International Transport Forum

Rail Efficiency: Cost Research and Its Implications for Policy

Draft Discussion Paper

Prepared for the Roundtable:
Efficiency in Railway Operations and Infrastructure Management

18-19 November 2014,
International Energy Agency, Paris, France

Christopher Nash

Andrew S.J. Smith[1]

Institute for Transport Studies
Leeds University,
United Kingdom

1st Draft October 2014

[1] We also acknowledge the contribution of Phil Wheat to material on which this paper is partly based.

OECD

THE INTERNATIONAL TRANSPORT FORUM

The International Transport Forum at the OECD is an intergovernmental organisation with 54 member countries. It acts as a strategic think-tank, with the objective of helping shape the transport policy agenda on a global level and ensuring that it contributes to economic growth, environmental protection, social inclusion and the preservation of human life and well-being. The International Transport Forum organises an annual summit of Ministers along with leading representatives from industry, civil society and academia.

The International Transport Forum was created under a Declaration issued by the Council of Ministers of the ECMT (European Conference of Ministers of Transport) at its Ministerial Session in May 2006 under the legal authority of the Protocol of the ECMT, signed in Brussels on 17 October 1953, and legal instruments of the OECD.

The Members of the Forum are: Albania, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Chile, People's Republic of China, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Former Yugoslav Republic of Macedonia, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Italy, Japan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Montenegro, the Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom and United States.

The International Transport Forum's Research Centre gathers statistics and conducts co-operative research programmes addressing all modes of transport. Its findings are widely disseminated and support policymaking in Member countries as well as contributing to the annual summit.

Discussion Papers

The International Transport Forum's Discussion Paper Series makes economic research, commissioned or carried out at its Research Centre, available to researchers and practitioners. The aim is to contribute to the understanding of the transport sector and to provide inputs to transport policy design.

ITF Discussion Papers should not be reported as representing the official views of the ITF or of its member countries. The opinions expressed and arguments employed are those of the authors.

Discussion Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the ITF works. Comments on Discussion Papers are welcomed, and may be sent to: International Transport Forum/OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

For further information on the Discussion Papers and other JTRC activities, please email: itf.contact@oecd.org

The Discussion Papers can be downloaded from: www.internationaltransportforum.org/jtrc/DiscussionPapers/jtrcpapers.html

The International Transport Forum's website is at: www.internationaltransportforum.org

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Abstract

In this paper we first consider alternative measures of efficiency. We explain why simple partial productivity measures are inadequate as the basis of overall measures of efficiency, and outline two alternative approaches. The first is technical efficiency – the degree to which output is maximised for a given level of inputs – and the second is cost efficiency, the degree to which costs are minimised for a given level of output. Cost efficiency implies technical efficiency but also allocative efficiency – choosing a cost minimising mix of inputs. We explain why we prefer to measure cost efficiency, both in terms of what governments and regulators are interested in and in terms of practical data problems.

We then examine applications of cost function analysis to two areas. The first is rail privatisation in Britain. British experience has seen a large increase in traffic, but also a similar increase in costs. We review attempts to understand and explain both the increase in passenger train operating cost and infrastructure cost using cost function analysis. The second is European rail reform. Countries in Europe have adopted a wide variety of approaches to rail reform, and studies using a mix of European and other countries should be able to shed light on the important question of what works best in different circumstances. Finally we consider how efficiency analysis techniques need to develop in future to address current weaknesses and tackle new challenges.

1. Introduction

Studies of rail efficiency usually have one of two motivations. Firstly they may aim to identify which railways are efficient and which are not, in order to draw lessons as to the level of improvement that may be required. An example of this is the benchmarking studies conducted on behalf of the British rail regulator in deciding on the financial requirements of Network Rail, the infrastructure manager discussed below (Smith et al, 2010). Secondly studies may seek to draw policy conclusions about which policies regarding industry structure, competition and regulation will be most beneficial. That is the approach we emphasise in this paper.

In sectors of the economy in which markets are a reasonable approximation to perfectly competitive, a reasonable measure of overall efficiency may simply be the profitability of the firm. Under perfect competition, prices are not influenced by the individual firm and therefore the more profitable the firm, the more it has been able to minimise costs of production and to produce the most valuable combination of goods in the eyes of consumers.

But rail is a long way from being a perfectly competitive industry. In some sectors (e.g. coal, commuters) rail operators still have considerable monopoly power, whilst for this and other social reasons rail prices are often regulated by governments, who also play a key role in specifying passenger sector outputs. If the aim is to examine the efficiency of railway management, these factors must be allowed for. To the extent that railway managers (at least in the European passenger sector) have limited control over their outputs, the key issue is whether they produce them at minimum cost.

Economists are very used to distinguishing between technical efficiency and allocative efficiency. Technical efficiency is measured by whether output is maximised for a given level of inputs (or conversely inputs are minimised for a given output). The standard economic approach to examine this is to estimate a production function, although in recent times a non parametric method – data envelopment analysis – has been most commonly used. This approach is considered further in the next section.

Allocative efficiency considers whether the correct mix of inputs is used to minimise cost for a given level and quality of output. Cost efficiency, which is the product of technical and allocative efficiency, and thus takes both into account, is the subject of the following section. We then examine evidence on both the efficiency of the approach to railway reform taken in Britain, and the relative efficiency of the variety of approaches taken around Europe before seeking to reach conclusions.

2. Technical efficiency

2.1 Introduction

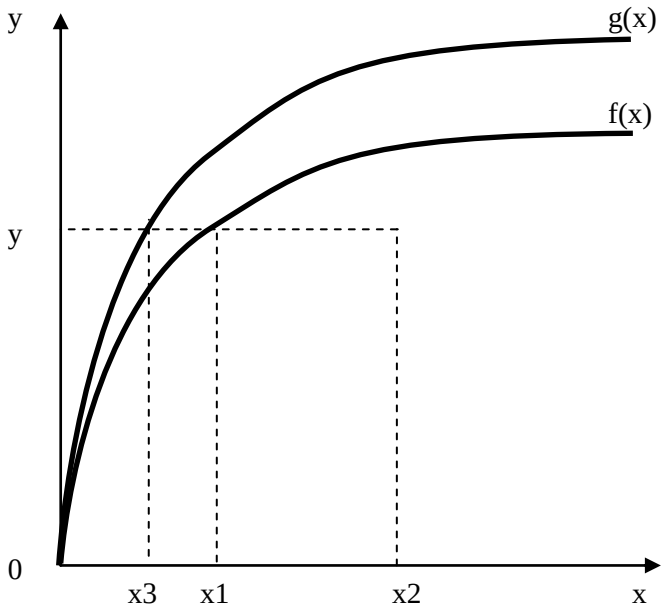
The material in this section is drawn partly from Smith (2004), and Nash, C.A. and Smith, A.S.J. (2007).

Before turning to discuss more advanced measures it is first worth considering why it is important to go beyond simple partial productivity measures - such as train kilometres per member of staff or per locomotive - which are often used as measures of technical efficiency. The first point to note is that they are partial measures and cannot therefore lead to conclusions about overall efficiency (unless the same railway is more efficient on every measure). Further, they are often impacted significantly by capital substitution effects (where capital is substituted for labour, therefore improving labour productivity). In a multi-input, multi-output environment such as the railways, the concept of total factor productivity (TFP) provides a more informative indicator of technical efficiency.

TFP is a measure of the ratio of all outputs to all inputs (with the different inputs and outputs weighted in some way). Under perfect competition, prices may be used as the weights so the measure returns to one of the ratio of the value of outputs to the value of inputs, or in other

words profitability. Of course, the underlying assumptions of constant returns to scale /density and marginal cost pricing are highly restrictive, especially in the context of the world's railway systems, which are characterised by high quasi-fixed costs and are heavily regulated. These restrictive assumptions mean that TFP measures are unable to distinguish changes in technical efficiency from underlying technical change and from changes in TFP resulting from scale or/ density effects or departures of prices from marginal cost. Figure 1 illustrates, for the single input, single output case, how inefficiency (the gap between a firm's input-output combination and the frontier) differs from productivity effects (moving along the production frontier).

Figure 1: **Single input production frontier**



One way of dealing with this problem is to estimate a cost function using econometric methods (see section 3 below). This approach allows the estimation process to calculate the extent of returns to scale, to/ density, the elasticities of cost with respect to the outputs, productivity growth resulting from technical change, and changes in efficiency. Data envelopment analysis (DEA) is an alternative index number method for deriving productivity and efficiency measures. It is a non-parametric approach, in which the efficiency frontier is computed using linear programming techniques (rather than estimated, using econometric methods; see section 3). The method permits the assumption of variable returns to scale and Malmquist productivity indices can be computed from such models that also allow the decomposition of TFP changes into technical efficiency, scale efficiency and technical change.

One challenge that the DEA approach faces is its difficulty in characterising the technology, since the introduction of numerous inputs and outputs into the DEA tends to result in all firms being on the frontier. It can thus be hard to distinguish between, for example, economies of scale and density. One way in which the approach can be augmented to deal with this criticism is through the use of a second stage, in which the efficiency scores from the first stage are regressed on a range of factors, such as train density and load factor and indeed policy variables, such as the degree of competition). However, the approach then starts to look much more like an econometric model, and it raises a question as to why adopt a two stage approach, and instead simply estimate an econometric model. A remaining, major weakness of DEA, however, is its inability to take account of random noise, meaning that measures of inefficiency can be overstated (or efficiency understated).

One supposed advantage, historically, of DEA has been its ability to deal with multiple inputs and outputs without recourse to potentially restrictive behavioural assumptions (for example, cost minimisation, as required in econometric cost function estimation). Another is the fact that it does not require the specification of a functional form for the underlying technology. However, both these advantages have been largely eliminated by the widespread use of flexible functional forms (e.g. the translog) and by the development of econometric methods for estimating distance functions (see Groskopf et al, 1997).

Overall, the most advanced approaches and widely used approaches in the academic literature for estimating technical efficiency are econometric distance functions (see for example, Kennedy and Smith and Coelli and Perelman, 1999) and DEA (see, for example, Canto et. al., 2010. However, as will be discussed in section 3, eonometric cost function / cost frontier estimation has numerous benefits, and this approach dominates the empirical discussion.

Importantly, whatever approach is adopted, a decision is needed as to what should be the relevant outputs and inputs, as well as associated output characteristic variables.

2.2 Outputs

At its simplest, transportation output may be regarded as the transport of passengers or freight. Thus measures such as passenger kilometres and freight tonne-kilometres are the usual starting point for output measurement. However, there are grave shortcomings with such simple measures of output.

Multiplicity of outputs is a common feature of transport firms. Strictly, an output needs to be described in terms of the provision of transport of a specific quality from a specific origin to a specific destination at a specific point in time. Thus an operator of rail passenger services running trains between ten stations ten times per day and offering two classes of travel is already producing 1800 different products. A large European railway will have literally millions of products on offer. Of course, it is not possible to provide cost or performance measures that separately identify each product.

This is only really a problem if the different products have significantly different cost characteristics, and traffic on them is growing or declining at different rates. For instance, if it costs a similar amount to transport passengers (per km) between London and Leeds and

London and Manchester, then performance measures will not be distorted by regarding these as the same product. On the other hand, failure to identify different traffic having very different costs will be very distorting. For instance, part of the rapid improvement in productivity of British Rail freight wagons in the 1980s was because of the decline and eventual abolition of movement of single wagonloads in favour of movement of traffic in full trainloads.

In passenger rail transport, longer distance, faster moving traffic and traffic moving in large volumes generally costs less per passenger-kilometre to handle than short distance traffic or traffic that must move slowly and in small volumes. This is because of the spreading of terminal costs and the economies of operating longer trains. Peaks in demand also lead to poor productivity by requiring the provision of a lot of resources that are only used for a small part of the day. Thus a fundamental distinction is between types of passenger traffic in rail such as inter-city, suburban and regional. Such peaks in demand are likely to be an issue in urban public transportation operations, with some services being more peaked than others. Likewise bus services can be provided at a lower unit cost (per passenger-km and bus-km) where there are large volumes as operators exploit economies of density (see section 5 below). Economies of density occur when adding more traffic to the same network leads to a less than prolirtionate increase in cost, whereas economies of scale occur when a given increase in both nework size and traffic leads to a less than proritonate increase in cost.

In any event, frequency of service is an important quality attribute. A transportation manager who was simply wishing to minimize costs – for a given number of passenger kilometres - might run one high capacity service per day, but this would not be very attractive to customers. No sensible transportation manager will provide the frequency of service that minimizes costs if a more frequent service will improve net revenue or benefits. This suggests that, unless a way can be devised of adjusting passenger and freight-tonne-kilometres for the quality of service provided, a more radical change to the output unit to train-kilometres rather than passenger- or freight-tonne-kilometres might be desirable (it will still be necessary to disaggregate train-kilometres according to their cost characteristics, as it costs much more to shift a 5000 tonne freight train than a two-car branch line passenger train). The use of vehicle-kilometres in place of train-kilometres may be a helpful further refinement of the train-kilometre measure, although this measure will still not correct for different weights of train. Certainly, to regard operations where rolling stock is grossly overloaded, as for instance in some developing countries, as therefore performing well - even if they are producing the service itself very inefficiently - seems mistaken.

Freight traffic is particularly complex because of the lack of a homogenous unit of measurement; at least in passenger transport we are always dealing with people. A tonne of freight may cost very different amounts to transport according to whether it is a dense product or not (for a dense product a single wagon will contain far more tonnes than for a product that is not dense) and the form it is in (bulk solids or liquids may be loaded and unloaded much more simply than manufactured goods, although the latter will be easier to handle if they are containerized). It follows that loaded-wagon-kilometres may be a better unit of measurement than tonne-kilometres, and that distinctions may be needed between trainload, wagonload and container or intermodal traffic. If tonne-kilometres are used, a distinction by commodity is important; for instance, a railway that has declining coal traffic and rapidly growing

intermodal traffic will almost certainly show declining productivity if tonne-kilometres are the measure.

2.3 Inputs

Providing a rail service requires locomotives, passenger coaches or freight wagons (or self-powered vehicles), track, signalling, terminals and a variety of types of staff (train crew, signalling, track and rolling stock maintenance, terminals and administration). While ultimately all may be regarded as forms of labour and capital, the length of life of the assets and government intervention over employment and investment will often mean that at a particular point in time an undertaking will not have an optimal configuration of assets and staff (see section 2.3 below). This renders attempts to measure inputs simply as labour and capital difficult, as measures of the value of capital stock will need to allow for excess capacity and inappropriate investment. An alternative is to simply look at physical measures of assets (e.g. kilometres of track, numbers of locomotives, carriages and wagons for railways), but this obviously makes no allowance for the quality of the assets.

2.4 Problems in measuring technical efficiency

A key problem in measuring technical efficiency is that of joint costs and economies of scale /and density. For instance, a single-track railway may carry both passenger and freight traffic, a passenger train first- and second-class passengers, and a freight train a variety of commodities. In this situation, only some of the costs can be specifically attributed to one of the forms of traffic; the remaining costs are joint. The result is that railways typically are characterized by economies of scope; i.e. the costs of a single railway handling a variety of types of traffic are less than if each distinct product were to be handled by a different railway. Moreover, most evidence suggests that railways are subject to economies of traffic density. Putting more traffic on the same route generally reduces unit costs and raises measures of total factor productivity, unless the route is already heavily congested.

The result is that apparent rises in productivity may be caused by diversification into new products or by increased traffic density rather than being relevant to the measurement of performance. Of course, under conditions of economies of density, running more services (and possibly different types of service) on the network does lead to a genuine improvement in productivity. The argument here, however, is that the improvement in productivity arises naturally as a result of the shape of the cost function, and not because of any improvement in working practices.

The operating environment will also exert a strong influence on railway performance through its impact on the nature of the traffic carried. This has already been considered above. However, geography has other influences as well; gradient, climate and complexity of the network are all likely to influence costs. The quality of the service delivered will also impact on costs, for example in terms of the rolling stock used (e.g. air conditioned trains; trains that give greater access for disabled users), and more widely the punctuality and reliability of services. Other factors, such as the extent of passenger information provided and the quality

of on-board catering services for rail will also affect costs. Particularly in the case of rail, the quality of the service will depend critically also on the capability of the infrastructure.

Of course, to the extent that the method used contains relevant measures of outputs and output characteristics such that it can capture some of these features of the technology (e.g. scale and density effects; quality; network complexity), then it should be possible to obtain measures of technical efficiency after having taken account of these effects. We consider that econometric methods, as opposed to DEA or partial or TFP measures give the best opportunities for getting at underlying technical efficiency; though as noted, DEA combined with a second stage econometric model is a useful alternative. As we will argue in section 3, we consider cost econometric models to be the most suitable to achieving the objectives set out earlier, namely assessing relative efficiency, and understanding the impact of industry structure.

2.5 Conclusion

There are many studies of railway technical efficiency using total factor productivity or DEA approaches. However, in addition to the problems of measuring outputs and inputs, there is a severe difficulty in terms of the data for this form of analysis. Typically, such approaches include physical measures of the labour input, combined with physical measures of the infrastructure and rolling stock measured in simple terms (track-km and number of rolling stocks). Usually other inputs are excluded which, given that there are varying degrees of subcontracting in the rail sector (rolling stock may be leased, maintenance and cleaning of rolling stock and stations may be contracted out, as may track maintenance and renewals and so on), this risks giving misleading results. For that reason as well as the technical reasons given above, we prefer studies based on costs, which should be comprehensive measures including all contracted out items. Moreover, it is costs ultimately that governments and regulators are most interested in, not measures of physical productivity.

3. Cost function estimation

3.1 Introduction

In addition to the reasons outlined for preferring the approach of estimating cost functions, there is an additional practical reason, namely that data is more reliable, although there remain problems of inconsistencies in treatment of costs such as depreciation and interest, particularly in international comparisons, but even in one country over time. The problem is not simply different assumptions about asset lives. In some cases, where assets are purchased with grants, no depreciation or interest is entered into the accounts. In some cases historic debts have been written off; in other cases interest is still charged on them. Getting consistent data remains a challenge.

The problems regarding how to measure inputs and outputs also remain, as does the issue of the operating environment. How these problems may be handled in cost function analysis is considered below. However, overall, the cost function approach does at least ensure that all inputs are considered and the allocative efficiency (or inefficiency) associated with using different input combinations accounted for. This compares to technical efficiency analysis which first of all does not, of course, take account of allocative efficiency, but perhaps more importantly for analysis of railways, often neglects some inputs (i.e. those represented by other costs; see section 2).

3.2 Cost function estimation

Derived from assuming cost minimisation in a production process, the cost function relates costs (C) to the level of outputs (Y) and input prices (P) and, where data is available over time, some measure of how costs change over time as a result of technical change. Thus

$$C_{it} = C(Y_{it}, P_{it}, t) \qquad (1)$$

It therefore automatically allows for one key issue in comparing costs between railways in different countries, and in a single country over time, namely different input prices.

There have been a vast array of forms proposed. Notable developments include the constant elasticity of substitution (CES, Arrow et al (1961)) and generalised Leontief (Diewert, 1971). The most widely employed cost function is the Translog (Christensen, Jorgenson and Lau 1971, 1973). The Translog nests the simpler, and widely-used Cobb Douglas function (see, for example Beattie and Taylor, 1985) as a special (restricted) case, however it is not derived from any production function using duality theory. Instead the Translog cost function is usually presented as a functional form which is a second order approximation to any cost function rather than being derived directly from economic theory[2]. The general form of the Translog cost function for m outputs and n inputs is represented as:

$$\ln C = \alpha_0 + \sum_{i=1}^{m} \beta_i \ln y_i + \sum_{i=1}^{n} \gamma_i \ln w_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{m} \beta_{ij} \ln y_i \ln y_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \gamma_{ij} \ln w_i \ln w_j + \frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{n} \delta_{ij} \ln y_i \ln w_j \qquad (2)$$

The function includes both first and second order terms in all variables. As such the Translog cost function (like the generalised quadratic) is called a second order flexible functional form whereas the Cobb Douglas (and the linear) are called first order flexible forms since they include only first order terms. Importantly, the use of second order approximations allows for elasticities and marginal costs to vary flexibly with the level of outputs and prices. In this

[2] This justification can also be applied to the generalized quadratic functional form which nests the linear as a special case.

sense the Translog does have appealing economic characteristics, such as the ability to deal with varying degrees of returns to scale and density as firm size varies.

There are a number of requirements that a functional form has to obey to be consistent with economic theory. Some, such as symmetry and homogeneity of degree one in input prices, can be imposed through suitable parameter restrictions. However others such as concavity in input isocurves cannot be directly imposed in the Translog cost function. Instead these restrictions have to be tested at each data point in a sample. As such, the function will not necessarily be globally consistent with economic theory, but the researcher should test whether it is locally consistent. This illustrates the general difficulty of choosing sufficiently flexible functional forms while maintaining requirement that the function has proper economic properties.

Finally we note that the Translog cost function is often estimated along with the factor share equations. Factor share equations are expressions for the proportion of total cost used by each input and are derived using Shephard's (1953) lemma as the partial derivative of the cost function with respect to each input price. Estimation can then proceed using Zellner's (1962) Seemingly Unrelated Regression (SUR) which is more efficient (in terms of estimation) than single equation ordinary least squares.

When it comes to measuring outputs, the need to distinguish between scale and density effects or the choice between passenger and freight tonne or train or vehicle km is only part of the wider issue of how to account for the heterogeneity of railway outputs. One way to deal with the heterogeneity in outputs is to group outputs (denoted y) into m groups and include a further set of r variables which characterise the outputs (denoted q):

$$C(y_1,\ldots,y_m,q_1,\ldots,q_r,w_1,\ldots,w_n) \qquad (3)$$

The move from potentially hundreds or thousands of outputs to a more manageable number of m outputs is obviously a simplification. However the inclusion of output characteristic variables is an attempt to reintroduce heterogeneity in outputs back into the model. Such variables may include revenue measures (such as passenger-km and freight tonnes-hauled) where available measures are adopted as output and vice-versa. As such it can become a little ambiguous as to what variables represent outputs versus output characteristics versus network size.

The inclusion of characteristic variables in the cost function specification has prompted new definitions of economies of scale and density to be proposed to allow for the possibility of characteristics of outputs changing along with the outputs or network size themselves. See Oum and Zang (1997) for a discussion. The ideas are similar to the discussion in Caves et. al. (1985) regarding the need to consider changes in unobserved network effects in EoS described above, however in Oum and Zang (1997) these relate to changes in observed rather than unobserved variables.

While this formulation does simplify the problem to a tractable level, the resulting function may be very complicated given the number of variables in and possible interaction and higher order terms for each. As a result the cost function may still not be suitably parsimonious.

Spady and Friedlaender (1978) developed a hedonic cost function, which restricts the cost function in equation (3) to:

$$C(\omega(y, q_1, \ldots, q_r), w_1, \ldots, w_m) \quad (4)$$

Two things distinguish equation (4) from (3). First there is only one output as opposed to n outputs (this is relaxed in some applications e.g. Bitzan and Wilson (2008) and Wheat and Smith (2014)). Second the output and output characteristic variables enter into their own function and this then enters into the general cost function. This is important as once a functional form is chosen for C( ) such as Translog, the use of hedonic output ($\omega(\cdot)$) imposes several restrictions on the model in equation (4). The benefit of this approach is a more parsimonious model; however this may represent an unacceptable simplification of the cost relationship.

It is perhaps surprising that there have not been too many applications of hedonic cost functions in transportation operations. One example in transport operations is that by Wheat and Smith (2014) who estimated such a function with three outputs (train hours, route length and number of stations operated) and many characteristic variables relating to the train hours output. This analysis provided rich insights into the impact of output heterogeneity on economies of scale and density (see section 4), whilst enabling a parsimonious model.

3.3 Stochastic frontier methods: introduction

The above discussion has focused on the relationship between costs, outputs, output characteristics and input prices. As noted earlier, a key motivation for policy makers is to understand the relative cost efficiency of transport operators. The cost function relationships discussed above can be augmented to allow the relative efficiency of companies to vary and for this degree of variation to be estimated.

The efficiency measurement literature cites three functions which may be estimated, depending on the appropriate behavioural assumption: cost functions, production functions or distance functions (the latter two are focused on technical efficiency; see section 2). Most applications in railways are based on cost functions, reflecting the fact that, due to the highly regulated environment in which most railways operate, it is appropriate to view railways as seeking to minimise cost for a given level of output (where the latter is more or less determined by government). In this section we focus on cost function relationships or, more precisely now that we are introducing inefficiency into the approach, cost frontier relationships.

The simplest econometric approach is to use the method of corrected ordinary least squares (COLS). This method proceeds by ordinary least squares (OLS), but then shifts the regression line down by the amount of the largest negative residual (for the cost function case), thus translating an "average" cost line into a cost frontier. However, like DEA, the COLS method is a deterministic approach which does not distinguish between genuine inefficiency and statistical noise when looking at deviations from the frontier. It is however, with suitable adjustments, widely used by UK economic regulators, in part due to its simplicity.

The alternative and more widely used method in the academic literature (and increasingly by economic regulators) is stochastic frontier analysis (SFA); see equation (5) below. The stochastic cost frontier model can be represented as:

$$C_{it} = f(Y_{it}, P_{it}, N_{it}, \tau_t; \beta) + v_{it} + u_{it} \quad (5)$$

where the first term ($f(Y_{it}, P_{it}, Q_{it}, \tau_t; \beta)$) is the deterministic component, and Y_{it} is a vector of output measures, P_{it} is a vector of input prices, N_{it} is a vector of exogenous network characteristic variables, τ_t is a vector of time variables which represent technical change and β is a vector of parameters to be estimated. C_{it} represents the cost variable to be explained. The i and t subscripts refer to the number of firms and time periods respectively. Whilst some applications may use only cross sectional data, most railway applications utilise panel data, and this type of data greatly expands the possibilities for increasing the richness of the analysis in a number of ways as discussed further below. The v_{it} term is a random component representing unobservable factors that affect the firm's operating environment. This term is distributed symmetrically around zero (more specifically assumed to be normally distributed with zero mean and constant variance). A further one sided random component is then added to capture inefficiency (u_{it}).

For cross-sectional data, it is necessary to make distributional assumptions concerning the one-side inefficiency term, and the estimation proceeds via maximum likelihood. This is a significant limitation as these assumptions may not be valid. For panel data, there are additional estimation possibilities. Before turning to the panel data approaches it is worth summarising the benefits of the econometric methods for studying the structure of railway costs and relative efficiency performance.

Compared to cost function (or average response function estimation) it is clear that frontier methods are a significant development since they explicitly allow for the possibility of variation in efficiency performance between railways and over time. Compared with the DEA approach, econometric methods provide estimates of the underlying structure of production / costs, for example, the elasticity of costs with respect to different cost drivers, such as traffic volumes - which DEA does not. In addition, through the development of stochastic frontier analysis, econometric techniques are also able to distinguish between random noise and underlying inefficiency effects. However, econometric approaches do require the choice of an appropriate functional form, and the more flexible forms (such as the translog) are not always straightforward to implement due to the large number of parameters to be estimated. In addition, the choice of distribution for the inefficiency term in stochastic frontier analysis is arbitrary. The precise method that researchers should use will therefore depend on a range of factors, and in many academic papers more than one method is used in order to provide a cross-check against the other approaches.

3.4 Stochastic frontier methods: panel data approaches

The existence of panel data offers a number of important benefits. First of all, by combining cross-sectional and time series observations it provides additional degrees of freedom for estimation. This may be very important, particularly if the number of companies for which data exists is small as it often is for economic regulators. Second, it provides an opportunity to simultaneously investigate inter-firm efficiency disparities, changes in firm efficiency performance over time, as well as industry-wide technological change over the period of the study. Third it can, for some models, permit the estimation of firm efficiency without recourse to potentially restrictive distributional assumptions. Finally, it offers the prospect of disentangling inefficiency from unobserved factors. This latter benefit may be particularly important for railways, where substantial differences exist between railways both within and between countries, but where it is hard to capture these differences in a set of variables to be included in the model.

One way of dealing with a panel is to treat each data point as a separate firm. In this case, each observation, including observations for the same firm over multiple time periods, is given a separate efficiency score. In the case of econometric estimation this assumption may not be appropriate, since it assumes that inefficiency is independently distributed across observations, even though it might be expected that an inefficient firm in one period is likely to retain at least some of that inefficiency in the next period.

The alternative and more usual approach, is explicitly to recognise the panel nature of the data set. Within this alternative, there are two further options. Firstly, to estimate the model using traditional panel data methods (fixed effects or random effects (GLS)); see Schmidt and Sickles (1984). Alternatively, Pitt and Lee (1981) offer a maximum likelihood version of the same approach. In both cases, inefficiency is assumed to be "time-invariant" and each firm is given one efficiency score for the whole period, rather than one score per firm for each period as in the simple pooled approach. The advantage of the traditional panel approach (fixed and random effects) is that it does not require distributional assumptions concerning the inefficiency term as in the maximum likelihood equivalent. This benefit does come at a cost though, as it requires the assumption that inefficiency does not vary over time, which is restrictive.

For long time periods, the assumption of time invariant inefficiency is clearly problematic, and a number of approaches which allow for inefficiency to vary, whilst retaining some structure to the variation, have been developed. Time varying models have been developed for both the traditional panel data methods (e.g. Cornwell, Schmidt and Sickles (1990), and the maximum likelihood approach (e.g. Battese and Coelli (1995); Cuesta (2000). Kumbhakar and Lovell (2000) describe these approaches in detail. A key distinction in the literature is between those models which make the assumption of independence in inefficiency over time (e.g. pooled SFA; Battese and Coelli (1995)) and those which permit firm inefficiency to change in a structured and not random way over time (Cuesta (2000)). The latter seem to have advantages from a regulatory and economic perspective

An important and relatively recent development in the literature has revolved around the problem of disentangling inefficiency from unobserved heterogeneity. In the standard panel literature, fixed and random effects are assumed to represent unobserved, time invariant

factors that vary between firms. As noted, in the efficiency literature, these models have been applied as efficiency estimation approaches, with the firm effects re-interpreted as inefficiency. This approach risks badging unobserved factors – genuine heterogeneity between railways – as inefficiency. Methods have therefore been developed in the literature to address this (Greene, 2005; Farsi et. al., 2005; Kumbhakar et. al., 2014; Colombi et. al., 2014). One version of Greene's approach includes a firm-specific dummy, to capture unobserved heterogeneity between firms, which is assumed to be time invariant (e.g. environmental factors, such as topography or climate) as well as the one-side inefficiency term (which varies over time). The decomposition therefore relies on the assumption that inefficiency varies randomly over time whereas unobserved heterogeneity is time invariant (as well as on the distributional assumptions of the model). The model is then estimated via maximum likelihood. This is one of the so-called "true" models, and there is also a random effects version of this approach.

The Farsi *et al*. (2005) approach separates inefficiency from unobserved heterogeneity by making the assumption that the former is assumed not to be correlated with the regressors whilst the latter may be (inefficiency being a function of the ability of management to control costs given the exogenous set of output requirements and input prices that it faces – hence this would not be expected to be correlated with the regressors). Finally, the approaches set out by Kumbhakar *et al*., 2014; Colombi *et al*., 2014 seek to go further and separate the model residual into four components: random noise, time varying inefficiency, time invariant inefficiency and time invariant unobserved heterogeneity. This model relies entirely on distributional assumptions to make this separation, which is a limitation. It further assumes that unobserved heterogeneity is uncorrelated with the regressors, which may not be valid. It is worth noting that these are relatively new approaches with relatively few applications as yet, and none, to our knowledge in railways.

3.5 Conclusion

To conclude, then, there exists a wide range of methods for estimating cost inefficiency in the literature, some of them relatively simple and widely used, particularly by regulators, and others more complex. However, some of the more complex methods are now entering the economic regulation sphere in the UK, for example, ORR has adopted a range of advanced methods, and others, such as OFWAT, have considered these approaches at least, though to date has fallen back on simpler methods, given the data and results obtained[3]. Panel methods offer much more scope for a rich analysis of the cost structure (economies of scale and density) and inefficiency and these are the most widely used in railways. The question of dealing with random noise and heterogeneity (observed and unobserved) remains a key issue for all regulators in railways and other sectors. Ultimately, the choice of technique will depend on a number of factors, including the number of data points, availability of cost driver data, model performance, economic theory and practical considerations. Usually, it is appropriate to run a range of approaches and compare the results and in some cases it will not be possible easily to choose between them. Economic regulators in that case tend to average the efficiency results across a range of models.

[3] Andrew Smith has been advising ORR and OFWAT on the use of these methods.

4. Rail privatisation in Britain

4.1 Introduction

Over the period 1994-97, the British rail system underwent the most radical transformation of any European railway. Infrastructure was separated from train operations, and the new infrastructure company (Railtrack) privatised by sale of shares. The freight sector was privatised essentially as two companies and open access for new entrants permitted. The passenger sector was largely franchised out in the form of 25 franchises, offered for a period in most cases of 7-10 years, on a net cost basis but with requirements as to what services should be operated and restrictions on the levels of some fares.

In the period since privatisation (1995 to 2010) passenger-km growth has been faster than in all other major European railways (Brown, 2013). To provide a few specific examples, between 1995-2010, passenger km increased 84% (Britain); 65% (Sweden); and 17% (Germany)[4]. Between 1995 and 2013 freight volume (in tonne km) increased by around 70%. Most studies conclude the main reasons for this growth,were not to do with privatisation, but that privatisation was a contributory factor. However, the story regarding passenger train and infrastructure costs is not so positive (Table 1)

Under strong regulatory pressure, Network Rail's costs have been falling in recent years, although the company is not improving efficiency as quickly as the regulator would like.

To understand the drivers behind these trends we will first consider studies of passenger train operating costs and then infrastructure costs to try to understand the reasons for this increase.

[4] Source: European Commission Transport in Figures.

Table 1: **Rail Industry Costs in Britain (Infrastructure and Passenger Train Operations): 1997 to 2012**

Costs (£b2011/12 prices)	1996/97	2005/6	2011/12
Infrastructure expenditure			
Maintenance	1.1	1.4	1
Renewals and enhancements	1.5	3.7	4.6
Other operating costs	1	1.4	1.5
	3.5	**6.5**	**7.1**
TOC costs less access charge payments	**4.2**	**5.8**	**5.9**
Total passenger rail costs	**7.7**	**12.3**	**13**
Unit cost measures (£)			
Total passenger rail costs per passenger train km	20.2	27	25.4
Infrastructure costs per passenger train km	9.2	14.4	13.9
TOC costs (excluding access charges) per passenger train km	11	12.6	11.5

4.2 British passenger train operating costs

There have been many studies of passenger train operating costs in Britain, including, Affuso, Angeriz, and Pollitt (2003); Cowie (2002a, 2002b, 2005 and 2009); Smith and Wheat (2012), Wheat and Smith (2014) and Smith, Nash, and Wheat (2009). Preston (2008) provides a review of, inter alia, previous cost studies of the British rail sector. The above papers have

utilised a variety of methods including non-parametric DEA (Affuso et al., 2003; Cowie, 2009, Merkert et al., 2009) and index number approaches (Cowie, 2002a; Smith et al., 2009), as well as parametric estimation of cost functions (Cowie, 2002b; Smith and Wheat, 2012, Wheat and Smith, 2013), production functions (Cowie, 2005) and distance functions (Affuso et al., 2003). Clearly the former methods (DEA and index number approaches) can only consider cost or technical efficiency and produce no estimates regarding the actual cost structure.

Importantly, of the British TOC cost studies, only four cover the crucial period after 2000 when TOC costs started to rise substantially, these being Cowie (2009), Smith et al. (2009), Smith and Wheat (2012) and Wheat and Smith (2014). Cowie (2009) covers the period to 2004, whilst the two further papers cover the period to 2006. Wheat and Smith (2014) provide analysis up to 2010.

An important issue is whether to include an infrastructure input in any analysis of train operating costs. Clearly the infrastructure input may be an important part of the transformation function and so should be considered for inclusion in any analysis. The four papers by Cowie all include some measure of infrastructure input in the analysis which is some combination of route-km and access charges paid by operators to the infrastructure manager (to form a price if applicable).

This in turn raises two important and related problems: (1) the infrastructure input is hard to measure (see the previous discussion for particular measurement issues in Britain); and (2) the inclusion of this input turns the analysis into an assessment of rail industry costs/production, rather than being targeted on the TOCs. In their study, Affuso et al. (2003) produce two models: one including the infrastructure input, and one not. The results differ as a consequence, although this problem is less severe during the early period after privatisation (which the study covers), since access charges and infrastructure costs were fairly stable during that period. Whilst there are good reasons for capturing the infrastructure input in a study of TOC performance, to capture the possibility that this input affects the TOC transformation function, Smith and Wheat (2012) and Wheat and Smith (2014) argue that, given the measurement problems noted above, infrastructure inputs are best left out of the analysis. The dependent variable in their paper is thus defined as TOC costs, excluding fixed access charges. Route-km is also included as an explanatory variable in their model, not as a measure of the infrastructure input, but to distinguish between scale and density effects.

The focus of Smith and Wheat (2012) was on the impact on cost efficiency of contract regimes following several renegotiations and temporary contracts being introduced following franchise failure. It used a panel data stochastic frontier framework which allowed efficiency to evolve over time (based on the model by Cuesta, 2000; see section 3). They also included dummy variables in the cost function to allow the extent of cost effects of different contract types to be directly estimated.

The focus of the Wheat and Smith (2014) work, in contrast, was how to best model the cost structure of the industry. This work utilised a hedonic cost function (see section 3.2) and the description of the data used is given in Table 2. In particular they defined three generic outputs (Route-km, Train-hours and number of stations operated) and defined nine characteristics of train services which go into the Train-hours output function (ψ_2). These

characteristics control for the heterogeneity in train service provision. They also define two inputs and associated prices.

Table 2. **Data used in Wheat and Smith (2014)**

Symbol	Name	Description	Data Source
Generic Outputs (ψ)			
$\psi_1 = y_1$			
y_1	Route - km	Length of the line-km operated by the TOC. A measure of the geographical coverage of the TOC	National Rail Trends
$\psi_2 = y_2 q_{12}^{\phi_{12}} q_{22}^{\phi_{22}} q_{32}^{\phi_{32}} e^{\phi_{42}q_{42}} e^{\phi_{52}q_{52}} e^{\phi_{62}q_{62}} e^{\phi_{72}q_{72}} e^{\phi_{82}q_{82}} e^{\phi_{92}q_{92}}$			
y_2	Train Hours	Primary driver of train operating cost	National Modelling Framework Timetabling Module
q_{12}	Average vehicle length of trains	Vehicle-km / Train-km	Network Rail
q_{22}	Average speed	Train-km / Train Hours	National Modelling Framework Timetabling Module
q_{32}	Passenger Load Factor	Passenger-km / Train km	Passenger-km data from National Rail Trends. Train-km data from Network Rail.
q_{42}	Intercity TOC	Proportion of train services intercity in nature	National Rail Trends for the categorisation of TOCs into intercity, LSE and regional. Where TOCs have merged across sectors a proportion allocation is made on an approximate basis with reference to the relative size of train-km by each pre-merged TOC
q_{52}	London South Eastern indicator	Proportion of train services into and around London (in general commuting services)	
q_{62}	$q_{42}q_{52}$	Interaction between Intercity and LSE proportions	
q_{72}	$q_{42}(1-q_{42}-q_{52})$	Interaction between intercity and regional (non-intercity and non-LSE services) proportions	
q_{82}	$q_{52}(1-q_{42}-q_{52})$	Interaction between LSE and regional proportions	
q_{92}	Number of rolling stock types operated	Number of "generic" rolling stock types operated	National Modelling Framework Rolling Stock Classifications
$\psi_3 = y_3$			
y_3	Stations operated	Number of stations that the TOC operates	National Rail Trends
Prices			
P_1	Non-payroll cost per unit rolling stock		TOC accounts for cost, Platform 5 and TAS Rail Industry Monitor for rolling stock numbers
P_2	Staff costs (on payroll)		TOC accounts (both costs and staff numbers)

Source: Reproduced from Table 1 in Wheat and Smith (2013).

Using the notation in Table 2, Wheat and Smith (2014) estimate the following system of two equations using non-linear least squares. This represents a full Translog hedonic cost function (S refers to the cost shares for each input):

$$\left\{ \begin{array}{l} \ln\left(\frac{C_{it}}{P_{2it}}\right) = \left\{ \begin{array}{l} \alpha + \sum_{l=1}^{3}\beta_l \ln(\psi_{lit}) + \delta_1 \ln\left(\frac{P_{1it}}{P_{2it}}\right) + \gamma_T t + \frac{1}{2}\sum_{l=1}^{3}\sum_{b=1}^{3}\beta_{lb}(\ln(\psi_{lit}))(\ln(\psi_{bit})) \\ + \delta_{11}\left(\ln\left(\frac{P_{1mit}}{P_{2it}}\right)\right)^2 + \sum_{l=1}^{3}\kappa_{l1}(\ln(\psi_{lit}))\left(\ln\left(\frac{P_{1it}}{P_{2it}}\right)\right) \\ + \sum_{l=1}^{3}\lambda_{Tl} t \ln(\psi_{lit}) + \varphi_{T1} t \ln\left(\frac{P_{1it}}{P_{2it}}\right) + \gamma_{TT} t^2 \end{array} \right. \\ S_1 = \delta_1 + 2.\delta_{11}\ln\left(\frac{P_{1it}}{P_{2it}}\right) + \sum_{l=1}^{3}\kappa_{lm}\ln(\psi_{lit}) + \varphi_{Tm} t \end{array} \right\} \quad (3.11)$$

Findings on the cost structure of passenger train operations

In this sub-section we present some results of the work undertaken to date to illustrate the richness and usefulness of the methods employed. As described above, returns to scale (RtS) and returns to density (RtD) can be defined specifically for operations (as distinct from infrastructure). RtS measure how costs change when a firm grows in terms of geographical size. RtD measures how costs change when a firm grows by running more services on a fixed network.

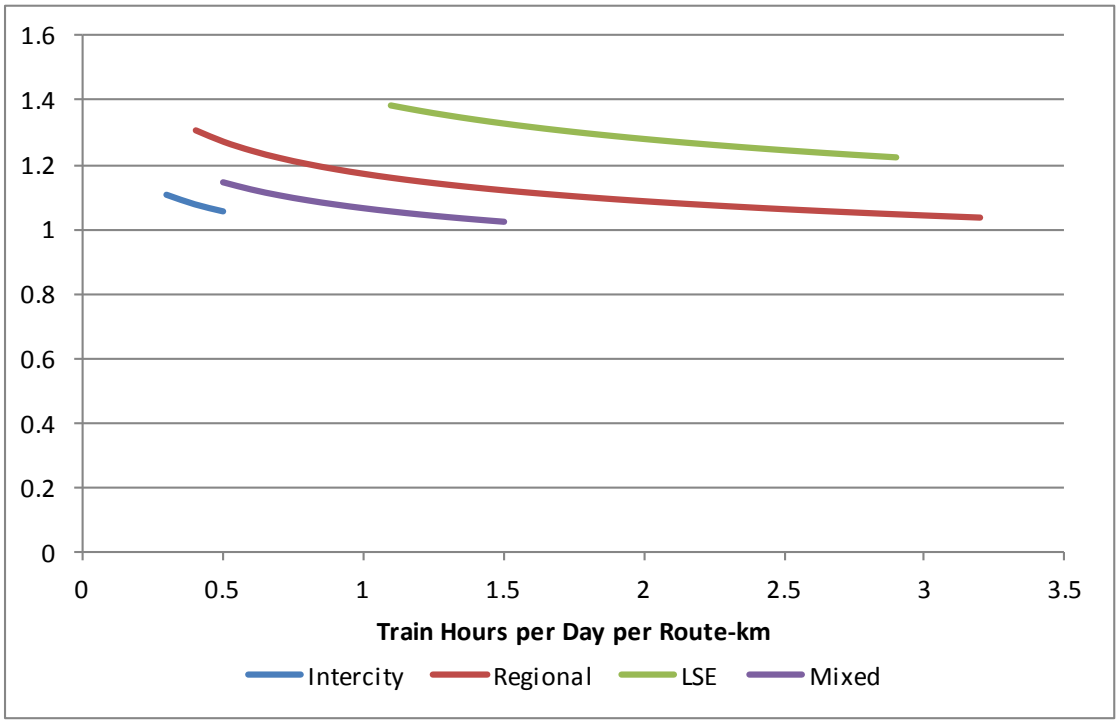
The DEA analysis yields few results with relation to economies of scale or density. Indeed the paper by Cowie imposes constant returns to scale. Merkert et al (2009) did estimate a variable returns to scale model and found that British and Swedish TOCs were below minimum efficient scale, while the largest German operators were above. Of the parametric papers, Cowie (2002b) estimates a cost model which provides evidence on economies of scale. Cowie finds evidence for increasing returns to scale and that these are increasing with scale. There is no attempt to differentiate between scale and density economies in the analysis.

Smith and Wheat (2012) put forward a model which yields estimates of the extent of both economies of scale and economies of density, where the primary usage output is train-km. They found constant returns to scale and increasing returns to train density. The policy conclusion of this finding is that whilst there would not be scale benefits from merging franchises, such mergers may reduce costs by allowing greater exploitation of economies of density (a single operator running trains more intensively down a given route); though see further discussion on economies of density and heterogeneity below. One limitation of the Smith and Wheat (2012) work was the inability to estimate a plausible Translog function. Instead, a restricted variant was estimated, selected on the basis of general to specific testing and on whether key elasticities were of the expected sign. This implicitly restricts the variation in economies of scale and economies of density.

Further work by Wheat and Smith (2014) estimated a Translog simultaneously with the cost share equations and adopt a hedonic representation of the train operations output in order to include characteristics of output in a parsimonious manner. This work provides new insights into the scale and density properties of train operations, since it allows RtS and RtD to vary

with the heterogeneity characteristics of output. Figures 2 and 3 summarise the findings on RtD and RtS.

Figure 2. **Returns to density for different TOC types holding other variables constant**



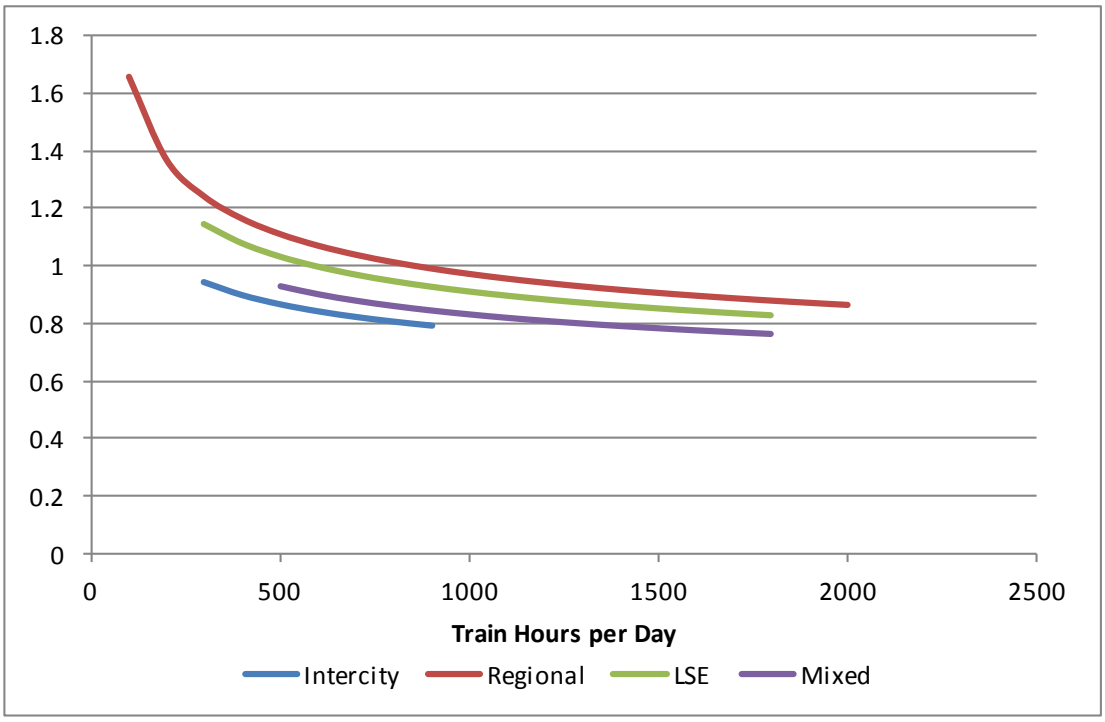
Source: Reproduced from Wheat and Smith (2013).

Figure 2 shows that all TOC types exhibit increasing RtD and that this does fall with density, although RtD are never exhausted within the middle 80% of the sample. At any given train hours per route km level, intercity TOCs exhibit the lowest RtD, while LSE (commuter services into London) exhibit the strongest (and indeed even at the 90th percentile density in sample the RtD estimate is in excess of 1.2). Intuitively, the curve for mixed TOCs is somewhere in-between the curves for intercity and regional. The policy conclusion from the analysis of RtD is that most TOCs should be able to reduce unit costs if there is further growth in train hours (on a fixed network) in response to future increases in passenger demand

Figure 3 provides a similar plot for RtS. This shows that for all of the central 80th percent of the train hours distribution, intercity (and mixed) TOCs exhibit decreasing RtS. LSE TOCs exhibit increasing returns to scale only for the very smallest in sample, whilst regional TOCs are the only TOC type to have an appreciable range of scale exhibiting increasing returns to scale. The results are consistent with a u-shaped average cost curve, although it would appear that most TOCs are operating at or beyond the minimum unit cost point. This finding has important implications for examining the optimal size of TOCs and is relevant to the recent

franchise policy change that has resulted in substantial franchise re-mapping, and in turn larger franchises.

Figure 3. **Returns to scale for different TOC types holding other variables constant**



Source: Reproduced from Wheat and Smith (2013).

The overall conclusion from this section is that modelling returns to scale and density in passenger train operations potentially requires a rich model to fully capture the effects. The initial work published in Smith and Wheat (2012) based on a restricted translog model suggested broadly constant returns to scale combined with fairly strong economies of density. This may suggest that there could be a case for making franchises smaller, which could help in reducing the risk of franchise failure, which has been a key problem in Britain (and Britain's franchises are already considerably larger, in general, than those elsewhere in Europe). However, if reducing the size of franchises also increases the degree of franchise overlap, then important economies of density may be lost in the process, so it is not a clear policy conclusion. Turning the argument the other way round, larger franchises, that result in reductions in franchise overlap and the exploitation of economies of density may reduce costs.

That said, Wheat and Smith (2014) develop a richer model, which takes account of service heterogeneity (in particular, in terms of train speed and TOC type) in relation to returns to scale and density. In that later paper it is found that the ability to exploit economies of density may be constrained by service heterogeneity. Likewise, the losses of economies of density from reducing franchise size might be smaller than indicated above. It is further found that

some franchises in Britain are operating at decreasing returns to scale, and may therefore be too large.

What the above research suggests is that it is possible to shed new light on the structure of costs of passenger train operations, and draw broad conclusions about the economies of scale and density of those operations. The most recent work suggests that there could be cost savings from reducing franchise size (because of scale diseconomies) and that losses in economies of density might be reduced by service heterogeneity. Whilst we can draw these general findings, we would however recommend more detailed, bottom-up modelling work be carried out on a case by case basis if a decision is required on franchise mergers / de-mergers. It may not be possible in an econometric to fully reflect the service and rolling stock possibilities that result from such changes.

Findings on efficiency variation

The purpose of this section is to illustrate the potential of the methodologies with respect to measuring efficiency of railway operators. There is an extensive literature analysing the efficiency and productivity performance of vertically integrated railways around the world (Oum et. al., 1999; Smith, 2006). More recently there has also been an interest in understanding the impact of vertical separation on total industry costs, mainly focussed on European evidence (Friebel, et. al., 2010; Asmild et. al., 2009; Growitsch and Wetzel, 2009; Cantos et. al., 2010 and 2011; Mizutani and Uranishi, 2013; Nash et. al., forthcoming; van de Velde et. al., 2012); although some studies considered evidence from North America (e.g.Bitzan, 2003). Overall, the results seem inconclusive, costs suggesting that much depends on the circumstances of the country concerned and the way in which the system is managed. However, we consider the recent work by Nash et. al., forthcoming; van de Velde et. al., 2012 to offer interesting new insights on the circumstances in which vertical separation and the holding company model might result in lower or higher costs. We review this work in more detail in section 5 below.

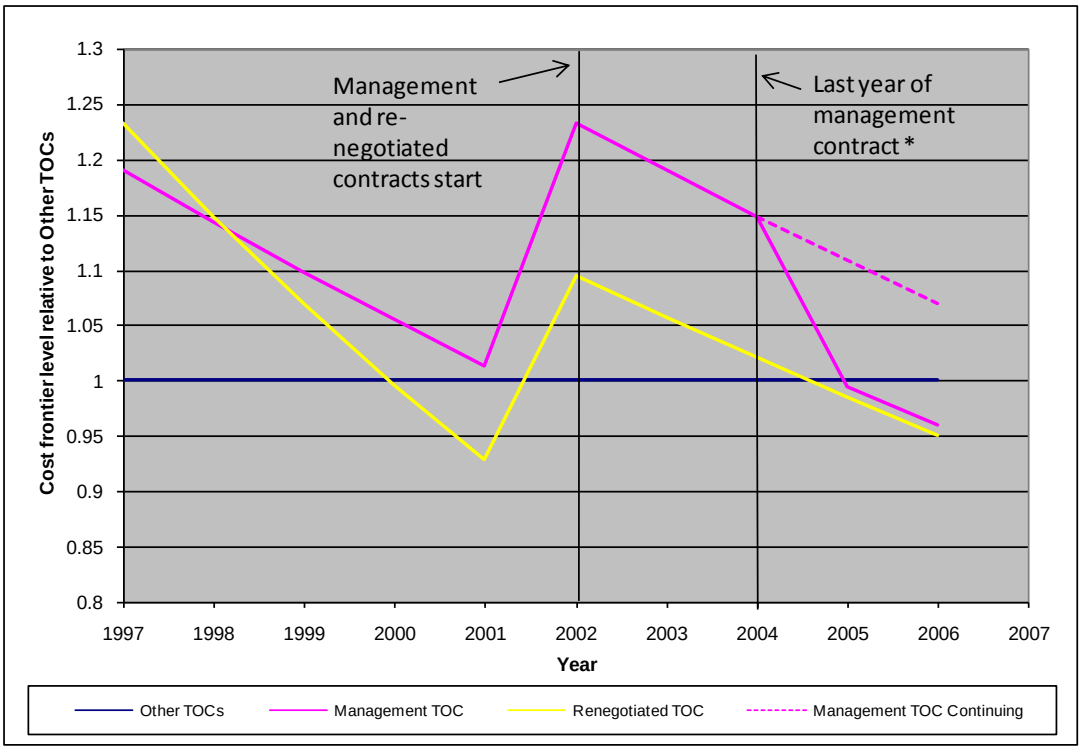
There have also been a small number of studies focusing on the impact of competitive tendering on one part of the rail industry, namely passenger train operations. In Germany and Sweden the experience of competitive tendering has generally been positive, with the evidence suggesting that savings in the region of 20-30 per cent can be achieved, alongside increased patronage (see Brenck and Peter, 2007; Lalive and Schmutzler, 2008; Alexandersson and Hulten, 2007; and Nash and Nilsson, 2009). Even here though, some franchises have failed to achieve favourable results. Kain (2009) describes the major problems that emerged in Melbourne, though the impact of the policy response is not described in any detail. Long-term passenger (and also freight) rail franchises have also been signed in Latin America, generally leading to radically improved performance, although in most cases re-negotiation has been required due to changed economic circumstances (in particular the severe economic recession in the late 1990s; see Kogan, 2006).

Turning to studies of British TOCs, Affuso et. al. (2002; 2003) and (Cowie, 2002a, 2002b, 2005) study the early years after privatisation (prior to the major cost rises) and all find improving productivity during this period. Only two studies cover the post-2000 period, after which costs started to rise. Cowie (2009) finds declining productivity growth after 2000, with the absolute productivity level falling post-2002. Smith and Wheat (2009) report productivity

levels falling as early as 2000 and not recovering over the remainder of the sample (to 2006). Smith et. al. (2010) reviews this literature. Thus Britain's franchising experience appears to be the outlier (at least within Europe), with costs rising rather than falling as in Germany and Sweden.

Smith and Wheat (2012) investigates the impact of the response of the franchising authority in Britain to franchise failure. Two approaches were adopted. First, most operators were placed onto annually-negotiated management contracts (similar to cost-plus contracts). The second saw some operators placed onto newly-negotiated short-term franchise arrangements. Figure 4 summarises the findings on the cost effect of different franchise contracts. This shows that the franchising authority's decision to place a large number of TOCs on management contracts for an extended period led to a substantial deterioration in efficiency relative to other TOCs. Furthermore, this effect was persistent and led to costs being considerably higher than other TOCs for several years. However, the relative inefficiency was eliminated by competitive re-franchising for those TOCs that were subject to this process during the sample period. The short-term franchise agreements which, once signed, retain incentives to reduce costs, saw a more positive pattern, with costs remaining in line with other operators (those that remained on their original franchise agreements throughout).

Figure 4. **Findings for TOCs in Britain that were subject to short term management or renegotiated contracts relative to other TOCs**



* Some TOCs saw re-franchising during this period and came off their management contracts. Other TOCs (dotted line in the above chart) continued on their management contracts to the end of the period.

Source: Reproduced from Smith and Wheat (2012).

The overall conclusion from this section is that whilst competitive tendering / franchising appears to have reduced costs (or at least subsidies) in Germany and Sweden (Nash, Nilsson and Link,) costs have risen in Britain. Whilst the management contracts put in place following numerous franchise failures explains part of the problem, Smith and Wheat (2012) also report a general upward trend in train operating costs in Britain (affecting all operators). It does appear that some British franchises may be inefficiently large.

However a major report on rail costs commissioned by the British Office of Rail Regulation (ORR) and the Department for Transport in 2011 (McNulty, 2011) concluded that a further factor leading to cost increases in Britain was a misalignment of incentives between infrastructure manager and train operating companies in the vertically separated structure of railways in Britain. They advocated closer working arrangements, including a high level Rail Delivery Group representing all parts of the industry, and alliances between the infrastructure manager and train operating companies or even leasing of the infrastructure to franchisees at the regional level. These issues are considered further at the European level in section 5.

4.3 Infrastructure costs

As noted above, infrastructure costs have risen even more than train operating costs. A major reason for this was the expansion of maintenance and renewals activity following the fatal accident at Hatfield in 2000 which was caused by a broken rail, but also the cost of the West Coast Mainline renewal and upgrade programme rose to some £8b from an original estimate of £2b. The result of these two events was that the privately owned infrastructure manager, Railtrack, was placed in administration in 2001, and was ultimately succeeded in 2002 by Network Rail, a company limited by guarantee, responsible to its members rather than shareholders and with its debts guaranteed by the government. The Office of National Statistics has recently ruled that, under current EU regulations, Network Rail is to be regarded as a public sector organisation.

How did these events affect the efficiency of the British infrastructure manager over time, and how does it compare with its European peers?

As noted above rail infrastructure costs rose substantially after the Hatfield accident. Whilst part of this increase was driven by the need to increase maintenance and renewal activity in the light of genuine concerns over the quality of the network, a wide range of evidence has shown that Network Rail's efficiency performance deteriorated sharply over that period. Strong regulatory action has therefore resulted, with Network Rail being tasked with making large efficiency gains in recent and the coming years.

Kennedy and Smith (2004) showed that Railtrack delivered substantial real unit cost reductions in the early years after privatisation (6.4% to 6.8% for overall maintenance and renewal activity between 1996 and 2002). However, these improvements were more than offset by the post-Hatfield cost increases, which resulted in unit cost increases of 38% for overall maintenance and renewal activity. Indeed, costs continued to rise after 2002 (see Smith et. al. 2009). A further finding of Kennedy and Smith (2004) is that there was scope for the infrastructure manager to reduce its costs by reduce intra-company performance differences.

ORR carried out a review of Network Rail's efficiency performance, in the light of the large cost increases, and commissioned a wide range of studies (2003 Interim Review of the company's finances). A key weakness of the 2003 review though was that the ORR's efficiency determination was ultimately based on two bottom-up consultant reviews of Network Rail's business plan (LEK, TTCI and Halcrow, 2003 and Accenture, 2003). These results were supplemented by internal benchmarking, which indicated the kind of savings that could be achieved if Network Rail implemented its own best practice consistently across the network.

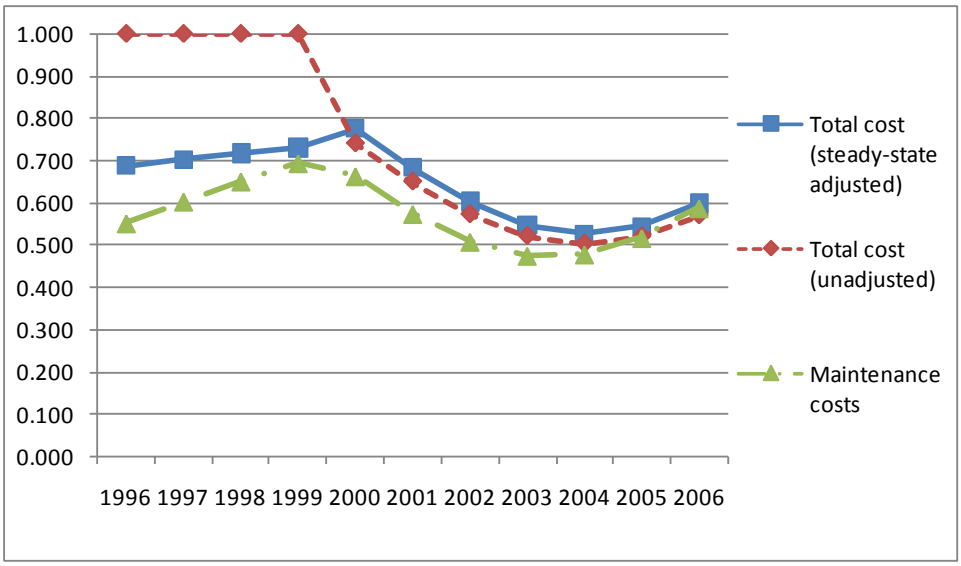
Ultimately then, the 2003 Interim Review was unable to provide a clear, empirically based assessment of Network Rail's relative efficiency position based on hard data from external sources. ORR nevertheless set a tough efficiency target of 31% over 5 years (2004-2009). However, costs were starting from a very high base. Thus, although costs then started to fall as Network Rail set about delivering its efficiency targets, by the time of the next periodic review in 2008, the scene was set to take the benchmarking approach a step forward by attempting international comparisons.

Two approaches were adopted during the 2008 review. The first used a panel of thirteen European infrastructure managers over an 11 year period. The data was provided from the Lasting Infrastructure Cost Benchmarking (LICB) project undertaken by UIC. The dataset included data on costs (adjusted based on PPP exchange rates), traffic volumes (by type), network length, and a range of other variables characterising differences between the companies (for example, extent of electrification, network density).

A structured inefficiency model (based on Cuesta, 2000; see section 3 above) was used that permits inefficiency to vary by firm over time, but in a structured way that recognises the panel nature of the dataset. The results are shown for Network Rail in Figure 5 (other companies cannot be shown for confidentiality reasons); see Smith (2012). Results are shown for maintenance only, and for maintenance and renewals, with the additional model variant to allow for Network Rail's renewals costs to be reduced downwards prior to modelling to allow for the fact that the company was renewing at above steady state levels in terms of renewal volumes. The overall message of Figure 5 is that Network Rail's efficiency deteriorated sharply after 2000, compared to its European comparators, leaving the company with an efficiency gap of around 40% by the end of the period. The analysis was carried out by the University of Leeds, with ORR and in conjunction with Network Rail and UIC.

In a separate, supporting study, ORR and University of Leeds, collected a new dataset comprising five other rail infrastructure managers in Europe and North America. This includes data on costs, outputs, and network characteristics at the regional level within each country. Thus, although the number of companies included was smaller than in the LICB dataset, the sample size was expanded via the use of regional data within companies (sub-company data structure). The dataset also allowed ORR to study within-country variations in inefficiency. The results broadly confirmed the results of the main study using LICB data (see Smith et. al., 2010; Smith and Wheat, 2012).

Figure 5: **Profile of Network Rail Efficiency Scores: Preferred Model**



It is further worth noting that the ORR carried out a range of other studies, principally based on bottom-up evidence. These confirmed the existence of a substantive gap, supported by examples of best practice in other countries (see Table 3).

Table 3: **Examples of European best practice**

Asset inspection and asset management.	In general best practice European railways undertake fewer track inspections but inspections are generally of higher quality. It is estimated that similar techniques applied in Britain could reduce foot patrolling inspection costs by around 75% and tamping expenditure by 20%
Recycling components	This is common European practice. In Switzerland, for example, rail, point motors, sleepers and signal heads are regularly refurbished then cascaded from higher to lower category routes. Cascaded rail on lines re-laid with steel sleepers could lead to savings. Additionally ballast cleaning (partial renewal) as opposed to traxcavation (complete renewal) could reduce ballast renewal cost in Britain by 40%
High output rail stressing	Stressing continuously welded rail by heating it rather than physically stretching it is a process discontinued in Britain in the 1960s and 1970s. Some European networks (using modern equipment) have re-introduced this method which doubles on-site productivity and, if applied to the renewals re-railing workbank in CP4, could lead to significant annual savings for Network Rail
Formation rehabilitation trains	Modern high output European plant is regularly used to undertake formation and also ballast renewals. If applied to Network Rail's CP4 category 7 and 12 track renewals RailKonsult estimate that it could reduce unit costs for both activities by around 40%
Lightweight station platforms	The use of modular construction polystyrene station platforms in the Netherlands could provide opportunities in Britain, given the substantial CP4 platform extension workbank. Analysis suggests a unit cost saving of around 25% in Britain

Efficient European re-railing techniques.	This particular study brought together many themes from the previous RailKonsult work by focussing upon the Swiss re-railing method. Bespoke plant, high output welding techniques and dedicated teams are applied routinely. Put together for basic re-railing work alone this method is around 40% more efficient than current Network Rail practice
Use of dedicated teams	Contractors are widely used by most continental railways, as they are in Britain. However there is generally a greater degree of specialisation by activity in Europe (such as S&C renewal or tamping). This ensures a highly skilled and productive workforce dedicated to particular tasks in contrast to the situation in Britain where contractors are often not even dedicated to rail.

Source: Taken from Smith *et al*., 2010.

Although ORR carried out / commissioned a wide range of studies – all of which pointed in the direction of a large efficiency gap – it was the output of the LICB-based econometric model which was used to set Network Rail's efficiency targets. ORR chose to compare Network Rail against the upper quartile of the peer group, rather than the frontier, thus meaning that the starting efficiency gap for its analysis – based on the preferred econometric model from the analysis of the LICB data- was 37% rather than 40%. ORR also gave the company ten years to close the gap, with only two thirds of the gap targeted to be closed during the immediate control period (control period 4 (CP4); 2009-2014).

In the next periodic review (PR13), ORR shifted the emphasis of its approach to bottom-up methods. This was driven by a number of factors, but in part reflected increased doubts after 2008 about the quality of the LICB data and the commitment of the different companies to providing accurate information. A re-run of the sub-company approach was also attempted, but again it was considered that there was insufficient time to get enough certainty about the quality and comparability of the data received. Therefore, although Network Rail acknowledged the size of the efficiency gap resulting from the PR08 econometric modelling, emphasis switched in the PR13 review to bottom-up analysis. Whilst new econometric modelling with an updated LICB dataset was carried out and reported, in the process also applying more advanced techniques (including the more recent methods set out in section 3), the econometric modelling played a supporting role to the bottom-up analysis (thus reversing the approach taken in PR08; see ORR (2013).

Perhaps one of the lessons that may be learned here is that international benchmarking is inherently problematic because it takes considerable time and commitment from a group of countries to make the analysis credible and usable. In PR08 ORR had the advantage of a ready-to-go dataset, produced by UIC, and this enabled top down, econometric international benchmarking to play a more significant role than it has in other regulated sectors. A further factor at play in PR08 was the sheer size of the cost increases that had occurred and the scale of the cost challenge, and given the lack of domestic comparators, there was a strong need imperative to use this kind of approach. In PR13, with Network Rail acknowledging the size of the gap, with the gap closing already, and with new uncertainties arising in the LICB dataset, ORR has become much more cautious about top down international benchmarking, and arguably the need to rely on it has become less. ORR also considers that its leading role in

collecting new data for the sub-company modelling approach is questionable – and that perhaps this kind of work should be led and funded by companies rather than regulators.

One consideration going forward, however, is the extent to which ORR, and regulators in general, can avoid the use of top-down benchmarking. Further, if international benchmarking is to work, then it may require concerted efforts by regulators / governments across Europe working together to establish a common benchmarking framework against which all companies can be compared, thus also implying that data can be requested and audited by regulators and policy makers. Finally, a further opportunity for benchmarking remains the notion of internal benchmarking. Whilst not without its problems it remains a useful part of a regulators toolkit as it establishes the savings that could be achieved if best practice (within-country) is consistently applied. The existence of disaggregation into units that have managerial autonomy (at least to some degree) , as with Network Rail's routes, is of course a pre-requisite for such an approach, but these groupings / disaggregations do also exist in other railways.

5. European rail systems

The previous section considered the impact of reforms on costs in Britain and attempts to benchmark Network Rail (and to an extent the TOCs) with a view to challenging their costs and improve efficiency. In this section we consider wider European experience. Most past studies on the impact of reforms at the European level have applied data envelopment analysis to physical data; our problems with that approach have been outlined above. Moreover they have usually used the data published by the Union International des Chemins de Fer[5], data which has been shown to contain inconsistencies (van de Velde et. al., 2012). Moreover this source only contains data on UIC members, generally the incumbent but not new entrants, and in some cases covers their activities in a number of countries rather than just their home country. A rare example of the estimation of cost functions to study the impact of European reforms is Asmild et al (2008); they also went to considerable efforts to clean up and supplement the UIC data. They found that competitive tendering for passenger, open access for freight services and accounting separation of infrastructure from operations all improved efficiency, but could find no further effect of complete separation of infrastructure from operations. However, their data series ended in 2001 before many reforms took place.

A recent example of the use of a translog cost function with panel data from a large number of countries is Mizutani and Uranishi (2013). They used data for 30 railway companies from 23 OECD countries for the years 1994 to 2007, giving 420 observations. Whilst most of the observations were from Europe, they included the vertically integrated passenger railways of Japan, and also South Korea and Turkey. Where vertical separation had been implemented, they added together the infrastructure manager and the train operating company to form a

[5] Note this is published data as distinct from the confidential data from the LICB project described earlier.

single observation. The basic source of data was UIC, but this was supplemented as necessary by data from company annual reports.

Two separate models were estimated, one using passenger kilometres and freight tonne kilometres as outputs and the other total train kilometres, with the share of passenger revenue to total revenue, passenger load factors and length of haul and freight number of cars per train to reflect differences in the characteristics and therefore costs of the train kilometres. Factor prices for labour, track, rolling stock and other materials were estimated. Finally route kilometres, train kilometres per route kilometre and the percentage of line electrified were included as descriptors of the network.

The rail reform variables were dummies reflecting complete vertical separation (the holding company model being regarded as integrated) and horizontal separation of passenger and freight operations. It was found that whilst horizontal separation unequivocally reduces costs, vertical integration only reduced costs for densely trafficked railways; for most European railways it increased them. Given that there are no separate variables representing the degree to which competition is permitted or actually takes place, it must be assumed that these impacts are the net effect of any additional costs directly caused by vertical separation and of the impact of competition which in most of Europe must presumably be sufficient to outweigh these costs. The explanation given for the impact on costs varying with density is that given above, that the transactions costs caused by vertical separation will be much greater in densely trafficked networks than in less densely trafficked ones.

Van de Velde et al (2012) takes this work further by updating and improving the data set and introducing separate dummy variables for holding companies and complete vertical separation (this work also later published in the academic literature, see Nash et. al. forthcoming and Mizutani et. al, 2014). They also added in Britain, the country in which the most radical reforms had taken place, but which had been excluded from most previous studies due to lack of data. Finally, they introduce dummy variables representing passenger and freight market competition.

They confirm the previous finding that, compared with complete vertical integration, vertical separation reduces costs at low levels of density but increases them at high; at mean European density levels costs are not affected by the change. This effect is not likely to be one of pure transactions costs (negotiating and enforcing contracts), which have been shown to be a relatively small proportion of total systems costs (Merkert et al, 2012) but is more likely a problem of misalignment of incentives leading to poor integration of infrastructure and operations in circumstances (dense traffic) when this is particularly important. They find weak evidence (significant at 10% only) that the holding company model reduces costs compared with vertical integration, but this does not vary with density, so the holding company would be preferred to vertical separation at high levels of density but not at low.

Within the range of the data, the introduction of competition seems to have had no effect on costs. Horizontal separation of freight and passenger undertakings seems to have sharply reduced costs (perhaps because this has typically been associated with preparation of the freight undertaking for privatization), whilst a high proportion of revenue coming from freight rather than passenger tends to increase the costs of vertical separation (perhaps because planning freight services efficiently requires closer day to day working than passenger, since

freight services vary from day to day whereas passenger services are generally fixed for the duration of the timetable). The paper also provides qualitative evidence on the issue of how misalignment of incentives may raise costs and show how, whilst efficiently set track access charges and performance regimes are important, they do not provide incentives for railway undertakings to assist infrastructure managers in seeking the minimum cost solution to infrastructure provision. Only a complete sharing of changes in costs and revenues, as provided for in some of the alliances now being negotiated in Britain, will achieve that.

The conclusion of the studies in this section is that there is no one size fits all policy for European railways. Based on a mixture of qualitative and quantitative research the evidence suggests that vertical separation may perform less well than the Holding company model for intensely used networks, whilst being the structure of choice for less dense networks. Whilst this is in part intuitive, it is not totally clear why separation reduces costs for lightly used networks, particularly if there is little competition. It is further disconcerting that it has not been possible to find clear competition effects in the data. To date no research has yet been published to consider the impact of regulation on costs, though research is ongoing at University of Leeds in this respect. A final note must be that although we consider the cost function approach to be the best, and with the van de Velde et. al. (2012) study incorporating new data from CER members to supplement published data, there nevertheless remains work to be done on the data side to improve its comparability.

6. Conclusions

There are various measures of efficiency available, and all have their uses. Partial productivity measures may have their value in shedding light on utilisation of particular resources, but they are inadequate as measures of overall efficiency, such as are needed in econometric studies of the influences of alternative policies on efficiency of the rail industry as a whole. For this there are a number of possibilities, including total factor productivity measures and Data Envelopment Analysis, but for various practical reasons we prefer the econometric estimation of cost functions.

Detailed studies of Britain suggest that the reforms in Britain did not achieve lasting improvements in cost efficiency. In the case of the infrastructure manager, the events surrounding the placing of Railtrack into administration brought about a major reduction in efficiency from which the current infrastructure manager, Network Rail, is only gradually recovering. Surprisingly, competitive tendering did not achieve a reduction of train operating cost, which has risen substantially. There appear to be various reasons for that, but the use of management contracts to deal with cases of financial failure, the use of too large areas as the basis for franchising (from a cost and risk of failure point of view) and the loss of economies of density through overlapping franchises all seem to be factors in this (though it must be noted that splitting franchises to reduce their size increases the prospect of franchise overlaps). Most continental franchises are smaller and subject to less overlap of homogenous

services, and they appear to be more successful. However, it must still be noted there that the studies carried out other than in Britain are based on data on subsidies and not cost, and it may be that costs have been reallocated elsewhere in the system.

Studies of European-wide reforms suffer even more data problems than studies in a single country, but a recent study which sought to overcome these was Van de Velde et al (2012) using part published data and new data provided by CER members. They found that complete vertical separation only reduces cost in low density countries; at higher densities it raises cost. The reason for this is likely to lie in the misalignment of incentives, a factor considered important in Britain by the McNulty report (McNulty, 2011). The level of competition in both freight and passenger markets appears to be an insignificant factor in determining costs.

It is clear from all these studies that much still needs to be done to understand the determinants of rail cost efficiency. The most fundamental requirement, whether within or across countries, is for good quality consistent data on the variables we are seeking to measure (in cost analysis, a key area here is the costs of depreciation and interest). Better measures of policy variables are needed, including in particular the quality of regulation and the degree of competition. At the same time, the measurement of service quality is an important factor – as noted above, no sensible operator minimises costs wholly at the expense of service quality.

The data challenges are unlikely to be overcome based on published data. Experience suggests that some progress can be achieved by working with companies, however this takes time and commitment, and sometimes companies may start involvement in a benchmarking project, only to withdraw later. Long-term commitment to developing a framework and data collection exercise is required. Ultimately it may be that data issues cannot be fully addressed without concerted, European-wide effort by multiple policy makers and regulators to require data to be collected to a common set of definitions, with appropriate audits in place to achieve comparability.

Whilst the data will never be perfect and there will never be data on every possible variable that may drive differences between companies, methodological advances mean that, with enough data points (and with panel data), there are ways of disentangling inefficiency from unobserved factors that cause costs to vary between railways. Thus we consider that it is both possible and sensible to continue to develop our understanding of railway costs and efficiency performance using existing datasets, improved where possible, and applying state of the art methods.

Finally we note that the challenges facing policy makers seeking to compare performance are further hampered by the increased need for railways to improve the quality of what they deliver, whilst expanding capacity, reducing carbon, and also responding to the increased challenges posed by climate change. All of these raise new data and methodological challenges which researchers and policy makers will need to grapple with in the coming years.

References

Accenture, 2003. Review of Network Rail's Supply Chain. London.

Affuso, L., A. Angeriz, and M.G. Pollitt (2003): 'Measuring the Efficiency of Britain's Privatised Train Operating Companies', mimeo (unpublished version provided by the authors).

Alexandersson, G. and S. Hulten (2007): 'Competitive tendering of regional and interregional rail services in Sweden', in European Conference of Ministers of Transport Competitive Tendering for Rail Services, Paris.

Arrow, K. J., H. B. Chenery, B. S. Minhas, and R. M. Solow (1961),"Capital-Labor Substitution and Economic Efficiency," Review of Economics and Statistics, August, pp. 225—250.

Asmild, M., T. Holvad, J.L., Hougaard, D. Kronborg, D. (2009), 'Railway reforms: Do they influence operating efficiency?', Transportation, 36 (5), 617-638.

Battese, G. E. and Coelli, T. J. (1995). "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data," Empirical Economics 20, 325-332.

Beattie and Taylor (1985)

Brown, M. (1962), "The Constant Elasticity of Substitution Beattie, B.R. and Taylor, C.R. (1985), The Economics of Production, New York, Wiley.

Berechman, J. (1993), Public transit economics and deregulation policy. Amsterdam: North Holland.

Bitzan, J (2003) Railroad costs and competition. Journal of Transport Economics and Policy. 37 (2) 201-225

Bitzan, J. D. And Wilson , W.W. (2008). 'A hedonic Cost Function Approach to Estimating Railroad Costs'. *Research in Transportation Economics*, 20. 69-95.

Brenck, H. and Peter, M. (2007), Experience with Competitive Tendering in Germany, In European Conference of Ministers of Transport (2007) Competitive tendering for rail services, ECMT, Paris.

Cantos P., J.M. Pastor and L. Serrano (2010), 'Vertical and horizontal separation in the European railway sector and its effects on productivity', *Journal of Transport Economics and Policy*, **44** (2) 139-160.

Cantos P., J.M. Pastor and L. Serrano (2011), 'Evaluating European railway deregulation using different approaches', Paper given at the workshop on Competition and Regulation in Railways, FEDEA, Madrid, March 12.

Christensen, L., Jorgenson, D., and. Lau, L. (1971), Conjugate Duality and the Transcendental Logarithmic Production Function, Econometrica, 39, pp 255-256.

Christensen, L. R., Jorgenson, D.W. and Lau, L.J. (1973), 'Transcendental Logarithmic Production Frontiers', *Review of Economics and Statistics*, vol. 55, pp. 28-45.

Coelli, T., Perelman, S. & Romano, E. 1999. Accounting for environmental influences in stochastic frontier models: With application to international airlines. *Journal of Productivity Analysis,* 11**,** pp. 251-273.

Colombi, R., Kumbhakar, S.C., Martini, G. and Vittadini, G. (2104), 'Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency', Journal of Productivity Analysis, 42, pp. 123-136.

Cornwell, Snmidt and Sickles (1990)

Cowie, J. (2002a). 'Subsidy and Productivity in the Privatised British Passenger Railway', *Economic Issues* 7 (1), 25-37 38.

Cowie, J. (2002b). 'The Production Economics of a Vertically Separated Railway – The Case of the British Train Operating Companies', *Trasporti Europei,* August 2002, 96-103.

Cowie, J. (2005). 'Technical Efficiency versus Technical Change – The British Passenger Train Operators'. In Hensher, D. A. Ed (2005). *Competition and ownership in land passenger transport: selected refereed papers from the 8th International Conference (Thredbo 8)* Rio de Janeiro, September 2003. Amsterdam ; London : Elsevier, 2005.

Cowie, J. (2009). 'The British Passenger Rail Privatisation: Conclusions on Subsidy and Efficiency from the First Round of Franchises', *Journal of Transport Economics and Policy,* 43(1), 85-104.

Cuesta, R. A. (2000). 'A production model with firm-specific temporal variation in technical inefficiency: with Application to Spanish Dairy Farms.', *Journal of Productivity Analysis*. 13 (2), 139-152.

Diewert, W. (1971): ''An Application of Shephard Duality Theorem: A Generalised Leontief Production Function,'' Journal of Political Economy, 79, 481–507.

Farsi, M., Filippini, M. and Kuenzle, M. 2005. Unobserved heterogeneity in stochastic cost frontier models: an application to Swiss nursing homes. Applied Economics, 37(18): 2127-2141.

Friebel, G., M. Ivaldi, and C. Vibes (2010): 'Railway (De) regulation: a European efficiency comparison', *Economica,* 77, 77-91.

Greene, W. (2005), 'Reconsidering heterogeneity in panel data estimators of the stochastic frontier model', *Journal of Econometrics,* vol. 126, pp. 269-303.

Growitsch, C. and Wetzel, H., 2009. Testing for economies of scope in European Railways: an efficiency analysis. Journal of Transport Economics and Policy, 43 (1) 1-24.

Kennedy, J. and Smith, A.S.J (2004), 'Assessing the Efficient Cost of Sustaining Britain's Rail Network: Perspectives Based on Zonal Comparisons', *Journal of Transport Economics and Policy*, vol. 38 (2), pp. 157-190.

Kogan, J. (2006): 'Latin America: Competition for Concessions', in Jose Gomez-Ibanez and Gines de Rus (eds.) *Competition in the railway industry: an international comparative analysis*, Edward Elgar, Cheltenham.

Kumbhakar , S.C. and Lovell, C.A.K (2000). *Stochastic Frontier Analysis*, Cambridge University Press, Cambridge UK.

Lalive, R. and Schmutzler, A. (2008). Entry in Liberalized Railway Markets: The GermanExperience. Review of Network Economics, Vol. 7, no. 1, pp. 37-52

LEK, TTCI, Halcrow, 2003. Bottom-up review of Network Rail's business plan: 2003/04-2005/06. Report to ORR. London.

McNulty, Sir R (2011) Realising the potential of GB Rail: final independent report of the Rail Value for Money study. Department for Transport and Office of Rail Regulation, London.

Merkert, R. Smith, A.S.J. and Nash, C.A. (2009). 'Benchmarking of train operating firms - A transaction cost efficiency analysis', *Journal of Transportation Planning and Technology*.

Merkert, R., Smith, A.S.J. and Nash, C.A. (2012), 'The measurement of transaction costs – evidence from European railways', *Journal of Transport Economics and Policy*, **46** (3), 349-365.

Mizutani, F. and S. Uranishi, (2013), 'Does vertical separation reduce cost? An empirical analysis of the rail industry in OECD countries', *Journal of Regulatory Economics*. **48** (1), 31-59.

Mizutani, F. Smith, A.S.J., Nash, C.A. and Uranishi, S. (2014). Comparing the Costs of Vertical Separation, Integration, and Intermediate Organisational Structures in European and East Asian OECD Railways, mimeo (available from the authors on request).

Nash C A and Nilsson J E (2009) Competitive tendering of rail services – a comparison of Britain and Sweden Paper presented at the Thredbo 11 conference, Delft.

Nash, C A, Nilsson JE and Link H (2013), 'Comparing three models for introduction of competition into railways', *Journal of Transport Economics and Policy*, **47**, Part 2, May 2013, 191–206.

Nash, C.A., A. S. J. Smith, D. van de Velde, F. Mizutani and S. Uranishi (forthcoming): Structural reforms in the railways: incentive misalignment and cost implications, Research in Transportation Economics.

Nash, C.A. and Smith, A.S.J. (2007), 'Modelling Performance: Rail', in Hensher, D.A. and Button, K.J. eds., *Handbook of Transport Modelling, Second Edition*, Elsevier.

ORR (2013).

Oum T.H. and Yu, C. (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, W.G. (II) and Yu, C. (1999). 'A Survey of Productivity and Efficiency Measurement in Rail Transport', *Journal of Transport Economics and Policy*, 33 (I), 9-42.

Oum, T. H. and Zhang, Y. (1997). 'A Note on Scale Economies in Transport'. *Journal of Transport Economics and Policy,* 309-315.

Pitt and Lee (1981), Preston (2008).

Schmidt, P. and Sickles, R.C. (1984), 'Production Frontiers and Panel Data', Journal of Business & Economic Statistics, vol. 2 (4), pp 367-374.

Shephard, R. W. (1953). *Cost and Production Functions.* Princeton: Princeton University Press.

Smith, A.S.J. (2004),' Essays on Rail Regulation: Analysis of the British Privatisation Experience, Doctoral Thesis, University of Cambridge.

Smith, A.S.J. (2006). 'Are Britain's Railways Costing Too Much? Perspectives Based on TFP Comparisons with British Rail; 1963-2002', *Journal of Transport Economics and Policy*, 40 (1), 1-45.

Smith, A.S.J. (2012). 'The application of stochastic frontier panel models in economic regulation: Experience from the European rail sector', *Transportation Research Part E*, 48, 503–515.

Smith, A.S.J., and P.E. Wheat (2009): 'The Effect of Franchising on Cost Efficiency: Evidence from the Passenger Rail Sector in Britain', *11th Conference on Competition and Ownership in Land Passenger Transport*, Delft University of Technology, The Netherlands, 20-25 September 2009

Smith, A.S.J. and Wheat, P.E. (2012), 'Estimation of Cost Inefficiency in Panel Data Models with Firm Specific and Sub-Company Specific Effects, Journal of Productivity Analysis, vol. 37, pp. 27-40.

Smith, A.S.J, Nash, C. and Wheat, P. (2009). 'Passenger Rail Franchising in Britain – has it been a success?*' International Journal of Transport Economics*, 36 (1), 33-62.

Smith, A.S.J., Wheat, P. and Nash, C.A. (2010). 'Exploring the Effects of Passenger Rail Franchising in Britain: Evidence from the First Two Rounds of Franchising (1997-2008)', *Research in Transportation Economics,* 29 (1) 72-79.

Smith, A.S.J., Wheat, P.E. and Smith, G. (2010), 'The role of international benchmarking in developing rail infrastructure efficiency estimates', *Utilities Policy*, vol. 18, 86-93.

Spady, R. H. and Friedlaender, A. F. (1978). 'Hedonic cost functions for the regulated trucking industry'. *The Bell Journal of Economics*, 9 (1), 159–179.

van de Velde, D., C. Nash, A. Smith, F. Mizutani, S. Uranishi, M. Lijesen and F. Zschoche (2012),"EVES-Rail - Economic effects of Vertical Separation in the railway sector", Report for CER - Community of European Railway and Infrastructure Companies, by inno-V (Amsterdam) in cooperation with University of Leeds – ITS, Kobe University, VU Amsterdam University and Civity management consultants, Amsterdam/Brussels, 188 pp.

Wheat, P.E. and Smith, A.S.J. (2014), Do the usual results of railway economies of scale and density hold in the case of heterogeneity in outputs: A hedonic cost function approach, Journal of Transport Economics and Policy (online first, January 2014).

Zellner, A. (1962), 'An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias, Journal of American Statistical Association, 57, pp. 348-368.