

Competition Design and Efficiency in Railways

Graduate Thesis

by Gertjan Driessen

ID number: 1256874

Supervised by

Dr. L. Schoonbeek (RuG)

Dr. M. Mulder (CPB)



RuG

Abstract

This thesis is concerned with the effects of the design of competition on productive efficiency of railways. We adopt a two stage empirical approach. In the first stage, we use Data Envelopment Analysis to measure productive efficiency. Subsequently, we utilise tobit regression to identify the effects of the design of competition. The results indicate a positive relationship between competitive tendering and productive efficiency. In addition, we find that both third party access and managerial independence from the government tend to diminish productive efficiency.

Keywords: Railways, Competition (Design), Efficiency, Data Envelopment Analysis, Tobit Regression

JEL codes: C24, C67, L11, L29

Contents

Preface	7
Summary	9
1 Introduction	11
1.1 Background	11
1.2 Hypotheses and scope of research	12
1.3 Outline	13
2 Economics of railways	15
2.1 Natural monopoly	15
2.2 Externalities	18
2.3 Asset indivisibilities	19
3 Theory	21
3.1 Introduction to the relevant concepts	21
3.2 The relation between efficiency and competition	30
3.3 Design of competition in railways	39
4 Measurement of productive efficiency	47
4.1 Introduction	47
4.2 Frontier analysis	48
4.3 Survey of the literature on railway efficiency measurement	56
4.4 Data	61
4.5 Results	65
5 Relationship between competition design and relative productive efficiency	70
5.1 Introduction	70
5.2 Estimation method	70
5.3 Data	74
5.4 Results	77
6 Conclusion	81
Bibliography	83

Preface

This graduate thesis is intended to investigate the effects of competition design on productive efficiency in railways. The motivation for the choice of this topic originated from courses in Industrial Organisation. In the spring of 2005, I decided that I would like to combine my graduate thesis with an internship to explore this interesting subject a bit more. And so it happened. In August 2006, I started with my internship at the Netherlands Bureau of Economic Policy Analysis (CPB). At the CPB, I joined a research team consisting of Machiel Mulder, Mark Lijesen, and Didier van de Velde (TU Delft). The purpose of this research is to analyse the different restructuring options for railways. My part in this research was to write a separate chapter for this research consisting of an empirical study on the relation between competition and efficiency in railways. This thesis is based on the research I conducted at the CPB. During the internship, I had the special opportunity to experience the inspiring research culture at the CPB. Moreover, in October 2005, I was given the fantastic chance to visit a scientific conference in Stockholm. At this conference, I have met many of the great railway economists in the world and could briefly experience the wonderful city of Stockholm. In return for these good experiences, I wish to thank my temporary colleagues of the CPB and above all my supervisors Machiel Mulder and Mark Lijesen for their helpful comments and suggestions in the development of this thesis. I also wish to specially thank Ali Aouragh, Jeannette Verbruggen, and Arie ten Cate of the CPB who provided excellent research assistance, which is gratefully acknowledged.

Besides these people I am grateful to Bert Schoonbeek who supervised this thesis and was always readily accessible for useful comments on the submitted drafts.

Finally, I wish to thank my parents and my girlfriend who kindly supported me during this period.

Gertjan Driessen

Den Haag, March 2006

Summary

All over the world, railways are experiencing structural reforms. Although many people think that the introduction of competition could reverse the downturn of railways, no systemic research has been done to its effects. The aim of this thesis is to empirically investigate the effects of competition in railways. More specifically, we examine the effects of two alternative designs to create competition: competition on and competition for the tracks. We adopt a two stage empirical approach, in which efficiency scores are obtained in the first step and subsequently used in the second step to identify the effects of competition design.

The main results are, first of all, that competition for the tracks tends to improve productive efficiency. Secondly, competition on the tracks could diminish productive efficiency. Additionally, we find that more managerial independence from the government may lower productive efficiency.

Possible explanations for these results are that, while competition for the tracks does not disturb the efficient operation of a railway company, competition on the tracks might. That is, competition on the tracks implies sharing of traffic, which eliminates important economies of density. These are required to cover the high fixed costs of railways. In contrast, competition for the tracks does not create this problem. Rivalry to obtain the right to operate a particular part of the network induces firms to achieve maximum productive efficiency in this type of competition. Furthermore, increased managerial independence from the government, without effective competition and adequate regulation, might give the management additional room for slack and as a result diminish incentives for productive efficiency.

Although these results are very interesting and relevant for policy purposes, readers should keep in mind that they have to be interpreted with due caution. Data used in this study are constrained to the period 1990-2001. Consequently, the effects of the introduction of competition at the end of this period might not have been fully materialised. Additionally, in the empirical analysis, we only consider the effects of competition design on productive efficiency and therefore neglect other important aspects such as allocative and dynamic efficiency.

Nonetheless, the results are obtained by employing an established empirical approach, using the most recent data available. Taken together, in our opinion, this thesis provides exciting insights to the effects of competition design in railways.

1 Introduction

1.1 Background

Since the beginning of the nineties railways in Europe have experienced a period of structural changes. The European Union initiated this process by imposing a series of railway packages. The main reason for these reforms was to revitalise the poor performing state-owned monopolistic railways. Compared to other transport modes, railways have not been able to take advantage of the growing demand for transport. Consequently, railways suffered a substantial decline in market share in the past 25 years. Policymakers seek to solve these problems by creating competitive forces in the railway industry. The objective of the reforms is clearly stated in the preamble to the first directive (91/440/EEC):

“in order to render railway transport efficient and competitive as compared with other modes of transport, Member States must guarantee that railway undertakings are afforded a status of independent operators behaving in a commercial manner and adapting to market needs”.

To achieve this objective, the directives oblige the Member States to:

- separate managerial decisions concerning operations, company restructuring and investment from the state completely;
- separate infrastructure management from transport service provision (at least create separate organisations or accounts within one company);
- grant international operators the right of access and transit to national railway systems;
- create non-discriminatory access conditions for other companies than the state railway company.

While these directives have been implemented in most of the Member States, little is known about their economic effects. In particular, there are only a few papers on the effects of the various forms to introduce competition in the railway industry. The purpose of this thesis is therefore to contribute to the knowledge in this field by investigating the effects of several designs of competition both theoretically and empirically.

The period analysed in this study (1990-2001) is characterised by many changes in the railway industry. A number of different approaches to railway restructuring have been adopted in these years. This creates an excellent opportunity to examine the effects of the reforms on the performance of railway companies.

The remainder of this introduction addresses the hypotheses as well as the scope and outline of this thesis.

1.2 Hypotheses and scope of research

The main research question of this thesis is: *How does the design of competition influence the efficiency of railway companies?* Essentially two ways to create a competitive environment can be distinguished: competition on the tracks and competition for the tracks. Both designs of competition are examined in this study. In the empirical analysis, we focus on the effects of the design of competition on productive efficiency. To investigate the effects on allocative and dynamic efficiency empirically is beyond the scope of this thesis, as each of the two types justifies a separate study on its own. Nevertheless, we do address these two important concepts of efficiency theoretically.

The supposed effects of the design of competition on productive efficiency are captured by the following hypotheses:

- *Competition for the tracks (competitive tendering) improves productive efficiency as it challenges the position of the monopolistic railway company.*
- *The relationship between competition on the tracks (open access) and productive efficiency involves a trade-off, because on the one hand competition challenges the position of the monopolistic railway company, which improves productive efficiency, while on the other it deprives railways of economies of density that are vital for productive efficiency.*

In the theoretical part of this thesis we provide the theoretical foundation of these hypotheses. Additionally, we examine whether managerial independence from the government as prescribed by the EU is beneficial for productive efficiency. This is captured by the subsequent hypothesis:

- *Managerial independence from the government improves productive efficiency since it provides the management of a railway company more freedom to pursue profit maximising activities.*

These three hypotheses are investigated in the empirical analysis of this thesis. In this analysis, we control for both structural and exogenous variables. Empirical evidence on structural measures could provide useful insights to the discussion on the future of the railway industry. For example, the Netherlands are going to evaluate the current structure in the near future. In the mean time, the First Chamber of the Dutch Parliament has asked the government to analyse the effects of the different options of railway restructuring. Furthermore, this study would contribute to the small existing literature in this field of economics.

1.3 Outline

This thesis is structured as follows. Chapter two provides a brief overview of the main economic characteristics of railways. Chapter three addresses the theory regarding competition and efficiency both generally and with respect to the designs of competition we focus on empirically. Chapter four contains the measurement of productive efficiency. Chapter five presents the results on the relationship between competition design and productive efficiency. Chapter six concludes and discusses policy implications and limitations of the results.

2 Economics of railways

This chapter briefly introduces the economic characteristics of railways (Seabright et al., 2003). Like other network industries such as telecommunication and electricity, railways are characterised by the presence of a network infrastructure. This element makes network industries different from other sectors. The prominence of high fixed sunk costs associated with the infrastructure, led many economists typically regard railways as the textbook example of a natural monopoly.¹ Yet, in recent decades, this notion has been challenged by the development of a number of new ideas to the economic analysis of this industry. In particular, the development of the theory of contestable markets shed new light on the proper concept of a natural monopoly. Baumol et al. (1982) considered a sub-additive cost function. In short this notion implies for railways that, whereas duplicating rail infrastructure is generally inefficient, the cost of operating rail transport services and rolling stock once the network has been deployed can be efficiently provided by more than one company. As consequence, infrastructure and railway services can be dealt with in different ways when considering the restructuring of the railway industry.

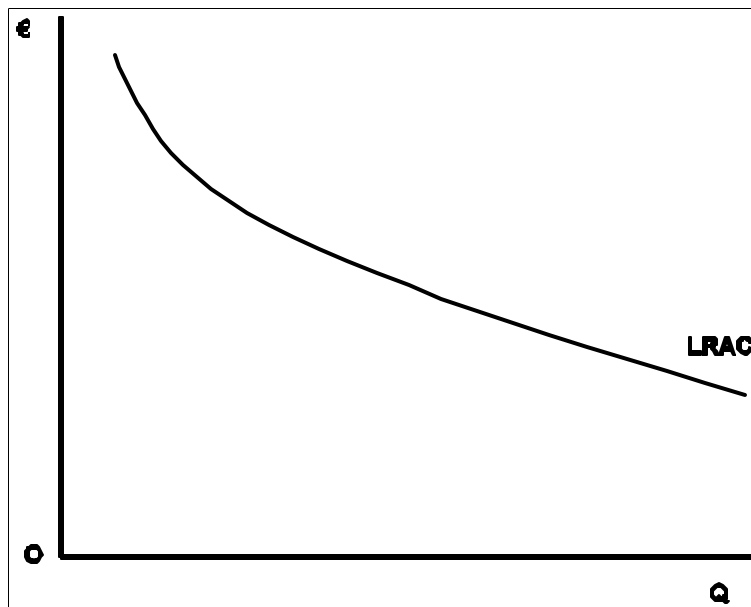
In what follows we briefly describe the most important economic features of railways. We begin by explaining the nature and consequences of the natural monopoly aspect of railways in the following section. This section also discusses the concepts of economies of scale, density and scope. After this, section 2.2 addresses externalities of rail transport. Section 2.3 analyses the existence of asset indivisibilities within the production process of railways.

2.1 Natural monopoly

Viscusi et al. (2000) define a natural monopoly as an industry where the production of a particular good or service by a single firm minimises cost. The key characteristic of such an industry is that the long-run average cost curve declines for all outputs. As a result, no matter how large market demand is, a single firm can produce it at least cost. This case is depicted in figure 2.1.

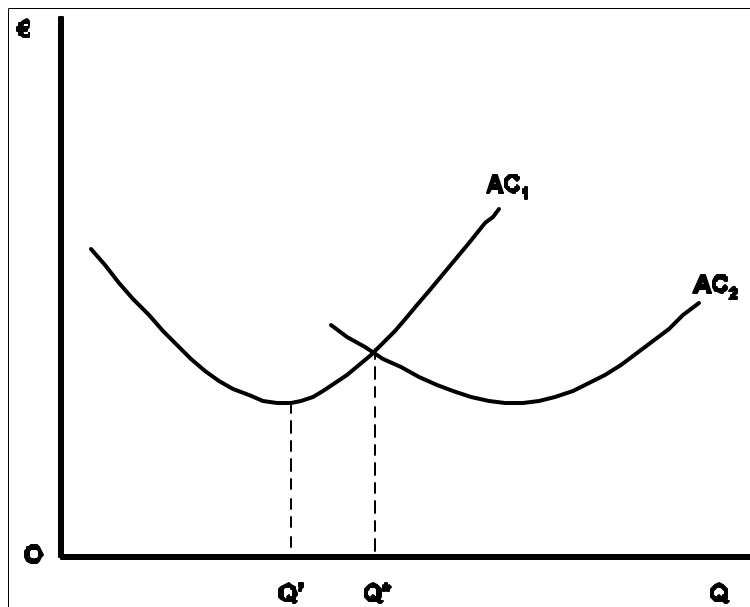
¹ Sunk costs are costs that cannot be eliminated, even by termination of production. Unlike ordinary fixed costs that can be eliminated by cessation of production, fixed sunk costs contribute to entry barriers.

Figure 2.1 Cost curve of Natural Monopolist



More realistically, a natural monopoly can be defined by using the concept of subadditivity (Campos and Cantos, 2000). As already mentioned above, this refers to whether it is cheaper to have one firm produce total industry output, or whether additional firms would yield lower cost. Figure 2.2 illustrates this notion. In this figure, the average cost functions are presented for two firms (AC_1 and AC_2). This example is more realistic, because average cost does not fall continuously as output increases. The AC_1 declines until the minimum of average costs for firm 1 is reached at the output level, Q' . Up to this level, average cost decline as output increases and economies of scale are said to exist (Kessides and Willig, 1998). It then begins to increase (indicating diseconomies of scale) until the intersection of AC_1 and AC_2 at Q^* . This output level defines the range of subadditivity. For all outputs less than Q^* , a single firm yields least-cost production. Beyond Q^* it is cheaper to have two firms producing total industry output. Notice that, despite the presence of diseconomies of scale between Q' and Q^* , it would be in society's interest to have a single firm produce in that range. So, economies of scale (declining average cost) are not necessary for a natural monopoly. Hence subadditivity is the best way to define a natural monopoly.

Figure 2.2 Minimum average cost curve for two firms



Turning now to multiple-product natural monopoly where subadditivity implies that whatever the combination of outputs desired, it is cheaper for a single firm to produce that combination. Railways are characterised by a multi-output structure: passenger and freight services. Typically, these outputs are measured by the amount of kilometres a passenger or tonne of freight is transported. Both outputs are produced by a large portion of the same inputs (common costs). In this case the interdependence among outputs is important. The term economies of scope has been proposed by Panzar and Willig (1981) to measure interdependencies in the production of multiple outputs. In short economies of scope mean that it is cheaper to produce a certain combination of outputs within a single firm than it is for separate firms to produce the required outputs. Thus, also in the multi-product case, subadditivity is necessary for natural monopoly to exist. In railways, economies of scope have been said to exist vertically, i.e. between the production of infrastructure and transport services, as well as horizontally, i.e. between the production of freight and passenger services. These interdependencies are frequently used to argue in favour of a vertically and horizontally integrated railway firm. Whether they actually exist remains an empirical matter. In the empirical analysis, we control for the structure of the industry to account for economies of integration.

The main reason for designating railways as natural monopoly is the high fixed sunk costs associated with the track network. Current estimates suggest that one kilometre of railway track

is costing between € 6 million and € 10 million depending on topographical conditions (Di Pietrantonio and Pelkmans, 2004). Total cost of infrastructure amount to approximately 50 per cent of the railway industry. As a result, duplicating a network or building alternative routes is mostly uneconomic. Only in North America, parallel routes can be found that connect the same origin-destination combination. In this special case, large volumes of traffic seem to justify duplication. So, generally it is justified to consider the infrastructure in railways as an essential facility. That is, track networks are necessary for operating railway services and duplication is not a reasonable economic option.

Due to high fixed costs of infrastructure and operation, railway services tend to be subject to economies of density (Seabright et al., 2003). That is, when holding the length of the network constant, unit costs of railway services decline as output increases. Note that the size of the network has little to do with it, because it is traffic density (i.e., market demand) that is the source of economies of density. For this reason it can be cost-minimising for a single firm to serve a network. More specifically, very large volumes of traffic are required to recoup the fixed costs, which represent a large portion of total costs (Kessides and Willig, 1998). This implies a high minimum efficient scale of operation relative to market demand. Hence the case for natural monopoly to avoid inefficient duplication of services on track.

2.2 Externalities

Externalities are another important economic characteristic of railways.² Positive network externalities arise when the value of a railway network increases with the length of the network. For instance, in a well-developed network, extending the system to more locations within the same area causes relatively low incremental costs due to the small distances which have to be covered. The railway system features the hubs-and-spokes architecture that is common to other transport networks.

Prominent negative externalities of railway transport are congestion, accident costs, and the impact on the environment (e.g., noise and pollution). Studies suggest, however, that these negative externalities of railway transport are much lower than those of other modes of

² According to Seabright et al. (2003) an externality can be defined as “an effect by which an agent is affected by another agent, the effect being not channelled through the market.”

transport (Seabright et al., 2003). As a result, policymakers often argue in favour of transferring traffic from other modalities (mostly road) to the railway industry in order to obtain an overall improved inter-modal balance.

2.3 Asset indivisibilities

The last prominent economic characteristic of railways we consider is the existence of asset indivisibilities within the production process of railways. That is, capital units (e.g., rolling stock, stations) can only be expanded in discrete or indivisible increments, whereas demand may fluctuate in much smaller units (Campos and Cantos, 2000). This can easily result in imbalances (either excess or shortage) in capacity, since supply is much less responsive than demand. This lumpiness may give rise to problems. For instance, the costs of an additional unit of traffic may be insignificant when there is excess capacity, but may be substantial when shortages exist. Consequently, attributing costs and deciding on investment and prices can be difficult.

3 Theory

This chapter examines the literature on the relationship between competition and efficiency. Both general and railway specific insights are discussed. Before we analyse this relationship in the first place, all the relevant theoretical concepts are introduced in section 3.1. After this, section 3.2 considers how competition influences the various types of efficiency. Three types of efficiency are dealt with, notably, allocative, productive, and dynamic. Finally, section 3.3 discusses the relative merits of the different designs of competition in railways. The insights derived in this chapter form the theoretical foundation to the hypotheses postulated in the introduction. These are tested in the empirical analysis of this study.

3.1 Introduction to the relevant concepts

In this section, we address the concepts of competition and efficiency. Particularly, we explain what economists imply when they discuss these popular, but rather vague notions. We start with the concept of competition. Thereupon, we assess the different types of efficiency.

3.1.1 Competition

In economics, competition is an important concept. However, neither a coherent definition, nor a robust measurement of competition exists (Boone, 2000). In this study we adopt the definition described by Stigler (1987). He defines competition as: *“a rivalry between individuals (or groups or nations) and it arises whenever two or more parties strive for something that all cannot obtain”*. This is a very broad definition which encompasses many types of rivalry (e.g. market trading, racing, auctions etc.). Note that this definition refers to behavioural aspects of competition as opposed to other concepts which refer to states or situations. Furthermore, it abstracts from the welfare effects of competition. So the desirability of competition is not identified by this definition and the definition is used strictly in a positive sense to avoid potential confusion.

While the proposed definition is useful to understand the idea of competition, it is unhelpful in the parameterisation of the concept. For this we need to know what ‘more competition’ means. A number of parameterisations are commonly used. In general, two ways in which competition can be intensified can be distinguished. On the one hand, competition is measured as an increase in the number of rivals in the industry. On the other hand, a more aggressive interaction between existing firms is related to more intense competition. A good example of more aggressive behaviour is a price war between competitors. Unfortunately, most of these measures are not monotone in competition. For instance, it turns out that it is not always the

case that a rise in competition reduces price cost margins, industry wide profits or concentration (see Boone, 2000). Luckily, a coherent measure that is monotone in different parameterisations of competition exists. Boone (2000) demonstrates that the profits of an efficient firm relative to the profits of a less efficient firm are always increased due to intensified competition. Furthermore, increased competition reduces the profits of the least efficient firm active in the market. Thus examining relative profits turns out to be the most robust parameterization of competition. The focus of this study is on the effects of the designs of competition. That is, we do not consider the extent of competition directly. Instead, we parameterise the form of competition.

Standard microeconomics typically focuses on the extreme cases of monopoly and perfect competition (Cabral, 2000). In contrast, Industrial Organisation is primarily concerned with the intermediate case (i.e. between one and many firms) of competition. This is also known as oligopoly theory. Both branches provide insights into the mechanism of competition. The focus is on the structure of the industry. In contrast to the concept above, this view refers to a certain state of a market. Static models give an idea of where the dynamic process of competition will converge in equilibrium. Although this view is rather narrow (polar cases from a static perspective), it still gives a good understanding of the basics of competition.

The first extreme case we discuss is the model of monopoly. The characteristics of this model are as follows. In the first place, one firm is active, which sells only one good. This monopolist chooses his price to maximise profits. Secondly, the market power of a monopolist, its ability to set prices above marginal costs, depends on the elasticity of demand.³ More specifically, the higher the elasticity of demand the lower the monopolist's market power (i.e. the lower the relative mark-up it earns). Furthermore, in equilibrium, monopolists with a higher marginal cost set higher prices than monopolists with lower marginal costs. While the situation of the monopoly model is rarely seen in the real world, the model provides a good approximation to industries that are close to monopolies. Moreover, most utility markets, for instance railways, are typically characterised by one large dominant firm.

³ The elasticity of demand depends on the characteristics of the good and whether suitable substitutes exist.

The second case we analyse concerns perfect competition. This model is based on number of assumptions. First of all, each firm is so small that it does not have a significant impact on other firms. Second, it is assumed that the products supplied by the many firms are homogenous. Third, there is perfect information. Fourth, all firms have access to the same production technologies. Finally, any firm may costlessly enter or exit the market as it wishes. As in the monopoly situation, firms have the objective to maximise profits. However, due to the absence of market power, all firms set their price equal to marginal costs and earn zero profits. They are so-called price takers. The model of perfect competition provides an approximation to the behaviour of industries close to the market of perfect competition.

In reality, the extreme cases of monopoly and perfect competition are seldom witnessed. In other words, the real world is often an intermediate case. As described above, Industrial Organisation deals with these cases. When there are a few competitors in a market, the market is designated as an oligopoly. The most important characteristic of these markets is the strategic interdependence between competitors. That is, each firm has to worry about its rivals' (re)actions. This type of decision making is modelled by economists as a game. The payoff for the firm in this model depends on its own actions as well as on the actions of other firms. This creates possibilities for strategic behaviour. The type of equilibrium we consider below is a Nash equilibrium. In this equilibrium no player can unilaterally change its strategy in a way that improves its payoff. We now briefly discuss the two basic models of oligopoly competition: the Bertrand model and the Cournot model.

Bertrand competition refers to the case in which firms independently and simultaneously choose the price at which they want to sell their product. Furthermore, in the basic model firms are identical and have no capacity constraints. Consumers address the firm which sets the lowest price. Accordingly, the firms will end up selling their product at marginal cost, and get zero profits. This result follows from the one-shot game the firms play. In this game the firms undercut each other until prices are equal to marginal cost. Strikingly, competition is just as tough as in the perfect competition case, despite the fact that only two firms are required instead of many small ones. This result is also known as the Bertrand paradox (Tirole, 1988).

When firms choose quantities, rather than prices, the one-shot game is known as Cournot competition. In this game, a firm chooses its output level given the output level of its rivals.

Consequently, firms cannot capture the whole market by simply expanding their outputs, since rival's output is taken as given. Due to this mechanism the incentive to compete aggressively is significantly weaker than in Bertrand competition. As a result, in equilibrium, prices are above marginal costs and positive profits are obtained. In contrast to the Bertrand case, the outcome in Cournot equilibrium varies with the number of firms that exist in the industry. This ranges from the monopoly case (with monopoly profits), where only one firm exists, to the perfect competition outcome (with zero profits), when the number of firms gets infinitely large. Therefore the number of firms matters for the equilibrium outcome when quantity competition exists.

Summarising, competition is a concept with various interpretations. It can be described in a static and dynamic way. Nonetheless, ultimately all notions of competition have their roots in the broad concept of competition as rivalry.

3.1.2 Efficiency

“Just as justice is to law and health is to medicine, efficiency is a central concept in economics”

Luís M.B. Gerbal (Gabral, 2000)

As made clear by the citation above, efficiency is a crucial concept in economics, in particular, microeconomics. There are several meanings of the word. However, they generally relate to how well an economy allocates scarce resources to meet the needs and wants of consumers. In this subsection, we introduce three notions of efficiency commonly used in economics (Motta, 2004).

Welfare economics

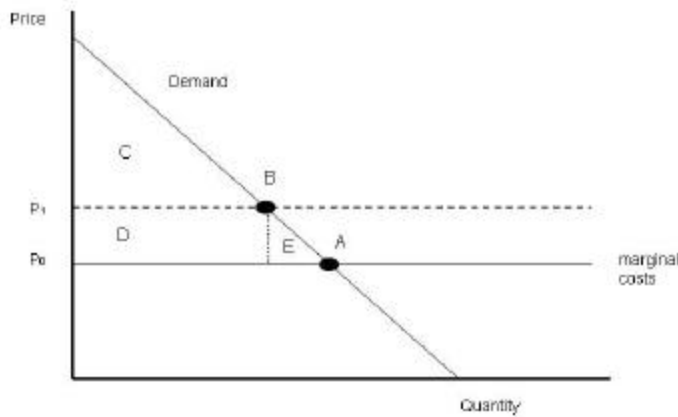
Before we introduce these concepts, it is imperative to explain the main analytical tool of economic welfare analysis. It is a measure which aggregates the welfare, so-called surplus, of the different groups in the economy. Total surplus (total welfare) is the sum of consumer surplus and producer surplus. Consumer surplus is defined as the difference between the consumer's valuation (willingness to pay) for a certain good and the price which has to be paid for it. The gains from trade for producers are called producer surplus. It is measured by the

profit the producer makes by selling the good in question. For both surpluses, individual surpluses (each firm or consumer) aggregate. In the standard case with a homogenous product and constant marginal costs, it automatically follows that, *ceteris paribus*, as the price increases, consumer surplus falls and producer surplus increases. Moreover, the increase in profits by firms (the rise in producer surplus) does not fully compensate for the reduction in consumer surplus. Consequently, total welfare is lower than before the increase in the price. This mechanism is graphically represented in figure 3.1.

In equilibrium *A*, price (P_0) equals marginal costs of production. As a result firms do not have any surplus, since profits are equal to zero. Consumer surplus, on the other hand, is equal to the sum of *C*, *D* and *E*. Now suppose that the price rises to P_1 and the market moves to equilibrium *B*. In this case, consumer surplus has reduced to *C* and Producer surplus (revenues minus costs) is equal to *D*. Note that, relative to equilibrium *A*, total surplus shrinks as *E* is lost. Although producers are better off, this is not good enough to compensate the lower consumer surplus. Hence, total welfare is lower due to the price increase. The welfare loss for any price above marginal costs is named the Deadweight loss.⁴

⁴ Here we do not consider prices below marginal costs, since they are irrational from a producer's point of view. If the price would be below marginal costs, producers would decide not to produce, as they would then incur a loss.

Figure 3.1 Consumer's and producer's surplus, and the deadweight loss



Allocative efficiency

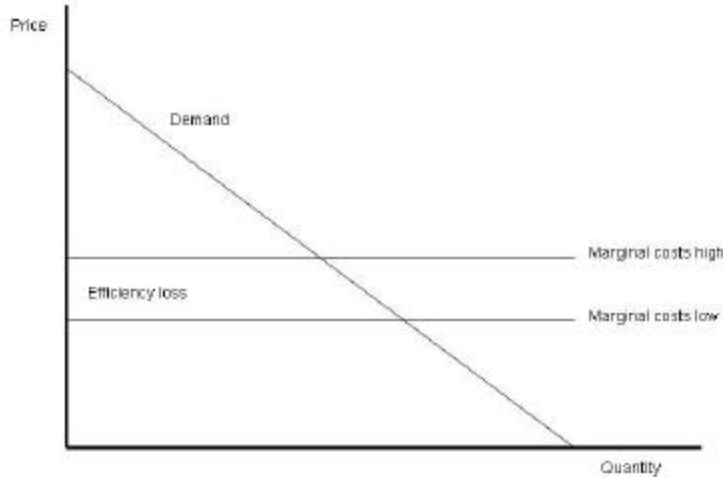
The first concept we explain is that of allocative efficiency. When we consider allocative efficiency, we assume that technologies (costs) are given, and that the most efficient technology available is used. In short, allocative efficiency requires that resources are allocated to their most efficient use. That is, when the value consumers place on a good or service (reflected in the price they pay) equals the cost of the resources used up in production. The condition required for maximum allocative efficiency is that price = marginal costs.⁵ From the preceding introduction to welfare analysis, it follows that allocative efficiency is measured by total surplus. So, if allocative efficiency is maximised, total economic welfare is also maximised. In this situation, no one can be made better off without making someone else at least worse off. This is known as a Pareto efficient equilibrium. Deviation from maximum allocative efficiency is represented by the deadweight loss described earlier. In the case that such an allocative inefficiency exists, total welfare can be increased by allocating resources from certain sectors in the economy to expand the production of the good in another industry. That is, for each expansion in output (until the optimal point), the gain in consumer surplus is greater than the loss of producer surplus. As we shall see in the next section, more competition, by increasing output, is a way to improve allocative efficiency.

⁵ Note that when perfect price discrimination is possible (which is unlikely due to information problems), then the profit of the monopolist would be equal to the sum of the area C, D, and E in figure 3.1. In this case no deadweight loss exists and allocative efficiency is also maximised (Motta, 2004).

Productive efficiency

Productive efficiency refers to how close the actual production cost is to the lowest cost achievable. Inefficiency in production results from using excessive amounts of certain inputs or from using the wrong input mix. This can be illustrated by figure 3.2. Economists usually explain technical efficiency by using two marginal cost curves. In figure 3.2., two marginal cost curves are depicted: one high marginal cost curve representing the inefficient production technique, and one low marginal cost curve representing the most efficient production technique available. The area between the two cost curves and the demand curve measures the extent of productive inefficiency associated with the high marginal cost technique. In the low marginal cost equilibrium, productive efficiency is maximised as producers minimise the wastage of resources in their production processes.

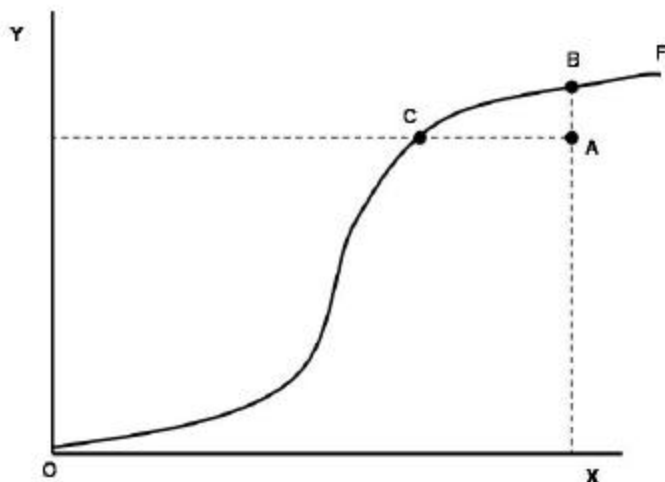
Figure 3.2 Productive efficiency loss



Productive efficiency is often interchanged with productivity. This is unfortunate because the terms are not precisely the same things. In order to illustrate the distinction between the terms, we examine a simple production process in which a single input (x) is used to produce a single output (y). The relationship between the input and output is defined by a production frontier OF , which is the line in figure 3.3. It represents the maximum output attainable from each input level. In other words, it reflects the current state of technology. Firms can either operate on that frontier or beneath the frontier. In the former case they operate technically efficient, whereas in the latter case they are technically inefficient. In this example, point A represents an inefficient point, whereas points B and C correspond to efficient points. The firm in A is technically inefficient because it can either produce more output with the same inputs (i.e. produce at point

B), or alternatively, produce the same amount of output with fewer inputs (i.e. produce at point *C*).

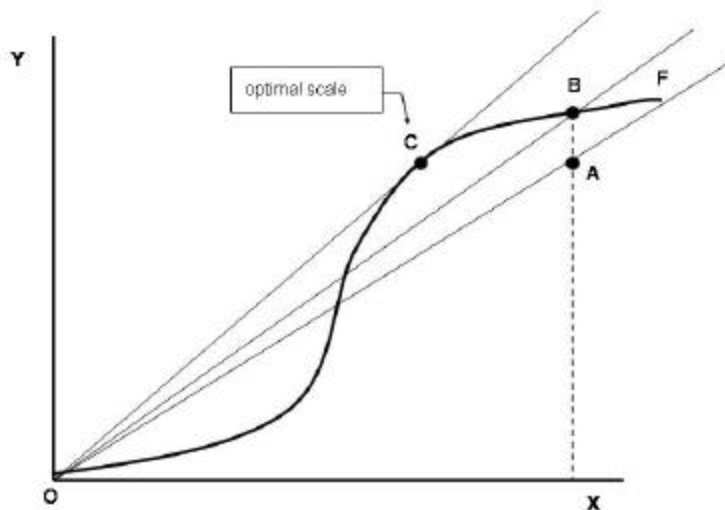
Figure 3.3 Production frontier curve and productive efficiency



Additionally, we add a ray through the origin to measure productivity at a particular point. Productivity is defined as the ratio of the output(s) that a firm produces to the input(s) that it uses. Hence a ray with a slope of y/x provides a measure of productivity. In figure 3.4 we have included three productivity rays besides the original production frontier OF . A move from point *A* to the technically efficient point *B* represents a rise in productivity, since the slope of the ray is greater in *B*. While moving from *B* to *C* does not improve the firm's technical efficiency, it does enable the firm to improve its productivity. In fact, because the ray from the origin is at a tangent to the production frontier at point *C*, the point represents the point of maximum possible productivity.⁶ Any other point on the production frontier results in lower productivity. The explanation for this result is scale economies. Point *C* is the point of optimal scale. Concluding, a firm that is technically efficient may still be able to improve its productivity by exploiting scale economies. Hence the level of productivity depends on two factors, namely, productive efficiency and the exploitation of scale economies. In the empirical part of this study, we use Data Envelopment Analysis to control for these scale effects.

⁶ Note that a steeper productivity ray will not touch the production frontier and is therefore not a possibility.

Figure 3.4 Distinction between productivity and productive efficiency



Dynamic efficiency

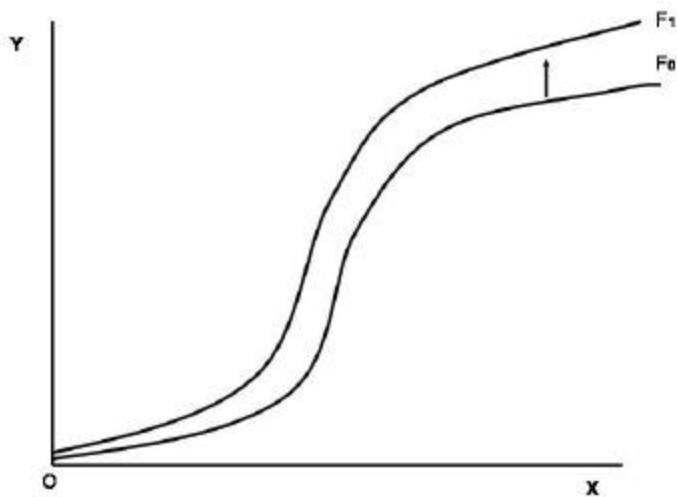
As the name suggests, dynamic efficiency contains a time component in contrast to the previous two notions of efficiency. The concept considers an additional source of productivity change, called technical change. This involves advances in technology which may be represented by an upward shift in the production frontier.⁷ Figure 3.5 demonstrates such a movement of the production frontier. In this figure, OF₀ depicts the production frontier in period 0, whereas OF₁ does the same for period 1. Technical change in this example allows firms in the second period to produce more output for each level of input relative to what was possible in the first period.

Despite the straightforwardness of the preceding example and unlike the previous two concepts of efficiency, dynamic efficiency is far more difficult to measure. Especially, measuring the dynamics of innovation and endogeneity of the market structure poses problems. The latter means that the market structure is not exogenously given, but determined inside the model. Moreover, it is very complicated to compare the relative magnitude of static inefficiencies (allocative and productive) with that of dynamic inefficiencies. As a result, dynamic efficiency has been given considerably less attention than static efficiency by economists. This bias is not

⁷ In general these advances in technology involve the rate of introduction of new products, as well as improvement in the production techniques of existing ones. The former concept is also known as product innovation, while the latter notion is called process innovation.

without harm. The following section will show that often a trade-off exists between static and dynamic efficiency. Therefore, maximum static efficiency does not automatically coincide with maximum dynamic efficiency. In particular, we find that too strong commitment to static efficiency can ruin the incentives to innovate, and therefore hurt dynamic efficiency which is vital for maximum welfare in future periods. In sum, dynamic efficiency stands for the present value of the future streams of static total welfare (= allocative plus productive efficiency).

Figure 3.5 Dynamic efficiency reflected in a shift of the production frontier curve



3.2 The relation between efficiency and competition

Where the former section has dealt with the notions of efficiency and competition separately, this section addresses the relationship between the two concepts directly. To deal with this subject, we use the key insights from the economic literature in this matter. The remainder of this section is structured as follows. First, we discuss the relation between competition and allocative efficiency. Then, we analyse the effects of competition on productive efficiency. Finally, we consider the relation between competition and dynamic efficiency. In this section, we implicitly employ the concept of competition suggested by Boone (2000) when no alternative concept is explicitly used.

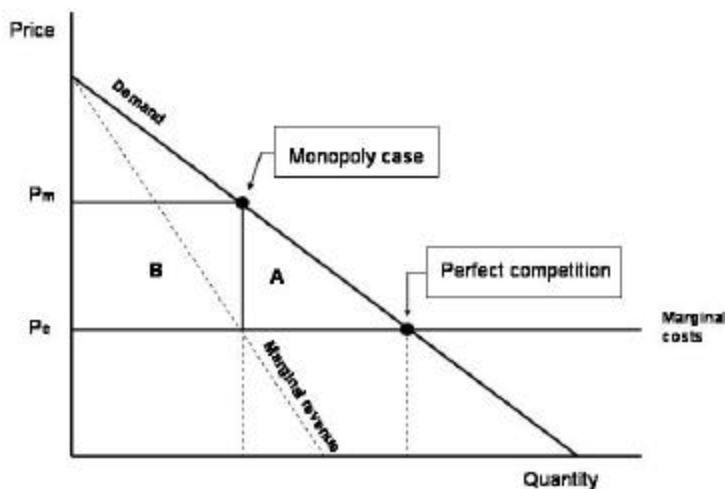
3.2.1 Competition and allocative efficiency

In a world with perfect competition there is maximum allocative efficiency. Meaning that all resources (e.g. capital, labour) are allocated in the most optimal way given their respective prices. If perfect competition exists, firms and households treat prices as given. Allocative efficiency is measured by the sum of consumer and producer surplus. Section 3.1.2

demonstrates this fundamental concept. Furthermore, we have seen that any deviation from the optimal point, which corresponds to marginal willingness to pay = marginal cost, brings about a decrease in allocative efficiency. In particular, a move from the perfect competitive equilibrium towards a monopoly is undesirable from a static economic welfare point of view. The welfare loss resulting from this change is depicted in figure 3.6. Relative to the perfect competitive case, a monopoly decreases allocative efficiency by the area of the triangle *A*, which is the deadweight loss for the economy. Note, however, that not everyone is worse off as producer surplus increases with respect to the monopoly case by the area of the square *B*. Still, net welfare is decreased since the gain in producer surplus is not enough to compensate the loss of consumer surplus.

The size of the deadweight loss is determined by the extent prices are above marginal costs. From this figure it is clear that the higher the price the larger the welfare loss caused by market power (i.e. ability to set prices higher than marginal costs). In the extreme case of monopoly, where market power is maximum, the welfare loss is the highest.

Figure 3.6 Welfare loss of a monopoly



Monopoly profits, the gain in producer surplus from a monopoly situation, create incentives for an industry's producers to lobby in favour of more protection and less competitive pressure. In fact, firms might even use these resources, called monopoly rents, in this lobbying process to keep or increase their monopoly power. These rent-seeking activities add to the welfare loss from a monopoly, because the resources used in these activities could instead be put in more

productive use. Posner (1975) therefore argues that the welfare loss of a monopoly can be as large as the area of A and B together. Note that for this result to hold the wasted resources in rent-seeking activities should not have any social value.

3.2.2 Competition and productive efficiency

Most economists have the vague suspicion or belief that competition exerts downward pressure on costs, reduces slack, provides incentives for the efficient organization of production, and even drives innovation forward. In other words, competition is said to improve productive efficiency.

While most economists believe that competition is good for productive efficiency, this trust is not supported by an undisputed theoretical foundation. Competition seems very well in practice. However, it is not so obvious how it works in theory. In fact, the question is rarely adequately pursued. Far more attention has been paid to the effect of competition on profitability. This may sound curious. After all it is productivity growth that is the cause of the “wealth of nations”. This link between competition and productive efficiency is an old and broad question. In some way, it seems reasonable to allege that a firm that does not face any competitive pressure will not make much effort to use the best available technologies, to improve its products and to innovate. Although this sounds very logical, it is important to provide this claim with solid arguments.

This subsection addresses such arguments. More specifically, we investigate three main approaches in the literature that relate competition to productive efficiency. The first approach employs the theory of competitive selection in which firms are treated as black boxes. In contrast to the first approach, the second approach does not treat firms as a black box. Indeed it assumes that managers of companies pursue other goals than profit maximisation. Finally, the third view concentrates on entry. Opposite to the typical belief of economists, this approach finds that free entry is not always optimal from a welfare perspective.

Approach 1: Darwinian economics

Standard microeconomic theory presumes that in the long-run under perfect competition price equals the minimum of long-run average costs. This mechanism works as follows (Cabral, 2000). When active firms make positive profits, new firms are attracted to the market and will decide to enter. If, on the other hand, active firms make losses, some firms will decide to exit the industry. This process continues and converges to a limit point in which each firm receives zero profits and there is neither an incentive to enter nor to exit. In this model of perfect competition, the distribution of firm size is either single-valued (assuming U-shaped cost

curves), or indeterminate (assuming constant returns to scale). This theory assumes that technology (i.e. costs) is the same for all firms and no barriers to entry (or exit) exist.

Although this theory seems quite plausible, the implied distribution of firms by this model is rather extreme. Moreover, empirical evidence is widely at odds with this view of industry dynamics (e.g., Du Reitz, 1975; Mansfield, 1962). For instance, in any given period and industry, entry and exit take place simultaneously. Furthermore, supranormal profits are persistent in the long run. In addition, the size distribution of firms is not concentrated on one size. That is, the average size of entrants and exiters is often much smaller than industry average size.

In order to let theory correspond with empirical evidence, some assumptions of the model of perfect competition have to be relaxed. Jovanovic (1982) provides a model in which the stylised facts described above are accounted for. He shows, both theoretically and empirically, that efficient firms grow and survive, while inefficient firms decline and fail. Also the existence of positive profit rates in the long run and simultaneous entry and exit are explained by his model. For this Darwinian result to hold, the author assumes that different firms have different degrees of productive efficiency. This in turn corresponds to different cost functions. Meaning that more efficient firms have a lower marginal cost schedule. Additionally, each firm is uncertain about its own efficiency. Once a firm enters the market, gradually more precise information on its true efficiency becomes available. This process introduces the selection mechanism which induces inefficient firms to exit the industry and efficient firms to gradually increase their output. As firms set prices equal to (expected) marginal costs, it follows that more efficient firms sell at a higher output. In equilibrium, each firm compares the expected benefits and costs from remaining active. Moreover, Jovanovic proves that the unique equilibrium is optimal from a welfare point of view. Competitive selection improves efficiency by selecting those firms with low marginal costs that maximise total surplus. Therefore, competitive selection makes both the firm and society better off.

Empirical work by Olley and Pakes (1996) gives strong support to the selection effect of efficiency. Using sophisticated econometric techniques to analyse the telecommunication industry in the US during the period 1963-97, the authors find that the selective process of entry and exit is the major driver behind this result. More recent literature confirms the role played by exit and entry in increasing efficiency (Disney et al., 2000).

Taken together, the role of competition in selecting the most efficient firms to survive seems to be supported both by theory and empirical evidence. Just as survival of the fittest is to evolution in biology, competitive selection is to productive efficiency in economics. However, we observe that theory in this branch does not explain why firms differ in their efficiency levels. In

fact, the distribution of efficiency levels is assumed rather than derived. These differences may result from a variety of factors. For example, some firms hold a sustained competitive advantage due to the abilities of its workers or management. In addition, there may be impediments to imitation of technology that allow some firms to perform persistently better than others. Further, success can be attributed to certain strategic or tactical decisions. Besides these preceding factors, some managers are more efficient in production than others due to incentive problems.

Approach 2: Incentives

Much of the analysis in economics treats the firm as a sort of black box, in the sense that it produces outputs from inputs in a predictable, mechanistic way to maximise profits. One can question whether this assumption is realistic as most modern corporations feature a management that is separated from ownership. In this setting, the manager's objectives may differ from those of the owners, because managers are motivated by other aspects than the firm's profits only. Therefore, we relax the assumption that firms are like black boxes which only aim is to maximise profits. In short, this approach deals with incentives that influence a firm's productive efficiency.

The literature on this approach is based on the determination of managerial effort (Nickell et al., 1997). The idea is that competition makes firms internally more efficient by sharpening incentives to avoid sloth and slack. In this category several types of incentives which influence managerial effort can be distinguished.

In the corporate finance literature it is commonly assumed that managers of large companies are mostly concerned with preserving their private benefits of control over the company. In the same time they wish to minimize effort (Aghion et al., 1997). Furthermore, most managers tend to be conservative. In the sense that they are likely to run their business 'quietly' without interrupting their habits. This risk-averse behaviour introduces agency problems. This involves managers (agents) not performing to their utmost best in maximising the profits of the firm. As a result, in practise, some firms appear to be doing more maximising than others (Nickell et al., 1997). Owners of companies (the principles) want to reduce the 'slack' that is created by this misalignment. Not only managers have the potential to capture rents in the form of slack or lack of effort. Also workers might produce inefficiently low levels of effort. Managers may want to share rents with workers to make their lives more comfortable (Smirlock and Marshall, 1983). In sum, shareholders care about profits, while managers and workers care about other things (i.e. their individual utility is determined by other aspects).

Although most people seem to think that competition would lead to higher effort, it is not really straightforward how. For instance, Jensen and Meckling (1976) note that the owner of a

monopoly has just as much incentives to prevent his managers from slacking as the owner of a competitive firm. Nonetheless, Nickell (1996) argues that the latter are in a better position to do so whenever managers have better information than the owners. A popular way to reduce these agency costs is to offer the manager an incentive scheme.⁸ In order to make this scheme work, the owners or the market have to monitor the managers' effort. Various studies suggest that these schemes will generate better incentives if the environment is more competitive (Holmstrom, 1982; Nalebuff and Stiglitz, 1983; Hart, 1983). This can be explained by the greater opportunities for comparison of performance. In a monopoly, managers can attribute poor performance to exogenous factors. This behaviour is less feasible when well performing competitors are present. In short, competition generates additional information not available in a monopolistic industry. However, opportunities for comparison only arise if both unobserved productivity shocks and managerial abilities between competing firms are nonnegatively correlated. Moreover, Meyer and Vickers (1995) show that this result holds so long as the former correlation is larger than the latter. If this is the case, and effort increases with competition, company performance will tend to improve due to competition.

This result has a number of popular practical applications. For example, payment by performance relative to others, and regulation of prices of monopolists in utility industries. This is also known as yardstick competition (Shleifer, 1985). Unfortunately this theory has received some criticisms (Scharfstein, 1988; Hermalin, 1992). Especially ambiguity of the effect on managerial effort to these incentives is a problem. Scharfstein shows that the result is reversed if managers are very responsive to monetary incentives.⁹ Consequently, competition may increase managerial slack. Despite this criticism, depending on the conditions, competition improves efficiency in many, but not all, circumstances (Vickers, 1995).

Alternatively, implicit incentives can arise from competition. These incentives are not the consequence of contract design, but occur from exogenous market forces. The key assumption in this type of models is that current managerial effort does not influence current earnings. Yet, it may affect future income due to the impact on the market's estimate of the manager's ability. As managers do not stay with the same firm forever, they are interested in creating a good reputation. Therefore, managers bother making effort because of the reputation effect as it improves future earning prospects. The labour market disciplines managers and provides them with the proper incentives. But how does competition help here?

⁸ The optimal scheme between shareholders and managers is one that balances the benefits from insuring the manager against risk, on the one hand, and the benefits from providing the manager with the right incentives, on the other hand.

⁹ The concept of risk aversion is central to this point.

Meyer and Vickers (1994) find that competition improves the possibility for comparison. Just as above, there needs to be a nonnegative correlation among the managers' ability levels and productivity shocks between competing firms. Once more the latter correlation has to exceed the former. The higher the correlation of productivity shocks between firms, the higher the precision with which the market can observe performance, and thus assess the ability of the manager. On the contrary, a high correlation of abilities among managers generates the possibility for agents to free-ride on each other's efforts.

The paper by Schmidt (1997) completely abstracts from any informational effects of competition. This paper focuses on the impact of competition on the probability of bankruptcy. The author shows that more competition, in terms of lower profits, will raise the probability of bankruptcy at any given level of managerial effort. As a result, managers have to work harder to avoid this unfortunate outcome. Increased competition lowers the leeway managers experienced before intensified competition. Especially, inefficient firms will have strong incentives to minimise costs and improve efficiency to avoid the disutility of their more likely fate of liquidation. This corresponds to the selection theory discussed earlier, however, in this case it is not differences in strategy or ability, but incentives to managerial effort that explain the survival of efficient firms.

Besides the 'threat-of-liquidation' effect, competition might also reduce effort, because, as profits decrease, the incremental benefit of inducing more effort decreases.¹⁰ In close spirit to a paper by Aghion et al. (1995), Schmidt shows that, under certain (rather strict) conditions, competition unambiguously increases managerial effort.¹¹ In an extension of his model, the author finds that managers have the strongest incentives to reduce costs in an oligopoly with few competitors if there is a substantial risk of liquidation.

In conclusion on the incentive approach, we can say that theory suggests that competition improves the incentives of managers to lower productive inefficiencies. Put differently, increased competition makes incentive schemes work better, enhances labour market discipline,

¹⁰That is, lower expected future profits diminish the owner's incentives to induce more effort from the manager. As a result, the payments of rents to the manager in order to increase his effort are lower.

¹¹These sufficient conditions require that the manager is not paid any rents in excess of his reservation utility. Hence, there is only the 'threat-of-liquidation' effect.

and increases financial pressure which induces managers to work harder. Note, however, that this theoretical foundation is not a very strong one. The next approach investigates the effects of an entry bias on the relation between competition and efficiency.

Approach 3: Entry bias

Opposite to the popular view of economists, free entry is not always beneficial to social welfare. Mankiw and Whinston (1986) show that, when the perfect competition model fails, then a tendency towards excessive entry may exist. The authors reasonably assume that firms have to incur a fixed (and sunk) cost when they enter the industry and do not longer act as price takers. In this case, a potential divergence between the private and the social incentives for entry of a new firm could occur. In short, the increase in total surplus due to entry is smaller than the profit earned by the new entrant. Hence, while entry is desirable from the entrant's perspective, this is not the case from a social perspective, because the increase in total surplus (i.e. decrease in the deadweight loss) does not compensate for the entry cost. The explanation for this divergence comes from the business stealing effect. This effect represents the profits "stolen" by the entrant from the incumbent firms. It is a transfer between firms that does not benefit total welfare, because the average cost of each firm increases. To summarise, while entry is good for allocative efficiency, it is bad for productive efficiency as less advantage is taken of scale economies, which exist due to the fixed entry cost.

On the contrary, if entrants introduce more product diversity, the bias toward excessive entry is reversed. That is, entry is more desirable for society than it is to the entrant, because the firm does not capture the resulting welfare gain in profits. There is a positive externality from entrant to consumers. Hence, this result contrasts to the homogenous product case depicted above. By extension, Boone (2003) argues that when consumers value variety, competition can be too fierce. In the sense that if competition would become less intense (i.e. less efficient firms are able to survive), and more firms could enter the market with differentiated products, welfare would be increased. In sum, when product differentiation is important, increased competition can lead to insufficient entry from an economic welfare point of view.

Ultimately, there are two opposing effects at work. First, there is the business stealing effect which leads to excessive entry if competition is soft. Second, the desire for product variety leads to too little entry than is socially warranted. Therefore the dominating effect of the two will determine in which direction the entry bias will be. The following subsection considers the relation between competition and efficiency in a dynamic perspective.

3.2.3 Competition and dynamic efficiency

The influence of competition on innovation is an old debate in the economic literature. This link is rather complex and surrounded by some ambiguity. Both the theoretical Industrial Organization and the more recent endogenous growth literature deal with this issue. The classic Industrial Organization theory on this theme, started by Schumpeter (1943), predicts that innovation should decline with competition, because more competition reduces the monopoly rents, and thereby the incentives that drive successful innovators. In addition, these monopoly rents can serve as a valuable source of research and development (R&D) funding.¹² Schumpeter argued that capitalism's main feature is the dynamic process of Creative Destruction. He commented that not static competition but "*competition from the new commodity, the new technology, the new source of supply, the new type of organization...-competition which strikes a decisive cost or quality advantage and which strikes not at the margin of profits and the outputs of existing firms but rather at their foundations and their very lives*".

Later, he observes that "*As soon as we go into the details and inquire into the individual items in which progress was most conspicuous, the trail leads not to the doors of those firms that work under conditions of comparatively free competition but precisely to the doors of the large concerns.*"

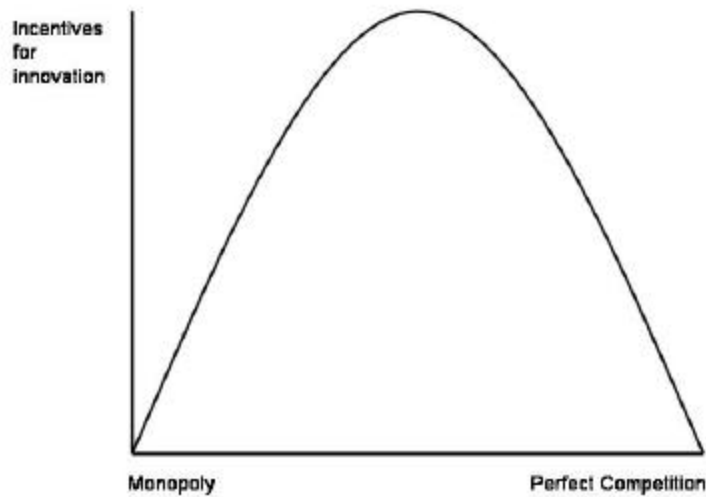
In contrast, following Arrow (1962), competing firms may have sharper incentives to innovate than monopolists as the pre-innovation profits are greater under a monopoly than under competition. The monopolist's comparative disincentive is also known as the replacement effect. Naturally, firms have to be able to protect their intellectual property from imitation by competitors. Otherwise the gain in profits from the innovation will be eliminated and all incentives to innovate will disappear.

Are the two preceding views inconsistent with each other? Scherer and Ross (1990) argue that there exists a middle ground environment in which some intermediate levels of competition are optimal for innovations and productive efficiency. This conclusion is also found in the

¹²This argument reasonably presumes that capital markets are imperfect. In reality, innovators are often unable to convince investors of the virtues of their potential innovation. The problem is that the firms risk losing the idea in convincing a capitalist. This problem is one of the major reasons why a large fraction of R&D investments are self-financed

empirical literature based on endogenous growth models. Aghion et al. (2005) find that the relationship between competition and innovation is an inverted U-shape (see Figure 3.7 below). All in all, the studies discussed above suggest that the existence and/or the prospect of enjoying some degree of market power has an important role in generating incentives to invest in R&D.

Figure 3.7 Inverted U-shape relationship between competition and incentives to innovate



3.2.4 Conclusion

This section has analysed the literature on the relation between competition and efficiency. In short, most of the literature suggests that competition exerts pressure on managers and firms to improve efficiency. Nevertheless, this relationship is not always a simple monotonic one. The next section examines the merits of the different designs of competition in the railway industry. It assesses whether the general theoretical insights obtained in this section will apply to the specifics of the railway industry.

3.3 Design of competition in railways

Why should competition be preferred rather than a monopoly in the production of railway services? As in other network industries, railways are characterized by the specifics of these industries such as scale economies due to high fixed (sunk) costs and the prominence of infrastructure which is an essential facility.¹³ This renders production by many small firms

¹³High fixed (capital) costs arise from the acquisition, installation and maintenance of tracks and stations; as well as the cost of acquisition, operation and maintenance of rolling stock.

uneconomic. Yet, some competition may be useful to disturb the quiet life of the incumbent, and thereby improve efficiency, as Hicks would say. Like we have seen in chapter 2, only the track network of the railway industry is subject to a natural monopoly. Therefore other parts of the industry, where natural monopoly is not an issue, can be opened to competition.

While most economists and policy makers agree that competition should be introduced in the railway industry, agreement on how this should be done is lacking. Several designs of competition are available. The aim of this section is to analyse the different designs of competition on their relative merits. In particular, we sketch the efficiency trade-offs associated with the different designs of competition. Likely this will depend on the various specifics of the railway industry, as within the railway industry several submarkets can be distinguished.

Before competition can be introduced, restructuring of the industry is often required. Consequently, industry structure determines to a great extent which type and degree of competition is feasible. Ultimately, this choice depends on the trade-off between allocative and productive efficiency.¹⁴ As we have seen in the preceding section, when economies of scale are important, free entry is often suboptimal. That is, the efficient scale of operation is large relative to the size of the market. In this case, the benefits of more competition, a rise in allocative efficiency, do not weigh up against the loss in productive efficiency. If on the other hand economies of scale are not too important, then competition can be expected to improve both types of efficiency. In chapter 2, we noticed that in many cases economies of scale are significant in the railway industry. Consequently, more competition does not seem to be the best solution to solve the misallocation of resources by the monopoly. A popular approach to tackle this dilemma is competitive bidding. It is the first design of competition we discuss.

3.3.1 Competitive bidding

Even when a monopoly seems to be the best solution to operate the railway system, competition can have its beneficial effects. Demsetz (1968) suggested that competitive bidding (known as competition for the market) should be evoked if several firms are candidates to operate a network. This ensures that the most efficient firm is selected. Following this suggestion, many countries have started to auction the rights to operate a certain market for a given duration. In

¹⁴ Regarding dynamic efficiency, we suspect that a middle ground exists as was described in the previous section. This involves the usual trade-off between static and dynamic efficiency, where from a static efficiency perspective perfect competition is most desired, but from a dynamic efficiency perspective an intermediate level of competition would be optimal.

The Netherlands, competitive bidding was introduced in 1999 when several regional lines were put out to tender. Competitive bidding of these concessions allows the government to keep the property rights of the track network, while the winning railway firm obtains the exclusive right to operate this network. In addition, the firm is permitted to behave as a profit-maximiser in so far as it respects some quality and environmental qualifications and fulfils certain redistribution obligations (for example, universal service and no price-discrimination).

The objective of maximising welfare is achieved when, in the first place, the bidding process has resulted in an efficient outcome. That is, the authority has skimmed most of the surplus that would otherwise disappear in the pockets of the monopolist (Loeb and Magat, 1979; Riordan and Sappington, 1987). The funds collected by the authority can be redistributed to satisfy certain distribution goals.¹⁵ In order to provide the monopolist with a strong incentive to minimise operating costs and thereby maximise productive efficiency, the firm has to be the residual claimant of the profits. Note that full implementation of this economic solution is hampered by social concerns such as the quality and safety of the railway product.

Similar to the agency problems within a firm, the authority (the principal) has to design a concession contract for the management (the agent) of the winning railway firm. Both can be seen as players in a game with strategic behaviour and private information as scarce resources. The latter, for example, being the information on the costs of the railway firm. The specific design of the contract is very important for the effective regulation of the firm. This hinges on the ranking of the different priorities of the government. From an economic viewpoint, maximising total welfare should be the sole priority, while other concerns might act as constraints that define the feasible set in which the most efficient solution is chosen. In short, the aim of the authority should be to select the most efficient firm via procurement and impose an optimal level of effort through a contract. Yet, it is often claimed that most politicians subordinate efficiency to income distribution and budgetary concerns.

One drawback of this type of competition is its high costs. In particular, organising auctions is costly (Laffont and Tirole, 1993). Both the procurers and the bidders incur substantial costs.

¹⁵ While equity aspects are interesting politically, they do not belong in an economic analysis.

Moreover, the whole process can be quite time-consuming. Complicated contracts have to be written down properly and terms and conditions have to be unambiguous. Furthermore, a contract can never be complete. There are many imaginable contingencies and many of these are unforeseeable. Generally, the longer the duration of the concession and the poorer the information to both parties, the more incomplete the contract will be. As a result of the incompleteness of contracts, contracts should specify what should be done when unforeseeable events occur. Parties can decide to renegotiate from the beginning or to negotiate only the new contingencies, to call for third party arbitration, etc. At least it should not discourage any effort to enhance efficiency (Crampes and Estache, 1997).

Another major challenge is to get the incentives right in the presence of asymmetric information. In general, the firm has an informational advantage over the authority. This tends to induce strategic behaviour by the firm. In particular, the firm is likely to pretend that it is less efficient (i.e., exaggerating costs), since this allows it to put forward less effort. The design of the payment to the firm for its services, which flows either from the market or from public subsidies, can solve this incentive problem. The optimal rewarding system gives the right incentives, in the sense that an efficient firm should truthfully report it is efficient and an inefficient one that it is inefficient. In this way the firm is rewarded for the information it discloses regarding its efficiency.

Normally, contracts are not signed for a very long term. Since, as time goes by, new and better information becomes available to the authority. For example, information on the behaviour and characteristics of the firm, and on potential challengers. This could lead to a renegotiation of the contract or a new bidding procedure to allow new firms to compete for the concession.¹⁶ Rebidding or renegotiation can have important efficiency gains over the traditional natural monopoly case requiring regulation. Prices may be adjusted in response to new circumstances and the relative efficiency of potential suppliers could have changed. Newbery (1999) argues that, “*Competition is more effective than regulating at cutting costs to improve productive efficiency, and aligning prices with costs to improve allocative efficiency.*” However, from a firm’s perspective long-run contracts are preferred that cannot be renegotiated, because otherwise large sunk investments cannot be recovered. Therefore, short-term contracts may

¹⁶ Shaw, Gwilliam and Thomson (1996) note that the average duration of rail concessions is about five to ten years when they refer only to services and up to thirty years when network investment and development are included.

induce underinvestment and short-termism in investment behaviour (Williamson, 1976). In particular, when R&D is important, firms need to be able to recoup their investments or else vital incentives will be deprived. A solution to this problem is to transfer the incumbent's assets at the end of the concession period to the new concessionaire at the right price. Valuating these assets can be difficult, because the worth of the assets depends on future conditions of the market. This is even more the case for nonmonetary and non-transferable investments such as the quality of past investment choices.

Whether auctions in reality correspond to the optimal theoretical outcome considered here remains an important question. When there are a finite number of diverse bidders, all monopoly rents cannot be recovered through competitive bidding (Riordan and Sappington, 1987). Competitive bidding requires multiple players to maximise the benefits from competition. For this reason, anti-trust authorities should take care that effective competition is not impeded by anti-competitive practises such as collusion among the bidders. There is ample evidence of collusion in procurement (e.g., Klemperer, 2002).

In sum, this subsection has discussed the main aspects of competitive bidding. A full treatment of this design of competition is beyond the scope of this research.¹⁷ That would require an investigation into the specifics of auction design and regulatory aspects such as tariff design. Depending on the significance of the caveats and drawbacks of competitive bidding, this design can be most favourable when economies of scale are so important to rule out competition on the tracks.

3.3.2 Competition on the tracks

When economies of scale are relatively unimportant and duplication of infrastructure is uneconomic, competition on the tracks can be the best solution to improve efficiency. For competition to be effective, competitors need to have non-discriminatory access to the essential facility of the incumbent. In general, two structural options are available to realise such access conditions. The first is the separation of infrastructure and train services so that the former incumbent is unbundled. The second method is to ensure fair third-party access to the network of the incumbent railway company. Both options are also known as open access. Regulation of

¹⁷The interested reader is referred to Campos and Cantos (2000).

prices, capacity and access are required to ensure the optimal use of the bottleneck facility.¹⁸ The rationale for regulation is twofold. Regulation is there to enable competition over the network to take place and to prevent the owners of the network from reaping excess profits (Klein, 1996).

Which of the two methods to ensure non-discriminatory access to the network should be preferred? Third-party access or vertical separation? The former requires no costly restructuring of the incumbent firm. Entrants need to have the ability to access customers over the network of the incumbent. A major concern is that the incumbent has an incentive to discriminate the entrants in favour of its own service provider. To avoid this, access regulation has to be installed. Note that this is easier said than done and often leads to regulatory inefficiencies (government failures). If favouritism cannot be prevented, effective competition is unlikely to arise (Pittman, 2000). Additionally, BTRE (2003) argues that, *“third party access may affect the efficiency of the incumbent’s rail business or other activities. For instance, sharing terminal space may reduce the efficiency of the incumbent’s shunting and marshalling activities. Further, in vying for capacity, capacity can become scarce and train scheduling can become less flexible. This may have the effect of reducing, the operational efficiency of other areas of the incumbent’s business.”* On the other hand, a major advantage of third-party access is that infrastructure and services remain integrated. In this way, possible economies of integration such as economies of scope, coordination benefits, and low transaction costs are not lost.

Vertical separation, alternatively, avoids the regulatory problems of ensuring equality of access. Now there is no reason for the infrastructure manager to discriminate among different train operators. Even so, efficient access conditions to the network have to be designed, which can be rather complicated. In contrast to third-party access, vertical separation leads to a loss of economies of integration. Without these, the incumbent loses a potential competitive advantage it had relative to its competitors. The welfare effects of separation ultimately depend on the trade-off between the benefits of a more equal level playing field and the costs of economies of integration lost.

¹⁸ Although the design of access regulation is a key aspect of open access regimes, a full treatment is beyond the scope of this study.

Railway investments often occur at the interface between infrastructure and rail services. As a result of this, a strong need for coordination between the users of the infrastructure and the infrastructure manager exists. In the case of separation, each component of the railway system optimises its own part, thereby neglecting the effects of its investments on the rest of the railway system.¹⁹ This calls for regulatory intervention which allows integrated planning and coordination to get the appropriate incentives for maintenance, improvements, and other investments. Experience in the USA and in the electricity industry indicates that such system-wide coordination is possible (Ordovery and Pittman, 1994).

While both restructuring options are a necessary condition to introduce competition on the tracks, they are by no means a sufficient condition. The success of both options heavily depends on the prospects of effective competition. High barriers to enter the market and a high minimum scale of operation relative to the market can hinder effective competition. When competition is inadequate, the benefits of competition will not be realised and the costs of restructuring are likely to exceed its benefits. Consequently, restructuring may not be economically worthwhile.

In railways, establishing optimal timetables can be rather complex when rights to use the train network are allocated among multiple parties. This is the case when there is competition among train operating companies on the same infrastructure. The question is whether an optimal timetable can be established through decentralised bargaining or whether a smart market is required that simultaneously generates the optimal set of paths through the network and the prices for all the paths contained in it. The value of each right to use a particular piece of track at a certain time is conditional on what happens with all neighbouring pieces of track. Consequently, a single optimising smart market could be needed. A way to solve this complex problem is to auction slots, that is a path through the network at a particular time. Currently, Sweden and the United Kingdom are investigating whether these smart markets can be created for their railway markets. Traditionally, railways are centrally dispatched to prevent the catastrophic cost of short-term supply/demand imbalances such as collisions. However, new computer and monitoring systems could reduce the need for centralized timetabling. Similarly

¹⁹ Especially in railways, the effectiveness of investments, and therefore the efficiency of rail services, depends on the exact point where vertical separation takes place. This point is also known in the technical engineering literature as the point where steel wheels meet steel track. In an examination of the railway literature Pittman (2005) argues that, " *a large portion of the discussion of possible technological improvements in the railway sector is focused on the ways that differences in rolling stock design and wheel design affect track wear, track maintenance requirements, and optimal track design, as well as the converse - the way that track design and maintenance may affect rolling stock.*"

to the internet, routes can be decided on a decentralized basis. Optimal use would be obtained if users of the railway system face prices that lead them to use the system optimally. In this way, prices continuously reflect demand and supply conditions. More specifically, the price system must reflect opportunity costs by time as well as location in order to create incentives to invest in extra capacity to relieve congestion. Further advances in telecommunication and computer-based smart markets could render timetabling in this fashion possible (Klein, 1996).

Where the size of the market is large in comparison to the minimum efficient scale of operation and several firms can operate at an efficient scale, competition in the market can be the most appropriate design of competition. Whether competition should be made possible through third-party access or vertical separation depends on the significance of economies of integration and the relative regulatory costs of both options. In order to optimally use the network with multiple firms, a timetabling system has to be developed that contains the concept of smart markets.

3.3.3 Summary

In this section, we discussed several designs of competition in the railway industry and their relative merits. In general, theory indicates that a priori every design of competition can be potentially beneficial to efficiency. However, it was also shown that all options have their drawbacks that could possibly offset these beneficial effects. The purpose of the following two chapters is to assess whether the theoretical insights derived in this chapter hold empirically. This empirical analysis contains two steps. In the first step we measure relative efficiency (chapter 4). The results from the first step are then applied in the second stage where we attempt to identify the impact of the various designs of competition on efficiency (chapter 5).

While this chapter has considered all concepts of efficiency, this empirical part focuses only on productive efficiency. There are two reasons for this. First, it is beyond the scope of this research to empirically investigate all concepts of efficiency. Second, data availability severely constrains the possibilities for examining the other two concepts of efficiency. As a consequence of the emphasis on productive efficiency in this study, we can only obtain econometric evidence on the relationship between the designs of competition and productive efficiency. So, in interpreting the results, the reader should keep in mind that the different designs of competition have effects besides that on productive efficiency, notably, allocative and dynamic efficiency.

4 Measurement of productive efficiency

4.1 Introduction

This chapter is concerned with measuring the relative productive efficiency of railway systems. Various approaches are available. Essentially, every method contains the conversion of inputs into outputs. However, the methods differ according to the assumptions they make, the data they require and the type of measures they produce. Whereas Total Factor Productivity (TFP) is most often applied to aggregate time-series data, Frontier approaches are frequently applied to data on a sample of firms. The box below discusses various efficiency measurement approaches.

Efficiency measurement techniques

TFP measures changes in total output relative to the change in the usage of inputs. It captures the effect of factor substitution when relative factor input prices change. Consequently, TFP includes two types of efficiency: 1) allocative efficiency (i.e., usage of inputs according to their relative prices) and 2) productive efficiency (i.e., converting inputs into outputs). TFP estimation is, however, subject to the problem of accurately measuring all inputs. It requires price indices for inputs and outputs. As a result, lack of adequate data makes it often difficult, if not impossible to perform a TFP analysis. In addition, TFP growth can be attributed to several factors (e.g., efficiency gains, technical progress), making it difficult to interpret its results.

Frontier analysis, on the other hand, provides a clear measure of relative efficiency among firms. In contrast to TFP, the method does not need the assumption that all firms are productively efficient. This ability to control for inefficiency is important if one wants to investigate the relative performance of railway firms. Furthermore, frontier analysis does not require information on prices to measure efficiency. For these reasons, frontier analysis is preferred to TFP measurement in this research.

Besides TFP and frontier analysis, partial productivity measures are often applied to measure efficiency. These measures generally relate a firm's output to a single input factor. As is clear from the previous chapter, variation in productivity arises from different sources: differences in efficiency, economies of scale and density, differences in network characteristics, and other exogenous factors that affect performance (e.g., climate, geography) and/or technological changes. To induce a 'pure' measure of productive efficiency, one must remove the effect on productivity caused by the differences in operating environment and exogenous factors (Oum et al., 1999). Partial measures do not take these external factors affecting productivity into account. Therefore, the results are to be regarded with due caution, as they can be inaccurate and/or incomplete. Another drawback of these measures originates from the fact that the productivity of one input generally depends on the level of other inputs used. That is, high productivity of one input may come at the expense of low productivity of other inputs. Using a partial measure, therefore, can lead to biased and unreliable results. Despite these shortcomings, the measures are widely used by both academics and industry analysts, probably because they are easy to compute, require only limited data, and are intuitively easy to understand. Even though partial measures have their virtues in terms of simplicity, we prefer not to use them due to their potential pitfalls.

In order to measure the relative efficiency of railway systems, we prefer to use the more sophisticated frontier analysis. The subsequent section addresses this technique. The theoretical foundation of the method is briefly illustrated. Hereafter, we discuss the two main techniques of frontier analysis and decide on which one of them to use in the remainder of this research. Additionally, this section provides a survey of the literature that has applied these approaches to estimate the efficiency of railways. In section 4.3, we discuss the data. Finally, section 4.4 concludes with the results. These results form the starting point for the regression analysis in the following chapter.

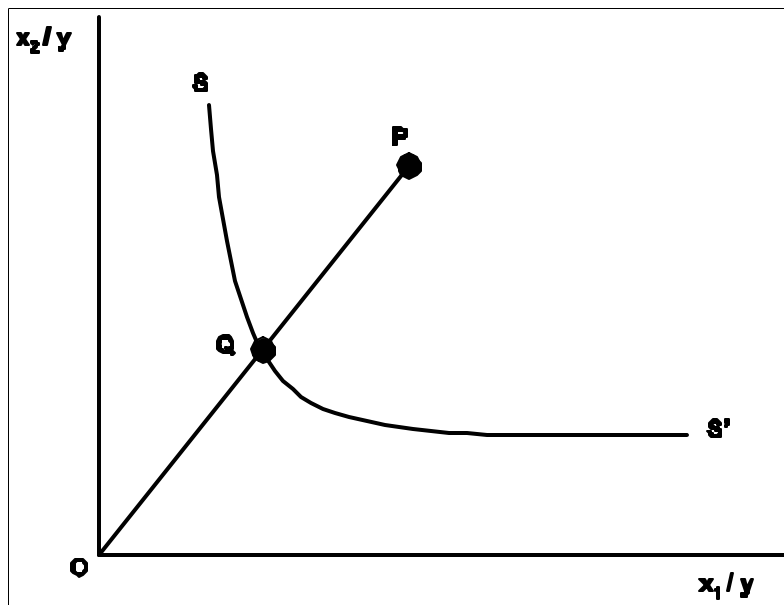
4.2 Frontier analysis

4.2.1 Analytical framework

In order to define a simple measure of firm efficiency which could account for multiple inputs, Farrell (1957) drew upon the work of Debreu (1951) and Koopmans (1951). Farrell suggested a measure of technical efficiency, also known as productive efficiency, which reflects the ability of a firm to obtain maximal output from a given set of inputs. The pioneer of modern efficiency measurement illustrated his idea in input/input space using an input-reducing focus. Hence the name input-orientated efficiency measures.

Figure 4.1 demonstrates this concept using a simple example borrowed from Coelli et al. (1998) involving firms that use two inputs (x_1 and x_2) to produce a single output (y). Constant returns to scale are assumed to allow the representation of the technology by using a unit isoquant. This unit isoquant, SS , represents the 'best-practice' production frontier on which a firm is fully efficient. It provides input-output combinations which are accessible to the firms under scrutiny. All firms must be either on this frontier or above and to the right of it. In the latter case, the productive inefficiency of a firm can be measured along a ray from the origin to the observed production point, holding the relative proportions of the inputs (or outputs) constant.

Figure 4.1 Farrell's definition of productive efficiency



If a given firm uses quantities of inputs, reflected by the point P, to produce a unit of output, the productive efficiency of that firm can be measured by the ratio QP/OP , which represents the percentage by which all inputs can be proportionally reduced without reducing output. As a result, the productive efficiency of a firm is usually measured by the ratio

$$PE = OQ/OP, \tag{4.1}$$

which is equal to one minus QP/OP . This measure is bounded between zero and one, and therefore provides an indicator of the degree of productive inefficiency of the firm. In the example, the firm in point Q is productively efficient as it lies on the efficient isoquant. Hence its efficiency measure is equal to one.

Note that this theoretical example assumes that the frontier is known. In order to apply this idea in practise, the efficient isoquant must be estimated from the sample data. To do so, Farrell proposed the use of either the construction of a non-parametric piecewise-linear convex isoquant or a parametric function fitted to the data. In both cases it is required that no observed point should lie to the left or below it. These two frontier approaches to efficiency measurement are the subject of the following subsection.

4.2.2 Efficiency measurement approaches

Both parametric and non-parametric approaches to efficiency measurement are frequently used for estimating frontier functions. The former are estimated by using econometric (statistical) methods, while the latter are assessed by applying mathematical programming. Essentially, two popular methods can be distinguished. On the one hand, a stochastic and parametric method, known as stochastic frontier (SF), is used to estimate the (unknown) underlying input-output production relationship using a functional form characterising the data. On the other hand, a non-stochastic and non-parametric mathematical programming technique, known as data envelopment analysis (DEA), is used. This method is less restrictive as it does not impose a functional form. Instead it takes the bounding observations as defining the best-practise efficient frontier.

A common feature of these two approaches is that information is extracted from extreme observations from a body of data to determine the best-practise production frontier. In contrast to other approaches which evaluate producers relative to an average producer, extreme point methods such as SF and DEA compare each producer with only the 'best' producers. Although this characteristic lies at the heart of frontier analysis, it also makes it vulnerable to outliers. An important assumption behind these two methods is that if a given producer is capable of producing Y units of output with X inputs, then other producers should also be able to produce the same if they were to operate efficiently. Another property of frontier analysis is that it is units invariant. That is, changing the unit of measurement does not affect the value of the efficiency measure. This advantage exists because productive efficiency is measured along a ray.

Unfortunately, there are no established methods or criteria for choosing between the SF and DEA (McMillan and Chan, 2004). Each is a viable approach to a common problem, namely assessing relative productive efficiency. Ultimately, the decision is a judgement call. We will now further outline the two methods and discuss their relative strengths and limitations.

Data envelopment analysis (DEA)

Mathematical programming to achieve the task of constructing the piecewise-linear convex hull approach to frontier estimation, proposed by Farrell (1957), did not receive wide attention until the seminal paper by Charnes et al. (1978), which devised the term data envelopment analysis (DEA). The authors developed the method "*to measure relative efficiency in situations in which there are multiple inputs and outputs and there is no obvious objective way of aggregating*

either inputs or outputs into a meaningful index of productive efficiency". Since then the literature using this method has vastly expanded. DEA has been applied in many situations (e.g., airports, universities, hospitals, banks).

The discussion of DEA begins with the linear programming problem of the standard input-oriented constant returns to scale (CRS) DEA model. In order to compute efficiency measures, we need to obtain a measure of the ratio of all outputs over all inputs, where we use weights for both the vector of outputs and the vector of inputs. These weights are obtained by maximising the efficiency for each firm, subject to the constraint that the efficiency measures must be less or equal to one and positive. In this way, for each firm a value of efficiency is obtained. If represented graphically, for a given set of firms, the efficient firms form the frontier that encloses the inefficient ones. Hence the name of the analysis - data envelopment analysis.

The programming problem can be mathematically formalised as follows.²⁰ First we need some notation. In this model, we assume that there is data available on K inputs and M outputs on each of N firms. For the i -th firm these are given by the vectors x_i and y_i , respectively. The data of all N firms are represented by the $K \times N$ input matrix, X , and the $M \times N$ output matrix, Y . As said above, the purpose of DEA is to construct a non-parametric envelopment frontier over the data points such that all observed points lie on or above the production frontier. In the case of an industry where one output is produced using two inputs, this can be represented by a unit isoquant (see Figure 4.1).

To find the weights, u for the $M \times 1$ vector of output weights and v_i for the $K \times 1$ vector of input weights, that maximise the efficiency measure of the i -th firm we employ a measure of the ratio of weighted outputs to weighted inputs, $u'y_i/v_i'x_i$. The relevant optimisation problem can now be specified as follows:

$$\max_{u,v} (u'y_i/v_i'x_i),$$

²⁰This example borrows from Coelli et al. (1998)

$$\text{st } u_j y_j / v_j x_j = 1, \quad j=1,2,\dots,N,$$

$$u_j, v_j = 0. \tag{4.2}$$

The first constraint implies that the ratio must be less or equal to one for all observed input-output combinations. The second constraint assures that the weights will always be non-negative. This solution process is repeated for each firm. Note that this ratio formulation has an infinite number of possible solutions. For instance, if (u,v) is a solution, then (au,av) is another solution, and so on. One can correct this problem by imposing the constraint $v_i x_i = 1$. Which in turn changes the optimisation problem into:

$$\max_{u,v} (\mu_i y_i),$$

$$\text{st } v_i x_i = 1,$$

$$\mu_j y_j - v_j x_j = 1, \quad j=1,2,\dots,N,$$

$$\mu_j, v_j = 0. \tag{4.3}$$

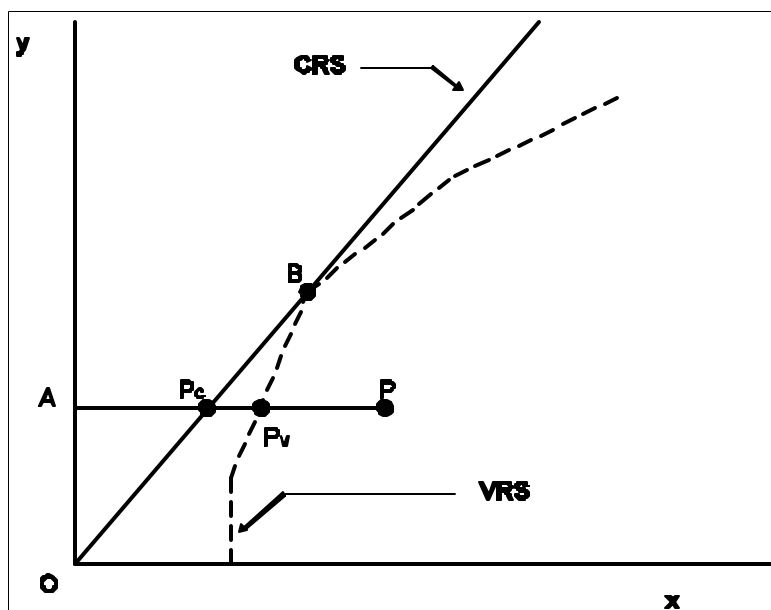
The notation has changed from u and v to, respectively, μ and v to reflect the transformation. In practise, solutions of this problem can be obtained by repeated application of linear programming software. While this model measures efficiency in terms of the potential proportional reduction in input use, other models measure efficiency in term of potential proportional output increase. Thus, DEA models may be input or output oriented.²¹ The choice of orientation depends on which quantities (inputs or outputs) the firms have most control over. In most cases input orientation seems most appropriate (Coelli et al., 1998).

DEA models that assume constant returns to scale (CRS) are insensitive to the specific orientation. However, when DEA analysis is extended to allow for variable returns to scale

²¹ An output-orientated CRS model is not presented here in the interest of brevity.

(VRS), efficiency scores can differ between the two orientations. The CRS assumption is only valid when all firms are operating at an optimal scale, which is defined as the region in which there are constant returns to scale in the relationship between outputs and inputs. In this region firms cannot take advantage of returns to scale by altering in size. If this is not the case, VRS are required to correct for scale efficiencies. Banker et al. (1984) were the first to propose such a VRS model. They modified the CRS linear programming model to account for VRS by imposing a convexity constraint. This constraint ensures that an inefficient firm is only compared against firms of a similar size. Since this convexity restriction is not imposed in the CRS model, a firm may be evaluated against firms which are substantially larger (smaller) than it. The VRS specification leads to a convex hull of intersecting planes which envelope the data points more tightly than the CRS conical hull. As a result, it leads to productive efficiency estimates which are greater than or equal to those obtained in CRS model. Figure 4.2 illustrates this difference by presenting both the CRS and VRS DEA frontiers.

Figure 4.2 Variable Returns to Scale versus Constant Returns to Scale DEA



In this example, one output is produced using one input. When CRS are assumed the productive inefficiency of the firm in point P is equal to the distance PP_c . Instead, under VRS this inefficiency is only PP_v . The difference between the two measures equals the scale inefficiency. In other words, the measure of productive inefficiency obtained from CRS DEA can be decomposed into scale inefficiency and 'pure' productive inefficiency, where the latter is measured when VRS is assumed. The special case, in which CRS efficiency is equal to VRS efficiency, is depicted by the firm in point B. In this point, firms are operating at the optimal scale. Most studies employ the more realistic and flexible VRS specification.

Now that we have discussed the basics of DEA, we turn to the method's advantages and disadvantages. One of the major strengths of DEA is its ability to handle multiple input and output cases. Unlike parametric techniques such as SF, DEA has no difficulty in accommodating the multi-output structure of railways that commonly produce both passenger and freight services. It does not require the construction of an aggregated index measure of output. Furthermore, DEA does not require the specification of a particular functional form for the production function and distribution of the data. Therefore it avoids the potential bias from selecting an incorrect functional form. DEA lets the data themselves dictate the profile of the frontier. It is much more flexible than SF in approximating the true production frontier (Oum and Yu, 1994).

Aside from these attractive features, DEA also has some disadvantages and points that require special attention. First of all, unlike SF, DEA does not take data noise (random shocks and measurement error) into consideration, because it is a deterministic approach. Empirical research always involves some degree of measurement error, data handling error, and stochastic errors. As a result, the efficiency estimates may be biased if the production process is largely characterised by stochastic elements. That is, any measurement error and random factors are reported as inefficiency. For this reason, DEA can be very sensitive to outliers in the data set. The second disadvantage of DEA is again a result of its non-parametric nature. Statistical hypothesis tests are not directly possible with this technique. However, a special type of regression analysis can be applied to the DEA results in order to identify the effects of particular variables of interest. Furthermore, DEA has no formal tests to assess the merits of including or excluding variables or the specific DEA model choice. Alternatively, one must rely upon the sensitivity of the results to the inclusion and exclusion of variables and judgement (McMillan and Datta, 1998). As a consequence, variable selection is a most critical part of DEA. The same point can be made for the selection of firms in the data set. Firms are expected to be relatively homogenous and employ a common technology to convert inputs into outputs.

A firm that is very different compared to the other firms in the data set, in the sense that it has unusual production characteristics, is rather likely to end up distorting the results.

All in all, this discussion of DEA has shown that DEA can be a powerful tool when used wisely. Yet, the same characteristics that make DEA powerful can also create some problems.

When applying DEA one should keep these limitations in mind. In addition to DEA, Stochastic Frontier analysis can be employed to measure productive efficiency. We now present a brief assessment of this alternative approach.

Stochastic Frontier Analysis (SF)

SF is a parametric method for estimating frontier functions and thereby measuring productive efficiency. SF involves the use of econometric methods to estimate the production frontier, and measures the efficiency of a firm using the residuals from the estimated equation. Consequently, the approach requires the specification of a particular functional form (e.g., Cobb-Douglas or translog) to describe the technology or efficiency frontier.

In reaction to the criticisms of traditional deterministic techniques, which do not take the influence of noise in the data into account, Aigner et al. (1977) developed SF, in which a two component error structure is incorporated. That is, deviations from the frontier are unravelled into stochastic and efficiency elements. Both elements are included as error terms in the functional form. As a result, one has to assume a particular distribution of these variables. Concerning the term representing statistical noise, most studies assume the residuals to be independently and identically distributed with mean zero and constant variance. Regarding the distributional form of the efficiency term, however, different options have been applied, for instance, half normal, truncated normal, and exponential. There is no a priori justification for the selection of any of these distributional forms. Therefore, the efficiency scores may be sensitive to distributional assumptions.

One of the main drawbacks of SF is that the approach is only well developed for single-output technologies. As a result it has substantial difficulties in accommodating multiple outputs. Aggregating outputs into a single measure is not a reasonable solution as it is very arbitrary. In contrast, DEA has no problems in handling multi-output technologies

Another disadvantage of SF compared to DEA is the need for an a priori imposition of the particular structures on the functional form and the distribution of the error term. Again this choice is rather arbitrary and could possibly influence the results.

The principal advantage of SF over DEA is its ability to take stochastic elements into consideration when measuring productive efficiency.

Conclusion

This section has analysed two methods to measure productive efficiency. The main advantage of DEA over SF is that it does not require specification of any functional form for production, avoiding the bias produced by specification errors. Furthermore, DEA is better than SF at assessing the productive efficiency of railway companies since it can easily accommodate the multi-product character of railways. Potential disadvantages of DEA are less problematic. For these reasons, we apply DEA in this research to measure the productive efficiency of railways.

4.3 Survey of the literature on railway efficiency measurement

In the former section, we described the two main approaches for measuring productive efficiency. Both approaches have been applied in the literature on estimating efficiency of railways. This paragraph provides a concise overview of these studies, concentrating on methodology as well as results. As a full treatment is beyond the scope of this paper, a selection has been made to provide the reader with a flavour of what is on offer.

The first study we address is by Gathon and Perelman (1992). The authors were among the first to apply SF to railways. They estimate a factor requirements frontier, for 19 European railways over the period 1961-1988, using a panel data approach, implicitly assuming the existence of complementarity (fixed proportions) between all the main inputs in production.²² ²³ Technical efficiency is assumed to be endogenously determined. Furthermore, special attention is devoted to an autonomy indicator representing managerial freedom with respect to authorities. They observe a high correlation between individual technical efficiency and this institutional indicator. Interestingly, higher load factors (measured as the number of passengers/tons by train) reduce efficiency. That is, for a given level of production (measured as the number of train-km), higher demand leads to more input requirements (for instance, more employees). This relationship points at diseconomies of density in the use of trains.

²² Gathon and Perelman consider this assumption to be plausible, since from an empirical analysis they find that labour expenses account for about 90% of the variable cost for all the railways. This result justifies, according to the authors, that the substitution possibilities between labour and energy are highly limited.

²³ A factor requirements function is a production technology in which a single input is expressed as a function of a number of outputs.

A few years later, Gathon and Pestieau (1995) decomposed productive efficiency into a management and a regulatory component. They estimate a translog production frontier to compute these efficiency indicators using displaced ordinary least squares. This method enables the authors to investigate for which part of efficiency slack management can be attributed, and for which part - beyond the control of the rail companies - the governments are responsible. The authors apply this procedure to 19 European railways over the period from 1961-88. Following Nishimizu and Page (1982), they argue that productivity gains are to be divided into two components, technical progress and efficiency changes, having different determinants.²⁴ Their results show that the Netherlands (NS) has the highest efficiency score in this period. In accordance with Gathon and Perelman (1992), the authors find that managerial autonomy is an important determinant of the government owned railway's performance.

Oum and Yu (1994) perform a comparative efficiency study of the OECD countries' railways. Like the last two studies, this study predates the reforms in Europe. Their data deals with the period 1978-89. The aim of this study is to identify the implications of public subsidy and the degree of managerial autonomy in technical performance. The authors estimate technical efficiency by using a DEA model, assuming constant returns to scale. Two alternative output measures are used: 1) revenue-output measures (passenger-/ton-km) and 2) available output measures (passenger/freight train-km). In order to estimate the effects of policy and other variables beyond the control of management, a Tobit regression model is applied. The main (new) finding of this paper was that railway systems with high dependence on public subsidies are significantly less efficient than similar railways with less dependence on public funds.²⁵ Again, the NS is among the most efficient performers.

²⁴ Efficiency is linked to the quality of management and to the institutional setting of railway operations, whereas technical progress is linked to R&D. Gathon and Pestieau estimate technical progress from the coefficients associated with a trend variable in their model.

²⁵ Following Nash and Rivera-Trujillo (2004), a word of caution is in order. It could be that the direction of causality is the other way around. That is, inefficient railways require high subsidies to survive, whilst high costs and low productivity might be the result of public service obligations to provide services such as peak commuter services which are costly but socially desirable.

Cowie and Riddington (1996) examine the methods of assessing rail efficiency. The authors note that as there is effectively no international trading and no common accounting practise, comparative international efficiency is best based on physical measures rather than value measures.²⁶ Cowie and Riddington, commenting on the studies previously mentioned, argue that there are clear reasons why the Dutch railways are more efficient than the Austrian railways (with a very low efficiency). According to them, this result follows from the high utilization of the infrastructure in the Netherlands (i.e., economies of density). Finally, the authors point at the problem that the results of the studies investigated in their article do not correlate with each other (except for the Dutch vs. Austrian regularity).

A paper by Cantos, Pastor and Serrano (1999) investigates the importance of output specification. Their results show that alternative output specifications lead to different results. Nonetheless, these differences can be brought substantially closer when output variables are corrected to account for the impact of the load factor. In this study, the authors use DEA.

Using data on 17 European railway companies during 1988-93, Coelli and Perelman (2000) estimate multi-output distance functions using corrected ordinary least squares. This approach is advocated because it avoids making unrealistic assumptions of firm behaviour, while at the same time it is able to handle the multi-output nature of railways. The distance function results are compared with those obtained from single-output production functions. They find that the results differ substantially between the two methods. As a result, the authors doubt the reliability of the single-output methods.

During the long period (1950s-1990s) of regulation, most railway companies improved their productivity level.²⁷ In contrast to this positive development, their financial state of affairs, defined as revenue over operating costs, usually got worse. These developments can only be

²⁶ As most European railways are not free to operate on purely commercial terms, the output measure should therefore not only reflect the physical nature of output, but also the public service obligations and the product they are actually selling (e.g., quality of service).

²⁷ This period stretches from the period of nationalisation in the 1950's to the deregulation measures undertaken in the 1990s.

explained by revenue growth falling behind the advances of productivity. Cantos and Maudos (2001) try to explain these facts by estimating both cost and revenue frontier functions (SF). In so doing, the authors are able to calculate the losses associated with both cost and revenue inefficiencies. The technique used to estimate efficiency is maximum likelihood. Their empirical analysis shows the existence of significant potential losses of revenue. They reason that this is caused by the strong policy of regulation and intervention by the governments in this period.²⁸ Consequently, Cantos and Maudos propose a light form of price regulation and a service adapted to market conditions if the companies' financial burdens are to be reduced.²⁹ They argue that it is time for a re-orientation from cost efficiency and productivity towards a policy focus on revenue. Policies such as concessions/franchises are regarded as positive, since they are compatible with the recommendations above.

There are relatively very few studies which extend efficiency analysis to the impact of rail restructuring in the 1990s. A study by Friebe, Ivaldi and Vibes (2003) investigates to what extent third-party access, independent regulation and the separation of infrastructure from operations affects railway performance of 11 European countries, over the period 1980-2000. Using production frontier analysis, the authors find that reforms that have efficiency improving effects are implemented sequentially, while reforms introduced in a package have at best neutral effects. Moreover, their results show that full separation is not a necessary condition for increasing efficiency. This result seems to conflict with the firm belief of many policy-makers. Interestingly, Friebe et al. (2003) find that all smaller countries, except for the Netherlands, have been able to keep or raise their efficiency levels. In addition, they argue that better data is needed. Especially, quality measures of output are not available.

²⁸ This argument can be explained as follows. Tightly regulated companies are not free to operate their companies on a purely commercial basis. As such, despite significant productivity improvements, companies may not increase their fares. The improvement in productivity may result in additional volume of traffic which reduces the average costs as result of economies of density, but this benefit may be insufficient to recover the investments generating the productivity improvements. As a result, the firm may face financial difficulties in spite of improved efficiency. Relaxing regulation, in particular regarding prices, would enable the firms to raise prices in order to increase financial returns. As traffic demand is highly inelastic, i.e. consumers have a high willingness to pay for train services, a rise in the price will hardly result in a reduction of traffic volume.

²⁹ Except for service that are socially desirable but not economically viable.

A recent study by Lan and Lin (2004) examines railway efficiency, effectiveness and productivity of 44 railway systems over the period 1995-2001, while controlling for environmental effects, data noise and slacks. In order to make these adjustments and to overcome the shortcomings of traditional DEA models, the paper proposes a four-stage DEA approach. Due to the non-storable nature of railway services, technical efficiency (a transform from outputs from inputs) and technical effectiveness (a transform of consumptions from inputs) represent two distinct measurements. That is, railway services' productive efficiency can be higher than its effectiveness (in terms of sales), because, once produced, outputs cannot be stockpiled for future sales. The authors find that the major decline of the rail industry (market share of passenger rail transport for the EU has declined from 32% in 1970 to 12% by 1999) should not be attributed to rail's poor performance in technical efficiency or service effectiveness; rather it is the consequence of higher level-of-service of other modes. In fact, their results indicate that the rail industry has a positive progress in recent years (1995-2001).

SF approach is applied by Rivera-Trujillo (2004). This methodology is justified by the arguments that it does not require a high availability and quality of data (which is complicated by rail reforms) and specific behaviour (e.g., cost minimisation). Furthermore, in order to take into account the multi-output characteristic of the rail industry, an input distance function is used. The study concentrates on freight transport during the period 1980-1999, which is the most important segment in North and South American railways. The countries included in the analysis are the United States, Canada, Brazil, Mexico and Chile. The results show that a great part of productivity improvement was due to technological change rather than technical efficiency change. Rivera-Trujillo notes that further research is needed to the selection and specification of the variables in order to obtain internationally agreed performance measures in the rail industry, as well as, on the whole period in which the recent rail reforms took place to determine their degree of success.

Summary

Technical efficiency can be measured by several approaches. Several of these methods have been applied in railways, although in different ways. As a result, it is difficult to compare the results of the studies. Measuring technical efficiency is a difficult task with many shortcomings. Therefore most studies should be taken with due caution and results are to be interpreted prudently. However, in general a picture emerges that the Netherlands (NS) had one of the most efficient railway companies in the world. Yet, this edge has deteriorated in the last decade. Other rail companies are "catching-up". One of the most relevant findings for this paper comes from Friebel, Ivaldi and Vibes (2003). They obtain the result that full separation of infrastructure from operations is not a necessary condition for improving railroad efficiency.

4.4 Data

In section 4.2, we have made the decision to use DEA to measure the relative productive efficiency of railways. The purpose of this section is to assess whether the data is suited to calculate meaningful DEA indices. For that purpose, we need to analyse the availability and the quality of the data. In particular, the comparability of railway systems is an important issue when applying DEA. Therefore, we need to end up with a relative homogenous set of countries. Furthermore, we have to select the appropriate input and output variables to accurately characterise the railway production process. This section starts with an overview of the data sources.

4.4.1 Data sources

The primary data source of this study is International Union of Railways (2003). The data from this source covers the period 1990-2001 for the railway systems of 52 countries from over the world. At present, it is the key source of information from which most industry analysts and academics obtain their information on railways. It is especially made to ensure comparability and consistency through the use of common definitions. However, in the end it is dependent on the quality of data provided by the individual railways. Supplemental data is received from the Norwegian Office of Statistics and a railway magazine from the Netherlands. Each railway system is represented by the main railway organisation in the specific country. When a railway system is made up of more than one organisation, we combine the operations to a single organisation. Despite the extensiveness of the data set, there are considerable data gaps and inconsistencies, severely limiting the number of railway systems available for assessment. In particular, the railway systems of the United Kingdom and Ireland could not be included as insufficient data was available. Nash and Shires (1999) find this issue a major source of concern. They argue that, due to institutional changes, railway operators are often reluctant to release details of their operations. This makes data collection difficult and reduces the scope for future research.

4.4.2 Country selection

Given the focus of this study, namely the analysis of the effect of the design of competition on productive efficiency, and the fact that we use DEA, we need to ensure that the group of countries is rather homogenous. That is, the railways systems have to be comparable in their production characteristics. From the initial set of countries, sixteen countries are available that have enough observations; that is, at most two observations are missing. Within this group, Japan, the United States and Luxembourg exhibit characteristics that make them somewhat different from the others.

Luxembourg is the smallest country in the data set. In Figure 4.3, this is reflected in the fact that Luxembourg's network size is almost ten times as small as the second smallest country (Denmark). As a consequence, Luxembourg has a unique scale that will always make it a fully efficient country. Even applying VRS DEA does not solve this problem, because Luxembourg has no countries, or in DEA terms peers, to be compared with. Including Luxembourg would not only limit the discriminatory power of the DEA analysis, it would also distort the analysis of the variation in efficiency scores in the subsequent regression analysis. Similarly, including the United States would be detrimental to the discriminatory power of the empirical analysis. Compared to the second largest country in the data set (Germany), United States are approximately six times as big in scale (see also Figure 4.3). So, both Luxembourg and the United States are left out of the data set in order to create a set of countries that are comparable.

Unlike Luxembourg and the United States, Japan is not an outlier in terms of scale. Instead, it is the country's composition of output that could pose a problem. Most countries in the data set are reasonably balanced, in the sense that they are not extremely biased towards either freight or passenger transport. Japan, is the most specialised in passenger transport. Figure 4.4 presents the composition of traffic of the countries in the data set. The table shows that over 90% of Japan's output is passenger transport. However, if we look at the statistics, it is clear that Japan is not the only country that is specialised in passenger transport. The second most specialised country in passenger services has a percentage of just 85%. As Japan is the only non-European country left in the data set, it could have a distinctively different environment in comparison to the European countries. Despite this special characteristic, we do not exclude Japan out of the data yet. Alternatively, we investigate whether including Japan distorts the DEA results presented in the following section.

Figure 4.3 Network size (in km)

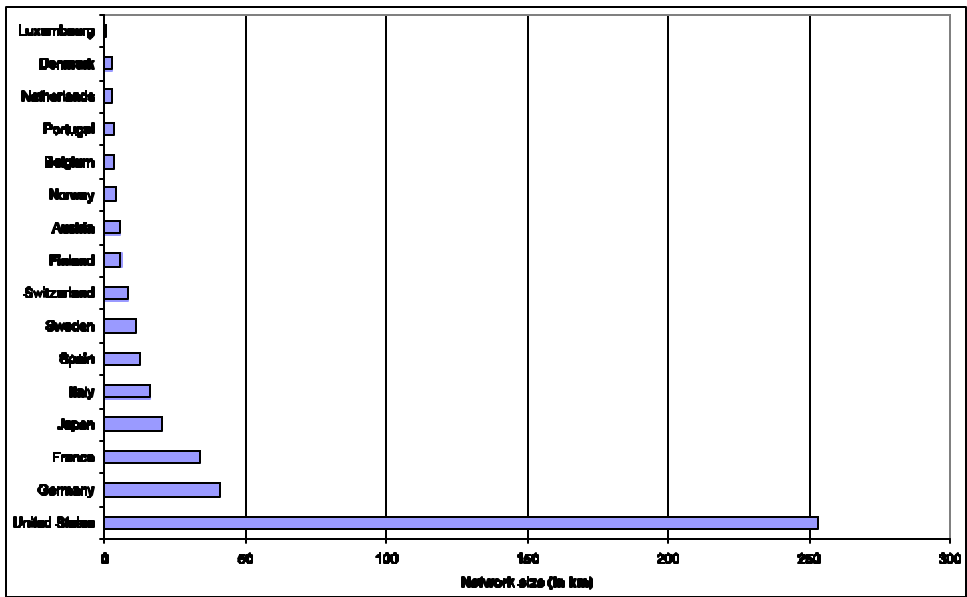
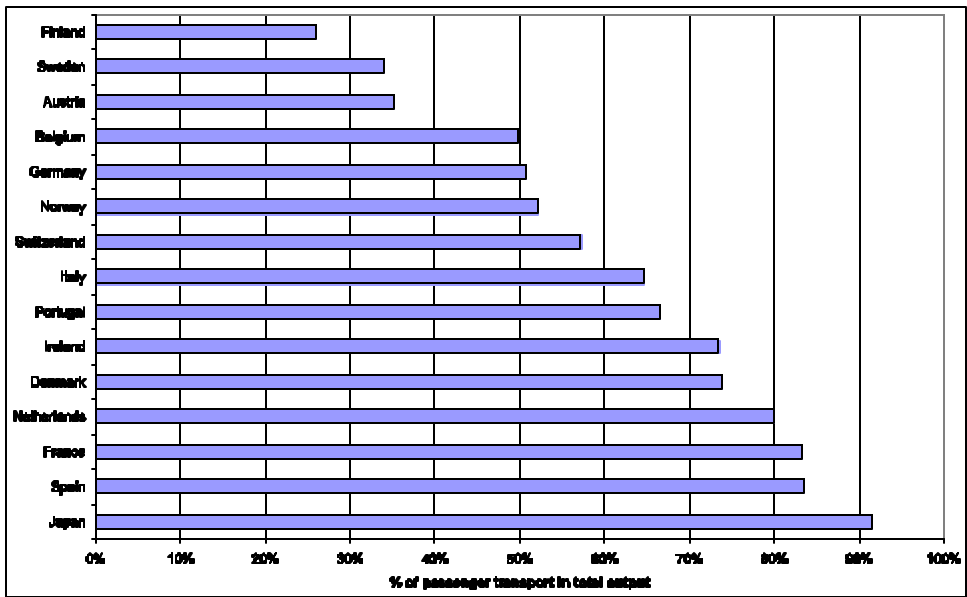


Figure 4.4 Share of passenger transport in total output (in %)



4.4.3 Input and output selection

The characteristics of the production process of railways are complex. As a consequence, measuring the performance of railways is also complex. In particular, the multiplicity of inputs and outputs poses some problems. With regard to output, railways produce the transport of passengers and freight. As a result, passenger-kilometres and tonne-kilometres are usually the starting point for measuring railway output. Although it would be simple to add these together to form a measure of output known as total traffic, this would be inappropriate, because the two

outputs require different combinations of inputs and have a different unit of measurement (people vs. freight). They tend to have inherently different cost structures (Productivity Commission, 1999). Fortunately, DEA is able to deal with the multiplicity of inputs and outputs. In fact, it is the reason why the method was developed in the first place. Concerning inputs, railway companies essentially use labour and capital to produce output (Affuso et al., 2002). Capital consists of rolling stock, tracks and stations.

As no common accounting practise exists among the different railway systems, comparative international efficiency analysis is best based on physical measures rather than value measures (Cowie and Riddington, 1996). Therefore, we use physical measures such as the amount of kilometres and employees instead of monetary measures like revenues and costs .

Ultimately, the choice of variables is constrained by the availability of data. This study uses two outputs (passenger-kilometres and tonne-kilometres) and three inputs (staff, track, and total rolling stock). Several previous studies have used the same measures of output and input. We prefer to use total rolling stock instead of the number of locomotives for the reason that the definition of locomotives is less comparable over the sample. Descriptive statistics of the input and output variables are presented in Table 4.1, where we have put Europe and Japan separately to facilitate possible comparison.

Table 4.1 Descriptive statistics of the data							
Variable	Symbol	Unit of measurement	Europe				Japan
			Mean	Min	Max	Standard deviation	Mean
Outputs							
Passengers kilometres	Fkm	Number of passenger kilometres (in millions)	19493	2104	74015	21607	182483
Freight kilometres	Fkm	Gross-hauled tonne-kilometres (in millions)	13583	1442	99914	18830	24288
Inputs							
Input of labour	L	Annual average number of staff	70425	6599	482269	87787	182483
Tracks	T	Total length of lines at the end of the year (in kilometres)	11238	2047	41718	11393	20198
Input of capital	C	Annual average number of rolling stock	55099	2992	438326	81061	48855
Dimensions:							
Countries: Austria, Belgium, Denmark, Finland, France, Germany, Italy, Japan, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland							
Period: 1990 -2001 (Denmark until 2000, Sweden until 1999)							
Source: UIC (2003)							

4.5 Results

This section presents estimates of relative productive efficiency for the 14 countries in our data set. The productive efficiency indices are estimated by using the Efficiency Measurement System (EMS) program by Holger Steel.³⁰ As was decided upon in the previous section, we show the DEA results including and excluding Japan. The purpose of this is to examine whether the inclusion of Japan affects the results.

³⁰ See EMS: Efficiency Measurement System User's Manual, available at www.wiso.uni-dortmund.de/lstfg/or/scheel/ems.

This study employs variable returns to scale DEA as most of the railway systems in our analysis are not operating at their optimal scale (see Preston, 1994). The variable returns to scale DEA estimates of both models are presented in Table 4.2 and Table 4.3. Several noteworthy aspects emerge from these results. First of all, most countries evolve over time towards a DEA index of 100%. This suggests that, in final years of the data set, the countries form a relatively homogenous group as their relative efficiency scores are identical. However, in the beginning of the nineties, significant differences were present between the countries. So, it seems likely that the railway systems have grown towards each other in terms of relative productive efficiency. Second, some countries, notably, Belgium, Netherlands, Japan, and Sweden, have the most efficient railway systems over the whole period (in terms of average DEA score over the whole period). Third, the difference between the results excluding and including Japan, is that a group of countries experiences a sizeable drop in their relative efficiency levels, because they are now compared to Japan. Looking at the two tables, it seems that especially countries such as France, Spain and Italy are affected by including Japan to the data set. Switzerland's results are also influenced to some extent. This signifies that Japan's input-output mix is comparable to these countries and that we need to check whether this Japan effect could disturb the robustness of the results in the second stage. That is, while Japan is interesting for the analysis, it may also be disturbing as it has a rather different railway system compared to the European countries.

Table 4.2 DEA estimates of productive efficiency, Europe, 1990 to 2001^a

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Austria	0.80	0.80	0.77	0.76	0.80	0.83	0.83	0.87	0.90	0.91	1.00	1.00
Belgium	1.00	1.00	0.99	0.96	1.00	0.96	0.95	0.96	0.96	0.95	0.97	0.94
Denmark	0.87	0.87	0.89	0.89	0.87	0.87	0.87	0.92	0.91	1.00	1.00	.
Finland	0.77	0.74	0.76	0.84	0.90	0.91	0.89	0.96	0.97	0.97	1.00	1.00
France	0.77	0.74	0.70	0.72	0.75	0.72	0.80	0.84	0.89	0.93	0.98	1.00
Germany	0.80	0.82	0.75	0.72	0.77	0.76	0.76	0.88	0.92	0.93	1.00	1.00
Italy	0.93	0.94	0.93	0.89	0.93	0.97	0.97	0.97	0.93	0.94	1.00	1.00
Netherlands	0.89	0.96	0.97	0.97	0.95	1.00	0.97	0.96	0.99	0.98	1.00	1.00
Norway	0.76	0.81	0.80	0.81	0.81	0.83	0.89	0.87	0.94	1.00	0.95	1.00
Portugal	0.67	0.68	0.68	0.84	0.88	0.90	0.89	0.89	0.88	0.74	0.91	0.95
Spain	0.51	0.50	0.52	0.59	0.59	0.64	0.68	0.75	0.82	0.88	0.93	1.00
Sweden	0.86	0.85	0.96	0.96	1.00	1.00	1.00	0.92	0.98	1.00	.	.
Switzerland	0.59	0.61	0.58	0.58	0.63	0.65	0.64	0.71	0.74	0.81	0.87	0.93

^a Variable Returns to Scale (VRS) efficiency scores

. = data not available

Source: CPB estimates

We have plotted the DEA results of both tables against each other in Figure 4.5. This figure allows comparing the efficiency levels of both models. Every country-year combination is represented by a point. From this figure, it appears that most countries have (nearly) equal efficiency levels in two models as most dots are on the 45 degree line. However, the group of countries described above has very different efficiency levels between the two models. These countries are represented by the points below the 45 degree line. So, including Japan does not only add another country to the data set, it also changes the results of a number of countries.

Concluding, the results indicate that it is sensible to investigate the sensitivity of the results to the inclusion of Japan in the following chapter. The next chapter contains the second-stage of the empirical analysis, where we seek to explain the differences between railways by considering the effects of various designs of competition.

Table 4.3 **DEA estimates of productive efficiency, Europe and Japan, 1990 to 2001^a**

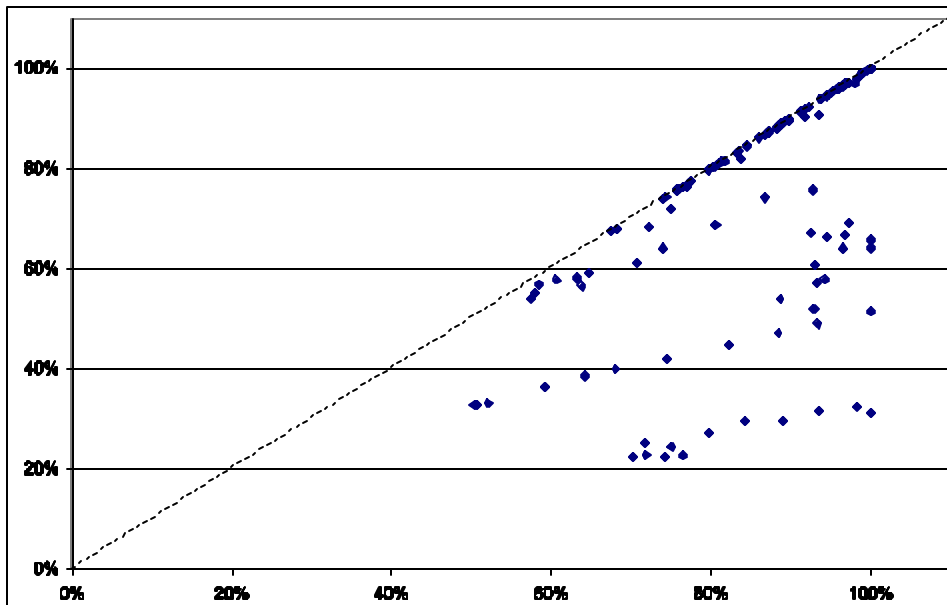
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Austria	0.80	0.80	0.77	0.76	0.80	0.83	0.83	0.87	0.90	0.91	1.00	1.00
Belgium	1.00	0.99	0.99	0.96	1.00	0.96	0.95	0.96	0.96	0.95	0.97	0.94
Denmark	0.87	0.87	0.89	0.89	0.87	0.87	0.87	0.92	0.91	1.00	1.00	.
Finland	0.77	0.74	0.76	0.84	0.90	0.91	0.89	0.96	0.97	0.97	1.00	1.00
France	0.23	0.22	0.22	0.23	0.24	0.25	0.27	0.29	0.30	0.31	0.32	0.31
Germany	0.80	0.81	0.72	0.68	0.76	0.76	0.76	0.88	0.92	0.91	1.00	1.00
Italy	0.52	0.58	0.57	0.54	0.61	0.67	0.64	0.69	0.67	0.66	0.66	0.64
Japan	1.00	1.00	1.00	0.99	0.97	0.99	1.00	1.00	0.98	0.99	0.99	1.00
Netherlands	0.89	0.96	0.97	0.97	0.95	1.00	0.97	0.96	0.99	0.98	1.00	1.00
Norway	0.76	0.81	0.80	0.81	0.81	0.83	0.89	0.87	0.94	1.00	0.95	1.00
Portugal	0.67	0.68	0.68	0.82	0.88	0.89	0.89	0.89	0.88	0.74	0.91	0.94
Spain	0.33	0.33	0.33	0.36	0.36	0.39	0.40	0.42	0.45	0.47	0.49	0.51
Sweden	0.86	0.85	0.96	0.96	1.00	1.00	1.00	0.90	0.97	1.00	.	.
Switzerland	0.57	0.58	0.55	0.54	0.58	0.59	0.57	0.61	0.64	0.69	0.74	0.76

^a Variable Returns to Scale (VRS) efficiency scores

. = data not available

Source: CPB estimates

Figure 4.5 Correlation between DEA results excluding and including Japan



5 Relationship between competition design and relative productive efficiency

5.1 Introduction

Using the DEA results, we now investigate the effects of the design of competition on productive efficiency. Contrary to the previous chapter, this chapter applies econometrics. In fact, we use a tool that is specially equipped to handle special data such as DEA indices. DEA scores are particular, because they are cut-off, or censored, at a value of 1. In other words, it is a dependent variable that is limited in its range. Nobel prize winner James Tobin was the first to consider the problems of this type of data (Tobin, 1958). He developed the Tobit model that treats censored observations differently than the rest of the observations.

This chapter is structured as follows. Section 5.2 describes the estimation method. Afterwards, section 5.3 presents the data. Finally, section 5.4 discusses the estimation results.

5.2 Estimation method

We use a limited dependent variable model to estimate the effects of the design of competition. Applying the usual least squares estimator in this case would result in biased results, since it fails in treating censored observations properly (Hill et al., 2001). The estimation method is further explained in subsection 5.2.1. After this, subsection 5.2.2 addresses the dummy variables that represent different designs of competition and other structural aspects of railway

systems. This is followed by subsection 5.2.3 that deals with the control variables. These variables allow controlling for certain environmental influences, which are to some extent beyond the control of the management of railway companies.

5.2.1 Tobit regression

The standard Tobit model is formalised as follows:

$$\begin{aligned}
 DEA_i^* &= z_i' \mathbf{b} + e_i, \quad i = 1, 2, \dots, N, \\
 DEA_i &= DEA_i^* \quad \text{if } DEA_i^* < 1 \\
 DEA_i &= 1 \quad \text{if } DEA_i^* \geq 1
 \end{aligned} \tag{5.1}$$

where e_i is assumed to be $NID(0, s^2)$ and independent of z_i . The first equation specifies a linear additive relationship between the unobserved DEA index of firm i , DEA_i^* , and observed characteristics, z_i . The unobserved dependent variable is commonly referred to as a latent variable and is indicated by an asterisk. The second line in this expression denotes that when the observations are not censored (from above) at one, the latent variable is equal to the observed variable, DEA_i . In contrast, when the observations are censored, as is the case in the third row, then all values are mapped to one. This model is also referred to as the censored regression model. Usually the parameters in Tobit models are estimated by the method of maximum likelihood (Verbeek, 2004).

Aside from their signs, the coefficients of Tobit models are not easy to interpret directly. One way to interpret the parameters is to consider the marginal effect of a change in variable k , z_{ik} , upon the expected DEA outcome. According to Verbeek (2004), this is simply given by the model's coefficient multiplied by the probability of having a non-censored outcome. The latter is expressed by the standard normal density function. Formally, this can be written as:

$$\frac{\partial E\{DEA_i\}}{\partial z_{ik}} = \mathbf{b}_k \Phi\left(\frac{z_i' \mathbf{b}}{s}\right) \tag{5.2}$$

The regression function to be estimated can be expressed as follows:

$$\begin{aligned}
 DEA_{it} &= CONSTANT + \mathbf{b}_1 VERT_{1it} + \mathbf{b}_2 VERT_{2it} + \mathbf{b}_3 THIRD_{it} + \mathbf{b}_4 TEND_{it} + \mathbf{b}_5 INDP_{it} + \\
 &\mathbf{b}_6 TIME_t + \mathbf{b}_7 AREA_{it} + \mathbf{b}_8 POPDEN_{it} + \mathbf{b}_9 GDP_{it} + \mathbf{b}_{10} TDENS_{it} + \mathbf{b}_{11} TSTRUC_{it} + \\
 &\mathbf{b}_{12} (DUMJAP_i) + e_{it}
 \end{aligned}$$

where DEA is the dependent variable and e is a random term, which is assumed to be symmetrically distributed with zero mean and constant variance. All other variables are introduced below. Most terms differ across countries and time and are thus indexed by it . Note that a time trend, $TIME_t$, is included to represent technological progress.

5.2.2 Dummy variables

Although it is always difficult to capture the complexity of the real world through discrete variables, dummy variables are useful instruments to test the effects of the different designs of competition on productive efficiency. Essentially, each dummy variable characterises a particular design of competition. Table 5.1 provides an overview of the dummy variables we use. Two dummy variables are designed to reflect the vertical structure of the railway system in a country. We distinguish between institutional (full), $VERT1$, and accounting (partial) separation, $VERT2$, to investigate whether there is a difference between the two options.

Table 5.1 Description of regression variables

Variable	Symbol	Description
Dummy variables		
Institutional (or full) separation	VERT1	If variable is 1, then infrastructure and services are institutionally separated; 0 if this is not the case.
Accounting (or partial) separation	VERT2	If variable is 1, then infrastructure and services are separated on an accounting basis; 0 if this is not the case
Third party access	THIRD	If variable is 1, then legislation is transposed that allows third party access to competitors (either freight or passenger) and competition has evolved to a significant extent; 0 if this is not the case. ^a
Competitive tendering	TEND	If variable is 1, then competitive tendering is used to procure regional railway franchises; 0 if this is not the case.
Managerial independence from the government	INDP	If variable is 1, then legislation is transposed that assures independent management from the government of railway companies; 0, if this is not the case. ^b
Japan dummy	DUMJAP	If variable is 1, then country is Japan; 0, if this is not the case
Control variables		
Total area	AREA	Measured in 1000 square miles
Gross Domestic Product per capita	GDP	Measured in constant prices (2000) 1000 US dollars PPPs
Population density	POPDEN	Measured in population per square mile
Traffic structure	TSTRUC	measured by passenger kilometres / total traffic in kilometres
Traffic density	TDEN	Total traffic in kilometres (in millions) / total length of lines in kilometres

^a Significant evolution of competition implies that competitors of the incumbent obtain sufficient and nontrivial large market shares. We work with a threshold value of 1%. Admittedly, this value is rather arbitrary, but required for this analysis.

^b Managerial independence as it is prescribed by the European Union by directive 91/440/EEC.

Competition ‘in’ and ‘for’ the market are represented by third party access, *THIRD*, and competitive tendering, *TEND*, respectively. Further, managerial independence from the government of the railway company is captured by a specific dummy variable, *INDP*. Finally, to identify the unique characteristics of the Japanese railway system, we use a special dummy variable for this country, *DUMJAP*.

Unfortunately, features such as horizontal structure (freight and passenger integrated/separated), ownership (private or public), infrastructure competition and yardstick competition, could not be included as a dummy variable, because of too little variance in these features in our data set.³¹ As a consequence, it is not (yet) possible to investigate the influence of these variables on productive efficiency in this research. With the help of better data sets, which encompass a larger set of countries and more recent data, academics should be able to investigate these aspects empirically.

5.2.3 Control variables

In order to arrive at accurate results of the effects of the designs of competition, we need to correct for particular environmental variables outside the control of the management of the railway firm. Environmental variables are aspects of the environment of a railway system that affect the efficiency of production of companies. It may happen that some managers are required to operate within a different environment from others. Moreover, the environment of railway companies changes over time. The purpose of control variables is to correct for these environmental influences when measuring the effects of the design of competition.

The control variables included in this analysis and their definitions are listed in Table 5.1. The first factor we control for is the influence of population density, *POPDEN*. High population density might facilitate (in terms of efficiency) a more intense use of inputs than would otherwise be the case. The second control variable is GDP per capita, *GDP*. Higher purchasing

³¹ Estimations including these variables did not deliver any meaningful results and are therefore not published in this study.

power, through income effects, could result in a higher mobility level of the customers of a railway system.

Besides these two factors, we also control for the effects of country size, *AREA* (facilitates economies of size), traffic structure, *TSTRUC* (to correct for the cost difference between producing freight and passenger services), and traffic density, *TDEN* (facilitates economies of density, in so far this is not captured by population density and income per capita). Note that traffic density is not fully beyond the control of management, since firms can influence the density of traffic by changing their supply of services. Nevertheless, social and regulatory objectives affect to a large extent whether and how often certain lines are to be operated, thereby influencing traffic density.

Although we purify the results from several important exogenous factors, an infinite list of other aspects cannot be dealt with due to data unavailability. To mention a few, hilliness, climate, topography, unionization, electrification, inter-modal competition and so on. These residual environmental elements are not accounted for. This might have an effect on our results. More on data in the following section.

5.3 Data

In this section, we describe the data we use in explaining the DEA results of the first section. In particular, the data of the dummy and control variables are discussed.

The data for the dummy variables are constructed through the use of a great variety of sources. For the exact list of materials used to construct this set, the reader is referred to the bottom of Table 5.2. The quality of the dummy data set is verified through a survey we did under a number of international experts of the railway systems considered in this study. Once more, note that there are certain limits concerning the extent to which one can interpret the results, because there are many reform specificities across countries that cannot be perfectly operationalised empirically (Friebel et al., 2003). We provide an overview of all changes in competition design for every country in our data set in Table 5.2. This table also includes aspects of structural design that we do not assess empirically due to lack of variation in the data.

In the previous section, the control variables were introduced. To give a sense of the range in scale of the data, we provide Table 5.3. This table gives the descriptive statistics of the five control variables included in this study. The sources of the data are U.S. Department of Justice (2003), OECD (2005), and UIC (2003).

Table 5.2 Structural design of the railway industry in European countries and Japan, 1990 - 2001

Country	Characteristic in 1990	Changes during 1990 – 2001
Austria	vertically and horizontally integrated, no competition, state-owned, public agency	- accounting separation in 1992 - independent management in 1992
Belgium	vertically and horizontally integrated, no competition, state-owned, public agency	- accounting separation in 1993 - independent management in 1991
Denmark	vertically and horizontally integrated, no competition, state-owned, public agency	- institutional separation in 1997 - independent management in 1999
Finland	vertically and horizontally integrated, no competition, state-owned, public agency	- institutional separation in 1995
France	vertically and horizontally integrated, no competition, state-owned, public agency	- accounting separation in 1995 - institutional separation in 1997 - independent management in 1997
Germany	vertically and horizontally integrated, no competition, state-owned, public agency	- accounting separation in 1994 - third-party access (freight) in 1994 - regional tendering in 1996 - independent management in 1993
Italy	vertically and horizontally integrated, no competition, state-owned, public agency	- accounting separation in 1998 - independent management in 1992
Japan	vertically integrated, horizontally separated, infrastructure competition, public agency	- yardstick competition in 1997
Netherlands	vertically and horizontally integrated, no competition, state-owned, public agency	- accounting separation in 1995 - horizontal separation in 2000 - third-party access (freight) in 1998 - regional tendering in 1999 - independent management in 1995
Norway	vertically legally separated, horizontally integrated, no competition, state-owned, public agency	- independent management in 1997
Portugal	vertically and horizontally integrated, no competition, state-owned, public agency	- institutional separation in 1997 - independent management in 1997
Spain	vertically and horizontally integrated, no competition, state-owned, public agency	- accounting separation in 1997 - independent management in 1994
Sweden	vertically institutionally separated, horizontally integrated, regional tendering, state-owned, independent management	- third-party access (freight) in 1996
Switzerland	vertically and horizontally integrated, no competition, state-owned, independent management	- accounting separation in 1999 - third-party access (freight) in 1999

Sources: Alexandersson and Hultén (2005); Berne and Pogorel (2003); Farsi et al. (2005); IBM (2004); Mizutani and Nakamura (2004); Nilson (2002); OECD (1998) OECD (2005); Nash (2004); Thompson (2003); UIC (2003); United Nations (2003); Van de Velde (1999); various websites of rail companies and DG Transport of European Union.

Table 5.3 Descriptive statistics of the control variables

Variable	Symbol	Unit of measurement	Europe				Japan
			Mean	Min	Max	Standard deviation	Mean
Total area	AREA	1000 Square miles	89.721	11.672	210.668	70.103	152,411
Gross Domestic Product per capita	GDP	In constant prices (2000) 1000 US dollars PPPs	23.287	6.765	37.113	5.645	24,913
Population density	POPDEN	Population per square mile	396	36	1224	325	824
Traffic structure	TSTRUC	passenger kilometres / total traffic in kilometres	0.583	0.227	0.897	0.199	0,910
Traffic density	TDEN	Total traffic in kilometres (in millions) / total length of lines in kilometres	3.045	1.153	7.100	1.361	13,342

Dimensions:

Countries: Austria, Belgium, Denmark, Finland, France, Germany, Italy, Japan, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland

Period: 1990 -2001 (Denmark until 2000, Sweden until 1999)

Source: U.S. Census Bureau (2003), OECD (2005), UIC (2003)

5.4 Results

This section uses Tobit regression to identify the effects of the designs of competition on productive efficiency, while at the same time netting out the impact of an array of exogenous factors. In what follows we present and discuss Tobit estimates.

Table 5.4 presents the regression results of two Tobit models.³² The difference between the first and the second model is the inclusion of Japan. In order to account for the possible particularities of this country a dummy variable is included in the second regression. Note, however, that even with a dummy variable, differences can be substantial between the results of the two models, because including Japan also affects the efficiency scores of some other

³²The estimations were done using Eviews 5.

countries as we have seen in chapter four. Both the regression coefficient and the marginal effect of each variable are reported in the table. The calculation of the marginal effect is explained in section 5.2.

Table 5.4 Tobit regression results

Model	(1) Europe			(2) Europe + Japan		
Dependent variable						
DEA efficiency indices						
Independent variables	Coefficient estimate	(Standard error)	Marginal effect	Coefficient estimate	(Standard error)	Marginal effect
CONSTANT	0.5827	(0.0493) ***	0.4643	1.0987	(0.0651) ***	0.8746
VERT1	0.0447	(0.0231) *	0.0356	-0.0005	(0.0301)	-0.0004
VERT2	0.0225	(0.0213)	0.0179	0.0854	(0.0282) ***	0.0680
THIRD	-0.0812	(0.0311) ***	-0.0647	-0.0773	(0.0417) *	-0.0615
TEND	0.0826	(0.0346) **	0.0658	0.2641	(0.0461) ***	0.2102
INDP	-0.0691	(0.0181) ***	-0.0551	-0.1495	(0.0239) ***	-0.1190
TIME	0.0211	(0.0033) ***	0.0168	0.0162	(0.0040) ***	0.01290
AREA	0.0002	(0.0001)	0.0002	-0.0018	(0.0002) ***	-0.0014
POPDEN	-7.75×10^{-5}	(5.45×10^{-5})	-6.18×10^{-5}	5.39×10^{-5}	(7.22×10^{-5})	4.30×10^{-5}
GDP	0.0016	(0.0014)	0.0013	-0.0012	(0.0018)	-0.0010
TDEN	0.0776	(0.0118) ***	0.0618	0.0261	(0.0154) *	0.0208
TSTRUC	-0.1331	(0.0481) ***	-0.1061	-0.5422	(0.0636) ***	-0.4316
DUMJAP				0.3929	(0.1477) ***	0.3131
Log likelihood		127.13			93.1	
Adjusted R-squared		0.67			0.82	
Number of observations		153			165	

Notes: Asterisks (*), (**), (***) represents statistical significance from zero at the 10%, 5%, and 1% level respectively.

The adjusted R-squared statistics indicate that both models perform well in explaining the variation of efficiency scores. The majority of coefficients are statistically different from zero at the 5% level of significance. In addition, multi-collinearity does not pose a problem, because the correlation among the explanatory variables is reasonably low. As we use a censored limited dependent variable, heteroskedasticity cannot be a problem. That is, the size of the variation in the residuals of the DEA indices is on average the same across countries.

Concerning control variables, the results show that:

1. the effect of both population density and income per capita are insignificantly different from zero,

2. traffic density is important for productive efficiency,
3. railways systems that concentrate on passenger transport have lower levels of productive efficiency.

The second result is in line with the literature. High utilisation of the track network is important for productive efficiency, due to economies of density. Also the third result is consistent with the literature. As aforementioned, one passenger-kilometre of output costs much more than one tonne-kilometre of output. Even though DEA corrects this somewhat, results show that a control variable is necessary. Further, when economies of density are accounted for by a traffic density variable, both population density and income effects do not matter for productive efficiency.

The results regarding the designs of competition that are common to the two estimated models can be summarised as follows:

1. competitive tendering improves productive efficiency,
2. third party access lowers productive efficiency,
3. increased managerial independence from the government is detrimental to productive efficiency.

The first result is in line with general expectations of this design of competition. In contrast, the second result is at first sight a bit awkward. The effect of third party access is to lower productive efficiency of the incumbent railway company. A possible explanation for this result could be that third party access disturbs the efficient operation of railway services and lowers economies of density of the incumbent. For instance, sharing terminal space and traffic may reduce the efficiency of the incumbent's activities. In addition, train scheduling can become less flexible due to competition on the tracks (BRTE, 2003).

Perhaps even more surprising is the third finding that more autonomy of management is bad for productive efficiency. However, if one looks at the national railway markets in Europe this effect might not come as a bolt from the blue. Most of the incumbent railway companies are state owned and do not face any competitive pressure. As a consequence, increased independence without sufficient competition and adequate regulation may deteriorate incentives for productive efficiency. For instance, agency problems could be harder to solve as

information asymmetry increases due to more autonomy. This point is also proposed by Vickers and Yarrow (1991), De Fraja (1991), and Caves and Christensen (1980).

Notice that the second and third result are in conflict with previous studies. This literature found that third party access (Friebel et al., 2003) and more independence (e.g., Gathon and Pestieau, 1995) are efficiency improving. The different definitions of third party access and managerial independence between those studies and the present study could possibly explain the divergence in the results. For instance, Gathon and Pestieau use an autonomy index based on a questionnaire created in 1990 for their managerial independence variable, while we use a dummy variable to represent whether a country has implemented legislation to create more managerial independence. In contrast to the paper by Gathon and Pestieau, our dummy variable is year specific. Moreover, we use different data and estimation method to measure productive efficiency. For example, Friebel et al. (2003) use SF, while we use DEA analysis. Lastly, the period under investigation by Gathon and Pestieau (1986 to 1988) is different from ours (1990 to 2001).

The results regarding vertical separation are not consistent between the two models. Whereas in the first model institutional separation is needed to get a positive effect on productive efficiency, the second model tells that accounting separation is sufficient. So, while the results suggest that separation could be beneficial for productive efficiency, the results disagree on which form of separation is preferred.

While the signs of the coefficients are mostly robust across the two models, the order of magnitude of the coefficients alters considerably when Japan is included. This Japan effect is caused by the drop in the DEA scores of some European countries explained in the previous chapter. For this reason, the results regarding the order of magnitude of the coefficients in the second model are disturbed and therefore to be interpreted with due caution.

6 Conclusion

In this study, we use a two stage approach to identify the effects of different designs of competition on relative productive efficiency of railways.

The first stage, using DEA, measures relative productive efficiency of the railways of a number of European countries and Japan. The results show a large but decreasing variation in the efficiency scores across countries over the period.

In the second stage of the empirical analysis, we applied Tobit regression to investigate the effects of various designs of competition on the efficiency results derived in the first stage. The main results are, first of all, that competitive tendering encourages productive efficiency, which is in line with the first hypothesis postulated in the introduction. Secondly, third party access tends to lower productive efficiency. This result suggests that the trade-off proposed in the second hypothesis leans towards the negative effects of this design of competition. Finally, increased managerial independence from the government delivers the opposite result to what was expected from the third hypothesis. Noticeably, the last two results are in contradiction with the small existing empirical literature on this issue. Factors that could possibly explain this divergence are differences in the definition of variables, data and estimation method. As regards to vertical separation, we find no unambiguous results as to which extent separation has to be implemented.

The results concerning the environmental variables indicate that both economies of density and the composition of output are important in explaining variation in productive efficiency. On the contrary, the impact of population density and income per capita is insignificantly different from zero.

What can be deduced from these results for policy purposes? First of all, the results indicate that the policy tendency towards competitive tendering seems to improve productive efficiency. In addition, the results show that introduction of competition on the tracks may not improve productive efficiency. However, the costs in terms of lower productive efficiency due to

competition on the tracks should be weighed up against the possible benefits in terms of higher allocative efficiency due to more competition. In particular, competition on the tracks should only be introduced where scale and traffic density allow for efficient duplication of services. Finally, providing the management of an incumbent railway company more autonomy could be an unwise step when competitive and regulatory forces may be insufficient to provide appropriate incentives for productive efficiency (e.g., De Fraja, 1991; Caves and Christensen, 1980; Vickers and Yarrow, 1991). For instance, agency problems could be harder to solve as information asymmetry increases due to more autonomy.

Limitations to this study are predominantly a result of data availability. Due to limitations to the data, we were not able to include an interesting country such as the United Kingdom. Furthermore, our analysis is constrained to the period 1990-2001. Consequently, the effects of recent structural measures may not be fully materialised. For instance, full separation of infrastructure and train services in the Netherlands did not occur until 2002. Another important issue is that we only investigate performance in terms of input and output quantities and therefore disregard important facets such as the quality of service and financial affairs. These are recommended avenues for future research.

Bibliography

Affuso, L.; Angeriz, A. and M. Pollitt, (2002), “Measuring the Efficiency of Britain’s Privatised Train Operating Companies”, Regulation Initiative Discussion Paper Series, no. 48.

Aghion, P.; Dewatripont, M. and P. Rey, (1995), “Competition, Financial Discipline and Growth”, mimeo, Universite Libre de Bruxelles.

Aghion, P.; Harris, C. and J. Vickers, (1997), “Competition and Growth with Step-by-Step Innovation: An Example”, *European Economic Review*, vol. 41, pp. 771-782.

Aghion, P.; Bloom, N.; Blundell, R.; Griffith, R. and P. Howitt, (2005), “Competition and Innovation: An Inverted-U Relationship”, *The Quarterly Journal of Economics*, vol. 120, pp. 701-728.

Aigner, D.J.; Lovell, C.A.K. and P. Schmidt, (1977), “Formulation and Estimation of Stochastic Frontier Production Function Models”, *Journal of Econometrics*, vol. 6, pp. 21-37.

Alexandersson, G. and S. Hultén, (2005), “Swedish Railways: From Deregulation to Privatisation and Internationalisation in A European Context”, Paper Presented at the Third Conference on Railroad Industry Structure, Competition, and Investment, Stockholm School of Economics, Stockholm, Sweden, October 20-22.

Arrow, K., (1962), “Economic Welfare and the Allocation of Resources”, in *The Rate and Direction of Inventive Activity*, National Bureau of Economic Research, ed. Princeton, NJ: Princeton University Press.

Banker, R.D.; Charnes, A. and W.W. Cooper, (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis", *Management Science*, vol. 30, pp. 1078-1092.

Baumol, W.J.; Panzar, J.C. and R.D. Willig, (1982), *Contestable Markets and the Theory of Industry Structure*, Harcourt Brace, New York.

Berne, M. and G. Pogorel, (2003), "Privatization Experiences in France", CESifo Working Paper, no. 1195.

Boone, J., (2000), "Competition", Center for Economic Research, no. 2000-104.

Boone, J., (2003), "Optimal Competition: A Benchmark for Competition Policy", Discussion Paper Series, no. 3766, Centre for Economic Policy Research.

Bureau of Transport and Regional Economics (BRTE), (2003), *Rail Infrastructure Pricing: Principles and Practise*, Report 109, Canberra ACT.

Cabral, L.B., (2000), *Introduction to Industrial Organization*, MIT Press: Cambridge, MA.

Cantos, P.; Pastor, J.M. and L. Serrano, (1999), "Productivity, Efficiency and Technical Change in the European Railways: a non-parametric approach", *Transportation*, vol. 26, pp. 337-357.

Cantos, P. and J. Maudos, (2001), "Regulation and Efficiency: the case of European Railways", *Transportation Research Part A*, vol. 35, pp. 459-472.

Campos, J. and P. Cantos, (2000), "Rail Transport Regulation", in de Rus, G. and A. Estache (eds.) *Privatisation and regulation of transport infrastructures: guidelines for policymakers and regulators*, The World Bank, Washington, D.C.

Caves, D.W. and L.R. Christensen, (1980), "The Relative Efficiency of Public and Private Firms in a Competitive Environment: The Case of Canadian Railroads", *The Journal of Political Economy*, vol. 88, pp. 958-976.

Charnes, A.; Cooper, W. and E. Rhodes, (1978), "Measuring the Efficiency of Decision Making Units", *European Journal of Operational Research*, vol. 2, pp. 429-444.

Coelli, T.; Prasado Rao, D.S. and G.E. Battese, (1998), *An Introduction to Efficiency and Productivity Analysis*, Kluwer Academic Publishers, Boston/Dordrecht/London.

Coelli T.J. and S. Perelman (2000), "Technical Efficiency of European Railways: a Distance Function Approach", *Applied Economics*, vol. 32, pp. 1967-1976.

Cowie, J. and G. Riddington, (1996), "Measuring the Efficiency of European Railways", *Applied Economics*, vol. 28, pp. 1027-1035.

Crampes, C. and A. Estache, (1997), “Regulatory Trade-offs in Designing Concession Contracts for Infrastructure Networks”, Policy Research Working Paper, no. 1854, The World Bank.

Debreu, G., (1951), “The Coefficient of Resource Utilisation”, *Econometrica*, vol. 19, pp. 273-292.

De Fraja, G., (1991), “Efficiency and Privatisation in Imperfectly Competitive Industries”, *The Journal of Industrial Economics*, vol. 39, pp. 311-321.

Demsetz, H., (1968), “Why regulate utilities?”, *Journal of Law and Economics*, vol. 11, pp. 55-65.

Di Pietrantonio, L. and J. Pelkmans, (2004), “The Economics of EU Railway Reform”, Bruges European Economic Policy Briefings, BEEP briefing no. 8.

Disney, R.; Haskel, J. and Y. Heden, (2000), “Restructuring and Productivity Growth in UK Manufacturing”, Queen Mary and Westfield College, Unpublished Manuscript.

Du Reitz, G., (1975), “New Firm Entry in Swedish Manufacturing Industries during the Post-War Period”, Doctoral Dissertation, Stockholm.

European Commission, (1991), *Directive 91/440/EEC*, Official Journal 24th August 1991.

Farrell, M.J., (1957), "The Measurement of Productive Efficiency", *Journal of Royal Statistics Society, Series A*, Part 3, pp. 253-290.

Farsi, M.; Filippini, M. and W. Greene, (2005), "Efficiency Measurement in Network Industries: Application to the Swiss Railway Companies", *Journal of Regulatory Economics*, vol. 28, pp. 69-90.

Friebel G.; Ivaldi, M. and C. Vibes, (2003), "Railway (De)Regulation: a European Efficiency Comparison, IDEI report, no. 3 on passenger rail transport, University of Toulouse.

Gathon, H.-J. and S. Perelman, (1992), "Measuring Technical Efficiency in European Railways: A Panel Data Approach", *Journal of Productivity Analysis*, vol. 3, pp. 135 - 151.

Gathon, H.-J. and P. Pestieau (1995), "Decomposing Efficiency into its Managerial and its Regulatory Components: The Case of European Railways", *European Journal of Operational Research*, vol. 80, pp. 500-507.

Hart, O.D., (1983), "The Market Mechanism as an Incentive Scheme", *Bell Journal of Economics*, vol. 14, pp. 366-382.

Hermalin, B.E., (1992), "The Effects of Competition on Executive Behaviour", *Rand Journal of Economics*, vol. 23, pp. 350-365.

Hill, R.C.; Griffiths, W.E. and G.G. Judge, (2001), *Undergraduate Econometrics, Second Edition*, John Wiley and Sons, New York.

Holmstrom, B., (1982), “Managerial Incentive Problems - A Dynamic Perspective”, in *Essays in Economics and Management in Honor of Lars Wahlbeck*, Helsinki: Swedish School Economics.

IBM and Humboldt University of Berlin, (2004), *Rail Liberalisation Index 2004*, IBM Business Consulting Services and Dr. Christian Kirchner, Humboldt University of Berlin, Berlin.

Jensen, M. and W. Meckling, (1976), “Theory of the Firm: Managerial Behaviour, Agency Costs, and Capital Structure”, *Journal of Financial Economics*, vol. 3, pp. 305-360.

Jovanovic, B., (1982), “Selection and Evolution of Industry”, *Econometrica*, vol. 50, pp. 649-670.

Kessides, I.N. and R.D. Willig, (1995), “Restructuring Regulation of the Rail Industry for the Public Interest”, Policy Research Working Paper, no. 1506, The World Bank, Washington, D.C.

Klein, M., (1996), “Competition in Network Industries”, Policy Research Working Paper, no. 1591, The World Bank, Washington, D.C.

Klemperer, P., (2002), “What Really Matters in Auction Design”, *Journal of Economic Perspectives*, vol. 16, pp. 169-189.

Koopmans, T.C., (1951), "An Analysis of Production as an Efficient Combination of Activities", in T.C. Koopmans (ed.) *Activity Analysis of Production and Allocation*, Cowles Commission for Research in Economics, Monograph no. 13, Wiley, New York.

Laffont, J.-J. and J. Tirole, (1993), *A Theory of Incentives in Procurement and Regulation*, MIT Press, Cambridge, MA.

Lan, L.W. and E.T.J. Lin, (2004), "Measuring the Railway Efficiency, Effectiveness, Productivity and Sales Force with Adjustment of Environmental Effects, Data Noise and Slacks", unpublished paper.

Loeb, M. and W. Magat, (1979), "A Decentralized Method for Utility Regulation", *Journal of Law and Economics*, vol. 22, pp. 399-404.

Mankiw, N.G. and M.D. Whinston, (1986), "Free Entry and Social Inefficiency", *Rand Journal of Economics*, vol. 17, pp. 48-58.

Mansfield, E., (1962), "Entry, Gibrat's Law, Innovation, and the Growth of Firms", *American Economic Review*, vol. 52, pp. 1023-1051.

McMillan, M.L. and D. Datta, (1998), "The Relative Efficiencies of Canadian Universities: A DEA perspective", *Canadian Public Policy - Analyse de Politiques*, vol. 24, pp. 485-511.

McMillan, M.L. and W.H. Chan, (2005), “University Efficiency: A Comparison and Consolidation of Results from Stochastic and Non-Stochastic Methods”, University of Alberta, Working Paper Series, no. 2005-04.

Meyer, M.A. and J. Vickers, (1994), “Performance Comparisons and Dynamic Incentives”, unpublished paper, Oxford University.

Meyer, M.A. and J. Vickers, (1995), “Performance Comparisons and Dynamic Incentives”, Discussion Paper no. 1107, London: Centre Economic Policy Research.

Mizutani, F. and K. Nakamura, (2004), “The Japanese Experience with Railway Restructuring”, in Ito, T. and A.O. Krueger (eds.) *Governance, Regulation, and Privatization in the Asia-Pacific Region*, The University of Chicago Press, Chicago, U.S.A., pp. 305-336.

Motta, M., (2004), *Competition Policy: Theory and Practice*, Cambridge University Press.

Nalebuff, B. and J.E. Stiglitz, (1983), “Prizes and Incentives: Toward a General Theory of Compensation and Competition”, *Bell Journal of Economics*, vol. 14, pp. 21-43.

Nash, C.A. and J. Shires, (1999), “Benchmarking of European Railways: An Assessment of Current Data and Recommended Indicators”, Institute for Transport Studies, Conference on Transport Benchmarking: Methodologies, Applications and Data Needs, Paris.

Newberry, D.M., (1999), *Privatization, Restructuring, and Regulation of Network Utilities*, MIT Press: Cambridge, MA.

Nickell, S., (1996), "Competition and Corporate Performance", *Journal of Political Economy*, vol. 104, pp. 724-746.

Nickell, S; Nicolitsas, D. and N. Dryden, (1997), "What Makes Firms Perform Well?", *European Economic Review*, vol. 41, pp. 783-796.

Nilsson, J.E., (2002), "Restructuring Sweden's Railways: The Unintentional Deregulation", *Swedish Economic Policy Review*, vol. 92, pp. 229-254.

Nishimizu, M. and J.M. Page, (1982), "Total Factor Productivity Growth, Technological Progress and Technical Efficiency Change: Dimensions of Productivity Change in Yugoslavia 1965-78", *The Economic Journal*, vol. 92, pp. 920-936.

OECD, (1998), *Railways: Structure, Regulation and Competition Policy*, Competition Policy Roundtables, no. 15, Paris.

OECD, (2005), *National Accounts*, Department of Economics and Statistics, OECD, Paris.

Olley, G.S. and A. Pakes, (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica*, vol. 64, pp. 1263-1297.

Ordover, J.A. and R. Pittman, (1994), "Restructuring the Polish Railway for Competition", Economic Analysis Group Discussion Paper EAG 93-10, September 13, 1993, Revised version in OECD (1994), U.S. Department of Justice, Antitrust Division.

Oum, T.H. and C. Yu, (1994), "Economic Efficiency of Railways and Implications For Public Policy: A Comparative Study of the OECD Countries' Railways", *Journal of Transport Economics and Policy*, vol. 28, pp. 121-138.

Oum, T.H.; Waters Li, W.G. and C. Yu, (1999), "A Survey of Productivity and Efficiency Measurement in Rail Transport", *Journal of Transport Economics and Policy*, vol. 28, pp. 121-138.

Panzar, J.C. and R.D. Willig, (1981), "Economies of Scope", *American Economic Review*, vol. 71, pp. 268-272.

Pittman, R., (2000), "Railway Competition: Options for the Russian Federation", unpublished.

Pittman, R., (2005), "Structural Separation to Create Competition? The Case of Freight Railways", *Review of Network Economics*, vol. 4, pp. 181-194.

Posner, R., (1975), "The Social Costs of Monopoly and Regulation", *Journal of Political Economy*, vol. 83, pp. 807-827.

Preston, J., (1994), "Does size matter? A Case Study of Western European Railways", unpublished paper, Institute of Transport Studies, University of Leeds.

Productivity Commission, (1999), *Progress in Rail Reform: an Assessment of the Performance of Australian Railways 1990-1998*, Melbourne.

Riordan, M.H. and D.E.M. Sappington, (1987), "Awarding Monopoly Franchises", *The American Economic Review*, vol. 77, No. 3, pp. 375-387.

Rivera-Trujillo, C., (2004), "Measuring Technical Efficiency in North and South American Railways using a Stochastic Frontier Model: An International Comparison", Institute for Transport Studies, University of Leeds.

Scharfstein, D., (1988), "Product-Market Competition and Managerial Slack", *Rand Journal of Economics*, vol. 19, pp. 147-155.

Scherer, F.M. and D. Ross, (1990), *Industrial Market Structure and Economic Performance*, 3rd edition, Boston: Houghton Mifflin Company.

Schmidt, K.M., (1997), "Managerial Incentives and Product Market Competition", *Review of Economic Studies*, vol. 64, pp. 191-213.

Schumpeter, J.A., (1943), *Capitalism, Socialism and Democracy*, Harper and Row, New York.

Shleifer, A., (1985), "A Theory of Yardstick Competition", *Rand Journal of Economics*, vol. 16, pp. 319-327.

Seabright, P. and M. Ivaldi, (2003), “The Economics of Passenger Rail Transport: A Survey”, IDEI Working Paper, no. 163.

Stigler, G.J., (1987), “Competition”, in Eatwell, J.; Milgate, M. and P. Newman (eds.) *The New Palgrave*, London, Macmillan.

Thompson, L.S., (2003), “Changing Railway Structure and Ownership: Is Anything Working?”, *Transport Reviews*, vol. 23, pp. 311-355.

Tirole, J., (1988), *The Theory of Industrial Organization*, MIT Press, Cambridge, MA.

Tobin, J., (1959), “Estimation of Relationships for Limited Dependent Variables”, *Econometrica* vol. 26, pp. 24-36.

Union Internationale des Chemins de Fer (UIC), (2003), *Railways TimeSeries data, 1990-2001*, Paris.

United Nations (UN), (2003), *The Restructuring of Railways*, New York, ST/ESACP/2313.

U.S. Census Bureau, (2003), *Statistical Abstract of the United States: 2003*, 123rd Edition, Washington, D.C.

Velde van de, D. (ed.), (1999), *Changing Trains - Railway Reform and the role of Competition: The Experience in six Countries*, Ashgate, Aldershot (Oxford Studies in Transport Series).

Verbeek, M., (2004), *A Guide to Modern Econometrics*, 2nd edition, John Wiley and Sons Ltd, Chichester, West Sussex.

Vickers, J. and G. Yarrow, (1991), "Economic Perspectives on Privatisation", *The Journal of Economic Perspectives*, vol. 5, pp. 111-132.

Vickers, J., (1995), "Concepts of Competition", *Oxford Economic Papers*, New Series, vol. 47, pp. 1-23.

Viscusi, W.K.; Vernon, J.M. and J.E. Harrington Jr., (2000), *Economies of Regulation and Antitrust*, third edition, Massachusetts Institute of Technology Press, Cambridge, Massachusetts.

Williamson, O., (1976), "Franchise bidding for natural monopolies - in general and with respect to CATV", *Bell Journal of Economics*, vol. 7, pp. 73-104.

