CHAPTER 5

THE RAIL TRANSPORT ECONOMIC REGULATORY ENVIRONMENT

5.1 Introduction

The aim of this chapter is to study the way in which concessioned railway systems are regulated from an economic perspective. One may raise the question as to why there is a need for economic regulation in concessioned rail systems. Chapter 3 identified the potential challenges in concessioned railway systems and the need for economic regulation. The approach to rail economic regulation was however, not addressed. This chapter on the rail transport regulatory environment therefore, investigates how rail transport is regulated under the concession regime from an economic perspective.

At the outset, the question can be raised as to what the objectives of the economic regulation of rail should be. On a broader level, various arguments are provided for the regulation of the entire transport system and some of the arguments are of doubtful economic logic. For instance, it is argued that official policy may cover different objectives that are inconsistent with policies designed to contain the negative externalities. The pursuit of policy to contain negative externalities may, however, run against the national policy that pursues the maximisation of the economic growth (Button, 1993: 243–245). It is known that the determination of policy objectives is done through the policy formulation process; consequently this chapter maintains that the objectives of the economic regulation of rail are taken as given policy areas.

In the concession environment, the objectives of economic regulation may be broadly described as firstly, to protect the users' interests regarding the prices and the quality of the service: secondly, to ensure that the concessionaire finances the activities if he operates profitably; thirdly, to promote efficiency; fourthly, to fulfil obligations that were decided by the policy-makers before the

concession was awarded; and fifthly, to ensure that the regulatory regime is sustainable and robust. The promotion of the efficiency objective is, however, in most cases at the heart of economic regulation. As a result, three aspects of efficiency need to be always borne in mind. The relevant efficiency aspects are static productive efficiency (improvement in the performance of existing assets); allocative efficiency (marginal cost reflectivity pricing); and dynamic efficiency (introduction of new technology and investment in new capacity) (Burns & Estache, 1998: 1–2).

A number of objectives specific to the economic regulation of rail can be identified as, firstly, the existence of financial constraints with a view to minimising the state financial burden of rail; secondly, the pursuit of internal efficiency in terms of providing rail services at the lowest possible cost; thirdly, the attainment of allocative efficiency; fourthly, the achievement of dynamic efficiency; fifth, the objective of equity such as providing rail transport to the population; and sixth, the objective of optimal allocation of capacity whereby railway capacity and its coordination with other modes of transport is considered (Campos & Cantos,1999: 16–18).

This chapter begins with the main factors that need to be taken into account in designing the appropriate economic regulatory regime for the rail industry. Thereafter, the mechanisms used in concessioned rail systems are studied in relation to economic regulation. Such mechanisms include the rate of return (ROR); the price cap, specifically the Retail Price Index less X (RPI-X) (the RPI in South Africa is known as Consumer Price Index); and quality regulation and instruments of quality control in the rail industry. The section on the economic regulation of infrastructure centres on access pricing of rail tracks (bottlenecks), particularly the efficient component pricing rule. Conclusions are provided at the end.

5.2 The characteristics of rail transport

Rail transport has a number of characteristics that needs to be taken into account when coming-up with a regulation framework. These characteristics are the multi-product nature of the rail activity, the cost structure of rail, the role played by the rail infrastructure and network, the existence of asset indivisibilities in rail industry, the organisation of rail transport as a public service and the existence of externalities in the transport system as a whole.

5.2.1 The multi-product nature of rail industry

Rail transport provides for the movement of different types of freight and passenger services. In the case of rail freight, for example, there is bulk export and general freight. Rail passenger transport includes long distance and commuter rail services. In the case of passenger services, even the same train unit may be composed of different categories of passenger coaches such as the metro and metro-plus, which are known in the rail commuter services in South Africa; hence the multi-product nature of rail activity. The multi-product nature of rail activity is evident in accounting and cost allocation procedures. It is not easy to allocate total operating costs between the different services operated. As an example, the costs of running a long-distance train unit include both fixed and variable costs. The variable costs may be attributed to the running of the train unit concerned, but the allocation of fixed costs of infrastructure to the train unit is not an easy task as infrastructure may be shared with other rail services. This presents a challenge to regulation. Furthermore, another aspect that needs to be considered in the multi-product set-up of rail activity is the subadditivity of cost functions. In Section 2.5 of this study, it was mentioned that subadditivity of costs means that the cost of producing output is more efficient with one firm than with more than one firm, regardless of how output may be divided between the different firms. This has two implications for the regulatory authority. Firstly, it is necessary to decide whether it is more efficient to use two enterprises: one for infrastructure and the other for the provision of rail services, in the concession

environment. Secondly, where rail infrastructure and the provision of services are separated, is it more efficient within the monopoly context to have one enterprise providing the rail services or to have more rail operators competing in the provision of rail services? These implications are connected to the advantages and disadvantages of separating the rail infrastructure from the provision of services or the rail industry organisational arrangements (Campos & Cantos, 1999: 4).

5.2.2 The cost structure of the rail industry

The second characteristic of rail transport is its cost structure. The rail costs are usually classified into train operational costs like energy, maintenance and depreciation of the rolling stock; track and signalling costs, which include operation, maintenance and depreciation costs; the terminal and station costs; and administration costs. In the case of the train operating costs, the rolling stock costs depend on the amount and the distance covered in the provision of rail services. In addition, energy costs depend on train kilometres while driver's costs also depend on the length of the route. Track and signalling costs depend on the amount of traffic (number of trains) and administration costs fluctuate depending on the size of the rail enterprise. From a regulatory point of view, the allocation of the different costs to various outputs involves a degree of arbitrariness. Although the allocation of rail costs to output is not an easy task, a distinction between the costs that are avoidable and those that are common is necessary (Campos & Cantos, 1999: 5). The scale and scope economies in the rail industry also create problems for regulation. The most notable problem of scale and scope economies is the fact that it is not easy to allocate costs in the rail industry (Kessides & Willig, 1995: 7).

5.2.3 The role played by rail infrastructure

From its inception, the rail industry was a monolithic enterprise that provided both rail infrastructure and rail services. In recent years, however, this traditional

organisational arrangement has been challenged, as in the case of Britain (see p. 65). The contestable market theory has provided a reason for challenging the monolithic structure of rail industry as, in terms of this theory, the cost function of the rail industry is identified as being subadditive. From a regulatory point of view this implies that rail infrastructure and the provision of rail services can be dealt with separately. Infrastructure can be dealt with as a natural monopoly and the operation of rail services can in principle be regarded as a competitive activity in a concession environment (Campos & Cantos, 1999: 5–6).

5.2.4 The existence of asset indivisibilities in the rail industry

The rail industry is very capital intensive with the existence of several asset indivisibilities in the provision of rail services. For instance, in the provision of rail passenger services, such things as the rolling stock, tracks and stations are required. These have huge financial implications. The lumpiness in the provision of rail transport facilities impacts on the investment and pricing decisions. As an example, consider the situation where capacity of rail track and trains is such that additional trains can be provided without purchasing new trains and the construction of additional track. In such a situation, assuming that there is excess demand for rail services, the rail transportation costs of additional traffic may be very small. The additional rail transportation costs may, however, be very large in the absence of spare capacity both for trains and track as new trains would have to be purchased and additional track be constructed. In a situation where there is no additional capacity in terms of trains and rail track, investment decisions may be delayed because of huge sunken costs that the provision of additional capacity especially for rail tracks may entail (Campos & Cantos, 1999: 6).

5.2.5 The organisation of rail transport as a public service

The historical development of rail transport as a public or social service is a characteristic that has determined its organisational arrangement. Rail transport

is regarded as extremely energy efficient. This reason contributed to rail's rapid growth as the first public transportation system. Military considerations, industrial and economic development played an important role in the public control of rail, which occurred with or without subsidies. Consequently, rail transport operators are required to fulfil public service obligations that are in some cases in the form of the servicing of unprofitable routes, the determination of timetables by the regulators or the provision of rail services to particular destinations. There are various other reasons that contributed to the organisation of rail transport as a public service. These include, the integrative role played by rail in overcoming geographical barriers to certain sections of the population; the supportive role played by rail in the economic development of underdeveloped areas; and in some cases rail transport guarantees minimum transport services for certain segments of the population (Campos & Cantos, 1999: 7). This implies that the economic regulation of rail transport also needs to take into account the historical factors that determined the organisation of rail, as well as the social and economic development role that is being played by the rail system of a country.

5.2.6 Externalities in the entire transport system

The transport policy goal of the public service obligation of rail is sometimes supported by the idea that rail transport contributes less regarding negative externalities compared with other modes like road transport. Empirical evidence also supports the view that negative externalities caused by congestion, accidents or environmental impacts like noise, visual impacts, pollution etc, could be greatly reduced if a large part of the road traffic were to be shifted to rail transport. The negative externalities of road transport arise from the fact that it does not fully pay the social costs it generates. In the absence of congestion and pollution pricing it may be preferable to lower the rail transport fares with a view to attracting more traffic to rail transport. This principle also needs to be taken into account in defining the role the economic regulatory body has to play in the concessioning of rail transport (Campos & Cantos, 1999: 7).

5.3 Mechanisms for rail economic regulation

These mechanisms are those that are concerned with the regulation of the price of rail services and those that regulate the quality of the service.

5.3.1 Price regulation

In the rail concession environment, the contract should set out the procedure according to which the concessionaire determines prices that are approved by the regulator. In general terms, the price regulation mechanism is set taking into account the degree of monopoly power entrusted to the rail concessionaire, the extent of government non-commercial objectives and the limitations that need to be recognised such as intermodal competition. The price regulation mechanisms that are studied are the rate of return (ROR) mechanism and price caps, specifically the Retail Price Index less X (RPI-X).

5.3.1.1 Rate of return mechanism

The rate of return (ROR) in the rail industry is used in countries such as Japan, Canada and the United States of America. The principle behind the ROR is to constrain prices of rail services so that the rail operator earns a fair return on invested capital (Campos & Cantos, 1999: 26).

Under the ROR mechanism, the regulator determines the revenue requirements based on the rail operator's accounting costs. These accounting costs may include operating costs, taxes, allowances for depreciation of assets and allowed returns. The allowed return is the estimate of cost of capital and is multiplied by a rate base including the undepreciated investment. After the revenue requirement is determined, the regulator determines the tariff structure in such a way that the aggregate costs are covered. The tariff structure is revised after a

period. The mathematical representation of the ROR mechanism is: Total Cost = Variable Costs + (ROR * Rate Base) (Liston, 1993: 26).

There are three characteristics that affect the definition of the rate base. Firstly, the treatment of investment that was made before the regulatory period. The treatment of past investment should be consistent and transparent so that investment in assets should not be expropriated by opportunistic regulatory behaviour. This was identified as the hold-up problem in Chapter 3 (see paragraph 3.4.1) of this study and could lead to the investor's fear that, after making an investment, the regulator might devalue such investment. Secondly, future investment and the expected operating expenditure as well as costs should be considered in the assets base definition to reduce the possibility of excessive investment and, thirdly, in as far as current investment is concerned, the challenge lies in determining the capital value of rail assets. In the rail industry, the existing assets, like stations, rail track etc, are sunk. In addition, such assets may have been financed before the concession process. Consequently, if such assets were to be evaluated at market value, because market value is lower than the replacement costs, such valuation would yield increases and excessive profits to the rail concessionaire at the expense of the users. If, however, the current assets were to be given a zero valuation, excessive gains would go to the users in terms of lower prices set by the regulator. In such a situation, the investor (concessionaire) would be reluctant to finance future rail assets, as he would earn a lower return on his investment. An appropriate method for addressing this would be to use average procedure that considers a financial projection of future rate base or to estimate the cash flow that the rail concessionaire would earn had the regulatory regime remained unchanged (Campos & Cantos, 1999: 26).

The advantages of the ROR mechanism are that, firstly, it allows regulators to limit the prices of rail services through close monitoring of the concessionaires' profits. It is important to note that the regulator approves prices of the services and not the ROR as this mechanism implies. Secondly, the prices administered combined with restricted entry in the provision of rail services allows second

best, that is, the cross-subsidisation of one service by another. Thirdly, by a deliberate cross-subsidisation objective, the regulator can achieve non-economic or social goals and, fourthly, it provides a rate hearings forum where users have an opportunity to air their views about prices and the quality of service (Liston, 1993: 27).

The disadvantages associated with the ROR mechanism are, firstly, that the cost-plus characteristic of the ROR mechanism induces the rail operator to produce at less than minimum cost: in other words, the incentive for productive efficiency is low. Secondly, if the return on capital is higher than the cost of capital, an input bias known as the Averch and Johnson effect (A-J effect) may result. The A-J effect results in overcapitalisation of assets in that the regulated concessionaire would be tempted to enhance the rate base and therefore the profits. In other words, ROR can lead to overinvestment as return on capital is guaranteed. Campos and Cantos (1999:27) point out that overinvestment may not necessarily be adverse in less developed economies whose capital needs are in most cases not fulfilled. Thirdly, in the multi-product situation and where the concessionaire competes with others, it may be difficult to detect predatory pricing behaviour. The relevant rail concessionaire may have an incentive to cross-subsidise its competitive services by allocating a greater share of common costs to the regulated services, Fourthly, rail concessionaires can easily capture (have great influence on) the regulatory body and therefore end up earning excessive profits. This arises because the rail concessionaire can initiate a price review if it is of the opinion that losses would result. Fifthly, ROR entails high administrative costs and time-consuming hearings when prices are to be adjusted (Liston, 1993: 27–28).

It can therefore be said that the ROR mechanism has one obvious flaw and one subtle disadvantage. The obvious flaw is that the regulated concessionaire has no incentive to operate efficiently because it knows that it will be able to recover increasing costs as the price of the service provided will ultimately be increased. Where price reviews take place frequently, as is done under the ROR

mechanism, the concessionaire pays no penalty for inefficiencies. The subtle disadvantage of the ROR mechanism is that it gives the concessionaire an incentive to overinvest in capital assets. As already mentioned this is known as the A-J effect. Furthermore, the ROR mechanism may be characterised by a low-powered incentive mechanism because the concessionaire benefits little from any efficiency gains that are made (Baldwin & Cave, 1999: 224–226).

5.3.1.2 The price cap mechanism

The price cap provides an alternative to the ROR mechanism. In the UK, the price cap mechanism is used in setting prices for the franchised passenger services as well as in regulating access prices for Railtrack (Campos & Cantos, 1999: 27–28).

The UK rail industry uses the Retail Price Index less X (RPI-X) in particular. The regulator determines the X, which is a percentage that reflects the efficiency improvement to be achieved by the rail operator, for example, franchised rail operators in Britain. The RPI-X mechanism in Britain was first applied in British Telecom (BT) in 1984 and was extended to most utilities (Armstrong, Cowan & Vickers, 1994: 165). The RPI-X mechanism is represented by the formula: $P_t = (RPI_{t-1} - X)P_{t-1}$ where RPI is the Retail Price Index of the previous year (Liston, 1993: 27).

If a firm is subject to the RPI-X regulation, it has to ensure that the weighted average price increase for its various services in a particular year does not exceed the RPI-X. The price increase in the provision of services is therefore decoupled from the industry-specific cost index. This has the advantage that the regulated firm is put in a situation where it cannot manipulate the prices when it is subjected to the retail price index. This mechanism further provides the users with a clear and predictable signal about the level of price increases (Armstrong et al., 1994: 168). According to Baldwin and Cave (1999: 226–227), the regulated firm is allowed to increase its price levels by the previous year's

inflation rate, that is, the RPI. The inflation rate is then varied by a percentage, X, to reflect the cost savings (efficiency gains) that the regulator expects to be achieved by the regulated firm. As an example, assume that the previous year's rate of inflation is 8 percent and the weighted average price change, that is the X, is considered by the regulator to be five. The firm subjected to this mechanism would therefore be allowed to increase its prices by three percent. The difference of five percent from the inflation rate would have to be recovered from the cost savings that the regulator expects to be achieved by the regulated concessionaire.

The regulatory period, specifically the review of X, is the main feature that distinguishes the ROR from the RPI less X. As a general characteristic, under the ROR mechanism, the price reviews are more frequent and endogenous as either the regulator or the regulated firm can request the price reviews. In the case of RPI less X the review is relatively long and the date of the next review is fixed in advance (Armstrong et al., 1999: 172).

The main aim of the RPI-X mechanism is the achievement of dynamic efficiency by allowing the regulated firm to share its efficiency gains with the regulator and therefore the users (Campos & Cantos, 1999: 27). This means that if the regulated firm has achieved cost reductions, the regulator can transfer such gains to the users in the form of lower prices after the review. The advantages of this mechanism are firstly, the incentive to minimise the costs, as the hearings to increase prices are not held frequently. Because of this, the regulatory link between increase in costs and increase in prices is severed. Secondly, the connection between profits and rate base is severed thereby removing the input bias (A-J effect) of the ROR mechanism. Thirdly, the RPI-X mechanism costs less to administer than the ROR and fourthly, the price cap regulation of monopoly services can assist in eliminating predatory pricing in competitive services where regulated and unregulated service prices are placed in different baskets (Liston, 1993: 29).

The price cap mechanism does have some disadvantages, however. Firstly, the regulated firm is the claimant of the gains below the capped price and as a result it has an incentive to reduce costs. The reduction in costs however implies that such cost reductions can be achieved by lowering the quality of the service. Secondly, predatory pricing may persist if competitive and regulated services are subjected to the same X and if the firm has common costs. Thirdly, the informational requirements of the price cap mechanism are also far from being simple. Fourthly, the price cap mechanism does not compel regulators to publish the RORs of the regulated firm, which may entail greater risk for regulatory capture. Fourthly, the absence of rate hearings deprives users of the opportunity to express their views on price increases (Liston, 1993: 29).

One of the disadvantages of ROR mechanisms that were identified is that it can result in overinvestment (A-J effect). Under a price cap mechanism, like the RPI-X, the regulated firm can, however, underinvest and as a result allow the quality of the service to deteriorate. The question that arises, therefore, is whether the social cost of underinvestment that may result from RPI-X regulation is higher than the social cost of the overinvestment that may result from the ROR mechanism (Helm & Thompson, 1991: 231–246).

There are three observations that are made regarding the social cost of underinvestment and overinvestment. The first is that the disbenefits of underinvestment are high where demand is inelastic and that disbenefts are low where demand is elastic. Secondly, the disbenefits of overinvestment depend on the capital intensity of the production process. Thirdly, under a wide variety of demand and cost conditions the disbenefits of underinvestment are greater than those of overinvestment. The third observation does not, however, hold when the production process is highly capital intensive and demand elasticities approach unity. The comparative costs of underinvestment and overinvestment are likely to be similar only if two conditions are met. The first condition is that prices are set at their efficient level in the short term. In other words, prices need to be used to ration capacity when there is underinvestment and to ensure maximum

utilisation when there is overinvestment. The second is that the users should be indifferent regarding the future path prices will take. In practice, however, prices are regulated more closely to their long-run level and quality reductions are used to ration demand if there is underinvestment (Helm & Thompson, 1991: 243–245). This means that quality reduction is used to ration demand in practice when demand exceeds the available capacity because of the capital intensity and the existence of asset indivisibilities with regard to the rail industry.

The issue that needs to be addressed in the adoption of a price cap mechanism like the RPI-X is the determination of X. The initial setting of X is important because if it is set too high, little in terms of surplus will be transferred to the users and the social losses will be too high. In a situation where X is set too low, the firm might be driven to bankruptcy as it might be unable to achieve the required break-even. Furthermore, if X is set too low it may render the firm unattractive to investors and, as a result, its service quality might deteriorate (Liston, 1993: 30).

Several factors need to be taken into account by the regulator in determining X. These include the cost of capital, the value of existing assets (asset base), the future investment programme, expected future changes in productivity, estimates of demand growth and the effect of X on actual and potential competitors (Armstrong et al., 1994: 183). There are many ways of dealing with these factors. The cost of capital and the value of existing assets can be obtained by using financial techniques. The future investment programme depends on expected productivity gains and estimated demand growth can be obtained from demand projections (Campos & Cantos, 1999: 28). In Britain the issue that emerged was the estimation of the capital cost of rail infrastructure. Rather than using the historic cost or replacement cost method a technique known as modern equivalent asset valuation was adopted. This method was used to estimate replacement costs taking into account the latest and most cost-effective technical possibilities, economies of restructuring and spare capacity. This, however,

required detailed knowledge about future operational requirements and such knowledge, as was available was not sufficient (Preston, 1996: 8).

The price cap mechanism is often perceived to be superior to the ROR mechanism. This is because the price cap is usually implemented when the private sector participates in the provision of services previously supplied by a public enterprise monopoly. In practice, the two mechanisms converge. In a multi-product firm, the price cap mechanism requires as much knowledge about the cost function as a ROR mechanism. Furthermore, if profit monitoring is envisaged, the regulator will need to know the same cost function under price cap as under the ROR regulation (Liston, 1993: 39–40).

5.3.2 Quality regulation in the rail industry

The price mechanisms used in the rail industry revealed some shortcomings. Where the ROR mechanism is used there is a risk of overinvestment and therefore excessive quality provision (gold plating), especially in the rail industry where the quality of service is dependent on investment. Under the price cap mechanism there is a risk of undersupply of quality (Baldwin & Cave, 1999: 252).

In a perfect market situation characterised by a large number of rail service operators and well-informed users of passenger and freight rail services, quality regulation would not be necessary. The competition between the operators of rail services would drive low quality service operators out of the market leaving behind only high quality service providers. In the absence of perfect market conditions, however, the disciplinary role exerted by competitive pressure does not exist. Consequently, poor and unreliable rail services might result owing to a lack of a market mechanism to look after the quality of service. In the rail industry, there are three main dimensions that define quality: the quality of service, safety and externalities and dynamic quality or investment (Campos & Cantos, 1999: 29–39).

In a rail concessioning environment, quality is defined in the concession contract. There are three elements that need to be considered in the design of concessions with a view to their later being incorporated. The first element that needs to be considered is the service standards like punctuality of trains, reliability, waiting time at stations and so forth. The second element is the flexibility with which scheduled services may be changed or new services introduced to accommodate changes in the level of demand. In this area rail transport is always at a disadvantage because of the need to co-ordinate timetables and operations with certain technical requirements. The third is the inter-modal coordination with other modes like the road-based modes (Campos & Cantos, 1999: 31).

The quality of the service dimension incorporates rolling stock, routes and services such as response to complaints, tickets and other aspects of the customer service department. The concession agreement should further include the passenger service requirements (PSRs) determined by the regulatory body and such PSR define the quality of service standards that the rail operator would be expected to fulfil. For instance, in Britain these PSRs include specifications of frequency of trains, stations to be served, maximum journey times, first and last train, weekend services, through services and load factors or peak train capacity especially for commuter services. PSR also include the limits to train cancellation (Campos & Cantos, 1999: 34). The concession agreement should also specify the penalties that the regulatory authority can impose on rail operators should they fail to meet their quality of service obligations (Baldwin & Cave, 1999: 248).

The safety and externality dimensions of quality regulation form part of social or external regulation. This, however, differs from the quality of service regulation in the scope of regulation. Non-compliance with this regulation affects not only the users but non-users of rail services as well. The social quality regulation in the rail industry relates to the regulatory approach that should be used. This relates to the need for the external safety regulatory function to be undertaken by a body that is independent from the rail industry. In South Africa, legislation has

been approved to establish the Rail Safety Regulator, which will be concerned with the assessment and assignment of risks. The rail industry generally has a good reputation for safety although some accidents do occur. Insurance against third party liability is important here. Insurance provisions in the concession agreement may include provisions that require operators to take out insurance against third party liability and that stipulate the type, level and identity of the insurers. Such provisions would require the approval of the regulatory body that set such minimum insurance requirements. The social quality is related to externality issues such as the environment. In the rail industry, most countries include in their regulation design and specification requirements that rolling stock should comply with to reduce, for example, noise (Campos & Cantos, 1999: 36–37).

The third dimension of quality regulation in the rail industry is the dynamic quality. The complete quality regulation requires identifying who will assume responsibility for deciding on investment in terms of fleet, track renewals, track and station maintenance or future investment obligations. Where the regulator assumes this responsibility, adequate mechanisms should be in place so that projects are not stopped before they are finished. Where the rail concessionaire undertakes investment, quality control should also be in place. This may involve monitoring the financial health of the operator so as to prevent cheating incentives (Campos & Cantos, 1999: 39). In the Argentinean case study it was mentioned that freight rail concessionaires came up with an investment plan during the bidding process; while in the subsidised commuter services, the authority specified the investment plan to be followed by the concessionaires. These measures were intended to regulate dynamic quality.

5.3.2.1 Instruments of quality control in the rail industry

There are various instruments that may be used to regulate the quality of the rail industry. The first is that the concessionaire may be required to publish performance results. Secondly, a measure of quality may be included in the price

mechanism. Thirdly, customer compensation schemes could be set up to compensate affected users in situations where quality standards are not met. The compensation scheme only works well, however, if quality failures can be easily quantified. Fourthly, minimum quality standards may be specified in the concession agreement and supported by legal sanctions such as fines or by the revision of the price cap when the RPI-X mechanism is used (Armstrong et al., 1994: 180–181).

In the rail industry, concessionaires are usually required to publish their performance results after a defined period and to report this information to the regulatory body. In a situation where the ROR mechanism is used, the concessionaire is required to calculate the asset base according to the specified method or to obtain an authorisation from the regulator for certain technological improvements in order to avoid the gold-plating risk. Where the price cap mechanism is used, the prices of the services that are controlled should be properly defined to avoid quality reductions that the concessionaire could use to increase profits, even though the same price cap is maintained. The practical difficulty associated with the compensation of affected users has led many countries to adopt the minimum service standards for the rail industry. These are backed by legal sanctions that include fines and withdrawal of the right to continue operating if minimum service standards are continually not met (Campos & Cantos, 1999: 39).

The quality regulation comprises three stages. The first stage occurs before the concessionaire actually enters the market. During this stage, the aim should be to minimise the conflict that may arise between the regulatory body and the concessionaire in the future. The services should be clearly defined, as well as the performance standards in terms of which the concessionaire's performance will be measured. The first stage should also specify the investment plans and the financing rules. The second stage is the market operation. During this stage, the quality instrument chosen must be related to the monitoring of the concessionaire's performance. This is the time when the concessionaire's

obligation to reveal information is put to the test and the auditing process for verifying information provided by the concessionaire takes place. The final stage occurs after the rail transport services have been provided. During this stage, compensation or punishment can be exacted. The scheme dealing with penalties and incentives needs to be graded accordingly since severe fines and large compensation (subsidies) could alter the behaviour of concessionaires in the market (Campos & Cantos, 1999: 42).

5.4 Economic regulation of infrastructure

One of the characteristics of rail transport is said to be the role that is played by the rail infrastructure. It has already been mentioned that rail transport has always been a monolithic enterprise, providing both rail infrastructure and rail services. The development of the contestable market theory, however, challenged the traditional rail organisational structure and in some countries rail infrastructure has been separated from rail operation. As a result, the provision of rail services can be considered a competitive rail activity.

Where rail infrastructure is separated from rail services, some countries have opted to retain rail infrastructure within the public domain with the establishment of a state-owned agency to manage it, like in Sweden's Banverket. In countries like France and Germany, independent state-owned enterprises were established to manage rail track while in the UK infrastructure was privatised as was mentioned in this study. Whether the infrastructure is in public or private hands, its regulation needs to outline, firstly, the minimum investment requirement; secondly, how the access prices are to be determined; and thirdly, where rail infrastructure and the operation are separated, the general rule is that the promotion of open access (on-track competition) should be encouraged (Campos & Cantos, 1999: 43). This general rule raises the question whether it is necessary to separate rail infrastructure and operation in an environment where competition is not envisaged. This is raised against the backdrop of current transport policy in

South Africa where on-route competition is not envisaged, especially where subsidisation occurs (Department of Transport, 1996: 23).

In a situation where integrated (infrastructure and operation) concessions are opted for, as was the case in Argentina, unintended competition may still occur in some segments of the rail network because of imperfect division of the rail network among the rail concessionaires. In such an environment, an approach for determining the access price will still be necessary.

One of the vexing problems of rail infrastructure regulation is how to determine the access price. In the context of integrated rail, access has two significant attributes. The first is that access is an intermediate good or service, that is, it is used as an input in the supply of rail services. Secondly, the provider of access uses this input not only to provide its own rail services but also to provide access to its rivals in the market. If the access owner charges its competitor higher access prices than it implicitly charges itself, the access price reduces the ability of the rival to compete with the access owner in the market. If, however, the access owner charges lower access prices to its competitor, it will amount to an implicit subsidy for the competitor (Baumol & Sidak, 1994: 172–173).

5.4.1 The problem of infrastructure costs

Proper allocation of costs was mentioned as one of the characteristics of the rail industry. Some approach must therefore be found that will assist in determining the access price to rail network.

Rail infrastructure costs consist of high fixed costs, low variable (avoidable) costs, and non-avoidable or common costs (Campos & Cantos, 1999: 43). Clarification of the marginal cost is essential. The marginal cost refers to the additional costs incurred as a result of additional units of output. The price that is set at a marginal cost satisfies the requirement of economic efficiency. In the case of the rail infrastructure, however, because of the existence of diminishing

long-term total costs, the infrastructure owner will suffer a loss when the access price is set at the marginal cost level (Baumol & Sidak, 1994: 176) because the latter is less than the total costs.

The second cost concept is the avoidable costs. Assume that a rail network comprises a number of rail services. The operation of such a network has total costs that are associated with the operation of that network. The effect of a change in train services on the total cost is known as the avoidable cost of that service. The third cost concept is the common costs. Still assuming the operation of the rail network, the addition of all the avoidable costs of each service across all the services may give an amount that is less than the avoidable costs of the system. This will always be the case when there are common costs that cannot be attributed to any of the services. Common costs may comprise earthworks, track, signalling etc and should ultimately be allocated to services such as goods or passengers (Kennedy, 1997: 60).

There are practical problems associated with infrastructure cost allocation. Two elements associated with total costs are identified and these are, firstly track usage costs, which are associated with short-run effects on maintenance and renewal costs of individual trains and, secondly, traction current costs. The other components of total costs are long-run incremental costs (long-run costs imposed by train operators on the infrastructure owner) and fixed costs. In the UK the track usage costs and traction current costs are estimated to be 9 percent of the total costs while the fixed costs are estimated at 91 percent. To allocate costs to short-term track usage and traction costs on the one hand, and long-run incremental and fixed costs on the other hand, however, presents practical problems. Firstly, it is not clear what the distinction is between short-term and long-run incremental costs. This is because short-term variable costs include track maintenance and track renewal costs. Track renewal costs might just as well be regarded as long-run costs. The distinction between short and long-run costs is, however, necessary to decide whether to charge access price on the basis of short or long-run costs. Secondly, fixed cost estimates are sensitive to time

considerations as well as assumptions about the life of assets and the way the assets would be depreciated during their life (Kennedy, 1997: 60).

As mentioned earlier, the avoidable costs that occur in the rail industry are the variable costs. In a situation of vertically integrated railways, econometric studies have shown that the marginal cost lies in the range of 60-70 percent of average total costs while the marginal social cost of infrastructure is estimated below 60-70 percent in the case where infrastructure is separated from rail operation. Consequently, in the determination of access prices, the problems of cross-subsidisation, cost recovery and the possibility of setting inefficient access prices arise (Campos & Cantos, 1999: 44–45).

When determining the access price to rail tracks, it is argued that the access price should be set equal to the marginal costs. This is, however, unacceptable. It is appropriate to use the marginal cost as the access price floor but to say that access price be equalised to the marginal cost distorts the legitimate principle of using marginal cost as a basis, because where diminishing return to scale is involved, such a pricing rule will form the basis for the insolvency of the infrastructure owner (Baumol, 1983: 348).

5.4.2 Efficient component pricing rule

The efficient component pricing rule guides the choice of efficient access prices for rail track and is variously known as the imputation requirement, the principle of competitive equality and the parity principle (Baumol & Sidak, 1994: 179). The efficient component pricing rule requires the consideration of access prices and their implications for non-discrimination among the users, the promotion of efficiency and adequate revenue for the infrastructure owner (Baumol, 1983: 350).

The consideration of non-discrimination among the users refers to the nature of the equity issue underlying the access price. In a situation of integrated rail

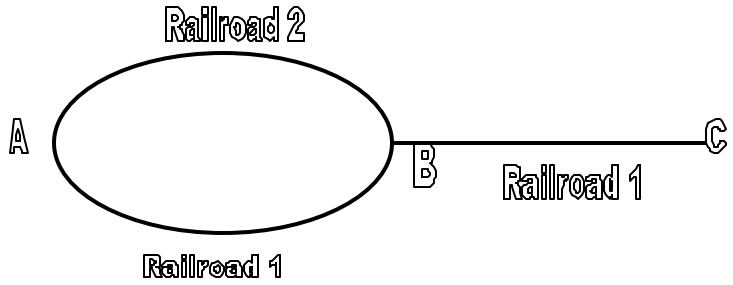
concession this means that it is necessary to analyse and compare the nature of the services supplied by the integrated concessionaire and the operator requiring access. This basically requires answering the question whether the infrastructure owner and the operator requiring access will compete in providing such a service if access is given. The promotion of efficiency refers to the setting of access prices in such a way that the more efficient rail operator ultimately serves the market. The consideration of revenue adequacy requires that the access prices be set in such a way that the infrastructure owner is compensated not only to cover the variable costs of access but also to make a contribution to the fixed costs (Baumol, 1983: 351–355).

In the UK, one principle of access charges that emerged from the infrastructure owner (Railtrack) is the obligation to behave commercially and to earn a return on capital. As a result, access charges made to franchised train operators were calculated as exceeding average total costs. Average total costs include both the operating and the capital costs. The other clear principle that emerged is that train operators are required to pay at least the avoidable costs of the infrastructure they use. The problem in the UK was, however, the allocation of various costs to various operators. It was said earlier that fixed costs were estimated to be 91 percent of total costs. There were two broad approaches for resolving this problem. The first was to charge train operators according to a tariff determined by allocating common costs according to some standard measure like train kilometres, gross tonne kilometres or some combination of these. It was, however, noted that such access prices would not be optimal as some train operators would be unwilling to pay such access prices and as a result the relevant traffic would be lost to rail and, furthermore, would lead to the remaining train operators having to pay more for access prices. The second approach was to use some form of Ramsey pricing, that is, to charge train operators according to their elasticity of demand (Dogson, 1994: 206–207). Ramsey pricing is, however, not always politically acceptable or easy to implement. Firstly, there could be distributional concerns, as equity considerations would be jeopardised. Secondly, Ramsey prices are not always

feasible. Discrimination among the users (franchised operators) in terms of their elasticity is not easy since the information about elasticity would in most cases be in the hands of the franchised operators. Such an approach raises information asymmetry constraints between the regulator and the franchised operators (Valletti & Estache, 1998: 8).

The following diagram is provided to throw more light on the access problem and the need for an efficient pricing rule. The diagram can be interpreted as applicable to a situation where the integrated infrastructure owner has to provide access to his rival in the provision of rail service.

Figure 5.1: Two alternative forms of the bottleneck case



Source: Federal Railroad Administration, US DOT (2000: 15)

In the above diagram railroad 1 is the owner of rail track from point A through B to C, while railroad 2 is the owner of a separate rail track from A to B. Point B may be thought of as a rail junction and point C as a large urban centre with many job opportunities. Point A can be interpreted as a large residential area. Rail segment BC is known as the rail bottleneck because for railroad 2 to serve point C with its rail service it will need access to segment BC from railroad 1.

Assuming that railroad 1 and 2 are competing in the provision of services, in the absence of regulation, railroad 1 may refuse to give access to railroad 2 or it may charge a very high access price for the bottleneck BC so that railroad 2 is unable to compete with it to serve point C. To come up with an access price that complies with the efficient component pricing rule that access prices be non-

discriminatory, promote efficiency and provide adequate revenue for railroad 1, a separation of the incremental costs each railroad would incur in the transportation of traffic to point C and the incremental costs on the BC segment is necessary. The incremental costs for the movement of traffic on segment AB can be referred to as competitive costs. A comparison of competitive costs between the two railroads over the AB segments would enable one to determine which railroad is more efficient. The cost of the BC segment can be referred to as the bottleneck costs. The bottleneck costs would be the same for both railroads 1 and 2 as they are borne by railroad 1. For railroad 1, the competitive costs would be equal to the avoidable costs if railroad 2 were to carry the entire traffic from A through B to C, and for railroad 2 the competitive costs would be those it would add if it were to transport all the traffic. The competitive costs therefore comprise above-the-rail operation costs for ABC traffic including wear and tear on rail track, maintenance costs and other variable costs on the AB segments caused by ABC traffic, plus any fixed costs on the AB segments that are solely caused by the movement of ABC traffic. The bottleneck costs are all incremental costs of ABC traffic on the BC segment excluding above-rail costs. The following notation can be given to the various costs: IC = average incremental competitive cost; IB = average incremental bottleneck costs, that is, cost on the BC segment excluding above-rail costs; C = average contribution to bottleneck costs, that is, surplus above incremental costs; P_f = final price to users for ABC movement; P_b = access price for the use of the rail bottleneck (BC segment). The subscription 1 and 2 are used to indicate railroad 1 or railroad 2 for the final price and costs for the use of each. As a result, the final price and the costs incurred by railroad 1 are: $P_{f1} = IC_1 + IB + C_1$ and for railroad 2 are: $P_{f2} = IC_2 + P_b + C_2$. In terms of the efficient component pricing rule, the access price to the bottleneck in the above diagram is $P_b = P_{f1} - IC$ or $P_b = IB + C_1$ (FRA: US DOT. 2000: 8-10).

The access price determined in terms of the efficient pricing rule protects railroad 1 from losing any part of its revenue that contributes to the fixed costs because railroad 2 will have to pay railroad 1 an access price equal to the average incremental bottleneck costs (IB) plus an average contribution to bottleneck costs

(C_1). In the absence of access, C_1 is paid by railroad 1 as its average contribution to the bottleneck costs. This satisfies the requirement of revenue adequacy for the rail bottleneck owner set by the efficient component pricing rule. Furthermore, the access price allows railroad 2 to undertake the movement of traffic ABC if its average incremental competitive cost (IC_2) is less than the average incremental competitive cost (IC_1) of railroad 1. Thus the access price determined in terms of the efficient pricing rule also satisfies the requirement of efficiency, as users would gain by being served by a more efficient railroad. The further requirement of non-discrimination is also satisfied in that the rail bottleneck owner railroad 1, is prevented from setting an access price that is higher than IB + C_1. Both average incremental bottlenecks cost (IB) and the average contribution to bottleneck costs (C_1) are in any case borne by railroad 1, and railroad 2 cannot be expected to pay (contribute) more than railroad 1 implicitly charges itself for the movement of ABC traffic (FRA: US DOT, 2000 : 10).

The efficient component pricing rule is therefore a necessary condition for economic efficiency in setting access charges and it also promotes the public interest. Access prices that do not follow this rule create an incentive for inefficiency, the cost of which users have to pay (Baumol & Sidak, 1994: 181).

Furthermore, assume that in Figure 5.1 railroad 2 is the less efficient operator of rail services from A to B. In other words, the average incremental competitive cost (IC_2) is greater than the average incremental cost (IC_1) of railroad 1. In such a situation, railroad 2 will lose money if it undertakes the movement of ABC traffic as it will still have to pay for access to railroad 1, which is more efficient in transporting the ABC traffic. Railroad 2 will thus be prevented from providing for ABC traffic, not because of improper pricing of access, but because of its inefficiency. Assume now that the average incremental competitive costs of both railroads are equal ($IC_1 = IC_2$). In such a situation, it would not matter which railroad undertakes the ABC traffic. In a situation where railroad 2 has equal average incremental competitive costs, it would be indifferent and has no

incentive to undertake the movement of ABC traffic (Baumol & Sidak, 1994: 185).

From a regulatory perspective, the efficient component pricing rule provides a ceiling for access price. As long as the integrated rail infrastructure owner is prevented from being a deterrent to the more efficient competitor, the requirement for productive efficiency would be met. If, however, the infrastructure owner chooses an access price that is lower than the access price set in terms of the efficient component pricing rule, such a rail operator would be unnecessarily subsidising the other railroad and therefore giving away money (FRA: US DOT, 2000: 14).

The explanation so far shows that the efficient component pricing rule is crucial in a situation where the access owner competes with the railroad seeking access. Such an approach omits the situation where the access owner does not compete with the railroad seeking access, such as when a passenger rail operator seeks access to a rail bottleneck owned by a rail freight operator. In this situation, Valletti and Estache (1998: 26) offer an example applicable to the telecommunication industry in which it is easy to find unregulated access agreements between mobile cellphone operators and fixed line phone operators. In this situation, as both parties do not compete fiercely against each other, they can benefit by access charges to terminate calls from fixed networks to mobile users. Such agreements are always agreed upon in bilateral settlements although this does not necessarily imply that such terms are in the interest of society as a whole. This can be interpreted as meaning that in the rail industry it may not be necessary to regulate access for a non-competing rail service. Such agreements in South Africa exist where, for instance, commuter rail services are provided using the infrastructure of rail freight operators and vice versa, and it may not be necessary in such cases to follow the efficient component pricing rule. Settlements for access charges in this environment can be negotiated at industry level.

A further issue that needs to be considered in access pricing of rail infrastructure is the question of inter-modal competition, that is, competition between rail and road transport. Modal choices could be distorted owing to different cost coverage ratios and the use of different cost calculations. It was mentioned earlier that negative externalities in road transport arise because road transport does not pay the full social costs it generates. To overcome this, a multi-modal integrated approach needs to be applied. The ultimate goal of rail infrastructure access and its pricing should promote efficient use of transport services while allowing rail infrastructure owners to make sufficient return (Campos & Cantos, 1999: 48–49).

5.5 Conclusion

In the concession environment, the promotion of economic efficiency is at the heart of economic regulation. The design of a rail economic regulatory regime requires recognition of the characteristics that distinguish the rail industry. Two types of price mechanism used in the rail industry, that is, the rate of return (ROR) and price capping, particularly the retail price index less the X (RPI-X), were researched. This chapter also investigated the regulation of service quality in the rail industry as well as quality control instruments. The economic regulation of rail infrastructure studied includes the determination of access prices with emphasis on the efficient component pricing rule.

The rail economic environment is characterised by a number of factors. Two characteristics can be singled out: the problem encountered in the proper allocation of costs in the rail industry and the role played by infrastructure. The second characteristic is associated with the question of whether infrastructure should be integrated with the provision of rail services or separated in the concession environment. The emphasis on these two characteristics does not in any way mean that other characteristics are unimportant.

The ROR mechanism can be characterised as a cost-plus mechanism. It provides little incentive for the concessionaire to reduce costs. Where the policy objective is to cross-subsidise services the ROR offers the advantage. It also has the advantage that it offers users an opportunity through rate hearings to express their views on price adjustments and the quality of the service that it is offered. The major disadvantage, however, is that it can lead to overinvestment. This may therefore be regarded as good in developing countries where there is lack of capital. Furthermore, the ROR can easily lead to the capture of regulatory authority.

The retail price index less X (RPI-X) mechanism has positives and negatives as well. On the positive side is the fact that it decouples prices of services from the industry specific index, thus it is not a cost-plus mechanism. As a result, the regulated concessionaire is unable to manipulate the prices to its advantage. One of the major negatives of this mechanism is that the concessionaire can underinvest in the supply of services. Since the major disadvantage of the ROR mechanism is overinvestment and that of the RPI-X mechanism is underinvestment, the question that arises is which of the two has greater social costs. Observations show that the social costs of underinvestment and overinvestment depend on the demand elasticities of the relevant services as well as capital intensity in the provision of services. The comparative social costs of underinvestment and overinvestment are, however, likely to be the same if some conditions are met.

The main dimensions of quality are the quality of service, safety and externalities as well as dynamic quality. In the rail concession environment, the quality of service is usually specified in the concession agreement. The quality of service is supported by legal sanctions if quality standards are not met and incentives if such standards are exceeded. The specification of quality standards supplements the shortcomings that are found in the price mechanisms.

In the rail concession environment the main question that arises is whether

infrastructure should be separated from the provision of rail services or integrated. The general rule is that where separation is envisaged on-track competition should be promoted. In the integrated concession environment, because of the imperfect division of rail network among concessionaires, unintended competition can occur in some segments of the rail network. An outline of how the access price will be determined will therefore be necessary. In the absence of regulation, the bottleneck owner may act strategically with a view to throwing the rival operator out of the market.

The efficient component pricing rule provides a basis for determining access prices in such a way that discrimination among rail operators is avoided, efficiency is promoted and the infrastructure (bottleneck) owner is adequately compensated for providing such access. Efficient component pricing rule is very applicable where a service provider also owns some of the rail tracks used in service provision by competing operators. Various other access pricing systems can be used for charging track usage as cost related charges (almost similar to efficient component pricing rule) and usage related charges (such as by RAILTRACK).

In the absence of competition between the rail services provided, such as freight and passenger services, it may not be necessary for the economic regulator to intervene. Access prices in that environment can be settled at the industry level through the negotiation process.