insignificant compared with levies on new vehicles. They are not a very good measure of road use since much depends on the vehicle's age, type and make. The total revenue from taxes on lubrication oils and grease is likewise insignificant and this type of tax is not really suitable as a road-user charging instrument. There are no dedicated road fund levies on these commodities in South Africa at present, and the customs, excise and import duties that do exist are regarded as general government revenue.

6.9.1.4 *Mass-distance tax*

This form of tax is intended to supplement fuel tax to compensate for the shortfall in revenue recovered from heavy vehicles. Whereas this deficit can be made good from heavy vehicle licence fees, these do not allow for distance travelled, so that an articulated truck travelling 10 000 kilometres per year pays the same licence fee as one travelling 90 000 kilometres. While fuel tax will to some extent compensate for this difference, distance travelled should somehow also be taken into consideration to ensure an equitable distribution. From a theoretical point of view, mass-distance tax ensures a much better allocation of financial resources. Progressive rates are established for vehicles according to mass (or damage caused to roads). The appropriate tax is then based on the distance travelled by a particular vehicle. The obvious drawback of mass-distance tax is that it involves an additional cost burden to both government and the taxpayer. On the one hand, auditors, inspectors and law enforcers have to be appointed to combat tax evasion; and on the other hand, this form of tax involves additional administrative costs and expenditure since transport operators have to install distance meters. The question is therefore whether the extra cost burden is justified by the savings achieved in terms of equitableness and economic efficiency. Mass-distance taxes are not currently imposed in South Africa.

However, there is no reason why mass-distance charges or simplified charges on progressive distance should not be earmarked for a road maintenance fund.

6.9.2 Tax on vehicles

6.9.2.1 *Duty on new vehicles*

Tax on new vehicles is levied in the form of import duty designed to protect and promote the local vehicle manufacturing industry, and excise duty (or value-added tax), which is imposed on domestically produced vehicles. Excise duty is a steady source of revenue that can be varied to encourage or discourage use of a particular type of vehicle (eg in

the interests of safety, to combat air pollution, or to promote fuel conservation). Both these types of tax apply to new vehicles only, which distinguishes them from annual licence fees, which are payable by owners of all vehicles, old and new alike. Although the main purpose of import tax on motor vehicles may be to protect domestic vehicle manufacturers, it is customarily regarded as general government revenue.

6.9.2.2 Licence fees

Licence fees are charged per period (usually one year) for the right to operate a vehicle on public roads. Licence fees charged vary according to various factors, including weight, size, engine output, type of vehicle and value, but not according to the distance travelled. Next to fuel tax, licensing is the most popular method of road-user taxation. Licence fees cover not only the administrative expenses of the licensing authority, but are usually applied to cover fixed road costs. However, because they do not vary with use, they cannot be used to discourage marginal trips.

In South Africa, licence fees vary between provinces, and are based on tare mass, with little regard for the real road costs associated with vehicle type. Admittedly there has been an adjustment to the licence fees of heavy vehicles, but they still do not reflect the relative costs brought about by heavy vehicles.

6.9.2.3 Axle or wheel tax

This type of tax, if based purely on the number of axles or wheels, is counterproductive because in some cases, heavy vehicles with a large number of axles cause less damage to roads than those with fewer axles. This type of tax does, however, distinguish to some extent between light and heavy vehicles. This form of tax is not applied in South Africa, although toll fees are levied on the number of axles on a vehicle.

6.9.3 Tax on place of use

6.9.3.1 Supplementary licensing

Supplementary licensing is a system whereby vehicle entry into a designated area during specified hours is restricted to vehicles displaying an appropriate supplementary licence disc. This type of road-user charge in effect constitutes a restrictive measure well suited to densely populated urban areas.

The main advantage of supplementary licensing as a method of combating traffic congestion lies in the power of the licensing authority to control the issue of permits and therefore to some extent also traffic levels in the restricted area. The system also has the advantage of generating revenue, although it should be pointed out that revenue maximisation (by allowing the unlimited issuing of permits) will defeat the primary restraint objective.

6.9.3.2 Toll systems

Toll systems involve a charge for the use of a facility. The purpose of toll systems is twofold:

(1) to finance costly facilities on the basis of “the user pays” principle (eg bridges, tunnels and high-quality roads).

(2) provide a more equitable method of charging users of a facility according to the damage caused by use.

Disadvantages of toll systems include the following:

(1) The construction and administrative costs involved in toll levies are relatively high. Thus toll levies are higher than the cost of road use because they also have to provide for the recovery of these costs.

(2) If a sufficient number of other forms of road-user charges and indirect taxes exist to cover the total expenditure on roads, there would be an over recovery or double recovery of road-user costs.

(3) Toll systems discriminate against people living near the toll facility who have no choice but to use the facility, as opposed to residents of other areas who enjoy free access to the road system without having to pay toll fees.

A general principle of toll roads should be the availability of alternative routes.

A further principle in toll fixing is that the charge should be smaller than the saving that can accrue to the user by using the facility. Obviously, toll fees should not be higher than the *perceived* cost of using the best alternative route. Since the operating costs of heavier vehicles are higher, differential charges can nevertheless be used to some extent. Also, by equipping roads with devices to determine axle mass, overloading of vehicles can be controlled.

This method of road financing is fast gaining popularity in South Africa.

6.9.4 Taxes imposed by local authorities

6.9.4.1 Property and land taxes

Property tax is commonly regarded as a general source of local government revenue. The use of local taxes to finance roads is generally advocated because these taxes are evenly distributed, fairly simple and inexpensive to administer and roughly proportional to wealth. Inequities that may exist are attributable mainly to varying population density and family incomes. Road financing by local authorities must, however, compete for funds with other local bodies providing health services, amenities and other essential services.

6.9.4.2 Service charges

More than half of local government revenue in South Africa derives from charges levied by service departments – especially electricity, public transport, water and sewerage departments. These levies are usually proportional to the service provided and cross-subsidisation is therefore minimised. Since it is sound economics to finance such facilities as far as possible out of the charges levied for each service, it cannot be recommended that cross-subsidisation be used to finance road provision. However, road provision could also be regarded as a service and in this case, parking fees or levies would be appropriate instruments for recovering road costs.

6.9.5 General revenue sources

General revenue sources that are applied for road construction include subsidies, loans or direct allocations. In the case of a direct allocation by the treasury, the money may derive from an earmarked (or dedicated) fund, in which case the revenue can be used for a specified purpose only, or from central government general revenue. Funds may be earmarked by creating a trust fund. This has the advantage that road authorities have a secure source of income as opposed to the alternative where taxes are paid into the

general state revenue account, and government then allocates funds to road authorities on an annual basis.

In conclusion, we briefly review the use of loan funds for capital projects. Traditionally, local authorities have made extensive use of loan capital. Loan financing can be advantageous if handled judiciously, because it facilitates the provision of a higher quality infrastructure at an early stage, regardless of whether or not inflation is a consideration. In an inflationary situation, the use of loan funds is further justified by the fact that the loan is effectively repaid with "cheaper" money, assuming that the applicable interest rate is lower than the inflation rate.

When constructing a facility that is expected to last 20 years or longer, the advantage of loan financing is that the costs are spread over time, in such a way that the current generation does not subsidise the next generation to any great extent.

Government obtains the general funds from which road provision is financed from different sources representing revenue from all kinds of economic subjects. These sources range from company tax to personal income tax and value-added tax. The revenue deriving directly from South African *road users* (it need not necessarily emanate directly from road use per se) includes the following: company tax payable by professional carriers, value-added tax on vehicles and other transport inputs, fringe benefit tax on vehicles supplied to employees for their private use, customs, excise and import duties on goods associated with the transport process and fines payable for traffic offences and other contraventions.

6.10 Conclusion

In this study unit we discussed the provision of road infrastructure and who is responsible for the costs involved.

In most cases the authorities are responsible for the provision of roads, while road users only pay for them indirectly in the form of taxes and levies.

The demand for roads depends on the demand for road transport which, in turn, also has unique characteristics of which the peak phenomenon (peak hours, peak periods such as holiday times) is the most important. The total cost of road transport comprises infrastructure costs, community costs and road haulage costs.

The South African policy on infrastructure provision is spelt out in the document entitled *Moving South Africa: the action agenda (a 20-year strategic framework for transport in South Africa)* (South Africa 1999). A holistic approach is followed in which investment in transport focuses on prioritised customer groups, namely urban passengers, rural passengers, tourists and long-distance passengers.

As far as urban passenger transport is concerned, the policy focuses on corridor development, public transport facilities along corridors and better utilisation of road space by means of control measures and the pricing of externalities. A distinction is made between strategic and supporting corridors. In the case of rural communities, the policy is aimed at the establishment of a framework for prioritising roads on the basis of development needs, development potential and other social criteria. Regarding tourism, an integrated approach is followed in which the Department of Transport is intensely involved in tourism strategies. The focus is on coordinated infrastructure development between tourist attractions and payment for use by tourists. The provision of infrastructure for road freight transport falls within the framework of a seamless transport system in which all freight transport modes and networks are integrated. The policy on infrastructure development also focuses on the upgrading of the transport system in order to promote the export of high-value manufactured products. As in the case of urban passenger transport, the focus is on corridor development with a strategic network of dense corridors supplemented by a supporting network. The principle of user charging for infrastructure and externalities is emphasised throughout the policy.

The benefits of road-building projects can be divided into two broad categories, namely road-user benefits or savings and nonroad-user benefits.

Road-user benefits comprise the following:

- reductions in vehicle running costs
- time savings
- a reduction in the accident rate

To quantify road user benefits, it is necessary to identify the following components of anticipated traffic:

- existing traffic
- diverted traffic
- converted traffic
- normal growth traffic
- development traffic
- generated traffic

Nonroad-user benefits are indirect gains derived by the community from new or improved roads, and can be classified as follows:

- economic benefits for:
  - the general public
  - land owners
  - roadside enterprises and advertisers
  - utility companies
  - goods consignors and consignees
- community benefits which primarily revolve around accessibility and mobility

The revenue sources a government may consider for road financing fall into five main categories:

- tax relating to **vehicle use**
  - fuel tax
  - tyre tax
  - levies on vehicle spare parts
  - mass-distance tax
- tax on **vehicles**
  - duty on new vehicles
  - licence fees
  - axle or wheel tax
- tax on **place of use**
  - supplementary licensing
  - toll systems

- taxes imposed by local authorities
  - property and land tax
  - service charges
- general revenue sources

Each of these road cost recovery methods has advantages and disadvantages. Some are more justified than others because they adhere to the principle of user payment.

6.11 Self-evaluation questions

(1) Briefly discuss the features of the demand for roads.

(2) What is meant by total road transport costs?

(3) Describe the authorities' role in road provision.

(4) Discuss in detail South Africa's policy on the provision of road infrastructure.

(5) Discuss the benefits (savings in road-user costs) of a new road for users. Differentiate clearly between the various user categories and explain how you would calculate the benefits for each traffic component.

(6) Kromfontein and Reguitspruit are currently linked by a single tar road 50 kilometres in length. Since the discovery of oil in the past three years, the area has experienced unprecedented economic development. The result has been an enormous increase in traffic with the concomitant traffic congestion on the existing road. The relevant road authority has asked you to conduct an economic study on the possibility of a new road by evaluating the advantages of such a road.

Compile a preliminary report in which you explain clearly to the road authority all the factors you would take into consideration in such a study. Where possible, also indicate how the various benefits of such a road can be quantified. (No calculations are required.)

(7) “It is an acknowledged economic principle that users should where possible bear the full and actual cost of their use to ensure that scarce economic resources are equitably allocated to users or consumers.”

(8) Elaborate on this statement by referring to the allocation of road costs among users and to practical methods of recovering these costs from them. Which road cost recovery method(s) do you regard as equitable? Briefly substantiate your answer.

(9) In your opinion, should the following sources of revenue be earmarked for the financing of road provision? Give reasons for your answer in each case.

(a) a levy incorporated in the retail price of fuel

(b) value-added tax

(c) vehicle licence fees

(d) fines for traffic offences

STUDY UNIT 7

Planning and investing in seaports

UNIT OUTCOMES

After working through this study unit you should:

- understand the important role ports play in regional development
- be able to indicate the levels at which port planning should take place
- be able to explain why governments are the major investors in ports

KEY CONCEPTS

- Hinterland
- Planning in ports
- Transport integration
- Port investment

7.1 General

No better introduction can be used to indicate the impact of a port on its surroundings than the following statement of philosophy presented by the Port of Rotterdam-Europort (Taylor 1974:139).

> *In the same way as a painting is not confined by the edges of the canvas and the surface of its paint, but has in addition a life of its own, a fourth dimension as it were, so a port is more than its quays, more than the depth of its water, more than is visible to the eye.*

Although a port is a business just like any other business, the effect it has on its environment is of national importance. This effect is felt not only in the areas immediately surrounding the port, but also in those areas where cargo is received for export and to which cargo arriving via the port is delivered. These areas, known as the hinterland of the port, may be hundreds of kilometres away from it. Ports also play a crucial role in the economic wellbeing of a country, and this economic wellbeing is influenced by all the changing concepts of transportation available for the movement of products, the changing and fluctuating pressures on the country's economic viability and the forces exerted by world currencies.

Fundamentally, ports grow by virtue of the trade they can attract and maintain. Prior to the Second World War, this was relatively straightforward because ships were reasonably consistent in their habits of picking up and transporting commercial goods. Changes in financial and commercial considerations, the cost of factors of production, and changes in distribution and upward trends in transport costs have altered all of this. Nowadays, goods that are more reliable require less handling and can therefore be processed more quickly, minimise the total cost of transportation and thus give the country a relative advantage when trading with other nations.

Whatever individual transport modes are used to move a country's products is immaterial; it is ultimately the transport system as a whole that affects the transport costs involved. In each of the transport modes available, the systems design and size as well as the related supporting services (labour and equipment) are likely to remain constant for some time. This applies especially to ports – hence the importance of regarding ports as part of a nation's integrated transport system and not as an isolated unit.

In this study unit, we shall be discussing the planning of and investment in ports from this perspective. We first examine the influence of ports on their environment, how ports should be planned within a national framework, and finally, investment in ports.

7.2 The interaction between a port and its environment

7.2.1 Introduction

The earlier understanding of a port as a place to which ships report to load and discharge cargoes – a point of transfer between land and sea – is outdated. The interaction between a port and its environment is such that it attracts and leads to the development of cities, industries and business. A new dimension has therefore been added. Current views lean towards the attraction of industrial development to port areas, known as MIDAS (Maritime Industrial Development Areas) as is evident in most ports around the world. This development is not confined to the immediate port area but extends to its hinterland.

7.2.2 A port and its hinterland

7.2.2.1 The concept of the hinterland of a port

The hinterland of a port refers in broad terms to the land side of the port from where commodities are received for export or to which commodities are sent which have been received at the port. A cost component should be included to accommodate competition between ports. For our purposes, the concept of the hinterland of a port can be defined as follows:

(1) *Natural hinterland*. This is the hinterland of a port where costs are sufficiently favourable to preclude the possibility of goods being diverted to another port which may have other cost advantages.

(2) *Competitive hinterland*. This is the hinterland where the lower cost of transport to one port is offset by another port's other cost advantages. The two ports therefore have to compete as effectively as they can for traffic from the hinterland.

The hinterland is therefore an economic rather than a geographical concept because a port can have as many hinterlands as the different types of commodities it handles. However, the following three factors have a major influence on the efficiency of a port as part of the transport system used:

(1) the pattern and nature of export cargo generation and the demand for import cargo in the hinterland

(2) a port's installations and equipment, which control the level and variety of cargo throughput that a port can handle

(3) the efficiency of surface transport networks in the hinterland used to assemble and distribute cargo

The volume of traffic handled by a specific port is determined by *inter alia* the choice made by importers, exporters, despatchers, receivers or other interested parties about specific consignments of goods. Their choice of port is not an irrational decision but is based upon economic considerations. Each consignment will incur a different total cost depending on the port through which it is sent, and the port with the lowest total cost will be selected. The total cost incurred by a consignment sent through a port is made up of the following cost elements:

(1) transport tariff from point of origin to port

(2) transport utilisation costs

(3) all port tariffs including storage and penalty rates

(4) port utilisation costs

(5) freight tariffs and costs

7.2.2.2 Utilisation costs

Bear in mind that a distinction is made between tariffs and utilisation costs. The total cost incurred by users of a service is derived not only from the tariff (or rate) they pay for this service, but also from costs incurred as a result of the use of the service or facility. These costs, which are termed “utilisation costs”, concern the qualitative factor of transport in general. Port utilisation costs, for example, consist of the following:

(1) the speed of handling of vessels and goods

(2) the availability of ports at all times and in all weather conditions

(3) the reliability of ports

(4) the suitability of port services and facilities for the specific consignment

(5) the capacity of ports

(6) the linking ability of ports

(7) attention to the products which flow through ports

(8) the safety of ports

These factors, which have mainly to do with the time and intrinsic utility of the infrastructure and moving assets, therefore affect the costs incurred by the dispatcher of the goods.

The planning of seaports and the subsequent transport development can therefore have a profound effect on the economy of the immediate hinterland and the entire country, of which they are part. A port that develops a reputation for good service and maintains its facilities to this end becomes attractive to shipping, and as a result, may well indirectly encourage the setting up of ancillary industries associated with the trade of the port, from which other transportation modes develop.

This also applies to instances where new ports are introduced or existing ones modernised/expanded, as evidenced by the development of Rotterdam, both before and after the Second World War. This port has brought a major and densely populated area of the

continent of Europe into direct contact with overseas countries and areas and has had a profound effect on the expansion of Western Europe (its hinterland). It has also promoted the flow of raw materials and foodstuffs inwards and manufactured products outwards to an extent that was regarded as highly unlikely prior to the Second World War. Much the same can be said of the port of Durban, although probably to a lesser extent. The fact remains, if a port is located in a position that has direct and deep-sea water approaches, and is supported by good inland transport systems, wide economic and social influences occur in its hinterland.

The corollary is equally true, namely that one form of economic influence attracts others, and associated forms of transportation develop, opening up large areas of sociotechnical expansion which, in itself, creates new and expanding population belts. This has complex advantages for the development of both the port itself and the hinterland it serves in meeting changing patterns of trade, which, in turn, stimulate far-reaching technical and economic changes in the inland transportation of goods.

As far as containerisation, in the modern sense, is concerned, ports designed for this kind of traffic will, because of their very nature, influence road and rail links. Hence these types of ports should continue to establish and expand their facilities so that the transport modes can encourage shippers to use the routes they serve. The cycle of events can only continue with complete integration. Infrastructure in this sense is of vital importance, and is determined largely by the equivalence between the various land/sea outlets and the economic development of the area they serve. The former is related to the attraction of ships to the port and the latter to the commercial/economic and financial expansion that occurs in the area served.

7.2.3 A port and its immediate surroundings

The influence of a port on its immediate surroundings can be seen in terms of the agglomeration factors that accrue. Agglomeration factors are the economic advantages that occur when an enterprise is grouped with other enterprises. The tendency to agglomerate cannot be attributed to a single cause but is the result of the convergence of economies of scale, economies of localisation and economies of urbanisation.

Economies of scale apply inside ports, and refer to situations where the average total cost (unit cost) declines as the quantity produced increases. Such economies originate from the indivisibility of machinery and other factors of production, and for various other reasons. Economies of scale induce enterprises to concentrate their activities geographically.

Localisation economies refer to the benefits that accrue to enterprises in a single industry because they are located near one another. A typical example is the car manufacturing industry. In a port, the fishing industry, boat repair industry and other similar industries are examples.

Urbanisation economies refer to the benefits that arise because of the geographical closeness of a variety of completely different industries. Because of the sheer size of the common investment, these industries might, for example, enjoy better public services at cheaper rates and avail themselves of a host of private services such as easy borrowing, access to research and development centres and better business services.

Although localisation and urbanisation economies are quite different, they both result from complementarity and linkages between enterprises. Historically, agglomeration economies tend to be cumulative so that development becomes bound by an irreversible historical pattern. They also tend to discourage firms from moving away individually.

The three types of economies need not climax at the same point in time – in fact they rarely do. Once an enterprise has grown to the point where it suffers from diseconomies of scale, it may elect to stay because it enjoys localisation and urbanisation economies. In such a case, the port acts as a magnet because it induces the firm to settle close to it.

This is also one of the main reasons why most of the largest cities in the world are situated next to ports.

7.2.4 Ports and sociopolitical considerations

If it is accepted that a port is a place to which ships report to load and discharge their cargoes, then it should follow that as ships develop in size and character and the fewer the number of constraints upon their movement, the more they are able to support the port. Since shipping encourages trade, or follows its expansion, areas of manufacture or raw material processing will use those facilities that provide better connections and effective distribution. Where this occurs a snowball effect results in trade patterns which encourage economic expansion and population increases because of the resultant employment opportunities.

This, in turn, has a sociopolitical effect because a country now has to balance its geographical economic structure by stimulating areas that have the potential for economic development and avoiding areas that do not. All of this influences ports, which then have to maintain themselves in the environment that has been created and utilise their facilities, or indeed, introduce other facilities to support this socio-technical, commercial and political planning.

Using the Rotterdam complex as an example, it is interesting to note that whereas in 1938, 20 000 ton vessels were considered to meet the trading needs of the port, in 1970, 200 000 and 250 000 ton vessels were not uncommon, with an increase in annual trade tonnage moving through the port from 42 to 226 million tons in the same period. Although much of this expansion was undoubtedly due to the exceptionally favourable economic and technological developments in Western Europe, it would not have been possible without a suitable sea outlet, and Rotterdam, as a port, may not have grown so rapidly if these developments had not taken place.

A seaport has to connect its services to the various other means of transport and its customers in the hinterland it serves. It has to adopt a flexible approach to optimum efficiency, either by accepting limits beyond which it cannot move financially, or supplementing its services resulting from corresponding development in the hinterland traffic routes.

The experiences of the major ports of Western Europe – particularly those that were reconstructed after the Second World War – are useful examples of the way in which certain ports, and the cities of which they are part, have worked to promote not only themselves, but also the regional economy in which they operate.

7.2.5 Ports and industrial/population considerations

A seaport and its development have to be planned within the national context since the benefits or difficulties involved stem from environmental and industrial influences. In Britain, for example, the main established industrial/population belt of the UK is situated more or less to the northwest, from the River Thames area to the Manchester/Liverpool areas, with a high concentration in the mid-midlands conurbation. In South Africa, most of the industrial/population belts are situated next to the seaports, or along their connection with Gauteng.

High intensity light/medium/heavy industry has been established along these belts for years, and the populated areas have been affected by overcrowding. Strips of heavy industry are found in those areas with a predominance of raw materials, a case in point being coal production in Mpumalanga. Much of the remainder of South Africa is less industrially developed.

A study of the present and projected road systems of the country (South Africa specifically, although this is also applicable in general) indicates that the authorities are thinking about servicing the established industrial/population areas and possible expansion outwards

from them, bearing in mind that trading patterns are likely to require a broader balance of internal development than has traditionally been the case, which will create a need for different sea outlets. This is sociopolitical planning, which is strengthened nowadays by the necessity in national development for an awareness of environmental needs. The obvious question here is how ports will fit into this pattern, where the emphasis should be, and what the priorities will be in their planning.

7.3 The level of planning of seaports

7.3.1 Generalisations about the planning of ports

The increasing trend in the planning of ports in a national context is, and will probably continue to be, towards the integrated transport concept. Whereas, traditionally, ports were places that received cargo, nowadays the receiving and distribution function is influenced by a variety of factors that were not always considered as vital in the planning of a port as the volume and nature of traffic and cargo envisaged. It is the kind of goods and their value and the types of ships likely to be encouraged to load this type of cargo which determine the desired structural investment, facilities and services. Not all types of ships work at the same speeds because many require specialist handling. These factors should be considered within the framework of a country's envisaged economic development programme and the resultant comprehensive port arrangements. Berth and quay arrangements should be planned accordingly.

Planning the use of mechanical/electric mobile and fixed equipment should not be done in isolation. The economic use of this equipment depends on the way in which the trade of the port permits its adaptation to systems of packaging and unitisation of goods. The acquisition of additional equipment or modernisation of existing equipment should also be given attention and liaison with the manufacturers of equipment is therefore necessary. In most forms of transportation, specialised equipment is used for particular purposes. Such arrangements suit both the customer and the carrier. Seaports would be at a disadvantage if their equipment for transferring goods from ship to land, or vice versa, were functionally irreconcilable.

7.3.2 The planning of ports

Because of the influence that ports have on the region and/or economy of a country, their planning needs to take place at national, regional and port level.

7.3.2.1 Planning at national level

In the South African context, port planning at a national level is done by the National Ports Authority. Of particular importance is the Maritime Industrial Development Areas concept (MIDAS), which is a way of thinking about supplying infrastructure which is able to accommodate certain commodities and industries in a port, for example:

Commodity	Industry
Ferrous and nonferrous ores	Steel and metal manufacturing
Grain	Food processing
Fertilisers	Agricultural and domestic needs processing
Chemicals	Chemical and detergent processing for industrial and domestic use
Timber	Furniture, paper and packaging manufacture
Fishing	Food processing and fertilisers

Seaports in the modern sense of the word are profitable industrial locations. The advantages of industries which are port sited not only have a direct effect on the port in question through the importation of raw materials in large ships, but also promote more acceptable distribution facilities for the industries that settle in port areas. Industrial estates influence other "economic", commercial and socio-technical development. The continuing availability of increasingly large-scale bulk raw materials in large carriers may well influence a wider acceptance of the MIDAS concept in supplying port infrastructure in the future.

In developing countries, the use of the MIDAS concept is less likely to be restricted by physical constraints, and port planning may even be more directly influenced by the extent to which a country's industries are situated nearer to a port.

To meet the special needs of maritime industrial development seaports, numerous factors have to be considered in association with national and local government, especially those related to

(1) the provision of adequate sea and land access

(2) the assurance of adequate transport facilities

(3) adequate and unrestricted space for development (The National Ports Authority has advised that a minimum of 500 hectares be set aside.)

(4) considerations about port trading resulting from the growth of secondary industries and services according to the MIDAS concept

(5) focusing on the continuing work of economic and maritime research associations

Understandably, as the above forms of development occur, ships will start to support the resultant flow of trade, and port facilities will grow accordingly.

The National Ports Authority should also take the following factors into consideration:

(1) National government strategy on the development of regional areas, both in terms of industrial and population values, should be noted. Of significance here are a number of economic and commercial considerations related to envisaged trade probabilities and the ability to attract and use them to the best possible social/financial advantage of the region. Included here will be estimates of transportation facilities, in terms of both size and type, which can be borne by the finances available, with due consideration of the physical geography of the country and its suitability to all or some of the transportation facilities considered necessary.

(2) Another important factor is manufacturing and processing probabilities. In older established countries that already have industrial and commercial areas, national economic planning may well have to decide whether it is preferable to allow these areas to remain where they are and service them from the ports by using adequate transport, or to diversify these areas by developing major sites for these industries nearer or adjacent to the ports. This will allow them direct accessibility to the raw materials that have to be imported, and result in faster distribution of manufactured products from a nearby port. This is a notable development in the steel, chemical, timber and motor manufacturing industries in the UK.

The planning of seaports within the framework of broad national strategies also influences social considerations. Whereas previously a port would have had less influence on social development in the region, nowadays, its greater industrial involvement will be associated with a number of secondary and service industries which may offer employment to the local population.

This form of development compels the national government to balance its social responsibilities towards the population, and the port through its involvement not only becomes a major roleplayer, but also benefits from a healthier hinterland.

7.3.2.2 Planning at regional level

Planning at regional level will be done by the regional governments in consultation with the National Ports Authority. Local government strategy will focus on the extent to which they are able to support a national plan by providing the land which a port may consider necessary for its development and the time periods within which local government can introduce or improve road access, housing, industrial sites, local services and water availability, such as in the reclamation of shore areas.

With regard to both the local and national level, the planning of a seaport becomes a direct part of the national and local integrated development – hence the traditional role of providing a place for the "interchange of goods" takes on a wider meaning. In terms of direct liaison, such forms of planning require wider associations with appropriate national and local committees. Thus the port's staff need to have specialist knowledge and experience to make a positive and objective contribution in discussions with these committees on matters in which they represent the port's interests.

7.3.3 Planning within ports

It is not only the influence of the port on its environment that should be planned but also the port itself, specifically the total distribution network, technological developments in terms of quays and handling equipment, staff and labour deployment and environmental considerations. We will briefly look at each of these factors.

7.3.3.1 Ports and their distribution systems

Since the development of a seaport is obviously closely related to that of both the hinterland and the ships it serves, the planning function is dependent on a number and variety of statistical information pointers which have considerable bearing on the basis of its business. Port planning therefore needs to consider and anticipate the probabilities of technological development, together with:

(1) trends in commercial/financial developments.

(2) political and social national and international trends.

These trends need to be considered in the light of both national port coverage and individually located harbours.

The statistical information necessary for any planning exercise should include the following:

(1) national (government) publications on the movement of goods, indicating the balance of exports against imports in terms of quantities and quality, modes of transport, countries of origin and destination and customs procedures

(2) systems of documentation pertaining to international traffic

(3) comparisons of trends in the types of commodities forming shipping loads, that is, bulk, container, unitised, conventional, vehicular, dangerous, as well as passengers

(4) information on ship design and shipbuilding trends

(5) information on the latest developments in mechanical handling aids

(6) the prospect of industrial development, particularly where this has a bearing on seaports

(7) trends in water/tide movements and siltation because special arrangements may have to made for dredging and other engineering factors may have to be considered

(8) information on international and national navigational and conservancy arrangements

7.3.3.2 Technological developments

Seaports are changing from a labour-intensive to a capital-intensive industry, and technological change affects both labour and equipment in this respect. In the latter case, maritime research has focused on developing different types and sizes of ships able to handle cargoes most efficiently, so that larger tonnages can be carried in fewer instead of large numbers of vessels.

As far as planning is concerned, it is unlikely that the different types of ships that operate today will keep the same design. Future changes will probably reflect the type of "cargo" to be carried. It is therefore important to keep abreast of possible economic/commercial changes in world markets and in the technological design of ships, because in the past, the former influenced the latter. This brings us to obsolescence.

Modern design and planning of ports require a completely different approach from that used in the construction and development of older established ports in the later 19th century and early 20th century. Solidarity of structure in both site and equipment was primarily the order of the day. Much of the design, however, showed foresight.

These days, rapidly changing technology is a crucial factor that affects the dynamic design approaches to all forms of port structures and equipment. Thus the return on capital invested and depreciation are two important considerations. With regard to the latter, the introduction of designs that could become obsolete in the near future should be avoided. Any piece of equipment (albeit a heavily capitalised one), such as a portainer crane, or a roll-on/roll-off terminal, could well be unsuitable in its original design in 10 years' time, if ship patterns continue to change as rapidly as they have in the past 10 years. It is therefore imperative to evaluate trends in the design of waterfront structures with the economic use of structural materials and design for obsolescence in mind.

7.3.3.3 Ships and handling equipment

As far as seaports are concerned, ship design influences the provision of adequate sea approaches, appropriate berthing arrangements, suitable wharves, berths and quays and correct mobile and fixed equipment. It therefore follows that in the planning of seaport facilities, the authorities should take cognisance of the types of ships they wish to attract.

The above also applies to the equipment provided by the port. Container (portainer) cranes, straddle carriers, fork-lift trucks, various types of trailers and more conventional cranage systems are manufactured in a competitive market which means that the question of design is just as important as that of capabilities and areas of use. Reference to manufacturers' specifications is therefore necessary when deciding on the most economic types of equipment needed to handle the ships and their cargoes in the port.

Overcapitalisation on equipment should be avoided, since return on investment is conditional on the extent to which the equipment is likely to be used.

Since planning is an ongoing exercise, the question of documentation in programming and computerising traffic reception and delivery to and from ships, and in the port itself, is a technological one that needs to be addressed. Furthermore, where hydrographic research facilities are available, model simulations of water movement and estuarial conditions afford useful information on the citing of facilities for ship reception. This form of research applies to the building and maintenance of breakwaters, locking systems, ship moorings, towage arrangements and dredging needs.

7.3.3.4 Staff and labour deployment

Technological development and research with regard to staff and labour deployment have indicated that it is now becoming more important that planning in this area should be aimed

at ensuring a balanced personnel complement. Nowadays, sophisticated equipment requires labour to be more technically oriented than previously. Relatively fast-moving mobile equipment calls for training in basic mechanics, since different conditions apply in the light (unloaded) and heavy (loaded) use of this equipment, while misuse of such equipment is expensive in terms of maintenance.

There is a need therefore to plan the extent to which a port requires conventional as opposed to advanced equipment. This involves forecasts and estimates of the envisaged traffic flow and optimisation of appropriately capable labour with the requisite skills.

7.3.3.5 Transport network considerations

Planning for a port's development cannot be carried out in isolation from the transport network since both are involved when endeavouring to achieve effectiveness in the port itself. The transportation facilities which move goods to and from the port should also be taken into consideration. A port's capacity to move goods of different sizes and quantities and the systems used to do this have a profound effect on the facilities a port has to provide for distribution efficiency. For example, freight train arrangements in a national railway system will require different reception equipment and facilities compared with those of heavy/light road transport, which in turn will differ from the methods used to move goods in, say, water canals.

Reliability in the timing of such forms of transport is normally beyond the control of the port – hence the need for flexibility in planning systems. Indeed, as has been indicated earlier, a port operates in a wide area of unpredictability. Forecasting and planning should take into account all the associated transport services, particularly those that operate from outside the port. Estimates of their probable effectiveness will provide the measure of optimum operational systems in the port.

7.3.3.6 Planning for development in a port

Planning, either for a new port or the development of an existing one, should take cognisance of the natural growth of the area or the country which it is to serve, and also recognise the competitiveness of other ports so that the balance of traffic availability is not unduly disturbed. This will ensure that capitalisation of equipment and facilities becomes more meaningful in the transportation system of a country. Market research into trading patterns and the amount of space available should also determine whether a port should be either mainly “specialised” (ie unit transport through containerisation and/or roll-on/roll-off services), or be more diversified and thus cater for different trading patterns requiring both large and small ships.

On the basis of this market research, the features of a port can then be determined in terms of the extent of capital equipment, services and facilities, quay space, warehousing and back-up areas required, while capital expenditure can be more realistically proportioned. Whereas heavy sophisticated cranage (portainers and transporters) and mobile equipment are necessary for accommodating large deep-sea vessels (container and bulk carriers), less expensive but equally versatile equipment will do for smaller short sea-route ships or general carriers.

Furthermore, port planning should take cognisance of the continuation of trade, especially with regard to large-capacity specialised vessels. This points to the need for negotiations with port users for some kind of guarantee that support for services will be maintained for a reasonable period of time to ensure that the port/quay services are adequately utilised. Experts in the industry regard periods ranging from three to 10 years as reasonable. Without such reasonable assurances, a port is unlikely to remain economically viable.

7.4 Transport integration in port planning and development

Port planning recognises the current trend of shipping companies forming large consortia, especially in the container trade. Here the power or strength of the group is such that the demand for services that suit them is a matter to which the port in question should give serious attention. Retaining the trade controlled by these shipping company groupings can only be achieved by means of close cooperation with them.

With regard to bulk cargo, there is the matter of "port-based" processing units – the infrastructure utilised by consignees to process raw materials for market distribution direct from the ship. The question of how far the port area can be profitably used in this way requires careful consideration.

These general statements about ports in an integrated transport system point to the following planning considerations:

(1) Ports designed to handle different classes and types of ships in the port structure should have appropriate layout and facilities to supply specialised berths for particular forms of traffic and to provide for interchangeability of cargo on a controlled basis.

(2) Ports designed or organised to handle specific industries – involving either specialised or conventional cargo – should have appropriate facilities.

(3) Ports designed for the provision of communications and connections with hinterland transportation services should have sufficient space, reception and distribution services.

(4) Ports should be designed and organised in such a way that the provision of adequate and suitable conservancy arrangements – water depth, dredging needs and access routes – are ensured.

(5) There should be adequate fixed and mobile equipment, that is cranes and mechanical handling devices.

(6) Suitable engineering and maintenance services should be provided and controlled.

(7) Transit sheds, warehousing and cargo-stacking provisions must meet the requirements of the types of trade for which the port is suitable.

(8) Marketing services and the costing structure should be coordinated. This will not only attract trade, but also obtain realistic returns on the services provided.

(9) The organisational and administrative structure should be sufficiently flexible in its human and other resources to make operational changes as circumstances dictate.

7.5 Summary of basic pointers in seaport planning

7.5.1 Commercial/economic considerations

Port development planning must be related to the overall economic and commercial considerations of a country and to its environmental and social needs. The modernisation of ports and equipment, the rearrangement of services and the introduction of new works, such as a new quay, are expensive capital projects. Planning to meet either eventuality or desirability should be based on decisions influenced by favourable commercial, financial and economic probabilities. On balance, where modernisation is likely to serve the needs over reasonable time periods, say five to 10 years, this may be the more profitable approach.

7.5.2 Determining priorities and flexibility

Planning should always make provision for the determination of priorities in respect of the envisaged short, medium and long-term needs. Whether for existing systems, new works or improvements, the design/planning approach should be flexible to cater for possible adaptations.

7.5.3 Trading patterns and transportation

The pattern of trading developments in a country's economic structure and the internal transportation facilities available – or their suitability for development – should have a bearing upon the location of a port, primarily assuring that the prevailing water and weather conditions are suitable.

7.5.4 Specialised traffic

The continuing development of specialised traffic, that is heavy/large size bulk carriers, container vessels, roll-on/roll-off ships and multipurpose conventional carriers, poses particular problems in traffic estimation and thus in the planning of all types of ports, especially in developing countries, where existing industries suggest enlargement or the introduction of new industries.

7.5.5 The distribution factor

Distribution of cargoes in the hinterland will depend upon costs, geographical conditions and associated transportation availability. This will need to be looked at in terms of the preference for one (or two) general purpose ports, or a number of smaller, “feeder” ports. This will also apply to a country able to export its economic wealth, bulk or manufactured, via more advantageous routes.

7.5.6 Expertise

The planning of seaports involves knowledge and expertise from a variety of disciplines. While civil engineering, both constructional and hydrographic, plays a prominent part, equally important are the views of maritime, commercial, financial, legal, operational and transportation interests. Where research model facilities are available, the hydrographic considerations can be simulated and trends in water depths, silting and current/tidal flows can be forecast.

7.5.7 National/political considerations

Planning will be conditioned by national/political considerations and local commercial possibilities.

7.5.8 The “network analysis” approach

A form of the “network analysis” approach is desirable in any major planning exercise – and this is certainly true in port planning. All the probabilities that may have an influence on siting, usage and development of a port can be graphically represented over time periods. The overall picture thus generated promotes flexibility of action and those priorities which are critical for optimum efficiency can be emphasised.

7.5.9 The feasibility approach

The use of a "network analysis" approach in a feasibility study provides realism. Such studies can be applied to minor and major planning, the latter involving wide and comprehensive examination of costing analysis, technical needs, commercial/economic factors, trade patterns and local conditions, as would be necessary in the planning of a new seaport. In modernisation schemes, or with the introduction of particular forms of equipment, such a study is likely to be more specific and conditioned by existing arrangements that will include any new features. Whatever arrangements apply, a feasibility study calls for professional expertise. In many cases it may also require consultancy services, backed up by a knowledge of local conditions.

7.5.10 Human resources planning

Port planning involves not only physical requirements, but also staff and labour, both with regard to the number and qualifications of employees. Changing technological and commercial needs call for reasonably periodic human resources planning exercises in established ports, as evidenced in Western industrialised countries because the ports industry has become more capital as opposed to labour intensive.

Mobility of labour, adapting to the usage of technical equipment and maintaining the correct balance between professional, technical and nontechnical personnel are important features of modern seaport efficiency. In-depth human resource planning is necessary in developing ports where, despite the probability of there being a numerical advantage with some forms of personnel, the need to determine training priorities at all levels is paramount.

7.5.11 Berth arrangements

Berth planning in a port involves specific considerations relating to the types and volumes of cargo likely to pass through the port. General or conventional berths will require a variety of services, while berths for specialised cargoes will require more specific services.

In case of the former, berth dimensions and the dimensions of supporting transit sheds, warehousing facilities and backup areas should be related to the "ship" accommodation available. Linear dimensions of 200 to 250 metres are probably adequate for most general cargo carriers, but for a common user berth to accommodate a number of vessels, large and small, the total length should be at least 100 metres for smaller vessels, thus allowing for the berthing, movement and "shifting" of the total number of ships likely to meet the trading patterns.

7.5.12 Statistical support

Berth planning is an exercise that is dependent on estimates which can be assisted by published statistical information for similar ports. It would be misleading to provide examples here because there are numerous differences between countries and ports. The important point here is that berth provision and backup facilities are interrelated and should be planned accordingly.

7.5.13 Summary

Ports cannot be planned in isolation. Such planning needs to be coordinated with national and regional planning, since the economic influence of a port is far reaching. Furthermore, a port as such no longer holds the same established position as a "homeport" for shipping lines. The modern trend is for ships to dictate the type of port to which they prefer to go,

and planning in this regard requires careful consideration of both the seaward aspects of ship attraction and hinterland requirements. General cargo, which used to be the more prominent feature of freight carriage, no longer has the same significance and is becoming a part of a “bulk concept”. Unit loads and port facilities and services need to be designed with this in mind.

7.6 Port investment

7.6.1 Introduction

Investment involves the employment of capital in order to obtain fixed assets such as buildings and equipment that can be used to provide services for a number of years. Investment plays a key role in the economic development of a port and a region and therefore requires careful consideration to prevent misallocation of scarce resources. Whatever the size of the investment, the most important questions when investing are whether or not the project is economically worthwhile, which investment alternative to select, when to undertake the project and the size of the project. The investment should realise specific objectives and satisfy predetermined criteria since the sums of money invested in ports are usually substantial.

7.6.2 The role of the government in port investment

Governments in general have played a decisive role, especially in the context of port investment in the past, which implies a substantial degree of control over the process of port development. This control may be either active or passive, depending on the role of the government in national economic development. Passive control merely implies providing finance, according to government norms, and does not involve the initiation or close scrutiny of port development projects. Active control, on the other hand, involves detailed attention to the timing, structure and administration of port investment programmes. The form and scope of active government control depend on many factors such as the number of ports in a given national system, the role of the ports concerned in the national economy and the level of development in the wider context of international maritime trade.

In many cases the financial resources of port authorities are insufficient to cover large-scale developments, which brings us to the matter of subsidies. Some governments have been prepared to invest substantially in port growth even if this meant a low or negative return on the investment in the short run, while others have adopted a more parsimonious approach and have regarded ports as largely self-contained economic units that should yield a profit.

These contrasting points of view are exemplified in the development of widely differing principles which underlie the pricing policies of ports and investment programmes in Britain and Europe. The European doctrine views a port as part of the social infrastructure and assesses its value in terms of the progress of industry and trade, rather than in the accounts of the facilities. Hence the justification for existing or proposed investment falls outside the ambit of a port (economic approach). The British view is that, notwithstanding the benefits to the hinterland, a port should stand on its own and not incur a loss but, at best, a reasonable profit (financial approach).

Subsidies can take numerous forms. If we break down the cost of a typical port into the three categories of capital, labour and land, the following types of direct and indirect subsidies are found:

7.6.2.1 Direct subsidies

(1) Construction subsidies may include the initial costs of constructing breakwaters, locks, berths, quays, transit sheds and warehouses. A further consideration may be the direct subsidisation of the continuous costs of maintaining the real estate, or the dredging and widening of channels.

(2) Capital equipment subsidies subsidise the initial purchase price of tug boats, lighters, cranes, straddle carriers and computers and/or the continuous costs of maintaining and replacing equipment.

(3) Direct labour subsidies may take one of two forms, namely:

(a) a production per head subsidy – where staffing levels are used to achieve the existing level of productivity

(b) a unit labour cost subsidy – arrived at by dividing the total wage bill by the throughput of the port

(4) Land subsidies may take various forms, for instance:

(a) the purchase price of the land, including any land reclamation

(b) rental of land from the government or government agencies by means of a lease

(c) the cost of maintaining the land, which is usually slight but may assume significant proportions when environmental standards have to be upgraded

7.6.2.2 Indirect subsidies

These subsidies normally relate to the state's taxation policies which could influence various aspects of port profitability. Indirect subsidies may relate to any or all of the following:

(1) *Capital equipment*. Taxation policies may allow ports generous tax deductions on investment in capital equipment. Ports may also be permitted to use accelerated depreciation techniques, or to depreciate their equipment at replacement value.

(2) *Real estate*. Generous grants of land, and light taxation on real estate, may provide an incalculable boost to a port's ability to expand and develop.

(3) *Wages*. The wage bill, probably the most crucial factor in the year-to-year operating costs once a port is fully operative, may differ greatly from country to country. Wage pressures in countries with highly developed social services are probably kept lower than might otherwise be expected, by national policies on wage controls, income taxes, social security payments and benefits, the level of unemployment, and so on.

7.6.3 Port investment objectives

The objectives of investment differ widely from one port to the next, and are dependent on factors such as port ownership, port control and the role of government in port investment. Port objectives may be stated in extremely wide or more specific terms. Objectives stated in wide terms normally influence an area or region far more than the port itself and, according to Fränkel (1987:83), may be expressed in the following ways:

(1) net national, regional, or local benefits such as income generated by a particular port investment.

(2) transportation cost savings, and the resulting impact on the transportation costs of trade and industry.

(3) indirect economic benefits, including secondary and multiplier effects.

(4) the ability to generate employment opportunities and reduce unemployment.

(5) the impact on local, regional or national economic growth.

Other port objectives may be equity or environmental quality objectives. Equity objectives relate to the distribution of income and wealth, whilst environmental objectives relate to the environment. Environmental objectives can be expressed in terms of a percentage change in

(1) air, water and noise pollution

(2) safety (employees, community users)

(3) community acceptance

Objectives stated in wide terms normally have a socioeconomic or political incentive and may be regarded as long term or strategic.

Objectives stated in specific terms are more useful for medium to short-term investment decision-making since they permit definition of tactical and operational decisions aimed at achieving the stated objectives. Investment planning approaches using narrower objectives assume quantifiable monetary benefits or losses in economic and/or financial terms. Some of these objectives may be:

(1) maximisation of port profit

(2) maximisation/minimisation of port employment

(3) maximisation of port facilities and resource utilisation

(4) minimisation of port costs per unit of output or time

7.6.4 Port investment criteria

Port investment decisions are not based only on objectives but must also satisfy various criteria. The criteria are usually represented by a measure that can be computed or evaluated, for example an analysis of economic and/or financial flows over the expected or proposed life of the project. Two of these criteria are as follows:

(1) In traditional benefit/cost analysis, various techniques are used to measure the benefit of the investment. The aim of this criterion is to determine the economic benefit of the project.

(2) The consumer surplus approach entails the analysis of public investment decisions. The difference between the users' willingness to pay and what has actually been paid is known as the consumer surplus. The aim of this criterion is to determine the increase in consumer surplus brought about by the effective allocation of resources.

The investment criteria that are used will correspond with the objectives to be realised. The investment criteria used in port planning do not differ from those that are generally used, but they are usually subject to more rigid constraints such as:

(1) one or more objectives which are often dynamic (They change over time.)

(2) a series of constraints and rules, either explicit or implicit

7.6.5 Evaluation methods and techniques

The evaluation methods and techniques used in port investment analysis usually fall into the following two categories:

(1) Standard benefit/cost analysis includes the following methods:

(a) the estimated social rate of discount and rate of return

(b) net present value (NPV)

(2) The expected value or utility of the project is maximised. Several methods can be used to calculate the "profit" that can be obtained from investing in different alternatives. These methods are:

(a) the discounted payback period

(b) the capital recovery factor (CRF)

(c) present worth (PW)

(d) the minimum average annual cost (AAC)

(e) the internal rate of return (IRR)

The choice of method when making investment decisions depends upon the formulated objective(s). Each method has particular advantages in a certain situation.

Where a socioeconomic or political objective is being pursued, benefit/cost analysis methods are preferable because they evaluate the impact of the investment on a region as a whole. Since most port investment projects are the responsibility of public or semipublic authorities it is preferable to use a broad definition of benefits and costs.

When an operational or tactical objective is being pursued, it is better to use the method that evaluates and compares the profitability of certain pieces of equipment or projects in order to achieve the specific objective(s). This financial approach is aimed at generating the highest profit from the investment and any of the methods mentioned in point (2) above may be used for this purpose.

7.7 Conclusion

Planning and investment in ports has a profound influence on the economy of a country as well as the region from/to which the port receives/sends cargo. Thus all the roleplayers (including the government) who will be affected by changes in a port should provide input. The whole process should therefore be coordinated by means of the National Ports Authority. Although governments are the main providers of funds for investment, they should encourage ports to become financially self-sufficient by providing most of their investment funds themselves and using government funds for major projects only. To determine whether investment is viable, relevant investment criteria and methods should be used to ensure that only those investments with an acceptable rate of return, whether socially or financially oriented, are implemented.

7.8 Self-evaluation questions

(1) Explain the interaction that takes place between a port and its environment.

(2) How should ports be planned? Explain by means of practical examples.

(3) What factors should be taken into account in port planning? Name and discuss any five factors.

(4) Why is it necessary to use an integrated transport network in port planning?

(5) Briefly name and discuss the various types of subsidies that government may use to assist ports.

(6) What investment criteria can ports use, and why do they use them? Explain in detail.

STUDY UNIT 8

Planning and investing in airports

UNIT OUTCOMES

After working through this study unit you should be able to:

- discuss the importance of airport system planning
- distinguish between regional and national airport planning
- explain what is meant by the term “integrated airport system planning”
- discuss the influence of competition between airports

KEY CONCEPTS

- Airport master plan
- Regional airport planning
- National airport planning
- Integrated airport planning
- Competition between airports
- Investment in airports

8.1 Introduction

Airports are public entities which not only interact with many other public and private organisations but also influence the community's everyday life. Any airport development plans affect aspects of community life, for example there is the land that has to be set aside and the noise or automobile traffic generated by airports. In addition, an airport cannot be planned in isolation because it is part of an airport network, which in turn is part of the national transportation system.

Planning for airport development thus requires a great deal more than simply allocating the capital for the required improvements. The need for airport development should also be weighed against other social needs and plans. Because of the high costs and long lead time involved in building or improving airports, planning is the key to determining the facilities needed and creating the programmes for providing them in a timely manner and at the same time using resources wisely.

In this study unit we will look at *airport system planning*, and not simply *airport planning*, because airport plans have to be developed as part of a system which includes local, provincial and national transportation planning. Determining need and programming development at individual airports has become formalised in a process called *airport master planning*. While master planning in the full sense of the word is practised primarily by large airports, even the smallest airports make use of some elements of the process to prepare for future change.

At a level above airport master planning is *regional system planning*, which involves the development of all the airports in a metropolitan area. It often entails difficult political decisions on development priorities of competing airports. In some cases, this responsibility is handled by a regional or metropolitan planning agency, but many provincial governments have also taken on the task of developing a coordinated system plan for airports serving not only major metropolitan regions but also outlying small communities and rural areas in the region. In some cases, these state agencies prepare the plans themselves; in others, they provide technical assistance for local planning bodies.

The role of the national government in airport planning includes a broad range of activities. The most comprehensive activity is that of strategic planning. It is the task of the national Department of Transport to approve, on a project-by-project basis, specific development projects for which airport sponsors seek funds.

8.2 The planning process

8.2.1 Planning at local level

8.2.1.1 General

At local level, the centrepiece of airport planning is the *airport master plan*, a document that charts the proposed evolution of a specific airport to meet future needs. The magnitude and sophistication of the master planning effort depends on the size of the airport. At major airports, planning may be in the hands of a large department capable of producing its own forecasts and supporting technical studies. At such airports, master planning is a formal and complex process that has evolved to coordinate large construction projects (or perhaps several such projects simultaneously) that may be carried out over a period of five years or more. At smaller airports, master planning may be the responsibility of a few staff members with other responsibilities who depend on outside consultants for expertise and support. At very small airports, where capital improvements are minimal or are made infrequently, the master plan may be a simple document, perhaps prepared locally, but usually with the help of consultants.

An airport master plan presents the planner's (be it a committee, consultant or single person) conception of the ultimate development of a specific airport. It effectively presents the research and logic from which the plan was evolved and artfully displays the plan in a graphic and written report. Master plans are used to modernise and expand existing airports and construct new ones, regardless of their size or functional role.

8.2.1.2 Objectives of the airport master plan

The overall *objective of the airport master* plan is to provide guidelines for future development which will satisfy demand and be compatible with the environment, community development, other modes of transportation and other airports. Specific objectives within this broad framework are as follows:

(1) to provide an effective graphic representation of the ultimate development of the airport and of anticipated land uses adjacent to the airport

(2) to establish a schedule of priorities and phasing for the various improvements proposed in the plan

(3) to present the pertinent background information and data which were essential to the development of the master plan

(4) to describe the various concepts and alternatives which were considered in drawing up the proposed plan

(5) to provide a concise and descriptive report so that the impact and logic of its recommendations can be clearly understood by the community the airport serves and those authorities and public agencies charged with the approval, promotion and funding of the improvements proposed in the airport master plan

8.2.1.3 Local coordination

Airports begin with local initiative. The local community normally decide whether they need a new airport or should expand an existing one. The development and operation of an airport impact on the entire community – hence the emphasis in planning successful airports should be on integrating the airport into a comprehensive community transport plan. This entails two important goals, namely achieving compatibility with local community goals and residential and commercial land uses, on the one hand, and achieving integrated transportation on the other.

Achieving compatibility requires a team effort. Effective coordination of airport planning at local level requires involvement of individuals who are interested and knowledgeable about the community and the importance of such development. An airport master plan draws widespread interest from private citizens, community organisations, airport users, area-wide planning agencies, conservation groups, ground transit officials, and aviation and airport concessionaire interests. If these groups are not consulted during the development of the plan, it is highly unlikely that the public will accept it. It is therefore essential that the master plan team coordinate their efforts with and seek the advice of these interest groups during the critical stages of the development of the plan. This coordination will help pave the way for acceptance and, more importantly, will permit vital input from organised interests which will lead to the evolution of a well-integrated plan.

It is not only the community which needs to be considered. In the case of large-scale planning, effective coordination between members of the planning teams is also essential to the development of a successful master plan. A balanced effort is not easy to achieve because of the many disciplines involved in the planning. For large projects, input may be required from economists, financiers, scientists, architects, civil, mechanical, electrical and traffic engineers, pilots, air traffic controllers, airline and concessionaire advisers, and airport managers. And to put the airport in its proper perspective, the roles of the environmentalist, ecologist and urban planner also need to be considered.

This is why the coordination of the master plan effort is so important. It should keep the enthusiasm of the advisers in check in order to balance the studies and costs of various master plan elements. If it is successful, a viable master plan will be developed that will lead to the construction of a functional airport that blends pleasantly into the environment.

The role of the planned changes in terms of the total local transport system also needs to be considered. An airport does not exist in isolation. It is an element of the total transportation availability of an area. Its integration into an overall system is even more important in these days of intermodal freightage compared with the situation in the past. The long-range planning of any form of transportation in an area can most effectively be done in concert with the planning for all modes. Here it is specifically land transportation that comes to the fore.

Land transportation reflects on an airport in at least two ways. First, the less surface transportation there is in an area, especially when great distances are involved, the greater the

need will be for air facilities. (It costs considerably less to build a runway than to construct one kilometre of highway.) Secondly, businesses looking to locate in a new community generally require that the airport be less than 30 minutes away from the plant location. A review of access roads, peak-hour traffic patterns, bus and rail transportation, and other nearby airports should therefore be an important part of the airport master plan.

8.2.2 Planning at regional level

Regional airport planning takes as its basic unit of analysis the airport hub, roughly coincident with the boundaries of a metropolitan area. The planner is concerned with air transportation for the region as a whole and considers traffic at all the airports in the region, both large and small. The practice of regional airport planning is relatively new and has been instituted to deal with questions of resource allocation and use that often arise when the airports in a region have been planned and developed individually and without coordination between the affected jurisdictions. Regional airport planning seeks to overcome the rivalries and the jurisdictional overlapping between the various local agencies involved in airport development and operation. The goal is to produce an airport system that is optimal for region-wide benefits and costs.

Thus regional airport planning addresses one critical issue usually not dealt with in an airport master plan, namely the allocation of traffic among the airports in a region since this can be a sensitive issue. Questions of traffic distribution normally involve political as well as technical and economic issues, which can greatly affect the future growth of the airports involved. One airport may be quite busy while another is underutilised. If traffic were to continue growing at the busy airport, new facilities would have to be constructed to accommodate that growth. However, if some of the new traffic were diverted to an underutilised airport, the need for new construction could be reduced, which could improve service to the region as a whole.

Although an airport planning agency may decide that such a diversion is in the interest of a metropolitan region and might prepare forecasts and plans showing how this could be accomplished, it may not necessarily have power to implement these plans. Where airports are competitors, it is probably not reasonable to expect that the stronger airport will voluntarily divert traffic and revenues to the other. The planning agency would probably have to influence the planning and development process at individual airports so that decisions that reflect the regional agency's assessment of regional needs are made. One way to influence these types of planning decisions is through the distribution of development grants. Implementation, however, depends not only on control of airport development expenditures but also on the ability to influence the activities of private parties – in other words, the air carriers and passengers.

Much of the regional agency's success may depend as much on negotiation and persuasion as on legal or budgetary authority. Often compromises can be reached on a voluntary basis. For example, in San Francisco, the California Regional Airport Planning Commission has been working with the three area airports to help each develop a “noise budget” to comply with California's strict environmental laws. Because noise is directly related to the level of aviation activity, the noise budget plan, when completed, will affect future traffic allocation between these airports because its implementation will probably require some diversion of new traffic growth from the busy San Francisco International airport to the other area airports. Even where airports in a region are operated by the same authority, allocation of traffic between airports may still be difficult. For example, the Airport Authority of New York and New Jersey implemented a planning decision to increase activity at Newark by instituting differential pricing, improved ground access and other measures.

If regional airport planning authorities have planning responsibility for other transportation modes, they may also plan for the airport as part of the regional transportation system. When multimodal planning responsibility resides in one organisation, there is a greater likelihood that the planning agency will consider airport needs in relation to other forms

of transportation in the region. Also, the regional agency may try to improve coordination between the various modes, to ensure, for example, that airport developments do not impose an undue burden on surrounding highway facilities or that opportunities for mass transit can be utilised. However, two conditions have to be satisfied here, namely region-wide authority and multimodal jurisdiction. At this stage these conditions cannot be met in South Africa.

8.2.3 Planning at provincial level

Airport planning at provincial level involves issues that are somewhat different from those of local or regional agencies. Provincial governments are typically concerned with developing an airport system that will provide adequate service to all parts of the province, both rural and metropolitan. This is because the development of airports is often seen as an essential tool for economic development or making rural areas less isolated. The goal would normally be to develop at least one well-equipped airport in each metropolitan area. The issue here is not the allocation of traffic among airports serving the same community, but instead deciding how to allocate airport development funds among candidate communities to maintain a balance between various parts of the province.

Provincial plans typically encompass a planning period of 20 to 30 years; the year 2035 is currently a common planning horizon. Planning periods are normally divided into short, medium, and long-term segments (usually 5, 10 and 20 to 30 years respectively). In each case, estimates of future needs are developed by comparing existing facilities with projections of future traffic. The major feature of the plans is a detailed listing of the actions planned by the class of airport and type of improvement. The types of improvements most commonly cited are land acquisition (new sites or expansion of existing airports), pavement repair or improvement (runways, taxiways, aprons, roads, parking), installation of lighting and landing or navigation aids, and building construction (terminals, hangars, administrative facilities). Airport system improvements can be divided into three *levels of need*:

> **Level 1:** Maintain the airport system in its current condition – maintaining the system includes such projects as repaving airfields and replacing lighting systems.
>
> **Level 2:** Bring the system up to current design standards – bringing the system up to standard involves such projects as installing new lighting and widening runways.
>
> **Level 3:** Expand the system – expanding the system includes constructing new airports, building terminals or lengthening runways to accommodate larger aircraft.

This classification by three levels of need is a method of assigning priorities to different types of projects. The system can be somewhat misleading because it is not as hierarchical as it might seem, and the placement of a type of improvement at a particular programme level does not necessarily reflect the priority given a particular project. High-priority projects, that is those which the provincial government feels should be conducted as soon as possible, may not necessarily correspond with level 1 needs. An expansion project (level 3) at an extremely congested and important airport might be more urgent than bringing a little-used airport up to standard (level 2). Thus, if the available funds are limited it might only be possible to conduct level 1 and 2 projects and leave vital level 3 projects unfunded. This would be ridiculous, which is why the levels are not in hierarchical order but instead reflect the size of the investment needed.

While there are superficial similarities, different provinces' planning will vary greatly in scope, detail, expertise and planning philosophy. One province's system plan may basically be a wish list, prepared primarily because planning funds were available and the Department of Transport (DOT) required it. Another province, however, may regard this type of planning as a valuable working document that is kept up to date and serves as a guide in programming and distributing state funds.

Virtually all provinces' plans estimate costs of recommended improvements and identify funding sources. In many cases, airport planning is part of a general transportation planning

process, but methods of interaction and feedback between the modal agencies vary considerably. Some provincial agencies are involved in master planning activities for local airports, especially rural or small community airports that do not have the staff to conduct master planning on their own. Provincial agencies may provide technical assistance or actually develop local master plans. Some provinces also participate in airport planning for major metropolitan areas, although most impose this responsibility on the local airport authority or a regional body. In recent years, there has been an increase in provincial participation in planning at larger airports, a trend that could be bolstered by current financial policy.

8.2.4 A national integrated airport systems plan

8.2.4.1 The current status in South Africa

Airport planning at national level is the responsibility of the national Department of Transport (NDOT). The NDOT is primarily responsible for providing guidelines on the development of the vast network of publicly owned airports and establishing a frame of reference for the investment of national funds. These interests are set out in *Moving South Africa: the action agenda (MSA)*, a document that outlines the future role of the different modes of transport for the next 20 years. The MSA is a plan in the fullest sense of the word. It establishes priorities, proposes a level of funding and commits the national government to a specific course of action. It is, however, somewhat vague about the type and cost of airport developments that might take place during the planning period at those airports eligible for assistance. Unfortunately, at this stage, the MSA does not incorporate a national plan of integrated airport systems. For an idea of what such a plan envisages we shall now look at the situation in the USA.

8.2.4.2 National Plan of Integrated Airport Systems (USA)

The Airport and Airway Improvement Act of 1982 in the USA reflects a strong congressional commitment to airport planning. At regional and state levels, the Act dictates that one percent of federal airport development funds should be for planning. As such, the Act provides an opportunity for state governments and regional agencies to institute or expand their planning efforts. The Act called for refinement of the national airport planning process by instructing the Department of Transportation to develop a *National Plan of Integrated Airport Systems (NPIAS)* by September 1984. The description of this plan in the legislation made it clear that the intention was to expand and improve planning at national level. Specifically, the Act calls for "integrated airport system planning" which it defines as follows:

> *.. the initial as well as continuing development for planning purposes of information and guidance to determine the extent, type, nature, location, and timing of airport development needed in a specific area to establish a viable, balanced, and integrated system of public-use airports.*

Planning includes identifying system needs, developing estimates of systemwide development costs, and conducting studies, surveys and other planning actions, including those related to airport access, that may be necessary to determine the short-term, intermediate and long-term demands that an airport must meet.

The policy declaration points out several ways in which the planning effort should be integrated. It states the following:

> *.. it is in the national interest to develop in metropolitan areas an integrated system of airports designed to provide expeditious access and maximum safety ... [and it is in the national interest to] encourage and promote the development of transportation systems embracing various modes of transportation in a manner that will serve the [provinces] and local communities efficiently and effectively.*

From this it is evident that the legislation requires a plan that is integrated in two ways:

(1) geographically, in the sense that all airports in a region should be considered together

(2) intermodally, in the sense that planning for an airport should be part of the planning for the regional transportation system as a whole

The requirements of the Act brought the airport planning process closer to metropolitan and regional transportation planning than ever before.

8.3 The need for integrated airport systems planning

Airport planning as it is practised today is performed largely by government agencies. It is mainly a political process, where value judgments and institutional relationships play as much a part as technical expertise. On the whole, airport planners have been reasonably successful in anticipating future needs and devising effective solutions. Still, mistakes have been made, sometimes because of poor judgment or lack of foresight, and at other times because of certain characteristics of the planning process itself. In effect, the process and methods employed predispose planners towards solutions that may be "correct" for a single airport but perhaps not for the community, region or airport system as a whole. As a result, airport plans have taken on a rigidity that is inappropriate in the light of changing conditions or a narrowness of focus that does not make best use of resources.

Airport planning at local, regional, provincial and national levels in South Africa is not well coordinated and integrated. To some extent, this arises naturally from different areas of concern and expertise. At the extremes, local planners are attempting to plan for the development of one airport, while NDOT may be trying to codify the needs of several airports that may have asked for aid. Local planners are more concerned with details and local conditions that will never be of interest to a national planning body.

The lack of common goals and a mutually consistent approach is also evident in provincial and local planning. There is also a lack of coordination between airport planning and other types of transportation and economic planning. This is particularly evident in the case of land use, where airport plans are often in conflict with other local and regional developments. Even though the airport authority may prepare a feasible plan, lack of information about other public or private development proposed in the community (or the failure of municipal authorities to impose and maintain zoning ordinances) allows conflicts to develop over use of the airport and surrounding land. This problem can be especially severe where there are several municipalities or local jurisdictions surrounding the airport property.

An additional problem is the lack of integration between airport planning and planning for other modes of transportation. An airport is an intermodal transportation centre, where goods and people transfer between the ground and air modes. It forms an important link in the total transportation system of a region. The land transportation system providing access to the airport can be a significant contributor to congestion, delays and the cost of airport operation. Yet airport operators have little authority or influence over decisions on transportation beyond the airport property line.

8.4 Airports and competitors

8.4.1 General

There is a widespread – and costly – belief that traffic will flow naturally wherever capacity is provided.

That this belief cannot be true should be obvious because the construction of airport capacity, runways, terminals and the like does not in itself attract traffic. Airports exist solely to enable people and goods to reach desired destinations by air. Alternatives to these

transport services are, however, also offered by other modes of transport. The trade-off that the user of the services makes between the various alternatives he or she can use thus determines the demand for airport services. Hence the demand for airport services exists only in so far as air transport provides a better combination of services than its competitors. We must therefore try to understand how the planning and design of airport systems influence this competition and thus the demand for air transport.

8.4.2 Factors that affect airport competition

8.4.2.1 Speed

Speed is the most appealing aspect of air transport. More precisely, it is not the absolute speed we might reach at some moment, but the overall rapidity of movement from place to place that is important. The capability of travelling hundreds and even thousands of kilometres in a matter of hours opens up a wide range of opportunities which would not otherwise be available.

The advantages of speed are obtained at a price. The fare for air transport between two cities is generally higher than that of the nearest substitute. This differential in costs is naturally offset by the value of higher speed. Frequently, the trade-off between greater speed and extra cost weighs against the former as can be seen in the fact that practically no-one, for instance, would pay the full cost of using the Concorde supersonic aircraft – hence the decision of the British and French governments to subsidise this service. The assumption that passengers are automatically attracted to the speed of the aircraft is therefore not true.

8.4.2.2 Total travel time

It is not only the airfare that should be considered – the total time (and cost) of travelling from point of departure to point of arrival should be looked at. Air travel has significant advantages in this regard. Compared with going by rail or automobile, travellers can save on hotel bills and the cost of meals along the way. For business trips exceeding 600 kilometres, these savings can easily make up for the higher cost of the airfare.

Conversely, the remoteness of an airport from and infrequent service to a particular destination may result in substantial delays in air travel and lead to lower use. Although these costs and delays are peripheral to the air journey, they can have a crucial influence on the level of air traffic.

8.4.2.3 Configuration of the airport system

The configuration of the air transport system largely determines these costs and the delays in access to air travel. The location of the airport determines how long it will take and how much it will cost to reach air transport services. Also, the number and size of airports in a metropolitan area and the services rendered by each influence the frequency of service available to any destination and thus the delays passengers will encounter in waiting for a departure.

8.4.3 The competitive position of an airport

8.4.3.1 General

The influence of the airport system on the level and distribution of traffic of a specific airport is difficult to quantify because its effect is not direct, but indirect. The design of the air transport system enhances – or detracts from – the competitive position of an airport with respect to other modes of transport, other local airports and even airports in other cities. To clarify these effects, the ensuing discussion focuses on the relationship between the

airport and one form of competition at a time, even though they usually appear together. We will first consider a single airport sharing the market with ground transport. Next we will examine the division of traffic among several airports serving the same metropolitan area. Finally, we will explore the competition between cities for long distance traffic which might use their airports as an interchange for other destinations. Each of these competitions may alter the traffic at an airport by up to a third.

8.4.3.2 *The single airport*

The single airport serving a city competes principally with other modes of transport. The volume of traffic depends on its comparative advantage in providing air transport services. Here, its edge in offering access to rapid transport will be counterbalanced by the potentially higher cost of this service, by the remoteness of the airport, and by the possible lack of frequent departures. Remember, the airport itself has no control over departures to destinations – this is the function of the air carriers that make use of the airport.

(a) Transcontinental trips

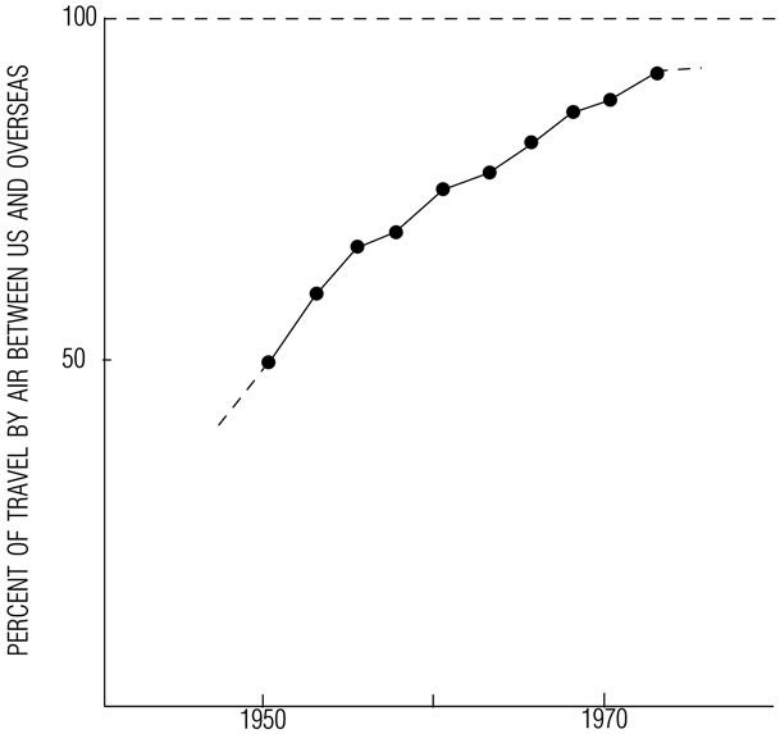
Airports now capture virtually the whole market for long distance trips. The speed of air transport applied over long ranges saves passengers days of travel time, producing substantial savings in meals and lodging while in transit. The saving in days is also valuable

- to a person on business because more time is available for productive work
- to holiday-makers because of the extra precious days of vacation

In overseas travel, for example, air transport has virtually eliminated the demand for passenger ships, as shown in figure 8.1.

Figure 8.1

Increasing dominance of air transport for overseas travel



Source: De Neufville (1976:62)

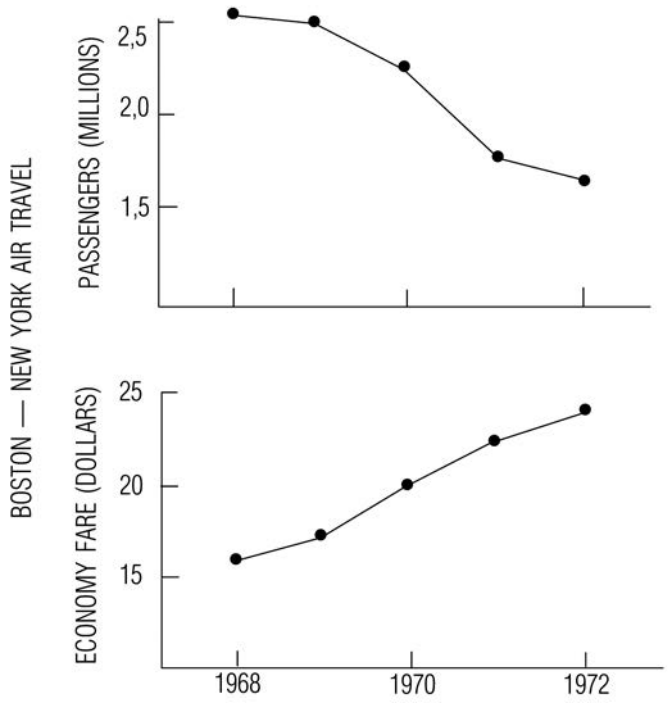
Much the same can be said of transcontinental travel, although the evidence is less obvious since many people do make the trip by land – some because they find it cheaper, and others because they wish to visit friends or relatives or do some sightseeing on the way. In Europe and Japan, the continuing governmental protection and subsidy of the railroads further masks the appeal of air transport. Increasingly more cargo also goes long distances by air. This traffic is becoming increasingly important for some airports and could at some time conceivably represent a major part of their business. In view of the inefficiency of having airplanes carry dense, bulky materials, however, air cargo will certainly remain a minuscule fraction of the total tonnage sent by rail, truck and ship.

(b) Total cost of the trip

The airport's share of the passengers going shorter distances is sensitive to the total cost of the air trip. The statistics on travel between Boston and New York, a distance of some 320 kilometres, illustrate this. In the early 1970s, a combination of events including a massive review by the US Civil Aeronautics Board led to a rise in the basic fare and the removal of family and youth discounts. The resulting increase in effective fares caused air traffic between Boston and New York to drop by about one-third. As shown in figure 8.2, this effect has persisted, relieved only partially when the scarcity of fuel for automobiles in 1973 encouraged travellers to switch to various modes of public transport.

Figure 8.2

Sensitivity of air travel over short distances to fare increases



Source: De Neufville (1976:64)

Any of the different methods airport planners might use to raise the cost of access to air travel would lead to similar results, in proportion to the amount of increased cost. This would include in particular, the construction of expensive facilities, whose costs would be passed on to travellers through higher parking fees or eventual fare increases triggered by higher charges for aircraft operations. A typical example would be the case of the passenger terminals at Tampa Airport in Florida. These magnificent structures cost about $6

per passenger in 1973, almost twice the amount typically paid by the airlines elsewhere in the USA. This sum is also equivalent to approximately one-fifth of the fare to Miami, some 320 kilometres away. Since the airlines could not absorb the extra charges indefinitely, they were ultimately passed on to passengers, which led to noticeably lower traffic. Any policy to raise taxes or make high profits by taking advantage of the airport's monopoly on landing facilities, as is widely done, can likewise be expected to reduce traffic.

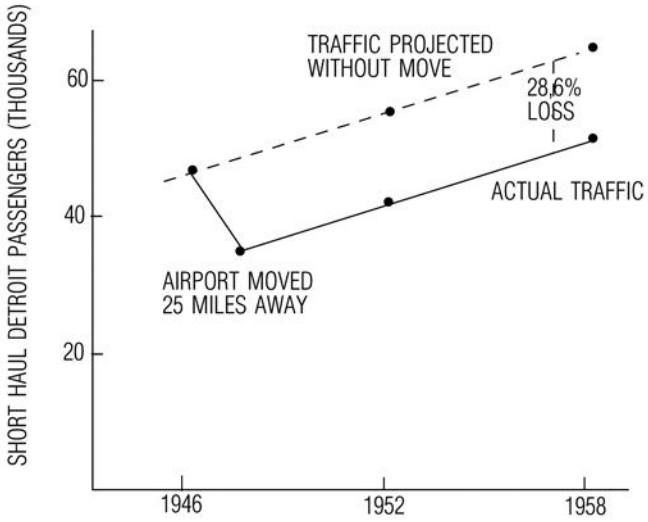
(c) Airport location

Planning decisions about the location of an airport can significantly affect the difficulty of reaching air services and thus the level of air traffic. This is because there are great pressures to choose a site far from the city. Modern jets require runways from one to three kilometres in length, which force airport developers to seek out large uninhabited areas. Noise is also a problem. Urban populations typically want to keep the noise and pollution of an airport as far away as possible. Hence the most acceptable sites for new airports are, from several points of view, far from the centre of the city. The new Montreal/Mirabel airport in Canada is some 25 miles from downtown, or some 20 miles further than the old Montreal/Dorval airport and the Maplin site for the proposed third London airport was some 35 miles further from London than either of the existing facilities at Heathrow or Gatwick. In South Africa, the airport at Cape Town is almost 30 kilometres from the city centre and the new King Shaka Airport is situated 35 kilometres outside Durban.

Any new airport is almost invariably much further away from town than the old one. This implies that air services through the new airport will be relatively less attractive compared with alternative means of transport, especially for trips over short distances. By moving airport operations to a new site, short-haul traffic may be reduced by as much as one-third. The case of Detroit illustrates this. In 1947 the city forced all commercial flights to shift their operations to a different airport 25 miles further away from the city centre. This led to an immediate drop in the demand for air transport to cities within 300 miles. Because the overall rate of growth in this traffic at Detroit for 10 years after the move equalled the rate of growth at all comparable airports in the vicinity, one can presume that this substantial decrease in traffic was also persistent, as indicated in figure 8.3.

Figure 8.3

Sensitivity of air travel over short distances to change in accessibility



Source: De Neufville (1976:65)

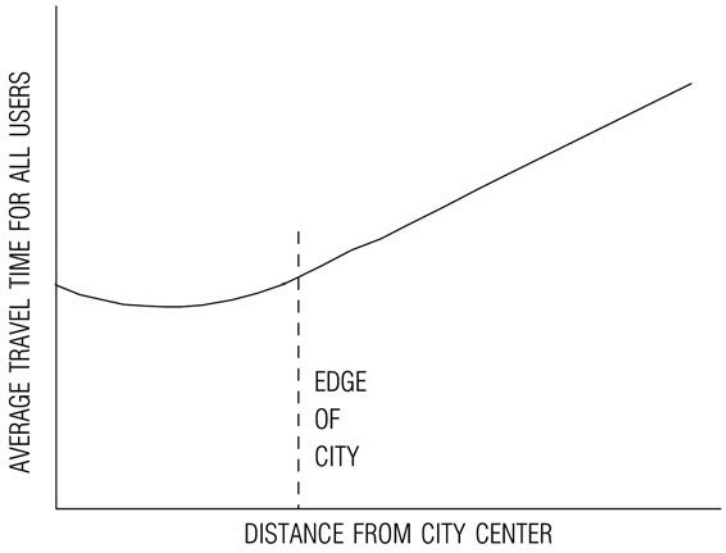
We should not, of course, assume that what happened elsewhere at another time and in a different environment applies directly to a new situation. However, similar decreases in traffic are associated with the opening of other remote airports, thus emphasising the effect of airport location on the level of traffic.

(d) Average travel time to airport

In thinking about the effect of new airport locations, one should remember that the distance between the airport and the city centre is only an indicator of the remoteness of air transport services from potential users. To the extent that many travellers live and work in the suburbs, suburban sites served by good highways are generally just as accessible as locations closer to the city centre. Figure 8.4 illustrates this for a hypothetical city.

Figure 8.4

Average travel time to airport only increases markedly for locations far from a typical metropolitan area



Source: De Neufville (1976:66)

People wishing to fly short distances have a strong preference for airports that are not far outside the city. The situation at Dallas illustrates this. When all the major airlines moved to the new Dallas/Fort Worth Airport in 1974, some 21 miles from the centre of town, the majority of their Texan customers for short flights deserted them, preferring service from Dallas/Love Field right in the suburbs. Southwest Airlines, the only carrier still serving Dallas/Love, thus registered a 40 percent increase in passengers that year. Siting a new airport at a distant location may shift traffic to alternative airports as well as different modes of transport. This brings us to the question of satellite airports.

8.4.3.3 Satellite airports

When an airport becomes congested, the planners' natural reaction is to try to expand its capacity. When this is impossible, the common sense solution is either to build a major new airport or, if this is out of the question, to develop some secondary facilities that might handle some of the traffic. Either way, this leads to the situation where one or more satellite airports are associated with the major air terminal for a metropolitan area.

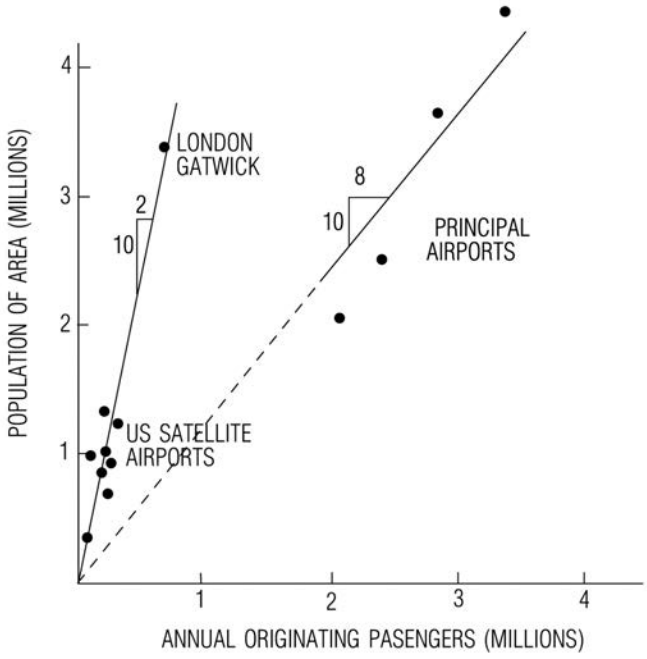
The concept of creating additional airports to accommodate excess traffic is a specific part of the airport system development. In any metropolitan area where it is a question of creating some relief for congested passenger airports, general aviation and pleasure flights usually operate out of special fields set aside for their use. In the Johannesburg/ Pretoria area passenger flights are mainly in and out the OR Tambo International Airport while general air traffic uses Rand, Lanseria and Wonderboom airports. Lanseria also caters for passenger flights, but Wonderboom Airport has been struggling to introduce scheduled passenger flights on a large scale. The question is how traffic will distribute itself between satellites and their principal airports. To answer this, the focus should be on the behaviour of airlines and their passengers.

(a) Choice of airport

The general idea is that each airport serves a particular territory and that the traffic at any airport therefore depends upon its sphere of influence in its "catchment area". The expression "catchment area" derives from a mental image of how rainwater flows downwards from a catchment area to a dam according to the physical laws of gravity since it has no choice about the direction in which it will go. As the UK Civil Aviation Authority once put it: "The traffic at an airport depends to a large degree on the total number of travellers using it, and hence on the extent of its catchment area." People, however, differ from water in that they can and do make a choice about which airport to use. Detailed studies show that people often deliberately avoid the airport that is closest to them in favour of a larger, busier facility. Around Cleveland (Ohio), for example, a large survey clearly demonstrated that over half of the air travellers from Akron (a metropolitan area of over 400 000 inhabitants) drove some 25 miles beyond their own airport to obtain service at Cleveland/Hopkins Airport. Examination of many "catchment areas" indicates that this is a general rule. Figure 8.5 shows this using both American and British data.

Figure 8.5

Principal airports attract a far greater share of the market for air travel than satelite airports



Source: De Neufville (1976:66)

Satellite airports typically attract only about one-quarter of the usual number of passengers from their "catchment area"; the remainder presumably go to the principal airport.

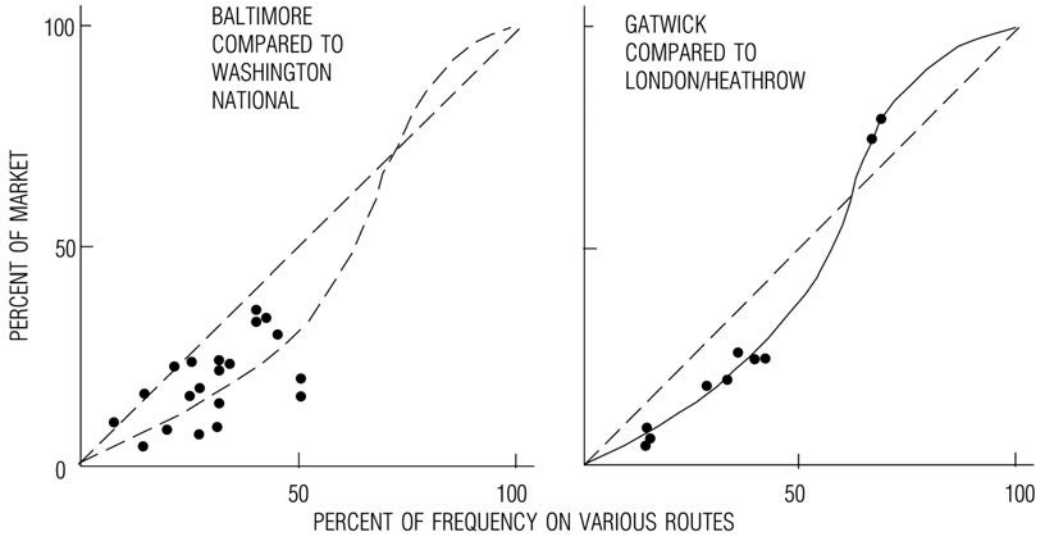
(b) Frequency of service

Frequency of service is often a crucial factor for a person contemplating which airport to use. The airport with more flights to a particular place will almost inevitably offer more convenient departures. Any residents of an area parking their cars at the airport would also be concerned about frequency of service on the return: they need the flexibility provided by backup flights in case they require extra or less time away from home. Conversely, a person travelling to a city with several airports often prefers to use the one with the greater service because it offers more possibilities for transferring to connecting flights.

Many passengers may, of course, attach little or no importance to frequency. For example, holiday travellers leaving on a charter flight may only be concerned about a single specific departure. This behaviour does not, however, contradict the general rule here, which is that frequency of service is a major factor in determining the attractiveness and use of an airport. The relationship between frequency of service and its attractiveness is generally represented by S-shaped curves of the type appearing in figure 8.6.

Figure 8.6

Low frequency service both on routes and the whole airport, causes satellite airports to attract lower market shares



Source: De Neufville (1976:66)

The figure shows how the relative frequency of service between two cities offered by a satellite airport, that is, its frequency share, affects the share of the total market it manages to attract. This phenomenon is also well documented for the competition between airlines on routes linking pairs of airports. The data support the widely held view in the air transport industry that competitors who are able to dominate a market reap substantial rewards and those who are unable to do so are at a constant disadvantage.

Figure 8.6 shows specifically that when a satellite airport offers about 30 percent of the flights from a metropolitan area to another city, such as London/Gatwick does to Edinburgh, for example, it obtains only 20 percent or less of the market. This implies that the airlines serving the satellite airports will either have to carry fewer passengers per plane or use smaller, less efficient aircraft. Either way, this places them at a substantial economic

disadvantage. Furthermore, airlines will find it difficult to overcome this handicap. Even if they increase their service on a particular route from a satellite airport, they will not be able to do anything about the fact that the major airport is inherently more attractive just because it offers more service overall, and thus more opportunities for connecting flights.

The economic handicap of operating from satellite airports has an obvious message for airlines in that they are much better off concentrating their service at the major metropolitan airports. This is exactly what they do, thus leaving the satellites with relatively little traffic. As a rule, satellite airports account for only five to 10 percent of the total airline traffic in a metropolitan area.

In the absence of any regulations forcing airlines to spread their service, competing airports in a metropolitan area only have equivalent levels of traffic when they cater to distinct markets. Thus New York/Kennedy and New York/La Guardia serve a comparable number of passengers, the one on shorter distance, domestic flights, the other on long distance and international flights. New York/Newark, on the other hand, competes with these airports and manages to attract only a fraction of the traffic they serve, even though it is just as accessible. Its position is rather like that of Oakland in relation to San Francisco. The situation is different with Miami and Opa Locka airports in Florida and Los Angeles International and Long Beach airports in California, each of which handle over 400 000 aircraft operations a year, the one handling commercial traffic and the other general aviation and pleasure flights.

These facts have significant implications for airport planning. They emphasise the futility of hoping that airlines will voluntarily spread their service to any great degree over two or more airports in a metropolitan area. Yet sometimes the public interest desires this to happen, either to reduce noise and pollution around a particular part of the city or to secure easier access to air transport for the inhabitants. The evidence then indicates that a policy to distribute traffic to satellite airports will only work if the government pressures the airlines to do this.

Many different kinds of regulations can be used to coerce airlines to serve satellite airports. In the USA, the federal government has placed quotas on the total number of operations allowed to use a principal facility. The effect of this policy is uncertain, however. It does encourage airlines to schedule more flights to satellite airports, but there is no control over which destinations will be served from the satellites. Worse, it is almost certain that all the least profitable and thus least important flights will be assigned to the secondary airports. This is what happened when the US Government limited the number of airline operations at Washington/National, and forced the airlines to serve more customers from Washington/Baltimore. This quota policy did reduce noise and congestion at the major airport, but did little to enhance the attractiveness of the service at the satellite.

Satellite airports can also be developed by forcing airlines to provide specific services at designated sites. This procedure is aimed precisely towards the desired objectives, but can be circumvented. The attempt of the US Federal Aviation Administration to develop Washington/Dulles by requiring all long distance traffic to use that facility has not really worked because domestic airlines continue to use Washington/National by transforming long-haul flights to short-haul flights by making stops at Chicago, Atlanta or other closer points. The British Government somewhat more successfully encouraged the development of London/Gatwick by transferring all British flights destined for West Africa and much of South America to that location. But passengers do not have to fly British or even travel through London. London/Gatwick still only accounts for a small fraction of the traffic through the London airports.

The French effort to develop the new Paris/De Gaulle Airport in combination with Paris/Orly was more drastic. It was successful to the extent that it did channel comparable levels of traffic through both locations. But it was unfortunate in that it noticeably decreased the quality of air services through Paris and thus seems to have reduced traffic, and in that it imposed high new expenses on the airlines. The French reasoned that Paris was a metropolitan area with nearly 10 million inhabitants, that many cities of five million or less

had substantial airports, and therefore that it was reasonable to split Parisian air services into two halves. The government simply enforced the split. The affected airlines were then immediately saddled with the extra cost of duplicate staff and of transporting crews between airports. Air France alone reportedly spent $2 to $5 million on this account in the first year. Travellers also found that connecting flights were less convenient from either airport, in that frequencies were roughly halved, and began to bypass Paris by making connections through other cities.

Planning for second airports is not just a question of organising the competition between airports in a metropolitan area. It also requires a thorough analysis of the services provided by airports to the entire air transport network and the competition offered by airports in other cities.

8.4.3.4 Airports and the air network

Airports are not just local facilities – they are part of the entire air transport network. They potentially serve a much wider market than the metropolitan area in which they are located. In addition to handling the traffic originating in and destined for their immediate region, they function as transfer points for passengers and goods coming from and going to distant cities.

(a) Transfer traffic

Transfer traffic can be extremely large. At Atlanta, for example, transferring passengers outnumber those originating in the city by almost three to one. At many major European airports they account for approximately half of all the passengers boarding the aircraft. Even at smaller airports handling three to four million passengers a year, transfer passengers may represent 20 percent of the total. Whatever planning decisions do to influence the attractiveness of an airport, transfers may change the total loads on the facility.

Transfer traffic is also volatile. Having no intrinsic reason to pass through any particular point, it may – and often does – appear and disappear rapidly. Its patterns are sensitive to the wide range of elements that constitute the environment for air transport. Political changes can be crucial. The independence of the Portuguese colonies, for example, reduced the need to reach them via Lisbon and thus lowered air traffic through that city. Similarly, easier East-West relations could divert the flow of air traffic between Europe and Asia from the Middle East to the former USSR, thus diminishing traffic at Bahrain, Teheran and other stopover points.

Aeronautical developments can be equally important. The introduction of modern, long-range jets completely reshaped the pattern of transfers across the world. Just as Gander and Shannon are no longer necessary stops across the North Atlantic, Denver is no longer a major stop for transcontinental traffic across the USA.

Here again, however, frequency of service is a fundamental consideration. Frequent departures increase travellers' chances of making an easy connection to another flight, and minimise the possibility that they will have to wait a long time for transport to their destination. This is the phenomenon that the Parisian airport authorities failed to recognise fully in planning for the new Paris/De Gaulle Airport. By reducing the service available, they drastically diminished the attractiveness of Paris as a transfer point and undercut its share of the market.

(b) Concentration of traffic

To plan airport systems properly, the concentration points of traffic must be anticipated. This forms the basis of understanding the basic forces that shape the development and evolution of the air transport network. This applies especially to the forces that influence airlines to change frequency of service at airports.

A transport network always represents a compromise between two major goals: the desire for short, direct connections between any two points, and the desire for frequent service.

If airlines scheduled direct, nonstop services between every point, many would of necessity either be extremely infrequent or considerably more expensive if smaller aircraft, with higher costs per passenger, were used. Although direct services are convenient if a flight happens to be leaving when you want it to, they also imply low frequencies, correspondingly long waits, as well as higher costs. To overcome these difficulties, airlines encourage travellers from smaller communities to proceed to their ultimate destinations via larger hub airports. These detours obviously increase the time some passengers spend flying, but there are compensatory advantages. By concentrating their traffic, airlines have more passengers on fewer links, can provide more frequent service and may also be able to use larger, more economical aircraft, and thus can reduce the overall cost and time of many trips.

At some point, the possible savings in time and money due to concentration of airline service equal the extra cost and travel time inherent in more indirect or circuitous travel. This trade-off is, however, offset by the fact that many travellers are actually not sensitive to frequency of service.

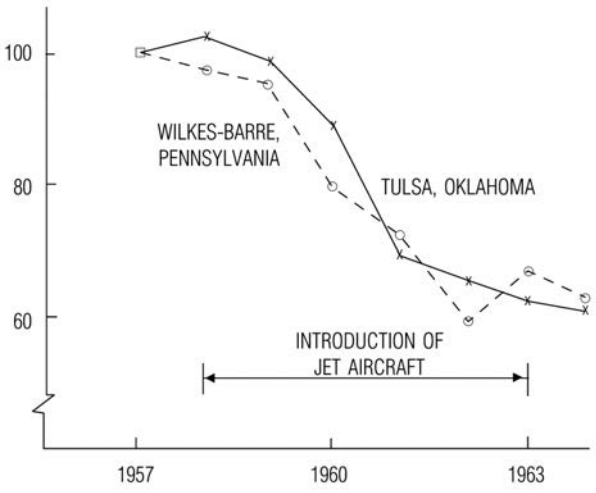
These forces determine the basic shape of the air transport network, and consequently the intensity of transfers at the hub airports. Finding the best pattern of service is an arduous process, since even a small number of airports imply an enormous number of distinct possibilities which can be overcome by means of computer-based methods.

Technological and demographic changes can shift the balance between the advantages of concentration of traffic and the disadvantages of circuitous travel. This then changes the percentage of travellers transferring. Anything that raises the overall volume of traffic, for example, makes it economical to offer more frequent service on direct flights, and this reduces the justification for concentration and decreases transfers. The introduction of larger aircraft, on the other hand, makes many direct flights unprofitable, increases the concentration of traffic, and therefore raises the traffic at hub airports while decreasing the number of flights at smaller airports.

The effect on traffic of introducing larger aircraft can be dramatic. When jets replaced smaller turbo-prop and propeller aircraft in the early 1960s, many smaller cities in the USA lost almost half their air service in a period of rapidly growing demands for air service in the country as a whole. Figure 8.7 illustrates this decrease.

Figure 8.7

Decrease in frequency of service accompanies the introduction of larger aircraft



Source: De Neufville (1976:76)

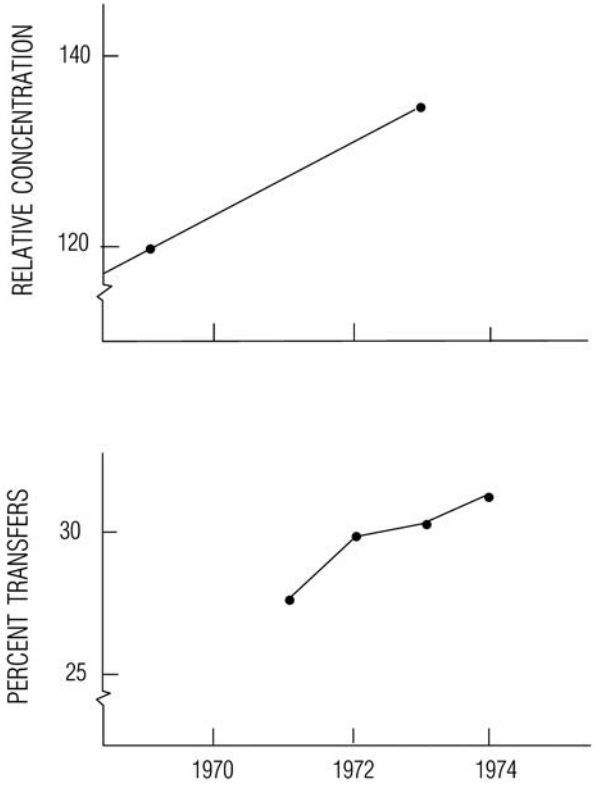
As the chairman of the US Civil Aeronautics Board put it:

> *the local carriers have been transformed ... their smallest aircraft are at least double the size of (those) they began with ... (They) have focused their energies on the ... higher density markets. The result has been that service to smaller communities has become less.*

Much the same result occurred in the early 1970s together with the introduction of wide-body aircraft. As figure 8.8 indicates, both the concentration on the air transport network and the percentage of transfers at major airports increased during this period.

Figure 8.8

Rise in transfer rates at hub airports accompanies decrease in connectivity of network



Source: De Neufville (1976:77)

Similarly, the traffic at major transfer points grew about twice as fast as traffic elsewhere as shown in table 8.1.

These observations emphasise that planning for any airport must consider the role of the airport in the air transport system and, specifically, the potentially rapid shifts in traffic due to competition from airports in other cities. This has not been done systematically in the past. Yet the continued failure to do so could prove terribly expensive in terms of wasted resources. As a recent study of the problem put it:

> *[O]ld methods of forecasting either national totals or individual airport traffic independent of service patterns will produce many mistakes in airport planning ... It is our fear that, using these methods of forecasting, excess capacity will be created in many smaller airports and too little capacity will be added at existing hub cities.*

Table 8.1

Passenger traffic grew faster at airports with higher percent of transfers during the introduction of jumbo jets

Airport Type	Location	Percent Transfer Traffic (1974)	Annual Percent Traffic Growth (1971–1974)
Transfer	Atlanta	73	13
	Dallas	55	10
	Chicago	47	7
Low Transfer or Terminal	Los Angeles	22	4
	San Francisco	21	5
	Detroit	15	5
	Boston	10	4

8.4.4 Summary

The fact that airports exist in a competitive environment underscores the idea that we should plan for systems of airports rather than for individual airports. As the preceding discussion indicates, the traffic at any location depends significantly on the development of services by other modes of transport and by other airports, both locally and further afield. Any planning process which fails to take this into account will almost inevitably find that its plans are inappropriate if not wasteful. Airport planning needs to be done on a national or at least a regional scale.

In this regard it should be recognised that the development of a realistic process for planning airport systems will be difficult. There is widespread reluctance to accept effective planning of a whole system. The industry has yet to fully acknowledge the costs of an individualistic, short-sighted view which overlooks the behaviour of the system.

Finally, we should be fully aware that the future performance of the air transport system is inherently uncertain. Quite apart from our own ignorance about how the system works, the detailed distribution of traffic between airports is highly volatile. These facts reinforce the recommendation that airport planning should be flexible in its approach to problems and formulation of solutions.

8.5 Investment in airports

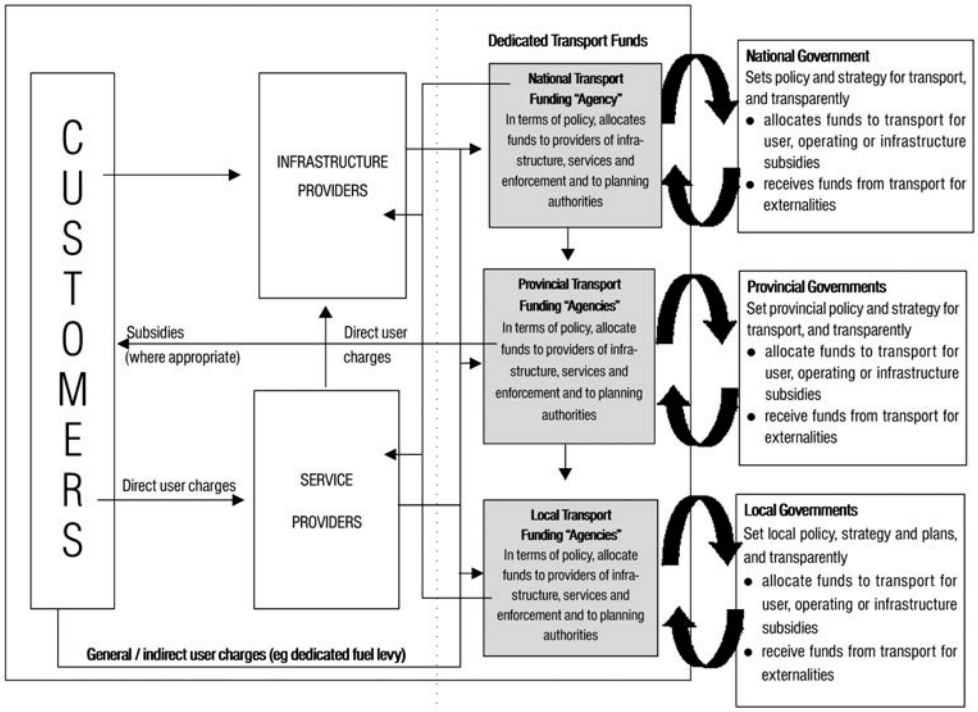
8.5.1 Financing by the government

Airports can be owned by either government organisations (also known as parastatals) or private entities. In both instances funds need to be available to finance the planned investments. The way in which infrastructure is financed in South Africa is depicted in figure 8.9.

Figure 8.9 shows that funds can be sourced from the national, provincial or local government. In all cases these funds will come from dedicated transport funds. The funds will be given to the service providers and/or infrastructure providers. They in turn will levy a charge on the users of the infrastructure – that is, the service provider and/or user of the infrastructure. This could take the form of passenger airport charges, landing charges for aeroplanes et cetera. If a loan has been granted, it will have to be paid back to the proper authorities. If a subsidy has been granted, then justification for the continuation of the subsidy should be provided.

Figure 8.9

Conceptual model of transport infrastructure funding



8.5.2 The bond market

The government does not have unlimited funds available for investment. Once the dedicated transport funds have been exhausted, alternative means of financing have to be found. The primary responsibility for financing the development plans then rests with the local operating agencies or authorities. There are many ways in which public financing of airport development can be accomplished. Finance can be raised from taxes, or by means of a bond or private financing (or a combination thereof).

A bond is basically a tender that is set out by the authority concerned to borrow a sum of money. We call this the issue of a bond. A bond issue can be sold competitively, with the airport accepting bids and selling the issue to the bond house that offers to buy it for the lowest interest rate, or the interest rate can be negotiated between the seller and a single buyer. Often airport sponsors use the service of a bond counsel who advises on the best way to market a particular bond issue. After a bond house has bought a bond, it resells the bond to commercial banks, insurance companies, pension funds and other large investors. Various types of bonds are found in airport financing and we will briefly discuss each.

8.5.2.1 General obligation bonds

General obligation bonds are issued only by provincial governments, municipalities and other local governments. The payments (interest and principal) to bondholders are secured by the full faith, credit and taxation authority of the issuing government agency. An advantage of general obligation bonds is that because they are a community guarantee they can typically be issued at a lower rate than other types of bonds. However, most governments limit the amount of general obligation debt that a local authority may issue to a specified fraction of the taxable value of all property within its jurisdiction. In addition, voter approval may be required before making use of general obligation debt.

In recent years, local governments have been under a great deal of pressure to finance all sorts of operations. The need to construct schools and other essential public works has required a considerable volume of general obligation bond financing. In numerous cases, local governments have reached statutory bond limits or desire to reserve whatever margin is left for more general functions of government. It is becoming increasingly difficult to obtain taxpayer approval for general obligation bond issues for airports.

8.5.2.2 Self-liquidating general obligation bonds

Self-liquidating general obligation bonds are also secured by the full faith, credit and taxation authority of the issuing government body, but there is adequate cash flow from the operation of the facility to cover servicing the debt and other costs of operating the facility. In other words, the bonds are self-liquidating (self-sustaining). Legally, the debt is not regarded as a part of the community's debt limitation. However, since the credit of the local government bears the ultimate risk of default, for the purposes of financial risk analysis, the bond issue is still regarded as part of the community's debt burden. Therefore this method of financing generally means a higher rate of interest on all bonds sold by the community. As a rule, the interest rate depends in part upon the bond's degree of “exposure risk”. Exposure risk occurs when there is insufficient net operating income to cover the level of debt service plus coverage requirements, thus forcing the community to absorb the residual.

8.5.2.3 Private financing

Specific facilities at airports such as hangars, fuel distribution systems and hotels are often built with private financing. Such facilities can be constructed with private capital on land leased from the airport. The obvious advantage of such an arrangement is that it relieves the community of all responsibility for raising the capital funds for the improvements involved. Tenant improvements in the terminal buildings are also typically financed by the tenant through its own funding source.

8.5.2.4 Revenue bonds

The debt on revenue bonds is serviced solely from the revenues derived from the operation of a facility that was constructed or acquired with the proceeds of the bonds. Revenue bond financing for airport improvements has become the most common financing method. Financing with revenue bonds provides an opportunity to make improvements without placing a direct burden on the taxpayer. Let us briefly look at how they work. We will use American airports as an example, since this method is seldom used locally.

After World War II, most of the larger airports began switching from general obligation bonds to revenue bonds as a method of financing new construction and improvements to existing fields. The first airport revenue bond in the USA was a $2,5 million issue sold in 1945 by Dade County, Florida, to buy what is now Miami International Airport from Pan American World Airways.

In the 1950s, the city of Chicago and the airlines that serve it worked out what has become the basic pattern for revenue bonds underwritten by airlines in the agreement that set up the financing for O'Hare International Airport. The airlines pledged that if airport income fell short of the total needed to pay off the principal and interest on the bonds, they would make up the difference by paying a higher landing fee rate. The historic O'Hare Agreement demonstrated that airports, backed by the airlines that use them, could raise the money they need in the financial market without depending on general tax funds. Airport revenue bonding thus became the accepted way to raise money for construction and expansion.

Revenue bonds are usually issued for 25 or 30-year terms, in contrast with the customary 10 or 15-year terms for general obligation bonds. Interest rates run slightly higher on revenue bonds than on general obligation bonds.

8.5.2.5 The market for airport bonds

Perhaps the toughest test of an airport's financial strength is its success in competing with other municipal enterprises for private investment capital in the bond market. While financially stronger airports tend to be most active in the bond market, even financially weaker airports can attract private capital, although they often have to use the taxation authority of the local government as security for bond financing.

Between 1978 and 1984, airports raised a total of $7 billion in new bond financing to pay for capital improvements. Since most municipal bonds are exempt from income tax, it is this key feature that makes this financing less expensive than most other sources of private money. Predictably, therefore, the vast majority of airport debt capital is raised in the tax-exempt bond market.

8.5.3 Bond ratings, interest cost and defaults

The competitiveness of airports in the municipal bond market can be gauged by three conventional indicators of investment quality, namely:

(1) bond ratings

(2) interest cost

(3) defaults

8.5.3.1 Bond ratings

Bond ratings is a system used by major investor services (such as Moody and Standard & Poor) to grade bonds according to investment quality. The top ranked bonds are as follows:

(1) *Best grade.* Bonds rated Aaa (by Moody) or AAA (by Standard & Poor) are graded best. Their exceptionally strong capacity to pay interest and repay principal offers the lowest degree of risk to investors in bonds.

(2) *High grade.* Bonds rated Aal or Aa (by Moody) or AA + or AA (by Standard & Poor) have a strong ability to pay interest and repay principal, but are judged to be slightly less secure than best-grade bonds. Their margins of protection may not be quite as great or the protective elements may be more subject to fluctuation.

(3) *Upper-medium grade.* Bonds rated Al or A (by Moody) or A +, A or A- (by Standard & Poor) are well protected, but the factors giving security to interest and principal are deemed more susceptible to adverse changes in economic conditions or other future impairments compared with bonds in the best and high-grade categories.

(4) *Medium grade.* Bonds rated Baal or Baa (by Moody) or BBB +, BBB- or BBB (by Standard & Poor) lack outstanding investment characteristics. Although their protection is deemed adequate at the time of rating, the presence of speculative elements may impair their capacity to pay interest and repay the principal sum in the event of adverse economic conditions or other changes.

Although investors have considerable confidence in airport bonds, ratings vary between the top and medium grades. A medium grade means that rating firms see the investment as having a measure of speculative risk. General obligation bonds generally draw the best ratings. Under this form of security, ratings are determined by the economic vigour of the

municipality or the entire province, and airports have little or no influence on the rating. Revenue bonds, on the other hand, draw ratings according to the fiscal vitality of the airport itself. Since more than 90 percent of all airport bonds (in terms of dollar volume) are secured with airport revenues, the criteria used by investor services to rate such bonds are central to their marketability.

Credit analysts at the major investor services rate an airport revenue bond according to a variety of factors, including the financial performance of the airport, the strength of passenger demand, and use agreements with the airlines serving the airport. Financial strength is viewed as a direct function of passenger demand at the airport, and credit analysts review both financial indicators and underlying patterns of passenger traffic.

Airline deregulation, which has freed air carriers from virtually all obligation to serve particular airports, has caused some shift in the relative weight credit analysts give to these different factors. In response to deregulation, nowadays the investor services place greater emphasis on local economic strength than on airport use.

8.5.3.2 Interest cost

Interest cost represents the payments by airports to attract investors relative to what other local authorities pay. The difference between interest cost paid by airports and by other public enterprises indicates that airports generally hold a strong competitive position in the municipal bond market. In deciding the price of a particular bond issue, underwriters identify a "ballpark" interest rate on the basis of general market conditions and then refine this estimate according to the credit standing of the airport in question. Two factors have great importance here: first, an airport's fiscal condition, and secondly, pressures on an airport to expand capacity which necessitates extensive capital developments.

8.5.3.3 Defaults

Defaults refer to the frequency with which a certain type of enterprise has defaulted on a bond issue, that is the enterprise could not repay it on time. To date, the airport industry has shown that it can cope with changes, and consistently make payments on its outstanding debts – hence its good credit rating.

8.5.4 Summary

Investment in airports is a comprehensive subject which is dealt with only superficially in this study unit. We have yet to look at the economic principles underlying investment in infrastructure. This is dealt with in another course. We have only looked at the various ways and means of financing airport investment.

8.6 Self-evaluation questions

(1) "Planning for one airport cannot be done in isolation." Discuss this statement in detail.

(2) What is meant by regional airport planning? Discuss some of the problems that might arise as a result of coordinating the plans of three airports in one metropolitan area.

(3) Summarise the objectives of an airport master plan.

(4) Discuss the importance of local coordination in developing an airport master plan and give some examples.

(5) Explain how competition between airports affects them. Use relevant graphs in your discussion.

STUDY UNIT 9

Rail transport investment

UNIT OUTCOMES

After working through this study unit you should:

- be able to identify practical ownership models of rail enterprises
- recognise the interdependence of investments
- be aware of the conflict of interest between the owners and operators of rail transport
- be able to identify the optimum maintenance strategy for rail transport
- recognise the necessity of track investment in terms of train operating performance

KEY CONCEPTS

- Rail track ownership
- Operating units ownership
- Vertically separated
- Vertically integrated
- Track design and maintenance
- Operations of trains

9.1 Introduction

When investing in rail transport, a distinction should be made between the variable (moving) components and the fixed components. Variable components refer to units such as locomotives and freight and passenger wagons and are generally known as rolling stock. The fixed components of rail investment are represented by objects such as the rail track, signalling system and buildings and are generally known as the infrastructure. An increase in the demand for rail transport will optimise the utilisation of the spare capacity of rolling stock, which means that the fixed component of rail transport will need to be upgraded or extended. Spare capacity in terms of rolling stock is usually available because of the movability of units, whereas a track, for example, can accommodate only a set number of trains. The interaction between rolling stock and infrastructure will therefore play an important role in investment decision making.

The worldwide tendency in rail transportation is towards privatisation. In South Africa, privatisation of railways is also an important issue. In certain instances, rail privatisation or concessioning (ie allowing operators to operate on certain sections as a private undertaking) is generally embarked on to solve financial crises. However, at Transnet this issue is approached differently. The argument is that financial, structural or other problems should first be solved in the enterprise as a parastatal. Only then will it be possible to place this state asset on the market and possibly fetch a "good price" (Chalmers 1999:28).

In this study unit we shall discuss planning and investing in rail transport in terms of privatisation. Track ownership models which have an important influence on planning, investment issues, conflicts between owners and operators, track design and maintenance and train operating performance will also be discussed. These discussions are taken from the research by Ferreira (1997:183–200). Although the discussions are based mainly on Ferreira's research and experience gained from the Australian freight rail sector, most of the conclusions have wider application.

9.2 Track ownership models

9.2.1 Models

Two main ownership models are emerging in practice, namely the *vertically integrated railway* with or without separate internal business units, and the *vertically separated railway* with track infrastructure managed and owned independently by multiple operators. The *vertically separated model* has been adopted or proposed in some countries, notably in Great Britain, Germany, the Netherlands and Sweden (Nash & Preston 1994; Jansson & Cardebring 1989). The European Union has a policy of moving towards the separation model (Nash & Preston 1994). A similar approach is under consideration for interstate freight in Australia, following the competition-related proposals adopted by federal and state governments (Hilmer, Rayner & Taperall 1993).

Figure 9.1 highlights the main features of these two models.

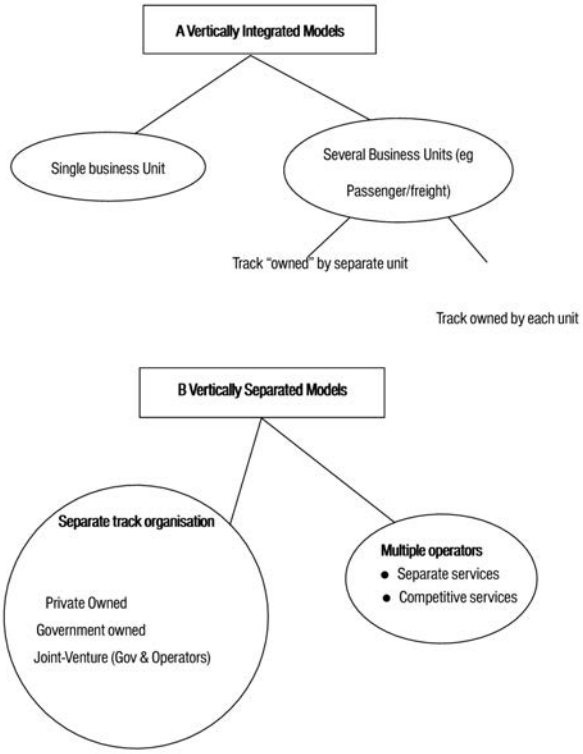
In the *vertically integrated model*, operators and track owners tend to have a customer-service provider relationship. The infrastructure provider exists to service the needs of its client(s). The latter may consist of several business units such as passenger services and various types of freight services. In some cases, each business group "owns" its own track segments, which are divided between operators on the basis of major user. User charges may be levied on non-main users using an internal cost transfer system designed to achieve accountability and "value for money" outcomes. It is argued that one of the drawbacks of the vertically integrated model is its inability to readily and fairly accommodate new entrants in the form of operating competitors who share a common track infrastructure. If existing railway systems are publicly owned, it is possible to open up track to new entrants through direct intervention by governments. However, the question of fairness in dealing with potential competitors would require strict contractual arrangements related to costs and service quality. The terms and operating conditions of track access need to be extended to train dispatching rules. This is particularly important in single line operations, where the train conflict resolution rules need to be seen to be fair and equitable for all operators as well as economically sound.

By contrast, the *vertically separated* model has been put forward as a way of increasing competition in the rail sector and of placing rail and road infrastructure investment and operations on an equal footing. The main stated aim of the separation of track from operations in the UK was to ensure

- competition in service provision, and hence
- improved customer service at lower costs

Figure 9.1

Rail functional models



Source: Ferreira (1997:186)

Since competition has not yet materialised in practice, the benefits of separation may turn out to be small relative to the costs of loss of coordination and transaction costs, such as contract specification and enforcement (Dodgson 1995). According to Bruzelius, Jensen and Sjostedt (1994), the vertical separation of railway functions in Sweden appears to have resulted in a lowering of the quality of service provided by the track owner.

This model has serious implications for the overall productivity of rail operations, given the nature of the railway business and the fact that the operators are not responsible for investments in the rail infrastructure. In addition, the bargaining power of new entrants to negotiate contracts with a monopoly track owner acting to achieve commercial objectives needs to be adequately safeguarded. The competitive pressures on train operators which are sought through this model are in danger of being absent to the infrastructure provider.

9.2.2 Options to achieve competition

Between the two extremes of total separation and total integration lies a range of options which may provide useful means to achieve competition at the train operators' levels while preserving the benefits of integration.

A hybrid model, which draws on the strengths of vertical integration while allowing for fair competition between operators, may be more desirable. In such a model, the track infrastructure could be a separate business entity owned by operators. Access by new entrants would be open, with charges and service contracts partially regulated to ensure fair play. Decisions about track investment would tend to be integrated with the overall investment plans of operators.

A variation of this approach is for a joint-venture company to own track infrastructure. The main shareholders in this company would be the operators and governments (national and/or regional). Such government involvement could be justified on three main grounds, namely:

- that it would tend to place road and rail investment on a more equal basis
- that it would ensure that new operators would be treated fairly with respect to service levels and price
- that public funds would provide for full cost recovery by supplementing access charges based on avoidable costs, since the latter usually cover only a small component of total track costs

In this model, new operators (large or small) would be given the option to join the joint venture. In this way discriminatory practices towards small operators, as well as unfair pricing policies, could be minimised without the need for heavy-handed regulation.

In this model the infrastructure-owning company would operate on a fully commercial basis and would be accountable to its shareholders using common commercial criteria. Equity funding would need to be sought on the basis of revenues from access charges and/or community benefits set by governments. The relationship with train operators would be a purely commercial one. In other words, access charges would need to cover short-term track damage, track capacity charges if applicable (eg peak period charges) and long-term investment requirements which are a direct result of train operations (such as increased track standards from new rolling stock). The advantage of such a model, relative to the fully integrated government owned railway, is that the potential for track managers to move to the most efficient maintenance practices is enhanced. Increased profit due to higher productivity can be passed directly to the operator or retained for future investment. This may not be the case for integrated systems where the in-house infrastructure provider can act as a monopoly able to pass costs on to the "parent" railway, and set its own standards. The latter are traditionally conservatively set and hence costly to sustain. In the vertically separated hybrid model, the train operators, as shareholders in the track-owning company, would be in a strong position to ensure that productivity improvements take place according to the access-charging agreements. The most effective organisational model to be adopted needs to take the following into account:

- the specific aims of the railway organisation(s)
- the existing levels of efficiency
- prices
- customer service

Freight railways which are efficiently run and operated according to international benchmarks seem to have little to gain by moving to a fully separated structure. Such performance comparisons, although fraught with difficulties (due mainly to differences in traffic densities, lengths of haul, terrain and outsourcing policies), suggest that freight railways could be run more efficiently. Thus it is likely that some form of vertical separation would see the entry of operating competitors on selected routes.

Activity 9.1

In your opinion, which privatisation model should be used for Spoornet in South Africa? Also mention the planning considerations that should be taken into account (consult study unit 2).

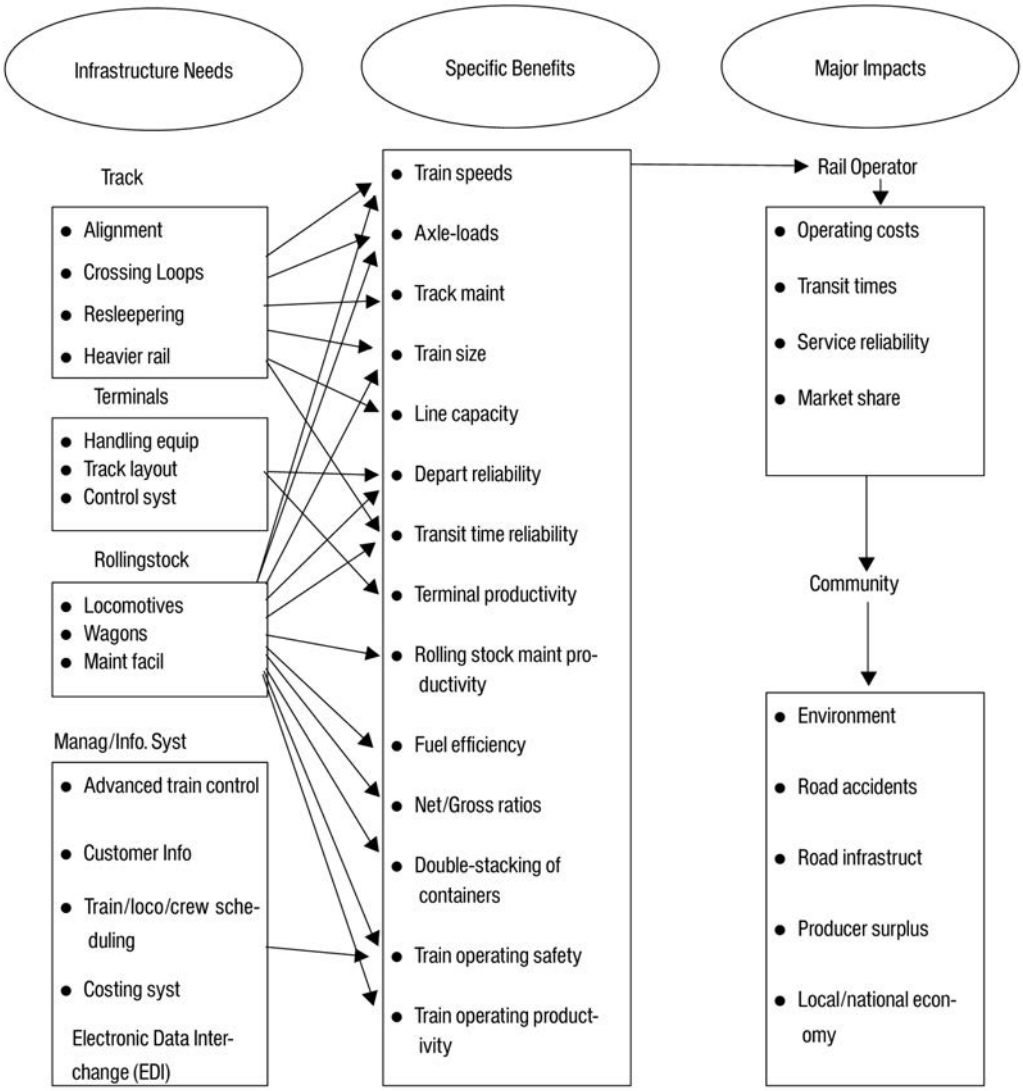
9.3 Investment issues

9.3.1 Interdependence of investments

There are strong interdependencies between functional units in railways because of the joint nature of the production functions and the complexity of the business as a whole. Railway track investment decisions are strongly interrelated with investment decisions about other types of rail infrastructure, as illustrated in figure 9.2.

Figure 9.2

Elements of freight rail investment appraisal



Source: Ferreira (1997:183–200)

Investment decisions about track capacity may have an impact on operating strategies and thus on the level of service provided. For example, track design standards and maintenance strategies have a direct influence on the following:

- maximum allowable axle loads
- train speeds
- track upgrading including double tracks and sidings or crossing-loops (short track sections on single lines to allow trains to cross and pass each other), which affects transit time reliability and line capacity
- terminal upgrading which results in a greater number of trains having to be coordinated with track infrastructure upgrading
- track condition which affects locomotive and rolling stock performance and maintenance costs
- train control technology which may require track-side investment and compatible locomotive cabin equipment

On the other hand, investment in new technology for wagon and locomotive fleets may have a direct impact on track standards requirements and on track maintenance costs. Thus track investment decisions should be part of an overall long-term strategic plan designed to achieve an organisation's goals and implemented over a certain period of time to maximise their benefit. In the vertically separated model, such decisions need to be taken in a spirit of cooperation between the main operators and the infrastructure supplier.

In Australia, for example, there is a considerable investment shortfall in rail infrastructure for the strategically important national links, which are almost exclusively made up of single-line track. These links require annual investment in the order of $A150 million over the 20-year period to 2015 (in 1995 dollars). This investment is needed to improve the level of service provided by operators (Bureau of Transport and Communications Economics 1995). The major benefits of such investment will be in the form of reduced train operating costs and increased business opportunities stemming from lower transit times and higher reliability of arrivals.

In a vertically integrated railway management model, the task of developing a medium to long-term investment plan in which the interdependence of projects is explicitly recognised is considerably easier.

9.3.2 Road and rail investment appraisal

If economic efficiency in the allocation of resources between road and rail is desired, then the same methods of investment evaluation need to be adopted. Road planning agencies currently evaluate road projects on the basis of social cost/benefit analysis (SCBA), potentially capturing the actual benefits of reductions in road vehicle operating costs, personal travel time, road accidents, congestion costs and environmental costs.

One of the main reasons for the vertical separation of Swedish railways was to allow road and rail to be placed on an equal investment and pricing basis (Jansson & Cardebring 1989). With respect to cost recovery in both road and rail sectors, Nilsson (1992) advocated reducing the relative price of rail to offset underrecovery of full marginal social costs from heavy road vehicles in Sweden. As a “second-best” approach, this is comparable to the argument for urban public transport subsidies (to offset farebox shortfalls) so as to reduce urban road vehicle congestion costs. Whatever railway organisational and ownership model is adopted, the major infrastructure projects for both road and rail should be subject to comparable economic evaluations, so as to fully quantify financial, economic and social impacts. In principle, the use of SCBA in railway project appraisal should not be dependent on the track ownership model adopted. In practice, in a vertically integrated

railway, economic evaluation tends to occur in an integrated fashion for track and operating infrastructure (such as rolling stock). This makes the results of such evaluations less comparable with road investments than would be the case if the operators were vertically separated from infrastructure providers.

9.4 Conflicts between owners and operators

Rail operators and the owners of railway infrastructure may have conflicting objectives because they have different stakeholders and levels of accountability. Railway services operated for profit will be concerned about reducing operating costs and increasing revenue (via growth in market share or freight rate increases). Market share increases are closely related to the level of service which each operator can offer. Transit times and reliability of arrivals play an important part here. Both these levels of service attributes are associated with track infrastructure design and maintenance standards. Therefore an operator's ability to perform efficiently and gain market share is closely related to its ability to strike an effective contractual arrangement with the infrastructure owner.

Railway infrastructure owners have to plan and manage their assets according to their overall strategic objectives. In the case of public ownership of railway infrastructure, there is an obligation to make investment decisions which cater for the interests of current service operators (sectional/private interest) and the community to whom the entity is accountable (collective/public interest). If, as in the case of plans in UK, the infrastructure is to be owned on a purely commercial basis, the owner has a profit-maximising strategy which will of necessity disregard the community costs and benefits of management decisions.

Whether infrastructure is privately or publicly owned, it is important to ensure that the owner has sufficient incentive to move towards the most productive maintenance methods and the most effective long-term track standards. This will require investment decisions to consider assets that may have an economic life of 50 years (eg concrete sleepers). Long-term commitments from operators will be required in such cases. There is a danger that existing low levels of track maintenance productivity in Australia (Bureau of Industry Economics 1993) will not be significantly altered if owners are left in a position to pass on costs to operators without short-term productivity incentives or long-term contracts.

If there is more than one operator, the infrastructure owner faces potentially different demands for track maintenance, track design and capital needs. More particularly, different market segments such as freight and passenger services will require different maximum speed and axle-load standards, which have implications for investment and ultimately for user charges. The owner will need to provide a “level playing field” so that each operator can gain access to track at the appropriate time and cost. The issue of time of access is important for a number of reasons. Conflicts of access to track are likely to occur between users who may be competing against each other in the marketplace. At present, such conflicts are resolved according to internal railway rules on traffic priorities. The question of limiting track capacity at peak times will involve a user-charging system which can take into account the risk of delays at such times, as discussed in a later section. The issues of track design and maintenance, track access costs and train operating parameters are discussed below.

9.5 Track design and maintenance

The allocation of track costs among users is a major issue in the light of the fact that there is still a poor understanding of the causes of track deterioration, despite considerable research efforts throughout the world (Hope 1992). At present, the effect of train speeds, axle loads and vehicle types on maintenance efforts is estimated without a great deal of precision.

In 1995, the Bureau of Transport and Communications Economics (1995) estimated that the penalty for failing to adequately invest in track infrastructure on the Australian main-line network, in terms of additional track infrastructure and additional track maintenance costs, would be about $A1 billion over the next 20 years (in 1994 dollars). According to the Bureau of Industry Economics (1993), track maintenance in Australia represented the most significant potential saving of total operating costs in 1991/1992. There are significant productivity gains to be realised in track maintenance through capital expenditure in both the maintenance task itself (eg mechanisation) and by moving to higher quality, lower maintenance track structures. The arguments for vertical separation would seem to be strengthened by the historically low underinvestment in track and inefficient equity. Track owners should be in a better position to fund major track upgrading on the basis of access charges and community benefits. The infrastructure owner should be responsive to operators' needs without interfering in train-operating planning issues.

There is also a need to ensure that maintenance of existing networks is undertaken according to a plan which maximises overall net benefits for all rail operators. The need for maximum resource productivity is coupled with the need to improve our understanding of the causes of track deterioration. Research in Australia has shown that there is insufficient knowledge in the Australian context of the forces generated by moving trains (in particular those with axle loads above 25 tonnes), and therefore of the consequent deformations (static and dynamic) of track components (Hagaman 1989; Murray & Griffin 1993; Muller 1985).

The model of vertical separation of track dictates that each user be charged track damage costs on an equitable basis. When several operators compete for the use of track owned by a separate business unit or company, it is essential to know the damage being caused by each user, both in the short and long term. Rail traffic users need to pay at least the avoidable costs they incur. The common costs, that is those which cannot be attributed to specific users, tend to be a significant component of total track cost (UK Department of Transport 1993). Costs related to track damage are typically less than five percent of total track infrastructure costs, with common costs making up around 50 percent. The remainder are various elements of long-run avoidable costs which can be allocated to services or groups of services.

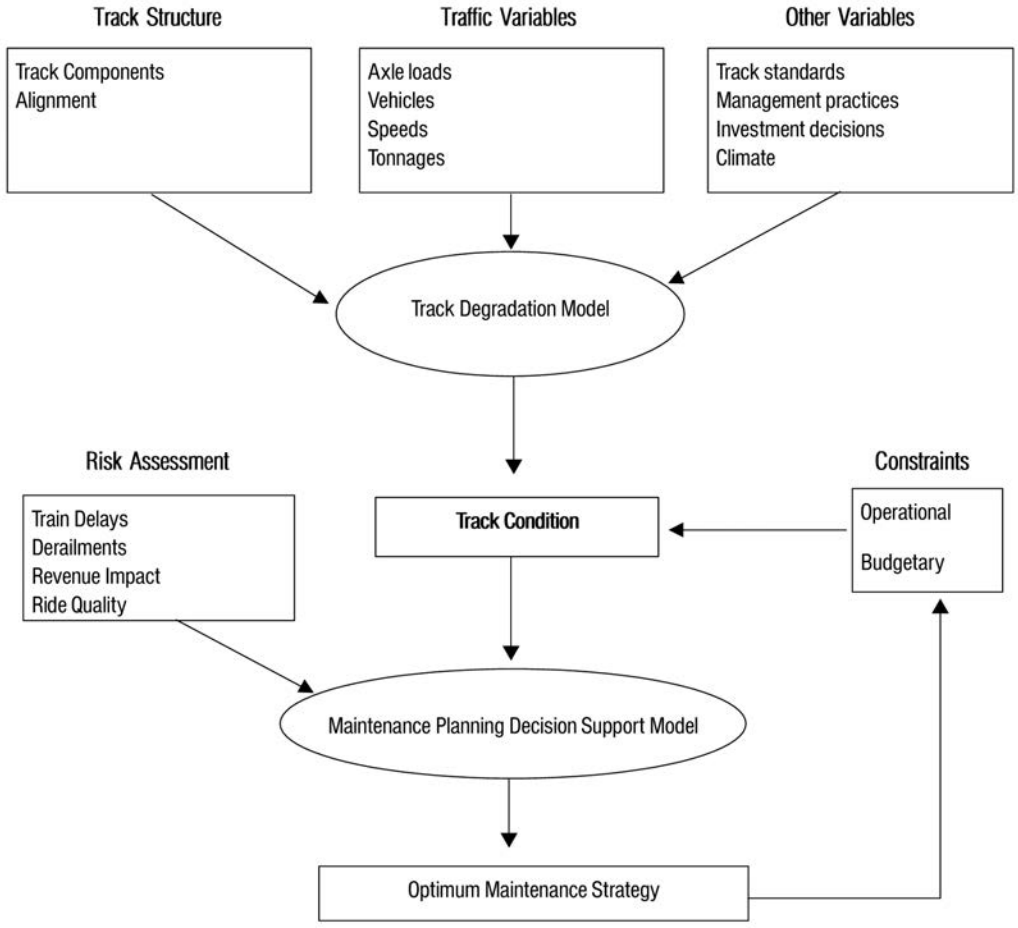
The infrastructure owner needs to maximise profit subject to satisfying the levels of performance required by each user. Such performance requirements may well be in conflict with optimum maintenance practices (eg scheduling maintenance windows in conflict with marketing needs). Binding contracts specifying in detail the outcomes desired by all parties may be difficult to negotiate and enforce in practice, as experience in the UK suggests (Dodgson 1995).

Figure 9.3 shows an optimisation model for conducting the track maintenance planning function in the context of the overall rail business. An important element of such a model is the explicit inclusion of risk variables to deal with the impact of track condition on train transit times, reliability of arrivals, accident/derailment potential and hence rail business revenue. These risks, which may result from suboptimal maintenance planning decisions, need to be part of the optimisation of the overall objective function. Thus the track investment and maintenance functions cannot be divorced from customer service and operations planning requirements.

As shown in figure 9.3, track maintenance needs to take into account the dynamic nature of the relationship between track condition and maintenance activity. Equations (1) and (2) highlight this relationship. For each line segment, the maintenance activity required at time period c is a function of observed track condition during time c – 1, traffic-related variables, risk-related variables and environmental factors.

Figure 9.3

Track maintenance optimisation parameters



Source: Ferreira (1997:183–200)

Track performance in time period c is a function of the amount of maintenance activity since the last major rehabilitation and of traffic and environmental factors.

$$TC_c = \int\left(\sum_{t=1}^{c} MA_t \sum_{t=1}^{c} TK_t, AX, S\right) + e_1 \quad (1)$$

$$MA_c = \int\left(TC_{c-1}, TK_{c-1}, \sum_{t=1}^{c} R_t, AX, S\right) + e_2, \quad (2)$$

where:

TC = track condition

MA = maintenance activity

TK = traffic task

R = set of risk-related variables

S = speed regime

t = time period

AX = axle-load regime

c = current period

T = ultimate planning horizon; and e_1 and e_2 are error terms.

In Australia, decisions related to investment in new track and the maintenance of existing track have been based mainly on engineering factors (Bell & Marsden 1991; Murray & Ferreira 1995). The concept of planning future maintenance schedules for an entire network based on predicted future traffic task by track segment has yet to be implemented in practice.

9.6 Train operating performance

9.6.1 Transit time reliability

When trains are scheduled on a rail corridor, the objective is to achieve a given level of customer service while minimising overall operating costs. Customer service in this context is made up of several attributes which include overall journey time and train arrival reliability.

In the context of freight movements, the benefits of improved reliability need to be estimated on a train-by-train basis. Each train is usually loaded with freight from a range of customers and origin-destination rows. The elasticity of demand with respect to transit time reliability will differ for each customer, commodity and origin-destination combination. However, reliability of arrivals is a critical performance measure for all rail markets. The ability of rail systems to compete effectively relies largely on this level of service attribute; price is obviously also important.

If operations are conducted on single-line track, where trains can only overtake or cross other trains at specified locations, transit time reliability is a function of a range of factors. The degree of “slackness” built into the schedule, the number and position of train conflicts, priorities for each train, terminal congestion, the number and nature of scheduled stops and train speeds are all influential variables. The calculated optimal schedule may only be optimal if no delays occur in the train operations. In contrast, little research has been conducted into the question of what happens to the schedule if unexpected delays occur. It is under such conditions that congestion-related access charges need to be estimated.

9.6.2 Track investment: train reliability nexus

When the flow of trains is near capacity, the system is unstable because of unexpected delays. A railway line with low risk of unexpected delays will be able to operate near capacity with very little instability. A schedule that is planned to minimise overall operating costs is not necessarily optimum from the viewpoint of overall net benefit to operators. This occurs particularly when the schedule is easily disrupted through small delays to individual trains.

Rail systems have a major interest in determining the risk associated with either a given schedule or additional train services. This is of decisive importance in vertically separated railways, with multiple train operators using common track corridors and potentially causing delays to one another. Delay risk may be further analysed to determine which trains are most vulnerable to delays, or which track segment(s) cause the most instability in the schedule. Such analysis is particularly relevant to new operators wishing to use congested track segments.

To minimise the delay risks associated with a given level of demand or to cater for additional demand without increasing the risk of delays, the following strategies are available:

- The source of track-related delays can be reduced through track investment and/or changes to maintenance practices.
- The maximum allowable speeds can be increased through investment in major track strengthening. The higher speeds have the potential to reduce conflict-related delays and improve journey time recover ability.

- Average train speeds can be increased through investment designed to alter track alignments, both vertical and horizontal.
- Another option is investment in additions to the number and length of sidings where trains can cross and pass each other. Conflict-related delays are directly affected by the number, length and location of such sidings.
- Investment in advanced train control and communication systems will allow trains to proceed at shorter headways, and with less stops required for safe train operations.

All these strategies involve additional costs to the track owner which need to be equitably passed on to users. The benefits of some of the above strategies usually extend beyond transit reliability gains. For example, in the case of track rehabilitation and upgrading, these benefits may include the following:

- reductions in overall transit times
- a reduction in accident risks
- lower track maintenance costs
- increased train productivity from higher maximum allowable axle loads
- reductions in rolling stock maintenance costs due to improved vehicle-track interaction
- improved locomotive productivity through the use of more modern equipment

Investments in nontrack-related areas will also result in reduced delay probabilities. Examples here are terminal infrastructure and information systems designed to improve freight-handling operations, thereby improving on-time departure performance, and new locomotives capable of higher maximum speeds and improved self-diagnostic capability to reduce breakdown incidents. In order to obtain maximum benefits it is usually necessary to combine a number of investment strategies into a coherent and complementary package of capital expenditure projects. For example, the gains in reliability from track upgrading projects can be augmented by investment in terminal infrastructure to allow faster loading/unloading of trains, and with regard to locomotives, investment to make higher train speeds possible. The costs arising from the lack of such coordination in the vertically separated model can be significant.

Activity 9.2

Explain the strategies available to minimise the delay risks associated with a given level of demand. Also mention how these strategies influence rail transport investment.

9.7 Conclusion

The issue of who should own the railway infrastructure has significant implications for investment and resource allocation for the land transport sector and for long-term rail profitability and performance. Operators may share the rail infrastructure with other rail systems responsible for passenger services and intrastate freight movements. Various organisations may therefore have different priorities when it comes to infrastructure upgrading.

Investment in individual elements of railway infrastructure should be integrated with the overall cost recovery strategy of the operator. Major railway projects should be submitted for

both financial and economic evaluation (as discussed in study unit 5), so that the interests of individual railway authorities and the community are considered.

Market share increases are closely related to the level of service each operator can offer. Transit times and reliability of arrivals play an important role here. Both these levels of service attributes are associated with track infrastructure design and maintenance standards. If there is more than one operator, the infrastructure owner faces potentially different demands for track maintenance, design and capital needs. The owner will need to provide a "level playing field" so that each operator can gain access to track at the appropriate time and cost. The issue of time of access is important for a number of reasons. Conflicts of access to track are likely to occur between users who may be competing against each other in the marketplace.

The European trend is towards the formation of rail infrastructure entities as separate businesses supplying services to operators, which draws increasing attention to track design and maintenance issues. This trend has also become evident in Australia, especially in the wake of political agreements to increase competition in the transport sector. However, one should not underestimate the difficulty of making the vertically separated railway work effectively in practice.

A hybrid model, which draws on the strengths of vertical integration, has been put forward. In such a model, the track infrastructure could be a separate business entity owned by operators. Access by new entrants would be open, with charges and service contracts to be negotiated between the parties involved. Some regulatory framework would probably be required to ensure fair play in the area of access charges. Moreover, the treatment of new/small operators would be difficult to ensure without the presence of an independent regulator. In this model, decisions about track investment would tend to be integrated with the operators' overall investment plans. A variation of this approach would be for a joint venture company to own track infrastructure. The main shareholders in this company would be the operators themselves and government. Thus the need for strict government regulation to protect small operators and avoid unfair access pricing practices would be avoided.

In this "hybrid" model, the company owning the infrastructure would operate on a fully commercial basis and would be accountable to its shareholders on the basis of common commercial criteria. The advantage of such a model, compared with the fully integrated government-owned railway, is that it enhances the potential for track managers to move to the most efficient maintenance practices. Increased productivity can be passed directly to the operator or retained for future investment. This may not be the case for integrated systems where the in-house infrastructure provider can act as a monopoly able to pass costs on to the "parent" railway and set its own standards. In the vertically separated "hybrid" model, the train operators, as shareholders in the track-owning company, would be in a strong position to ensure that productivity improvements take place in accordance with the access-charging agreements.

The most effective organisational model needs to take into account the specific aims of the railway organisation(s) as well as the existing levels of efficiency, prices and customer service. Freight railways which run and operate efficiently seem to have little to gain by moving to a fully separated structure.

9.8 Self-evaluation questions

(1) Explain and distinguish between possible practical ownership models for rail enterprises.

(2) Discuss investment in rail transport in terms of the ownership models.

(3) How does the conflict of interests between the owners and operators of rail transport influence the planning and investment of rail transport?

(4) The strategy for maintaining and upgrading both infrastructure and operating performance influences investment options. Fully discuss this statement.

STUDY UNIT 10

Transport policy and regulation

UNIT OUTCOMES

After working through this study unit you should be able to:

- explain the reasons for having a transport policy
- discuss the objectives of a transport policy
- explain the factors to be considered in a transport policy
- describe the elements of a sound transport policy
- identify the level of a transport policy
- discuss the South African transport policy
- explain the need for regulation

KEY CONCEPTS

- Transport policy
- Policy instruments

10.1 Introduction

The aim of this study unit is to explain the reasons for transport legislation and why it is required in the transport industry. This involves discussing why government intervenes in transport, the need for a national transport policy and regulations and the implementation of the policy by means of different kinds of legislation.

The authorities in South Africa have played a major role in developing the transport infrastructure that exists today and consequently in regulating transport. The present government has emphasised the importance of regulating transport by announcing that it has targeted transport as one of its five main priority areas for socioeconomic development (*White paper on national road transport policy 1996:1*). The role of government may therefore be defined in a transport policy dealing with various laws, rules and funding programmes aimed at controlling and promoting the different modes of transport such as road freight transport.

The transport industry in South Africa is controlled by a central authority within a comprehensive legal and administrative framework. The existing laws, regulations and ordinances are applied at national, provincial, regional and municipal level and involve transport services in general. Since the transport sector is only one element of the South African economy, it is important for a transport policy to consider the entire economic framework in which it functions. However, economic factors are only part of the picture – social, strategic and

political factors also need to be considered. The relative importance of these elements fluctuates from time to time. Thus a transport policy cannot be formulated and executed in isolation, and these fluctuating elements should be handled in a coordinated and integrated manner. Furthermore, a transport policy should not be static but dynamic, and therefore continually be reconsidered and, if necessary, revised.

Regulation is a means of implementing a transport policy. However, it is important to understand the need for a transport policy, the development of such a policy, and the South African transport policy itself. We shall be discussing all of these points in this study unit.

10.2 Government intervention in transport

Government authorities have a definite influence on the performance of the transport system because of their involvement. Their activities can have either a positive or a negative influence on transport costs and practices, and consequently on the economy as a whole. Investment in transport infrastructure can improve access to different land uses and reduce congestion. By constructing new roads, rail tracks, tunnels, transfer facilities et cetera, transport costs and transport times can be reduced. However, transport operations and cost can be negatively influenced by government taxes on vehicles and fuel, or by regulations controlling the maximum allowable axle masses, the maximum allowable dimensions of vehicles and the banning of heavy vehicles from some routes during specific time periods.

Governments become involved in transport for various reasons, most of which usually concern the protection of public interest by promoting the provision of sufficient and safe transport services while limiting any negative impact of transport on the community and environment. Government may intervene in transport for ideological reasons or become involved only when the transport market fails to produce the desired results.

Transport regulation, which may be either economic or noneconomic, can severely restrict transport operators' freedom of action. Noneconomic regulation takes the form of a multitude of technical standards aimed at the safe and efficient use of vehicles and infrastructure. Economic regulation, on the other hand, is aimed at restricting competition in the market as an instrument for guiding economic decisions in a certain direction. It is often implemented through authorisation procedures before a tribunal, for example, when applying for an operating permit.

In recent years, the trend in South Africa has been to move away from comprehensive regulation to a more market-oriented approach. The government has moved away from economic regulation towards safety regulation, and from quantitative to qualitative regulation.

10.3 Why do we need a transport policy?

A good starting point in examining the nature of the national transport policy is to consider the need for such a policy. The answer to the question of why we need a transport policy lies in the significance of transport in the everyday lives of people and involves the following elements (see study unit 1):

10.3.1 The importance of transport

Transport permeates every aspect of a community and touches the lives of all its members. The transport system ties together the various communities of a country, making possible the movement of people, goods and services. Transport therefore creates time and place utility because goods and people are moved in order to be at a specific place at a specific time. The physical connection that transportation affords to spatially separated communities gives people a sense of unity.

10.3.2 Economic benefits

In addition, transportation is fundamental to the economic activity of a country. It furthers economic activity such as the exchange of mass-produced goods between one location and locations where the need for these goods is greater. The citizens of a country would not enjoy the carry-over benefits of economic activity such as jobs and improved goods and services without a good transportation system. Transport therefore plays an integral part in the production process. Production is not possible if the inputs required cannot be delivered from their respective origins to the place of production. The production process is incomplete until transport has added the necessary utility of time and place to the outputs concerned. Thus transport costs can be regarded as an element of production costs. If transport costs decrease because of greater transport efficiency, the factors of production can be delivered to the producer at a lower price.

10.3.3 National defence

An efficient transportation system is also fundamental to national defence. In times of emergencies, people and materials must be deployed quickly to various parts of a country. Without an efficient transportation system, more resources would have to be allocated for defence purposes in many locations. Thus an efficient transportation system reduces the amount of resources consumed for national defence.

10.3.4 Public investment

Many of the transportation facilities in South Africa cannot be developed by private enterprises. For example, the capital requirements to construct a highway between Johannesburg and Cape Town are probably beyond the resources of the private sector. Efficient and economic highway routes require government assistance in securing land from private owners; if the government did not assert its power of eminent domain, routes would be quite circuitous and inefficient. Furthermore, public ownership and the operation of certain transportation facilities such as highways are necessary to assure access to all who desire to use the facilities.

10.3.5 Resource allocation

The purpose of a transportation policy is to provide direction in determining the quantity of national resources to be allocated to transportation and the quality of service that is essential for economic activity and national defence. The allocation of resources to transportation reflects both a market-allocation process and a political allocation process. Ideally, the political process should recognise the potential inadequacies of an unrestrained market that provides for each individual's basic necessities and then act to prevent market imperfections. Furthermore, the market process should operate within such constraints to efficiently provide the transportation a society desires and is willing and able to pay for. It is essential to understand this blend of marketplace interaction in the transportation area. The government's role is multifaceted and revolves around *loans and subsidies* and *economic and safety regulation*.

10.3.6 Decision guidelines

The national transportation policy provides guidelines to the many entities that have decision-making powers about transport and to the courts that make and interpret the laws that affect transportation. Thus transportation policy provides the framework for the allocation of resources to the different transportation modes.

10.3.7 Responsibilities

The responsibility of the government as a developer and owner of the transport infrastructure is mainly to

- ensure the safety of travellers
- protect the public from the abuse of monopoly power
- promote fair competition
- develop and maintain vital transport services
- balance environmental, energy and social requirements in transportation
- plan and make decisions

These responsibilities indicate the diversity of public needs that transportation policy must serve. However, one should keep in mind that some sectors of the transport industry in South Africa have been deregulated economically, such as the road freight industry by means of the promulgation of the Road Traffic Act 29 of 1989. Only the regulation of traffic on public roads (licensing, registration of vehicles etc) and certain requirements regarding the fitness of operators are applicable.

10.4 Developing a policy

10.4.1 Introduction

Formulating a public policy is a complex process in which the policymaker is confronted with a series of decisions or choices between different options. The analyst is rarely in a position to comprehensively analyse a complex policy question with the full assurance that all the positive relevant elements have received sufficient attention. Instead he or she is forced to choose between alternative methodological approaches. There are no simplistic models and the policymaker cannot optimise values but must balance conflicting values. There is also no specific solution to a policy problem, since practice and theory are not the same and there is no one single set of values. Values vary with people, groups, circumstances, time et cetera, and different people have different perceived benefits from the same policy. Thus it is not possible to arrive at a “best” policy – there is always a “trade-off” of interest, with bargaining being at the heart of the policymaking process. The nature of transport policy is hampered not only by the formulation process itself, but also by the basic objectives, which, for the transport sector, are normally determined by outside bodies. These objectives are sometimes conflicting.

10.4.2 Objectives of a transport policy

The first step in planning for the future of transport is to decide precisely what the system is supposed to do – in other words, what utilities it is expected to produce. Generally speaking, the public desire a system that will contribute to the economic and social development of the country. All economic activity, and national growth and prosperity, is affected by the adequacy, cost and efficiency of transport. The entire transport system should operate in perfect harmony. Nevertheless, the fact that one mode can supply certain desirable features of service which others cannot should be recognised. In fact, it is the wish of the public that such inherent advantages should be preserved.

Although adequate, cheap and efficient services are basic requirements, it is also important for a transport system to be economically and financially sound in order to meet the public's constant demands. For a transport system to be sound, all the different segments should

meet their full costs. This is a reasonable and essential prerequisite from the viewpoint of both the carriers and dependent economic interest groups. Tariffs must represent fair compensation for services rendered. They should enable each mode to find its proper field of operation, and treat each shipper and location equally in comparison with other shippers and locations. If this is not the case, the application of productive resources will be wasted.

The main elements of the objectives of a transport policy may therefore be summarised as follows:

- providing adequate transport facilities that will promote the interests of users
- creating a financially sound transport industry which will promote (or develop) the interests of transport suppliers

10.4.3 Factors to be considered in a transport policy

A great deal of thought is required before a policy can be adopted or implemented. Bowersox, Calabro and Wagenheim (1981:157–158) maintain that consideration should be given to priorities, ownership types, degrees of subsidies and implementation.

(1) *Priorities*. The matters to receive priority in the formulation of a new policy must be determined. For instance, if low costs are the main consideration, attention should be focused on the modes that have low variable costs, such as rail transport and pipelines. Unless priorities are determined, any policy to coordinate a network will fail.

(2) *Ownership*. A policy must provide for a certain type of ownership. If private ownership is preferred, competition should be encouraged and control relaxed.

(3) *Degree of subsidisation*. There are inequities in the subsidies given to different modes. They must be eliminated so that all modes receive the same treatment (which is an aim of the road freight transport policy). If all modes share equally in the total cost of operations, a good competitive environment will be created. If subsidies are more equitable, a fairer policy can be developed and applied.

(4) *Implementation*. The administrative format for the efficient implementation of a transport policy must be determined. A coordinated effort is required. Government should consider two major issues when implementing a transport policy. First, it must decide what degree of *regulatory control* it will exercise over each mode. If control is to be strict, the second issue concerns the *method of control*. To ensure coordination in the transport sector, control should be in the hands of one body.

When designing a policy, the responsible parties should remember that the policy must provide guidelines for the achievement of the objectives. Without implementation possibilities, policies are totally useless.

10.4.4 Elements of a sound transport policy

When a transport policy is designed, there are certain factors that will ensure a sound and effective policy. The following elements are important:

(1) *Regulation*. Prerequisites for a sound transport system include a comprehensive and scientific regulatory policy as well as adequate and efficient regulatory organisation and machinery to administer it. Legislators and administrators should consider only national and public interests, not the interests of pressure groups. The main aspects of a regulatory policy should be more thorough control over the whole field of transport, impartial treatment of all modes and uniform control.

It may happen that carriers do not receive equal treatment. This may be because equal treatment is unnecessary from a regulatory viewpoint or because it is impossible from an administrative point of view to afford everyone the same treatment. Impartial treatment of modes means that the law should recognise the economic characteristics peculiar to each in a manner that will permit and encourage every firm to operate in the field best suited to its abilities so that the public can derive optimum benefit.

Uniform control of all modes seems to be a prerequisite for an integrated national transport system. There may be valid arguments for separate regulation of a mode in its infancy, but once it has reached maturity, separate regulation becomes an obstacle in smooth functioning. If control is centralised in one body, it must be well organised and staffed with efficient personnel to ensure proper control over transport.

(2) *Consolidation*. Consolidation, with the resulting coordination within a mode, is a major means by which a mode can strengthen itself financially and competitively to provide better service. Responsibility for the failure to consolidate rests squarely on the shoulders of selfish interest groups. Consolidation may lead to enormous savings.

(3) *Coordination*. Coordination, that is cooperation between firms to provide a joint service, should be encouraged and enforced when such a practice will result in lower tariffs or better service. Coordination does not imply that a certain commodity or type of traffic will be transported by the same mode for all shippers under different conditions. Some shippers may be interested only in the tariff, whereas others who ship the same commodity or type of traffic may desire special services and are willing to pay for them. Where coordination is achieved through integrated transport firms, management determines by means of a cost study which method is cheaper or most efficient for any particular situation. Regulatory supervision of these activities is obviously necessary.

(4) *Finance*. No transport system can provide an economical and efficient service unless it has a sound financial basis. Financial activities must be organised and regulated to create a sound credit position. The transport industry should continue to expand its investments, and the funds should be supplied from the capital market on the basis of the competitive ability of the industry to attract capital. One means of improving credit is to reduce expenditure through consolidation, coordination and other means. There should also be sound accounting procedures, uniform wherever possible, to provide an accurate and economically sound picture of all expenditure.

(5) *Tariffs*. Tariffs and tariff relationships are at the heart of the transport problem. In the interests of national prosperity it is imperative to adapt tariff structures according to marginal costing principles. Tariffs and tariff structures cannot be changed rapidly, because of possible disruptive economic consequences. The revision of tariffs in a way that will remove discrimination and inconsistency nevertheless has several advantages: it brings tariffs more in line with the cost of service and enables each type of transport to operate where it is economically the best type of transport; it furnishes a fair return for the amount of productive resources necessary to provide an efficient transport service; and it helps to build a well-balanced economy.

(6) *Promotional policy.* This factor significantly affects the solution of most of the problems indicated above. Promotion is an important government task and must be ongoing. The proper development of the country requires new services, heavy investments and transport research – all of which only the government can assure. Government's promotion of transport should, however, be conducted more systematically than in the past, and the problems arising from promoting the different modes should be synchronised. Promotion and regulatory problems should be harmonised to ensure a sound transport system.

Consideration of all these elements in the design of a transport policy will result in a sound policy which should promote the interests of modes and users of transport.

10.4.5 Policy instruments

The following is a list of policy instruments most often used by government authorities to achieve their goals:

(1) *Taxes and subsidies*. The government may use its fiscal powers either to increase or decrease the costs of various forms of transport or service over different routes – or indeed the cost of transport in general. It may also influence the costs of transport inputs such as vehicles, fuel and tyres.

(2) *Direct provisions*. Local and central government are direct suppliers, via municipal and nationalised undertakings, of a wide range of transport services. They are also responsible for supplying a substantial amount of transport infrastructure, notably roads, rail tracks, airports and supplementary services, such as the police.

(3) *Laws and regulations*. Government (and to a lesser extent, local authorities) may regulate the transport sector legally, and an extensive body of law which in effect controls and directs the activities of both transport suppliers and users, has developed.

(4) *Competition policy and consumer protection legislation*. It is useful to distinguish between general industrial legislation, governing such things as restrictive practices and mergers, and consumer protection legislation, covering such things as advertising, which embraces all forms of activity in the economy and not just transport. They obviously also apply to transport.

(5) *Licensing*. The government may regulate either the quality or quantity of transport provision by its ability to grant various forms of licences to operators, vehicles or services.

In the case of road freight transport, in recent years South Africa has moved away from a system of quantitative regulation by means of road transport permits by introducing a qualitative system in the form of the Road Transport Quality System (RTQS).

(6) *The purchase of transport services*. Various nontransport activities of government require the use of transport services. Given its position as a large consumer, the government may be able to counteract the power of transport suppliers to an extent.

(7) *Moral persuasion*. In many instances this is a weak form, usually of an educational nature or offering advice on matters such as safety (eg advertising the advantages of wearing seat belts). However, it may be stronger when the alternative to accepting advice is for government to use its powers over others (eg refusing a licence or withdrawing a subsidy).

(8) *Research and development*. The government may influence the long-term development of transport through its own research activities. These are conducted in part by its own agents (eg Transportek at the CSIR) and in part through the funding of outside research. The Department of Transport can also appoint private consultants or universities to do research for it.

(9) *Provision of information*. The government, through various agencies, offers technical advice to transport users and provides general information to improve decisionmaking in transport. Many of these services are specific to transport (eg weather services for shipping) while others assist the transport sector less directly (eg information on trading arrangements overseas).

(10) *Policies relating to inputs*. Transport is a major user of energy, especially oil, and also utilises a wide range of other raw materials and intermediate products. Government policy on energy and input in these sectors can therefore have an important indirect bearing on transport.

10.4.6 Levels of a transport policy

It is obvious from the foregoing discussion that a transport policy is a complex matter in which policy decisions have to be made at several levels.

(1) *Policy at the highest level*. At this level, guidelines are formulated in terms of national objectives. Such guidelines are obviously described in general terms and aimed at general objectives. They should make provision for individual policy decisions at a lower level, but this will depend on fundamental principles. The role of the various levels of government should be indicated, and the necessary administrative measures and machinery for the implementation of the policy as a whole will have to be created.

(2) *Policy at provincial level*. South Africa is demarcated into nine provincial areas. A policy at this level is concerned with the determination of policy objectives on a provincial basis. The policy will be based on the achievement of specific goals for provincial development and will obviously be in line with general urban, regional and national plans. Policy at this level will make provision for the coordination of supervision over all matters relating to provincial transport.

(3) *Policy at regional level*. At this level (the various regional councils in a province), policy is concerned with the determination of policy objectives on a regional basis. Here policy will be based on the achievement of specific goals for regional development, and will be coordinated with general urban, provincial and national plans.

(4) *Policy at local level*. Policy at this level is concerned with the establishment of objectives and goals for the urban area to implement the overall policy. These goals will not be described in general terms since they are specific. They include the description of alternative methods/plans for the achievement of the goals. An important subsection of policy at this level is the traffic policy which is based on the system for traffic flow.

10.5 The South African Transport Policy

(This section is based on the *White paper on national transport policy [September 1996]*.)

10.5.1 Introduction

The most recent effort to formulate a transport policy for South Africa is contained in the *White paper on national transport policy (1996)* and *Moving South Africa: the action agenda (a 20-year strategic framework for transport in South Africa) (South Africa 1999)* (also see study unit 6). The *White paper* is the result of meetings and workshops attended by individuals and representatives of a large number of organisations, a steering committee and working groups – hence a broad public policymaking process was involved.

The formulation of a revised transport policy for South Africa was necessary to keep in line with the changing environment and the national policy. According to the *White paper on national transport policy (South Africa 1996:3)*, the *vision* for South African transport is a system that will:

> *Provide safe, reliable, effective, efficient, and fully integrated transport operations and infrastructures which will best meet the needs of freight and passenger customs at improving levels of service and cost in a fashion which supports government strategies for economic and social development whilst being environmentally and economically sustainable*.

10.5.2 Goals and objectives of the South African Transport Policy

The following are the broad goals of the South African Transport Policy (White paper on national transport policy [South Africa 1996:3–6]):

(1) *To support the goals of the Reconstruction and Development Programme (RDP) for meeting basic needs, growing the economy, developing human resources, and democratising decisionmaking*.

The role of transport in the RDP relates to the need to supply services to give people access to schools, shops and health care and to transport farming products in rural areas. The scarce resources needed for these services should be mobilised in the best interests of society. These services should also be affordable.

The Department of Transport will promote small, medium, and micro-enterprises delivering these services and will encourage public participation in the decisionmaking process on important transport issues.

(2) *To enable customers requiring transport for people or goods to access the transport system in ways which best satisfy their chosen criteria*.

Key customer groups and special groups, including the poor and disabled, need to be identified to determine their individual transport needs. The key customer groups should consist, first, of the users of passenger transport services for the purpose of commuting, education, business, tourism and private purposes, and secondly, the people using freight transport, all within the urban, rural, region and international environment.

Customer requirements such as mobility, maximum speed and choice of modes should be a top priority. A flexible transport system and transport process are therefore necessary.

(3) *To improve the safety, security, reliability, quality and speed of transporting goods and people*.

The quality of service in respect of speed, safety, suitability, comprehensiveness, reliability, frequency, regularity, liability, comfort and accessibility cost should be of high standard.

(4) *To improve South African's competitiveness and that of its transport infrastructure and operations through greater effectiveness and efficiency to better meet the needs of different customer groups, both locally and globally*.

Economic growth requires, *inter alia*, that a country should be competitive in world trade. Transport cost as an element of the production process therefore plays a crucial role in international and regional competitiveness and should be strictly controlled and monitored because the cost of the final product will decrease if transport costs decrease. However, care should be taken not to reduce transport cost by lowering the quality of the service (ie the frequency, reliability, regularity, accessibility of service etc), because the level of the quality of service, which is the actual requirement set by the transport user, will directly influence the demand for the specific service.

The price of diesel fuel will, however, directly influence the costs of public passenger and road freight transport. The price of diesel fuel in relation to petrol fuel should therefore be considered in terms of world practices.

Another area of importance in competition is the provision of an infrastructure. The presence of monopolies, policies to regulate them and the level of competition should be identified and evaluated.

(5) *To invest in infrastructure or transport systems in ways which satisfy social, economic, or strategic investment criteria*.

Total transport costs include the costs of transport operation, externalities and infrastructure. The latter are long term, and usually represent only a small portion of total transport cost. However, a sound financial base should be created for the maintenance of roads and the upgrading of transport infrastructure. Furthermore infrastructure should be built at the right places, thus serving the needs of the society and economy effectively.

There are usually conflicting priorities between the need for infrastructure for the society and for the development of the economy. Investment decisions should be based on analysing the return on such an investment, the ultimate aim being to optimise the use of scarce resources such as human, financial and material resources.

The above objectives should be achieved in a manner which is economically and environmentally sustainable, and which minimises possible negative side effects. A cost/benefit analysis should be undertaken for each proposed project and be quantified in both economic and sustainability terms. The effective use of scarce energy resources should be manipulated by differentiating between the prices of such resources. An example of the latter is the difference between the prices of leaded, unleaded and diesel fuel.

The key strategies to attain these goals will be as follows:

- *Integration*. Integration refers to modal, spatial, institutional and planning integration. When decisions are made, the integration of the stakeholders, such as the appropriate government departments, private sector and consumers, should be promoted. This will ensure minimum regulation by government and that the private sector will operate in a competitive environment.
- *Intermodalism*. Intermodalism implies not only coordination of and cooperation between different modes of transport in respect of operations, but also in terms of sharing information when an infrastructure is developed. Intermodalism as a strategy to attain the goals and objectives of the transport policy reduces the duplication of services, decreases ruinous competition, minimises total costs and maximises social and economic return on investment. However, the specific role of each mode of transport should be acknowledged in a hierarchical transport service, and ultimately the effective utilisation of the capacity available.

10.6 Transport regulation

10.6.1 Introduction

Transport regulation may be defined in terms of (1) economic regulation and (2) safety regulation.

Economic regulation relates to restrictions on the participation of carriers in the market and the tariffs charged. The aim of safety regulation, on the other hand, is to promote a safe industry for the participants and to protect the infrastructure which belongs to the state.

10.6.2 Safety regulation

Safety regulation is necessary to protect the lives of carriers' employees and passengers, users' property and the lives and property of others who may be harmed by the activities of carriers. The application of this type of regulation is sometimes more extensive than in the case of economic regulation. *Governments may lay down and apply certain safety measures even though exemption from economic regulation has been granted*.

Safety regulation applies mainly to the condition of equipment, operators' qualifications and operating procedures. Although control of these factors is not an economic matter, it does have economic consequences. Many of these measures lead to cost increases since carriers must meet certain standards. These requirements are important and should constantly be taken into account in the fleet planning process. However, since the efficiency of services is also improved, the final cost of services will be lower (Tally 1983:53–54).

The following are examples of elements of safety regulation in the road freight industry:

(1) *Regulating the condition of equipment*. Rules and regulations controlling the physical condition of equipment are found in every phase and type of transport since breakdowns may cause serious accidents. A carrier's equipment must meet certain standards and is subject to regular inspection. Governments usually determine the length, width, height and mass of vehicles to protect other road users and the road itself. Other factors that are subject to regulation include brakes, lights and the steering mechanism. They are regularly tested by road patrols to ensure that vehicles are in a roadworthy condition.

(2) *Regulating operators' qualifications*. Such regulations are common. A case in point is the requirement that all drivers of vehicles must have a valid driving licence. The qualifications that operators require are generally the highest in the public transport industry.

(3) *Regulating operating procedures*. There is a wide variety of safety rules for the operation of vehicles, for instance, vehicle speeds and the traffic signs to control the flow of traffic. An extensive system of operating procedures is prescribed for each mode of transport by various bodies and levels of government to achieve the objectives of transport safety and reliability.

It is clear from the above discussion that safety regulation has two main objectives, namely safety and reliability. These goals coincide with the general objectives of a regulatory system to protect the public and promote the best possible transport system.

10.7 Conclusion

It is apparent that the integrated nature of transport in the economic, social and geographic sphere of any country necessitates some form of transport policy and its implementation. The first step is to formulate such a policy, with due regard for all the relevant aspects. Secondly, the necessary legislation needs to be developed and passed through parliament to implement the policy, and thirdly, the actual implementation should take place by means of regulation.

10.8 Self-evaluation questions

(1) Is government intervention in road freight transport justified? Explain your answer.

(2) Explain the reasons for having a transport policy.

(3) Briefly discuss the development of a transport policy.

(4) Explain the various levels of the transport policy.

(5) Discuss the South African transportation policy.

(6) Discuss the elements of safety regulation.

BIBLIOGRAPHY

Anderson, A & Stromquist, U. 1988. The emerging C-society, in *Transportation of the future*, edited by DF Batten & R Thords. Berlin: Springer.

Banister, D. 1994. *Transport planning in the UK, USA and Europe*. London: Chapman & Hall.

Bell, F & Marsden B. 1991. Track maintenance to match the traffic task. Paper presented at the Eighth International International Rail Track Conference, Rail Track Association Australia, Sydney.

Bolan, RS. 1991. Planning and institutional design. *Planning Theory* 5/6, Summer/Winter:1–8.

Bowersox, DJ, Calabro, PJ & Wagenheim, GD. 1981. *Introduction to transportation*. New York: McGraw-Hill.

Bruzelius, N, Jensen, A & Sjostedt, L. 1994. *Swedish railway policy: a critical study*. Gothenburg: Chalmers University of Technology, Department of Transportation and Logistics.

Bureau of Industry Economics. 1993. *International performance indicators: rail freight update*. Research Report 52. Canberra: AGPS.

Button, KJ. 1993. *Transport economics*. 2nd edition. England: Edward Elgar.

Button, KJ & Hensher. DA. 2001. *Handbook of transport systems and traffic control*. Oxford: Elsevier.

Button, KJ, Tampere, CMJ, Viti, F & Immers, LH. 2010. *New developments in transport planning: advances in dynamic traffic assignment*. Cheltenham: Edward Elgar.

Chalmers, R. 1999. Spoornet. *Transport World Africa* 1(1).

De Brucker, K, De Winne, N, Peeters, C, Verbeke, A & Winkelmans, W. 1995. The economic evaluation of public investments in transport infrastructure: the use of multicriterion analysis. *International Journal of Transport Economics* XXII(3), October.

De Neufville, R. 1976. *Airport system planning*. Cambridge, MA: MIT Press.

Dodgson, J. 1995. Separating railway infrastructure and operations: the British experience. Paper presented at the Fourth International Conference on Competition and Ownership in Land Passenger Transport, Rotorua, New Zealand, July.

Ferreira, L. 1997. Rail track infrastructure ownership: investment and operational issues. *Transportation* 24(2).

Flere, WA. 1967. *Port economics*. London: Ward & Foxlow.

Fogel, RW. 1964. *Railroads and American economic growth: essays in econometric history*. Baltimore: Johns Hopkins University Press.

Fokkema, T & Nijkamp, P. 1994. The changing role of governments: the end of planning history? *International Journal of Transport Economics* XXI(2):127–145.

Freeman, PNW. 1981. The recovery of costs from road users in South Africa. DCom thesis, University of South Africa, Pretoria.

Fromm, G. 1965. Introduction: an approach to investment decisions, in *Transport investment and economic development*, edited by G Fromm. Washington: Brookings Institution.

Georgi, H. 1973. *Cost-benefit analysis and public investment in transport: a survey*. London: Butterworths.

Harvey, M. 1995. Assessing the adequacy of national transport infrastructure: a methodology. *Road and Transport Research* 4(1).

Hagaman, BR. 1989. Optimisation of 1067 mm gauge railway track under static and dynamic loading. MCom dissertation, Queensland University of Technology, Brisbane.

Hilmer, FG, Rayner, MR & Taperall, G. 1993. *National competition policy: report by the Independent Committee of Inquiry*. Canberra: AGPS.

Hirscham, AO. 1958. *The strategy of economic development*. New Haven, CT: Yale University Press.

Hogendorn, JS & Brown, WB. 1979. *The new international economics*. London: Addison-Wesley.

Homburger, SH & Kell, JH. 1984. *Fundamentals of traffic engineering*. 11th edition. Berkeley, CA: Institute of Transportation Studies, University of California.

Hope, R. 1992. Marpas relates track costs to traffic. *Railway Gazette International*.

Jansson, JO & Cardebring, P. 1989. Swedish railway policy: 1979–88. *Journal of Transport Economics and Policy* 23(3):329–337.

Moody's Bond Record. 1992. *Standard and Poor's rating guide*. New York: McGraw Hill.

Muller, JD. 1985. A report on the serviceability of prestressed concrete railway sleepers. MCom dissertation, Queensland University of Technology, Brisbane.

Murray, M & Ferreira, L. 1995. Rail track maintenance modelling: current practices and future needs. *Research Digest* 5(2).

Murray, MG & Griffin, T. 1993. In-service displacements of steel and timber rail sleepers. Paper presented at the Thirteenth Australian Conference on the Mechanics of Structures and Materials, Wollongong.

Nash, CA & Preston, JM. 1994. Competition in rail transport: a new opportunity for railways. *Working Paper 397*, Institute for Transport Studies, University of Leeds.

Nilsson, 1992. Second-best problems of railway infrastructure pricing and investment. *Journal of Transport Economics and Policy* 26(3):245–259.

Owen, W. 1964. *Strategy for mobility*. Washington: Brookings Institution.

Pienaar, WJ. 1981. The road engineer and economics. *Imiesa* 6(8):13–25.

Pienaar, WJ. 1985. Vervoerekonomiese bepaling van voertuigloopkoste vir aanwending in die evaluering van pad- en verkeersingenieursprojekte. DCom-proefskrif, Universiteit van Suid-Afrika, Pretoria.

Prest, AR & Turvey, R. 1965. Cost-benefit analysis: a survey. *Economic Journal* 75.

Saaty, TL. 1980. *The analytic hierarchy process*. New York: McGraw-Hill.

Saaty, TL. 1986. Axiomatic foundation of the analytic hierarchy process. *Management Science* 32(7):841–855.

Schiller, PL, Bruun, EC & Kenworthy, JR. 2010. *An introduction to sustainable transportation: policy, planning and implementation*. London: Earthscan.

Shahia, M, Cronjé, JN, Brits, A & Barendrecht, JW. 1995. *Road transport: only study guide for TRE3049*. Pretoria: University of South Africa.

South Africa (Republic). Department of Transport. 1999. *Moving South Africa: the action agenda (a 20-year strategic framework for transport in South Africa)*. Pretoria: Department of Transport.

South Africa (Republic). Department of Transport. 1996. *White paper on national road transport policy*. Pretoria: Department of Transport.

Tally, WK. 1983. *Introduction to transport. Cincinnati*, Ohio: South Western.

Taylor, LT. 1974. *Seaports: The introduction to their place and purpose*. Glasgow: Brown, Son & Ferguson.

UK Department of Transport. 1993. *Gaining access to the railway track: the Government's proposals*. London: Department of Transport.

Wells, AT. 1986. *Airport planning and management*. Blue Ridge Summit, PA: TAB Books.

Wilson, GW. 1966. Towards a theory of transport and development, in *The impact of highway investment and development*, edited by GW Wilson, BR Bergmann, LV Hirsch & MS Klein. Washington: Brookings Institution.

Wohl, M & Martin, BV. 1967. *Traffic systems analysis for engineers and planners*. New York: McGraw-Hill.