Econometric cost analysis in vertically separated railways

Phillip Edward Wheat

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Institute for Transport Studies

December 2013

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 5 is based on Wheat, P. and Smith, A. (2013). ‘Do the usual results of railway returns to scale and density hold in the case of heterogeneity in outputs: A hedonic cost function approach’. Accepted for publication in Journal of Transport Economics and Policy. My contribution was scoping, undertaking the analysis and drafting the paper. Dr. Andrew Smith’s contribution was general supervision and comments in line with that expected by PhD supervisors.

Chapter 6 is based on Smith, A. and Wheat, P. (2012). ‘Estimation of Cost Inefficiency in Panel Data Models with Firm Specific and Sub-Company Specific Effects’, Journal of Productivity Analysis. 37 (1) 27-40. My contribution was developing the analytical framework with Dr. Andrew Smith and the interpretation (section 1, 2 and 3). I was the lead in the estimation section 4 and was also the one who undertook the empirical example (section 5).

Chapter 7 is based on Wheat, P., Greene, W. and Smith, A. (2013). ‘Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models', Journal of Productivity Analysis. In print. I was lead contributor and drafted all of the paper. The co-authors share intellectual property for the ideas, helped formulate the paper and develop the ideas and commented on/revised the paper.

Acknowledgements

This research has been carried out by a team which has included Dr. Andrew S. J. Smith[1] and Professor William H. Greene[2]. My own contributions, fully and explicitly indicated in the thesis, have been made clear on the previous pages with respect to the joint papers that have been transformed to provide the basis for the three core research chapters in this thesis. The contributions by the above mentioned team members to the research are also detailed on the previous page. In line with the research degree regulations, I have, to the best of my ability, avoided including elements of the published paper that I did not write, in consultation with my supervisors.

In addition to the joint work declaration, I acknowledge and thank my supervisors (Dr. Andrew Smith and Professor Chris Nash) for their help and advice with respect to this work. I thank them for the support they have provided, particularly given my relatively unconventional study arrangements.

The work has been supported financially through a part-time Engineering and Physical Science Research Council studentship and also through my employment as a Senior Research Fellow at the University of Leeds. As part of this employment, I have been fortunate to work closely with the British Office of Rail Regulation which has actively supported much of the work in this thesis. I also acknowledge Network Rail, the Department for Transport and the Office of Rail Regulation for providing

[1] University of Leeds. Dr. Smith is also one of the supervisors for this thesis.

[2] Department of Economics, Stern School of Business, University of New York, New York City, NY, USA

access to data.

I am indebted to my wife, Alison, for her understanding and support especially during the writing up phase of this thesis, and my young son, Ben, for providing perspective.

Finally, following examination, this version of the thesis contains minor editorial changes up to 1st September 2014.

Abstract

This thesis is concerned with addressing the cost analysis challenges in vertically separated railways. Both the challenges in infrastructure management and passenger railway operations are considered.

A hedonic cost function is applied to better incorporate measures of heterogeneity of output into passenger train operation cost analysis. This allows for a richer understanding of the cost structure of the industry, through explicitly making scale economies a function of output heterogeneity, which in turn allows for tenders to be specified in a cost minimising manner. Three example tender re-mappings are considered for Britain. It is estimated that two out of three actually increase costs, reflecting that the model implies that for very heterogeneous TOCs, returns to density can not be exploited.

In addition, the thesis details methodological work in developing analytical frameworks to exploit a multi layer panel dataset comprising observations on regions of many individual infrastructure managers. As well as providing extra observations to estimate cost frontiers, the data structure permits estimation of a dual-level inefficiency model which separates sub-company persistent inefficiency from sub-company varying inefficiency. This decomposition gives an indication as to whether inefficiency predominantly varies within firm or between firms. The example shows statistically significant inefficiency variation at both levels, and importantly, failure to take into account the dual-level nature of inefficiency is shown to under predict inefficiency.

The thesis also develops new techniques to quantify uncertainty in inefficiency predictions from stochastic frontier models. This has application across the applied efficiency analysis discipline and not just in railways.

Overall, this thesis finds that robust cost and efficiency analysis can only be conducted through explicit allowance for heterogeneity in output (both observed and unobserved), ensuring sufficient data quantity and that data relates to the organisational level to which decisions are made and that, in any analysis, the impact of uncertainty is quantified.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CD	Cobb Douglas
CDF	Cumulative Distribution Function
COLS	Corrected Ordinary Least Squares
CSN	Closed Skew Normal
DEA	Data Envelopment Analysis
DfT	Department for Transport
GLS	Generalised Least Squares
HPD	Highest Posterior Density
HS intervals	Prediction intervals proposed by Horrace and Schmidt (1996)
IM	Infrastructure Manager
LICB	Lasting Infrastructure Cost Benchmarking
LR	Likelihood Ratio
ML	Maximum Likelihood
OLS	Ordinary Least Squares
ORR	Office of Rail Regulation
PDF	Probability Density Function
RtD	Returns to Density
RtS	Returns to Scale
SUR	Seemingly Unrelated Regression
TFP	Total Factor Productivity
TOC	Train Operating Company
UIC	International Union of Railways
UK	United Kingdom

1. Introduction

1.1. Thesis motivation

This thesis is concerned with addressing the cost analysis challenges in vertically separated railways. Vertical separation refers to separation in management and administration of different aspects of the production process. For the purpose of this thesis, vertically separated railways refer to railways where passenger and freight operations (the running of trains) are separate to the infrastructure. There can be other forms of vertical separation in railways such as contracting out maintenance and/or renewal activity by the infrastructure manager but this is not the primary area of interest of the thesis.

1.1.1. Policy Context

Railways throughout the world have undergone varying degrees of liberalisation over the last fifty years or so. Generally, such liberalisation has been motivated by falling market share and worsening financial performance. In the European Union several European Commission Directives have been enacted which aim to introduce competition (especially for freight and cross border passenger) and cost transparency within railways through separation of accounts between infrastructure management and train operation functions. In particular Directive 91/440 requires countries to have separate accounts for infrastructure management and train operations and price for access to their infrastructure on the basis of incremental (marginal) cost, be it with an allowance for non-discriminatory mark-ups. Importantly, different countries have adopted different organisational systems for their railways which still may conform to the Directive's requirements. At one extreme, Britain has opted for institutional separation between infrastructure and operations while Germany has essentially opted for divisions of activity (with separated accounts) within the same holding company. In recent legislation proposals, the European Commission has set out further steps to encourage competitive tendering in domestic passenger train operations (European Commission, 2013).

In Britain, privatisation and vertical separation took place between 1993 and 1997 resulting in the formation of a private (subsequently a private firm limited by government guarantee) monopoly infrastructure manager and franchised passenger railway services. This is one of the most radical reforms to railways undertaken within Europe and is certainly one of the fastest reforms (Jupe and Crompton, 2006).

There have been several studies which have documented the performance of the Great British Railway since privatisation. Demand for rail services has increased by 57% since privatisation (McNulty (2011) Total Passenger Journeys 1996/97-2009/10). Furthermore, Wardman (2006) concludes that this growth is in excess of that implied by conventional demand models as detailed in the Passenger Demand Forecasting Handbook, (Passenger Demand Forecasting Council, 2013), which is the accepted industry applied demand forecasting methodology in Britain, which take into such factors as income growth, generalised cost of rail and relative generalised costs of competing modes. Thus it seems that there is a positive privatisation effect, which has yet to be fully explained.

However, on the cost side, the result has been an increase in costs. Costs did initially fall after privatisation resulting in subsidy falling from £2.9bn in 1992/93 to 1.8bn in 2000/01 in 2012 prices (Smith, Nash and Wheat, 2009). However it is reasonably well documented in both the industry, academia and the wider press, that the costs of maintaining, renewing and operating the infrastructure have increased considerably which coincided both directly and indirectly with the Hatfield accident in October 2000. This accident was found to have resulted from substandard maintenance practices and prompted an increase in maintenance and especially renewal activity. Costs for infrastructure rose from £4.5bn in 1999/00 to £8.5bn million in 2003/04 in 2012 prices, an increase of 88% (Wheat, Smith and Nash, 2007). This increase was a contribution of three factors. The first is that because of the lumpy nature of renewals expenditure (in practice track is not installed at a uniform rate and to uniform capability over time), there may be a cost minimising 'need' for expenditures to be greater due to the natural renewal cycle. The second is that at or even before privatisation, expenditures were below steady state requirements so there is a need to temporarily do more maintenance, and especially renewal, to catch-up. Thirdly, the infrastructure manager could have become less efficient in conducting or planning its maintenance and renewal activities.

The industry structure requires that the Office of Rail Regulation (ORR) regulates the infrastructure manager such that it does not abuse its market power as a monopoly provider of railway infrastructure, while still being allowed to make a reasonable return on its assets. The latter requirement is important given that the infrastructure is subject to returns to density, which means that the marginal cost of train usage is below average cost. Some combination of mark-ups above marginal cost, two part tariffs and direct subsidy are required for the infrastructure manager to earn normal profit.

The ORR adopts, like other UK regulators, price cap (also known as RPI-X) regulation (Beesley and Littlechild, 1988) which requires the regulator to determine Network Rail's revenue requirement at each Periodic Review (typically every five years). RPI-X regulation encourages firms to operate in a cost efficient manner as the firm can retain any super-normal profit over and above that set by the price caps/direct subsidy level. Importantly, however, at the Periodic Review period, the revised price caps/subsidy must 'recapture' this over performance (or be relaxed in the case of underperformance) to prevent the infrastructure manager from abusing its monopoly position (or making an unviable loss in the case of previous under performance). Furthermore, the regulator needs to set a challenging profile of price caps and subsidy so as not to allow the firm to make excessive super-normal profits. Therefore the regulator needs to understand the scope for cost saving reductions available to the infrastructure manager. A key strand of research to inform this understanding is efficiency benchmarking. Within this is a need to understand the robustness of any analysis; quantifying uncertainty in efficiency predictions is important. Railways are not unique in being subjected to economic regulation. Many network industries are similarly regulated. Thus this thesis has many transferable aspects outside of railways.

As well as the infrastructure cost story, less well documented is that train operating companies costs (TOCs costs) have increased from £10.11 per train-km[3] in 1999/00 to £13.65 per train-km in 2005/06, a 35% increase, and a 15% increase since privatisation (privatisation completed 1996/97) (Smith, Nash and Wheat, 2009). The reasons for this increase are less understood. Smith, Nash and Wheat discuss three possible explanations. The first is that costs initially did not fall as much as expected because they had already been reduced during the pre-privatisation period. Perhaps as a response, there may have been some unsustainable cost cutting in the earlier years which required increased expenditure in later years. Secondly, franchising policy has been through at least three phases (Smith, Nash and Wheat, 2009). In particular between 2000 and 2004 several train operating companies were moved on to management contracts, as they were unable to meet their franchise obligations following too optimistic bidding for the initial franchises. The incentive properties of such contracts are known to be weak, because the contract is analogous to rate of return regulation, in the sense that the firm is allowed to make a set return on its costs, potentially irrespective of the level of the costs. Therefore there is little incentive for the firm to cut costs. Thirdly there may have been external factors such as more stringent health and safety regulations and better quality rolling stock which have increased costs. For example, the average fleet age was 13 years in 2006/07 versus 21 years in 2000/01. However, given the evidence available to Smith, Nash and Wheat (2009), it was difficult to determine which factors were most important.

The McNulty review was an independent study funded by the ORR and DfT to examine the “major problem of efficiency and costs” (McNulty, 2011 p. 8). It concluded that the railway in Britain had costs which were approximately 40%

[3] Excluding access charge payments, 2012 prices

higher than costs elsewhere. The study identified a lack of focus on cost reduction as a major reason for the high industry costs (McNulty, 2011 p. 37). In addition a key recommendation (McNulty, 2011, Chapter 7) is the need for the ORR to undertake benchmarking across both the infrastructure manager and passenger train operating companies. The government has accepted the main recommendation of the McNulty Study (Department for Transport, 2012). Thus in Britain there is a clear mandate for the continuation of cost analysis and cost efficiency analysis in railways.

1.1.2. Pressing issues

Given the above policy context, it is possible to identify several pressing issues within the sector:

- For passenger train operations, the use of competitive tendering requires the need to understand the returns to scale and density properties of the activity. Competition for the market (via competitive tendering) lessens the need for explicit economic regulation but tenders have to be designed in a manner to best exploit the cost structure within the constraints of passenger rail demand. The issue is becoming more pressing given the move to competitive tendering of passenger railway operations at the EU level and the current re-mapping of franchises in Britain where reducing cost is a major motivation. Understanding the scale and density properties of passenger train operations allows for tenders to be designed so as to minimise overall costs (of providing the passenger trains services). This includes understanding of both how costs are affected by the overall geographical size of a tender and also how costs are affected from combining TOCs that run on the same network, taking into account that they may be running different types of services.

In practice tenders may not be specified to minimise total cost if doing so would eliminate any (beneficial) competition (perhaps there are benefits to passengers for example from multiple train operators on one section of track). In such a case a full cost and benefit analysis needs to be undertaken, but this is not within the scope of this thesis. However, a key input into such an analysis is related to how (average) costs change between various mapping scenarios. Thus the motivation for the analysis is still present (and indeed the research does consider examples of such discrete changes).

- For railway infrastructure managers, economic regulation is much more important (than in operations) due to the cost structure of the industry leading to a single monopoly company for a country (or at least for a specific region). Furthermore, long asset lives makes competition for the market (through tenders) problematic to implement. At the EU level, there is a desire to make rail more competitive; this requires a reduction in railway costs (European Commission, 2013). At the level of Britain, the Office of Rail Regulation (ORR) has to promote efficiency with respect to the infrastructure manger (section 4(1)(c), Railways Act, 1993). Given the recent rises in infrastructure costs there is a pressing public interest in assessing the scope for cost reduction.

As such, developing models to determine the efficiency of infrastructure managers is very important. Given the single monopoly supplier in Britain, there are two broad ways of undertaking efficiency analysis. One approach is to seek international comparators. This however presents difficulties in both getting access to data (and over enough years) and validating that the data (particularly cost data) conforms to comparable definitions. A further approach is to undertake internal benchmarking of the firm. This is limited as it only reveals differences in performance within the firm and not systematic failings of the firm relative to best practice elsewhere.

There is therefore a need to utilise alternative dataset formulations. In this thesis, techniques to analyse multi-layer panel datasets are presented. Utilising data which includes multiple observations within a firm (geographically disaggregated) over a number of firms can be seen as a way of combining the two approaches. Utilising data of this type also allows for scale and density to be modelled at the level at which management decisions are made, which guards against aggregation bias (Theil, 1954). This approach is feasible and was used by ORR in the 2008 Periodic Review (ORR, 2008).

- More generally, there is a need to understand the uncertainty in efficiency predictions. Public policy is increasingly interested in understanding uncertainty associated with analysis which supports decision making. This is particularly important with respect to economic regulation given the large expenditure under consideration and further with respect to railways, where a large amount of public money is provided to the industry. Further, the Rail Regulator in Britain (ORR) has adopted relatively sophisticated statistical techniques (with respect to those used by regulators of other network industries) for their efficiency assessment. Thus understanding uncertainty in these predictions is important as the debate surrounding the usefulness of these techniques progresses (this was recommended by a peer review of work undertaken in the 2008 Periodic Review (OXERA, 2009)).

1.2. Thesis aims and objectives

The aim of this thesis is to apply appropriate econometric techniques to better analyse the cost structure of vertically separated railways - specifically the infrastructure management and passenger train operations activities - to inform regulatory bodies and policy makers. The specific objectives are to:

- Explore the use of a hedonic cost function approach to incorporate measures of output heterogeneity in the analysis of train operating companies (TOCs) costs;
- Provide new empirical evidence as to the cost implications of redrawing franchise boundaries, crucially drawing on the scale and density properties of the estimated model and how these vary with heterogeneity of the TOC's output;
- Explore via econometric analysis the exploitation of a multi-layered panel dataset to predict the inefficiency level of infrastructure managers;
- Provide new empirical evidence on the potential efficiency saving of the infrastructure managers in sample;
- Explore the most appropriate predictor of firm efficiency from parametric stochastic frontier models covering point and interval predictors;
- Provide a new method to incorporate the effect of parameter uncertainty into

predictors of firm efficiency and illustrate these concepts via application to TOCs in Britain.

Each of the three research chapters (Chapters 5 to 7) addresses two of the objectives in turn. Thus a general trend in the research is to both apply, and in some instances develop, innovative methods in railways analysis and also to provide new empirical evidence.

1.3. Structure of the thesis

The remainder of the thesis comprises three literature reviews (Chapters 2, 3 and 4) and then three core research chapters (Chapters 5 to 7). Finally, Chapter 8 concludes.

Chapter 2 provides the economic foundations for cost function analysis. It outlines the key features of a cost function in terms of its description of the underlying technology of production. In particular the cost function relates cost to the level of outputs, level of input prices and (in the case of panel data) some characterisation of how costs change (exogenously) over time – technical change. The cost function concept is further extended to situations in which firms do not successfully minimise cost and in these cases the cost function describes the minimum cost 'frontier' and (positive) deviations from the frontier are evidence of inefficient management. The overall conclusion from this chapter is that the cost function or frontier is an appropriate economic device for this thesis, as it allows for scale and density properties of the technology to be derived (the subject of Chapter 5), allows for cost efficiency to be incorporated (the subject of Chapter 6 and 7) and since outputs are exogenous, this is the usual approach in the literature.

Chapter 3 considers the econometric techniques necessary to estimate the models put forward in the research chapters. The conclusion from this chapter is that the estimation techniques used are well established and the properties of the estimators are known. Importantly inference can be conducted after estimation relating to the population parameters and the estimators are shown to have desirable properties (relative to other estimators), but, as this is the subject of Chapter 7, some challenges remain relating to understanding interval predictors for firm inefficiency.

Chapter 4 then considers relevant past applications to railways. This extracts railway specific issues relating to populating a cost function such as output definition and then summaries the empirical research undertaken. It highlights the limited empirical work done to date on passenger train operations and the well-developed literature on infrastructure cost analysis, albeit with the limitation of established data for comparing one railway with another for efficiency benchmarking purposes.

There are then three research chapters with content as follows.

1.3.1. Chapter 5 – Passenger Train Operating Company cost analysis

As identified in 1.1, a key research need with respect to passenger train operating company cost is to better understand the cost characteristics with respect to the size and output make up of tendered areas. This in turn informs policy as to the optimal size (in terms of minimum cost) of tendered areas given the mixture of services that are to be provided. It also provides useful results with respect to the cost implications of formulating franchises such that there is overlap between franchises (each providing a slightly different service).

The research in this chapter attempts to determine the influence of heterogeneity of output on the scale and density properties of franchises. The dataset has been assembled to include many measures of output and these are formulated as primary outputs and characteristics of output. There is then the issue as to how to enter such numerous measures into a cost function. If the standard Translog cost function is adopted then the model has many parameters (approximately 150 – the sample size is only 243). It is not feasible to adopt a general to specific methodology for such a large model and so the chosen solution is to adopt a hedonic output formulation; that is to nest a function which equates the various characteristics of output within a Translog cost function. This makes the estimated model manageable in terms of parameters.

While overall this methodology is not unique in the cost function literature, it is argued that adopting such methodology allows robust analysis of a relatively complex problem. This is the first time such a methodology has been applied to the study of railway operations. The methodology builds on literature, the innovation being that it allows the interaction of density measures and heterogeneity in a manageable and parsimonious way

1.3.2. Chapter 6 – Infrastructure cost analysis

It was identified in section 1.1 that there was difficulty in assembling sufficient data to undertake robust statistical benchmarking of railway infrastructure managers. A potential solution to this data problem is to utilize data on multiple railway undertakings but disaggregated into regions. This has several advantages over data at the company level. Firstly, this can often be a way to obtain more data points. Ultimately to get the same number of observations for regression, less countries and/or years of data need to be collected since for each year and country there are multiple observations by regions. This is of much practical importance given the issues with collecting data from multiple sources. Secondly, the data structure allows us to distinguish between efficiency variation at two geographical levels; variation which is systematic across firms but the same for all regions within a firm and residual variation across regions within a firm.

Chapter 6 outlines an econometric model to utilize the data structure and measure efficiency at dual-levels. This is a useful decomposition in itself, since it gives an indication as to whether inefficiency predominantly varies within firm or between firms which is useful in terms of identifying where efforts should be made to eliminate inefficiency. Furthermore, the empirical example also indicates that failure to take account of the dual-level inefficiency variation may result in under estimation of inefficiency. The dual-level inefficiency model is applied to data on five railway infrastructure managers, comprising firms from North America alongside European national infrastructure managers (IMs). Each IM in the sample is divided into a number of regions. The chapter also considers the Mundlak (1978) transformation of fixed effects as a means to control for the influence of time invariant unobserved heterogeneity in efficiency models of this sort, in addition to the usual measures of returns to scale and density, thus complementing the work in Chapter 5 which considers incorporation of observed heterogeneity.

1.3.3. Chapter 7 – Uncertainty in efficiency analysis

The final research chapter considers a key challenge that faces economic regulators; namely to understand the uncertainty surrounding the inefficiency estimates derived from their models. This is of crucial importance in utilising the model output in order to produce robust and defensible efficiency targets for the regulated firm. In Britain, this is even more relevant given the maturity of the regulatory process. The current Periodic Review of Network Rail is the fifth undertaken since privatisation and the third since the Hatfield accident that precipitated a large cost rise. The result is that top down benchmarking techniques are coming under more scrutiny. Players are becoming more informed as to their basis. At the same time, the 'efficiency gap' for the regulated firm relative to the best performing firm is decreasing, implying that greater accuracy is needed in predicting efficiency levels.

A set of statistical techniques relate to understanding uncertainty in firm specific predictions of inefficiency. Many empirical studies (and the study in Chapter 6) have simply reported point estimates for firm inefficiency following the methodology of Jondrow et al (1982). However, in cross sectional models, these point predictors are known to be inconsistent for the quantity of interest; namely the firm specific realisation of a random variable. The question then arises; how precise is the prediction of firm inefficiency? With this in mind and the general desire of practitioners to understand uncertainty in their estimates, it is perhaps surprising that interval predictors are not commonly reported in the empirical literature.

While a body of literature exists on such intervals, overall it is not clear as to what the properties and limitations are with respect to each innovation. The purpose of Chapter 7 is to clarify, and in places develop, the existing literature on the subject that has developed over the last two decades. The literature is decomposed into two themes. Firstly, the case where the parameters in the model are known (as opposed to being estimated) is considered. The more realistic case where model parameters are estimated is then considered. Through taking into account additional uncertainty due to estimation of parameters, an interval which is truly analogous to a prediction interval can be developed. Simar and Wilson (2010) have outlined a method using bootstrapping, however in Chapter 7, a method which samples from the asymptotic distribution of the parameters is proposed.

The approach is illustrated using a simplified version of the model used in Chapter 5 (now with focus on cost efficiency rather than scale efficiency).

1.3.4. Summary

There are common themes across the research chapters. One is the attempt to model the cost function/frontier component of any model as accurately as possible. In Chapter 5, the innovation is bringing in measures of heterogeneity in output directly into the cost function. In Chapter 6, as well as differentiating between returns to scale and returns to density, a benefit of utilising the regional data is that it allows returns to scale and density to be model at the level that managerial decisions are made. This in turn reduces any aggregation bias (which is inevitably present to some extent in micro economic analysis (Theil, 1954)). In addition the consideration of the Mundlak transformation in Chapter 6 is an attempt to control for the impact of unobserved time invariant heterogeneity. Finally accounting for the impact of uncertainty in estimation/prediction is a common theme. Clearly this is the purpose of Chapter 7, however the trade-off between flexibility and precision of estimates when determining functional form is the key motivator for the hedonic cost function approach in Chapter 5.

2. Economic concepts

2.1. Introduction

In this first literature review, the relevant economic concepts that are considered to apply to this thesis are examined. The cost function is the central economic tool that is utilised in this thesis. As such the structure of this review is as follows. Firstly, the motivation for the use of a cost function is established by relating it to the theory of firm production. Secondly, the variation in firm's costs is broken down into characteristics relating to returns to scale and density, technical change over time, input prices and efficiency, all of which influence firm's actual cost of production. The latter influence of costs, firm efficiency, requires a re-formulation of the cost function as a cost frontier which implies that the deterministic function conceptually represents minimum cost of production rather than actual cost. Finally it is discussed how the cost function device is appropriate for use in this thesis versus other devices such as production and distance functions.

2.2. The Cost Function

The first empirical application of the formal cost function as it would be recognised today (a function of output levels and input prices) can be traced back to Nerlove (1963). His seminal study on the costs of generating electric power was the first to utilise an analytic relation (developed by Shephard (1953)) between the structure of a firm's costs and the structure of the firm's production transformation function. In particular, subject to the behavioural assumption that the firm was cost minimising and the firm faced exogenous input prices and output levels, Nerlove showed that a Cobb Douglas cost function could be derived from a Cobb Douglas production function and input price information. Importantly the resulting (total) cost function was a function of both output levels and input prices. The Cobb Douglas cost function is termed “dual to” the Cobb Douglas production function. This work established the cost function as having meaningful economic interpretation which provided a strong motivation for estimating such functions.

More generally, the economic model for firm production is the transformation function, which is in itself a multi-output generalisation of the production function. This relates inputs to outputs, i.e. it shows the input required to produce an amount of outputs. (Chambers, 1988 p. 260). Thus it can be represented as:

$$0 = T(\mathbf{y}, \mathbf{x}) \tag{2.1}$$

where $\mathbf{y}$ is a vector of outputs and $\mathbf{x}$ a vector of inputs. The cost function is derived from this relationship by a further assumption that firms minimize costs of production of a given output, subject to the output being within the feasible set (contained within the transformation function) and that firms take prices ($\mathbf{p}$) for the inputs as given (they are input price takers). Thus firms choose the level of inputs to use in order to minimise costs.

Mathematically, the problem is:

$$\min_{wrt\ \mathbf{x}} C = \mathbf{x}'\mathbf{p} \text{ subject to (s.t.) } 0 = T(\mathbf{y}, \mathbf{x}) \tag{2.2}$$

Let $C*$ denote the minimum cost solution from this constrained minimisation problem then the solution can be represented as a function of the exogenous factors:

$$C* = C*(\mathbf{y}, \mathbf{p}) \tag{2.3}$$

Thus the cost function relates minimum cost to the outputs and inputs prices which are taken as given. The question then arises as to whether the cost function can be used to learn about the underlying production technology of the firm, such as returns to scale and density and the effects of technological change. If so, the cost function is said to be ‘dual’ to the transformation function. A set of ‘regularity conditions’ need to hold for this to be the case. (See Fuss and McFadden (1979) for full details of these conditions.) It is important to test or at least check that these conditions hold for an estimated cost function. These are discussed in the following sub-sections (2.2.1 and 2.2.3 with respect to output and input prices respectively) and in detail in Chapter 5 when the estimated cost function is checked against these conditions.

Before the properties of the cost function are examined, given the use of panel data (data over time for a number of cross sectional units) it is useful to explicitly account for changes in the cost function over time. In (2.3), these are implicitly contained within the functional form incorporating $\mathbf{y}$ and $\mathbf{p}$ (C(.)), so such a representation is still consistent with any underlying duality. Further cross sectional (i=1,…,N) and time (t=1,…,T) subscripts are added to relate the cost function to an (dual) indexed set of observations. Thus (2.3) can be expressed as:

$$C^* = C^*\left(\mathbf{y}_{\mathbf{it}}, \mathbf{p}_{\mathbf{it}}, \boldsymbol{\tau}_{\mathbf{t}}\right) \tag{2.4}$$

Where $\boldsymbol{\tau}_{\mathbf{t}}$ represents a vector of firm invariant (but time varying) variables which captures changes in the cost function over time. For example this could comprise a time trend and/or dummy variables suitably coded with time.

Some basic characteristics of the function are described below and these are illustrated in Figure 2.1. The subsequent sub-sections expand on each of these aspects.

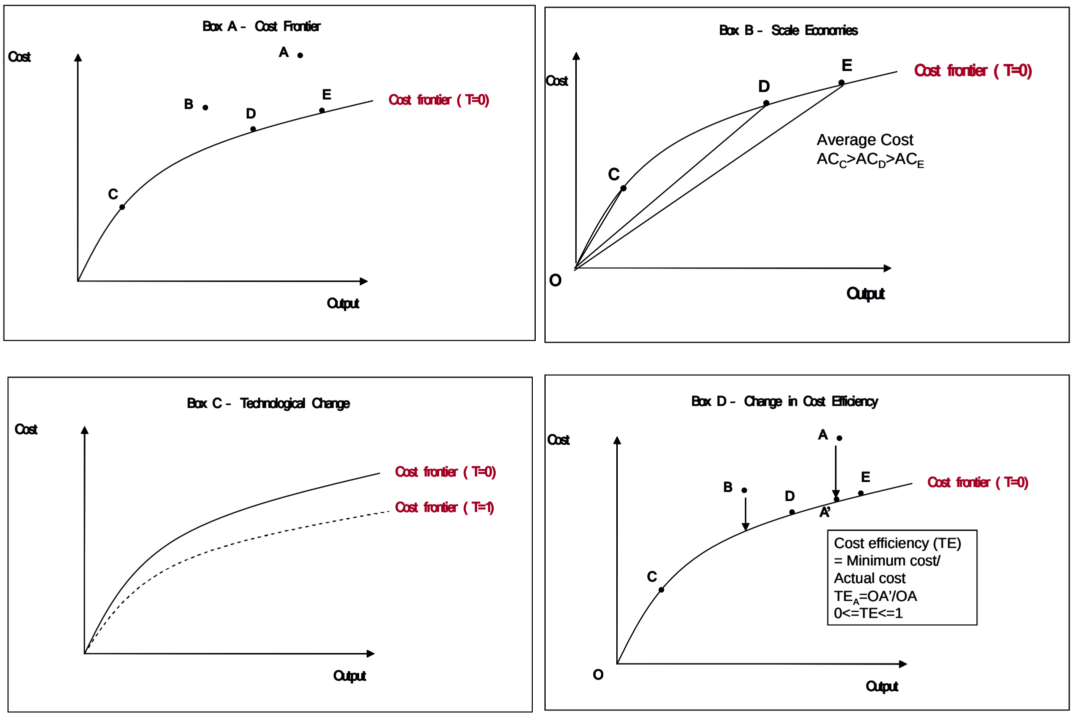
- Scale effects and other output changes (box b): differences in average costs of production due to the effect of returns to scale and density. TOCs are producing different output levels. This is a movement along a given minimum cost frontier.

- Technological change (box c): over time the best practice technology (hopefully) improves which means firms can achieve lower costs for the same output with all other things equal. This is a movement of the position of the cost frontier over time.

- Input price changes: changes in the price of inputs alters the cost of producing a given output. Like technological change, input price changes affects the minimum cost boundary and thus are represented as a shift of the minimum cost frontier.

- The efficiency of the firm (box d): the ratio of the firm's actual cost to the minimum cost possible if the firm adopted best managerial practice – i.e. the gap between the firm and cost frontier. As a result some firms do not produce at minimum possible cost and are thus above the cost frontier.

Figure 2.1 Economic concepts with respect to a cost frontier



2.2.1. Output and returns to scale and density

The (total) cost function requires that the following properties hold[4]:

C1 $C(\mathbf{y}_{it}, \mathbf{p}_{it}, \boldsymbol{\tau}_t) > 0$ for $\mathbf{p}_{it} > \mathbf{0}$ and $\mathbf{y}_{it} > \mathbf{0}$ (nonnegativivity)

C2 If $\mathbf{y}_{it} \geq \mathbf{y}_{jt}$, then $C(\mathbf{y}_{it}, \mathbf{p}_{it}, \boldsymbol{\tau}_t) \geq C(\mathbf{y}_{jt}, \mathbf{p}_{it}, \boldsymbol{\tau}_t)$ (nondecreasing in output)

C3 $C(\mathbf{0}, \mathbf{p}_{it}, \boldsymbol{\tau}_t) = 0$ (no fixed costs)

[4] The precise requirements depend on the actual requirements on the underlying production technology. The requirements given above, and 2.2.3 for input prices, are sufficient to describe a fairly general underlying technology (see Chambers, 1988, p. 9 and p 51 for exact definitions).

The above properties seem reasonable. However whether cost functions meet the above conditions depend on the exact functional form.

For some functional forms, the properties hold irrespective of the parameter values. For other functional forms certain (simple) sign restrictions on the parameters are sufficient to guarantee the cost function meets the properties for all levels of output (and input prices). In this case the cost function satisfies the property 'globally'. An example is in a Cobb Douglas model which, provided the coefficients on outputs are all non-negative, conforms to requirements C1 and C2.

For other functional forms it is not possible to provide restrictions that apply to all possible admissible values of output (and input prices). Instead properties have to be verified for a given sample, i.e. verified 'locally'. Many flexible functional forms can only be verified locally and further have to be checked for conformity post estimation rather than restricting the admissible parameter values pre estimation. Such an exercise is undertaken in Chapter 5 when the flexible Translog functional form is utilised.

Finally, note that sometimes a functional form does not conform to these properties, at least for some data points. Property C3 is, for example, impossible for a double logged functional form (such as Cobb Douglas or Translog) to conform to. This is because the logarithm of zero is undefined, so the functional form is only valid for positive output levels. Typically however this is not of interest to the analyst as they are interested in how costs behave in the region of their data (which is ultimately where statistics can provide most precision in any case).

Returns to scale refer to how costs change as output varies. Within a cost function approach, the measure of returns to scale measures how average costs change as outputs are increased by the same proportion (in the case of multiple outputs). If there are returns to scale then as outputs increase by a small proportion, costs increase by a lower proportion; thus average costs fall as outputs increase. Clearly findings on returns to scale have implication for the optimal scale of operations, at least from the perspective of minimising unit costs. The minimum efficient scale output point refers to the output level where the average cost curve is minimised.

In network industries, it is often useful to distinguish between increasing the size of a network and continuing to utilize the network at the same level as opposed to increasing the utilization of the network holding the size constant. Indeed, failing to allow for differences in the cost characteristics associated with scale and density can lead to model mis-specification and thus bias estimates. The returns to scale and density results may be of interest in their own right (as opposed to simply guarding against omitted variable bias) as is the case in Chapter 5 of this thesis.

Caves et al (1981 and 1984) outlined expressions for returns to scale and returns to density in cost functions. Caves et al's derivation was based on finding equivalent definitions of returns to scale and density for the cost function as for the production function via the duality theorem (as first explained by Nerlove (1963)). To distinguish between scale and density effects, models should include both traffic usage variables and variables to capture network size. In Caves et al's study the network size variable was the number of points served, however in other studies other measures such as track-km have been used (e.g. Farsi et al, 2005a, Gathon and Perlman, 1992, Coelli and Perelman, 1999,Coelli and Perelman, 2000). Caves et al showed that returns to scale (RtS) and density (RtD) can be computed as follows:

$$\mathrm{RtS} = 1 \Big/ \left(\sum_{i=1}^{m-1} \varepsilon_{y_i} + \varepsilon_S \right) \tag{2.5}$$

$$\mathrm{RtD} = 1 \Big/ \sum_{i=1}^{m-1} \varepsilon_{y_i} \tag{2.6}$$

Where ε_{y_i} is the elasticity of cost with respect to the ith output (i=1,…,m-1) and ε_S is the elasticity of cost with respect to the network size variable[5].

The Caves et al measures of scale and density both have the properties that unity represents constant returns and they are monotonic i.e. the greater the measure, the greater the degree of returns to either scale or density. In Chapter 5, extensions to these concepts are utilised which additionally incorporate measures of heterogeneity of output into the calculation. This extension is motivated from the literature on hedonic cost functions (Spady and Friedlaender, 1978) and discussed in more detail in Chapter 5.

[5] For notational convenience and consistency with other equations which do not distinguish between the network size variable and other outputs, the network size variable is treated as the m’th output and so only the first m-1 output elasticities are used in the RtD equation (which excludes this output).

2.2.2. Technical change

Technical change reflects how the underlying technology available to firms' changes overtime. In particular, it could be hoped that there are technical innovations, introduced over time, which result in more output for given amount of inputs. For a cost function, this is equivalent to costs falling over time for the same amount of outputs being produced and for the same input prices. Clearly given its temporal nature, it is only applicable to time series and panel data sets.

Technical change can enter the cost function in a number of ways. First there is an issue as to whether it is entered as a trend e.g. linear or quadratic or as a set of time dummy variables. The former approach has the advantage that the trend can be extrapolated outside the sample to give a prediction of how the cost function may change further into the future (of course this is subject to past trends being reflective of future trends). However, such a parametric function requires an assumption on functional form which may be deemed an undesirable. In contrast, time dummy variables allow for an unrestricted path of technical change. In panel data this leads to a 'two-way' panel model, with both time invariant effects (firm effects) and firm invariant effects (time effects). However the limitation of such an approach is not much can be said for the likely future path of technical change, unless there is some kind of extra auxiliary regression of the time effects post estimation.

Second, there is, perhaps the more important economic issue, as to how technical change should interact with output and input prices in the cost function. Does technical change influence the behaviour of costs with respect to output changes e.g.

change the returns to scale properties over time? Similarly, does technical change influence the relative choice of inputs? If the answer to both of these questions is no, then technical change simply is a scaling factor on cost. When technical change does not affect the cost minimising input ratios, then technical change is termed 'cost-neutral' (Chambers, 1988, p. 216). This is equivalent to the cost share equations being independent of time, which, in a Translog cost function, requires no interaction terms between the time trend and input prices in the cost function (Chambers, 1988, p 229).

2.2.3. Input price changes

Nerlove highlighted the importance of including input prices in all cost functions. Failure to include such prices implicitly assumes that the excluded input prices are constant across all observations[6]. If this assumption does not hold then the model is mis-specified. Also, if at least one input price is excluded it is difficult to impose the linear homogeneity in input prices restriction (see below for definition).

More formally, and continuing with the list of properties in 2.2.1, for a cost function to be a valid descriptor of the underlying technology (i.e. dual to a transformation function), the following properties with respect to input prices need to be adhered to.

[6] It is noted that panel data techniques can be used to relax (but not entirely eliminate) this strict assumption. For example inclusion of time effects may proxy for cross section wide changes in input prices. However in the context of performance measurement such time effects (and similarly for cross section effects) have specific interpretations such as technical change (or inefficiency) and using panel data techniques to control for input price effects is not ideal.

C1 (restated) $C(\mathbf{y}_{\mathbf{it}}, \mathbf{p}_{\mathbf{it}}, \boldsymbol{\tau}_{\mathbf{t}}) > 0$ for $\mathbf{p}_{\mathbf{it}} > \mathbf{0}$ and $\mathbf{y}_{\mathbf{it}} > \mathbf{0}$ (nonnegativivity)

C4 If $\mathbf{p}_{\mathbf{it}} \geq \mathbf{p}_{\mathbf{jt}}$, then $C(\mathbf{y}_{\mathbf{it}}, \mathbf{p}_{\mathbf{it}}, \boldsymbol{\tau}_{\mathbf{t}}) \geq C(\mathbf{y}_{\mathbf{it}}, \mathbf{p}_{\mathbf{jt}}, \boldsymbol{\tau}_{\mathbf{t}})$ (nondecreasing in prices)

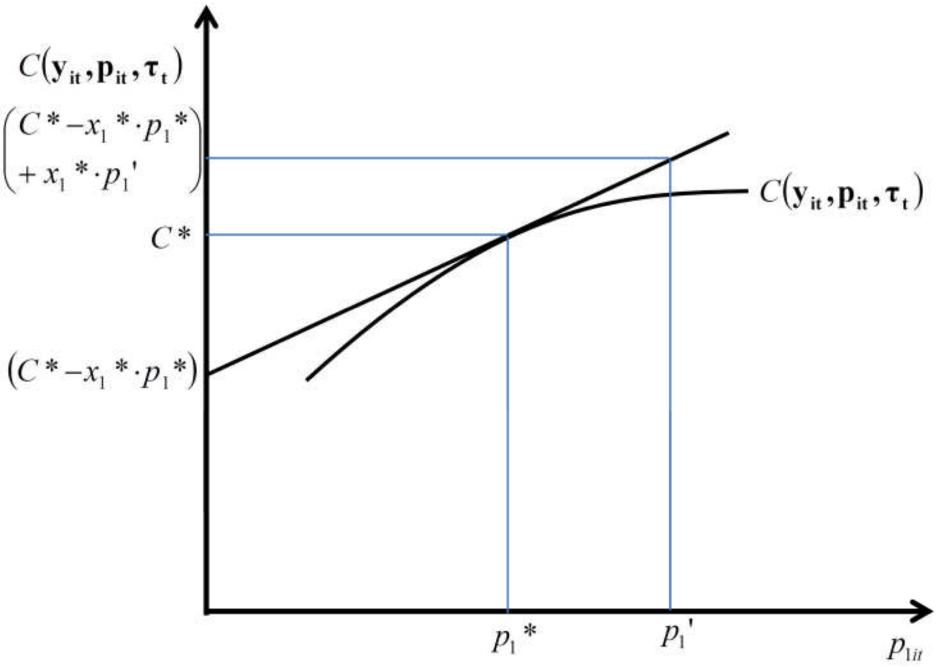
C5 $C(\mathbf{y}_{\mathbf{it}}, \lambda\mathbf{p}_{\mathbf{it}}, \boldsymbol{\tau}_{\mathbf{t}}) = \lambda C(\mathbf{y}_{it}, \mathbf{p}_{it}, \boldsymbol{\tau}_{\mathbf{t}})$ for $\lambda > 0$ (positively linearly homogenous)

C6 $C(\mathbf{y}_{\mathbf{it}}, \mathbf{p}_{\mathbf{it}}, \boldsymbol{\tau}_{\mathbf{t}})$ is concave and continuous in $\mathbf{p}_{\mathbf{it}}$

C1 is self explanatory. C4 simply states that if all prices are at least as large as another price vector, then cost must be at least as large. C5 is intuitive as if all prices increase by the same proportion and the output requirement is the same, then the optimal choice of input use must be the same, implying input quantities are the same, thus cost must increase by the same proportion.

C6 is a little more involved and a comprehensive description (and link to Shephard's Lemma) is given in Chambers (1988, p 53-55). Figure 2.2 provides intuition as to why cost functions need to be concave in input prices (a formal proof is contained in Chambers (1988)). The cost function is drawn with respect to one input price, holding all outputs and other input prices constant. Consider the situation where the input price increases from p_1* to p_1', again holding outputs and other input prices constant. In response to this change the firm could continue to utilise the input at the same level i.e. the firm's choice of inputs do not change. In this case cost would increase to $(C* - x_1 * \cdot p_1 * + x_1 * \cdot p_1')$ where x_1* is the cost minimising input level at p_1* for input 1. Alternatively the firm could change its input mix (substitute away from the input 1). In this case, cost must be less than or equal to $(C* - x_1 * \cdot p_1 * + x_1 * \cdot p_1')$, otherwise the firm is not cost minimising (it could have maintained the same input mix). Hence the cost function must be on our below the straight line through $(p_1*, C*)$; that is, concave.

Figure 2.2 The need for concavity in input prices



$C(\mathbf{y}_{it}, \mathbf{p}_{it}, \boldsymbol{\tau}_t)$
$\begin{pmatrix} C* - x_1 * \cdot p_1 * \\ + x_1 * \cdot p_1' \end{pmatrix}$
$C*$
$(C* - x_1 * \cdot p_1 *)$
$C(\mathbf{y}_{it}, \mathbf{p}_{it}, \boldsymbol{\tau}_t)$
$p_1 *$
p_1'
p_{1it}

Source: Reproduced (by own transpose with revised notation) from Chambers (1988, p53)

As with the properties on outputs, depending on the functional form adopted, the properties above can exist globally or locally. In the cost function used in Chapter 5, the Translog, C5 is imposed (globally) prior to estimation, but C6 and C4 are checked post estimation. C4 corresponds to the elasticity of cost with respect to each price being positive at all data points (equivalently that the cost shares are all positive, by Shephard's Lemma). C6 is verified for each data point, through computation of the matrix of second derivatives of input prices to verify if it is negative definite; a necessary and sufficient condition for concavity in prices. See Chapter 5 for more details.

Note, in Chapter 6 only a single input price is available which limits the extent to which the properties of the cost function can be imposed or checked. In particular, C5 can not be imposed. However C4 holds for the single input price (as the cost elasticity with respect to the price is constant – and found to be positive). C6 also holds in a similar ‘partial’ sense, that is the single second derivative is negative as required.

2.2.4. Efficiency

It is often assumed in neo-classical economics that firms, or more generally decision making units, are successful in maximising an objective function. This is often in contradiction with reality. One example was provided by Hicks (1935, p. 8): “people in monopolistic positions… are likely to exploit their advantage much more by not bothering to get very near the position of maximum profit, than by straining themselves to get very close to it. The best of all monopoly profits is a quiet life” (cited in Kumbhakar and Lovell, 2000). There are now many formal economic models which allow for a degree of sub-optimisation. These include a raft of principal-agent models where the management of a firm has a different objective to the owners and the owners have an informational asymmetry which prevents them from successfully monitoring managers (see for example Vickers and Yarrow, 1988). These models lead to a degree of slack in the production process, which prevents maximisation of the firm principal’s objective; this will subsequently be termed as the firm as exhibiting a degree of inefficiency.

Work on developing an analytical framework to measure the degree of sub-optimisation can be traced back to the 1950s with the work of Koopmans (1951) who defined technical efficiency, and Debreu (1951) and Shephard (1953) who developed a closely related concept of the distance function. Importantly, Farrell (1957) was the first to measure efficiency (the extent to which firms optimise) with respect to the cost minimising objective and decompose cost efficiency into technical and allocative components. In the following decades, the literature on measuring efficiency and, more generally, performance has grown considerably.

Applied to a cost function, efficiency refers to cost efficiency which reflects the ability of the firm to choose and combine its inputs at minimum cost. This can be shown to be the product of technical and allocative efficiency. Technical efficiency refers to the divergence between the output produced by the firm for a given amount of input and the maximum amount of output possible for the same input combination. Allocative efficiency refers to the divergence between the chosen ratio of inputs and the optimal ratio of inputs in terms of that which minimises cost. It is possible to decompose cost efficiency into technical and allocative efficiency, for example, by estimating a cost function along with input factor share equations in a Cobb Douglas model (Kumbhakar and Lovell, 2000).

In terms of formally stating the economic model to account for inefficiency of firms, it should be noted that the solution of the constrained optimisation in (2.2) yields the minimum cost frontier, while in practice firms may fail to optimise. Taking logarithms (the reason for which will simplify computation of cost efficiency described below), then actual cost (C)[7] can be expressed as a function of the minimised cost function (C* – now termed the cost frontier) and an additional component representing the difference between actual cost and minimum cost (u):

$$\ln C = \ln C^* + u \leftrightarrow \ln C = \ln\left(C^*\left(\mathbf{y},\mathbf{p}\right)\right) + u \tag{2.7}$$

By definition $u \in \Re^+$.

Cost efficiency is defined as:

$$Eff = \frac{C^*}{C} \tag{2.8}$$

Using the model in (2.7)

$$Eff = \frac{C^*\left(\mathbf{y},\mathbf{p}\right)}{C^*\left(\mathbf{y},\mathbf{p}\right)\cdot e^{u}} = e^{-u} \tag{2.9}$$

2.3. The appropriateness of a cost function for this thesis

The cost function is not the only economic device available for the task of efficiency analysis in railways. Others include the production function, distance function, revenue function and the profit function. All have been used in railway performance

[7] For ease of exposition, time and cross sectional sub-scripts are omitted. Further, and related to the lack of subscripts, explicit account for technical change is not made; again for simplification.

analysis. Each function makes assumptions as to what attributes are and are not under the firm’s control. Importantly the problem formulated in (2.2) assumes that firms minimize cost by choosing the level of each input. However, it also requires that input prices are outside of the firm’s control (they are price takers) and the same applies to the output level.

Oum et al (1999, p. 36) in concluding their review of the railway literature argue for the cost function over the production function given that railways produce multiple outputs and production functions can only accommodate a single aggregate output. The cost function treats output as exogenous which is likely to be most appropriate in the case of railways (at least in Great Britain).

For passenger train operating companies, it may be questioned as to whether output levels are truly exogenous to the firm, especially when we consider characteristics such as passenger load factor of the primary output (train hours). TOCs in Britain do have some ticket pricing discretion and thus they can grow output to some extent. However, this discretion is relatively minor compared to the overriding concern to produce train services at lowest cost subject to quality constraints. Ultimately, minimum service levels are set in franchise agreements. Similarly, it can be argued that TOCs have some price making power in input markets, for example, they negotiate staff pay with unions.

A similar argument can be made with respect to endogeneity of input prices in infrastructure cost functions, although, in practice, this is an academic point given the lack of available input price data for infrastructure maintenance and renewal activity in many applications.

However, for regulatory purposes it can be argued that measuring cost efficiency is the objective of any exercise and thus the cost function is the natural economic model to utilise. This argument can be applied to the infrastructure analysis where the objective of the analysis is to determine the potential cost savings feasible for a firm. Ultimately the regulator is interested in the overall cost of provision of the service and ensuring prices to users are minimised for a given output.

The alternatives to the cost function are not without flaw and it can be argued, *a priori* that the flaws are stronger in the alternatives than the cost function. A transformation function (or the rearranged variant called the distance function used with respect to efficiency measurement) could be estimated. This does not require an assumption of the firm's economic objective (cost minimization, profit maximisation etc.). However, it requires exogeneity of input and output levels (with the exception of the arbitrarily chosen left hand side output). Exogeneity of input levels seems at odds with the aspiration of competitive tendering, namely to introduce innovation into the provision of service.

A further alternative is the profit or revenue function. However such a behavioural assumption seems at odds with the general motivation for competitive tendering (to lower cost or subsidy) and in any case prices (both input and output) are still assumed exogenous (firms are price takers). The only alternative is to estimate a demand and supply system of equations, however the specification of this system is unclear and constructing data for the demand side which is compatible with the supply equation is non-trivial and beyond the scope of this thesis. This is because the 'raw' demand data measures origin to destination station trip data, a final output to use the terminology in 4.2.1, while TOCs produce intermediate outputs (train-hours). Mapping between the two requires a trip assignment model.

It is concluded that the cost function is preferred over the alternatives.

2.4. Summary

This chapter has reviewed the economic foundations of the cost function. The behavioural assumptions underpinning it have been discussed. It has been further shown that the cost function provides information regarding the underlying technology of the firm and can provide measures of returns to scale/density and technical change over time. The economic framework for considering the cost function as a cost frontier has also been outlined. The cost function still represents the minimum cost of production but now explicit allowance is made for a degree of sub-optimisation in terms of allowing the firm's cost to be on or above the minimum cost frontier. This allows the measurement of cost efficiency. Finally the chapter has motivated the use of a cost function/frontier over other economic devices. Ultimately the cost function provides the measures that are of interest; namely cost efficiency and scale and density properties of the industry given that output is (best) characterised as exogenous.

One important point to be made is the inter-dependency of components of the cost function. Clearly, an analyst cannot expect to measure cost efficiency correctly if outputs have been incorrectly specified and/or variation in input prices have not been taken into account, for example failing to take into account heterogeneity in outputs. These issues are returned to in Chapter 4, when the details of how to specify a cost function in the specific application to railways are considered. In particular, a common theme is that railway output is not homogenous and thus there is a need to explicitly account for the characteristics of output to accurately specify a cost function.

3. Econometric Methodologies

3.1. Introduction

The previous chapter was concerned with relevant economic concepts. Crucially, the cost function was introduced along with the concept of efficiency where firms can in reality produce at or above the minimum possible cost. In this chapter, issues relating to estimation of these economic models using real world data are considered; the review moves from the economic to the econometric research domain.

The basic precept of moving from the economic to the econometric domain is to recognise that economic models are abstractions from reality and thus there is a need to append an error term to an economic model to form an econometric model. This 'noise' captures the failure of the economic model to fully explain real world data. Estimation of the economic model parameters using sample data can then be undertaken. Of crucial importance to the properties of the estimators is the assumptions placed on the behaviour of the error term and the relationship between it and the data.

In addition to the usual desire to obtain the best estimates of model parameters, exercises which try to measure cost efficiency are concerned with measuring the realisation of the error term in the model for a given firm. Thus, unlike conventional econometric analysis, it is no longer sufficient that the error is simply 'well-behaved'; the analyst actually cares about its value (Greene, 2008). This presents unique econometric challenges.

This chapter is divided into two subsequent themes. In the first theme, comprising section 3.2, the general econometric techniques are reviewed which are relevant to ‘best’ estimating model parameters. The second theme considers the issue of measuring firm efficiency. This motivates the methods used in Chapter 6 and also provides background for the methodological development on uncertainty in efficiency predictions in Chapter 7. Section 3.3 introduces the econometric efficiency model. Section 3.4 considers cross sectional data applications and section 3.5 considers panel data applications.

3.2. General econometric concerns

In this section, the basic econometric analysis tools for cost function analysis are reviewed. For the purpose of this section, firms are assumed to be efficient, or at least, measuring efficiency is assumed not the objective of the exercise[8]. The techniques are thus most relevant to the returns to scale and density research for TOCs in Chapter 5, although it provides a useful background for techniques developed in the following sub section specifically for measuring efficiency.

It is not the purpose of this section to provide a comprehensive treatment of econometric theory. This is beyond the scope and there exist many text books that provide treatments at a variety of technical sophistications (e.g. Studenmund, 2011 and Greene, 2012). Instead, firstly, the relevant statistical properties of estimators of

[8] As discussed in 3.3, some of the general econometric techniques provide consistent estimates of (most) model parameters even in the presence of inefficiency.

model parameters, such as unbiasedness, consistency and efficiency are presented. Being clear on the meaning of these properties is important for justifying the choice of estimation method invoked in the research chapters and particularly important for the technical discussion of appropriate prediction intervals for cost efficiency in Chapter 7. Secondly, two general estimation frameworks are introduced, namely least squares and maximum likelihood, as they are used in some form throughout the research.

3.2.1. Properties of estimators

Consider a model which comprises a set of unknown parameters which are contained in a vector θ. One way to learn about what the values are of the elements of θ is to use a sample of data. An estimator is defined in Greene (2012) as:

"*An estimator is a rule or strategy for using data to estimate the parameter [(θ)]. It is defined before the data is drawn… A point estimate is a statistic computed from a sample that gives a single value for θ.*" p 1095.

It then follows that "Obviously, some estimators are better than others" (Greene 2012, p 1095). It is the identification of what properties of estimators makes them 'better' than others that is considered below.

Let an estimator of θ be denoted $\hat{\theta}$. By the definition above, $\hat{\theta}$ is a function of observed sample data. However a given (random) sample of data will not perfectly mimic the population. Thus for each sample drawn from the population, a different value of $\hat{\theta}$ will be computed. Thus $\hat{\theta}$ will have a sampling distribution i.e. a probabilistic distribution relating to the value of $\hat{\theta}$ from multiple resampling exercises. The properties of this distribution are used to evaluate various competing estimators, some of which are considered below.

3.2.1.1.Unbiasedness

"An estimator of a parameter θ is unbiased if the mean of its sampling distribution is θ. Formally, $E\left[\hat{\theta}\right]=\theta$." (Greene, 2012 p. 1096)

This would appear to be a very desirable property of an estimator. 'On average' the estimator corresponds to the true parameter value. However, it is important to note that the properties of the sampling distribution refer to the behaviour of the estimator through resampling. In practice, researchers generally are faced with a single draw from a distribution. Thus just as important (and perhaps more important) is to understand the spread of the sampling distribution, summarised by the variance of the estimator ($v\left[\hat{\theta}\right]=E\left[\hat{\theta}-\theta\right]^2$). Indeed an estimator that is biased (i.e. not unbiased) may be preferred to an estimator which is unbiased if it has a sufficiently smaller variance of its sampling distribution.

3.2.1.2.Efficiency (of unbiased estimators)

"An unbiased estimator $\hat{\theta}_{1`}$ is more efficient than another unbiased estimator $\hat{\theta}_{2`}$ if the sampling variance of $\hat{\theta}_{1`}$ is less than that of $\hat{\theta}_{2`}$. That is $v\left[\hat{\theta}_1\right]<v\left[\hat{\theta}_2\right]$." (Greene,

2012, p 1096)

An estimator that is unbiased and is efficient relative to all other unbiased estimators is known as the best unbiased estimator (BUE). Note that this is referring to unbiased estimators only. When biased as well as unbiased estimators are considered, there is a trade-off between bias and variance. One criterion is to choose the estimator with the minimum mean squared error but in practice this is difficult to compute.

It is often not possible to obtain measures of the properties described above for all estimators in finite samples. What is often easier (and indeed feasible) is to obtain large sample (asymptotic) properties of estimators and use these as the basis of determining which estimator is best. It is necessary to denote the estimator as $\hat{\theta}_{n`}$ to denote in what dimension its sampling distribution changes. Thus, the following properties concern how the sampling distribution of the estimator changes as n, the number of observations, increases.

3.2.1.3.Consistency

"The random variable x_n converges in probability to a constant c if $\lim_{n\to\infty} \Pr(|x_n - c| > \varepsilon) = 0$ for any positive ε… [short hand] we write $\operatorname{plim} x_n = c$" (Greene, 2012 p 1107)

"An estimator $\hat{\theta}_n$ of a parameter θ is a consistent estimator of θ if and only if $\operatorname{plim}\hat{\theta}_n = \theta$." (Greene 2012, p 1109)

Consistency of an estimator implies that the sampling distribution of the estimator converges to a point at the true parameter value as the sample size is increased. This is a very desirable property of an estimator and is generally a prerequisite property of a useful estimator. Ultimately it states that as more and more data is considered in the estimation, the probability of the estimate being any distance away from the true value diminishes. Thus the more data is available the 'better' the estimate (in a probabilistic sense).

It should be noted that in Chapter 7, consistency of predictors of random variables are considered. In particular the Jondrow et al (1982) predictor (described in 3.4) is an inconsistent predictor. As stated in the Chapter 7, however, the Jondrow et al (1982) sample predictor is a consistent estimator of the population expectation; however this is not the quantity of interest.

3.2.1.4.Convergence in distribution, central limit theorems, and the asymptotic distribution of an estimator

Knowing that an estimator is consistent for the quantity of interest is useful, however in finite samples there is a need to understand the spread of the sampling distribution and undertake statistical inference. Assuming that small sampling properties cannot be derived, it is necessary to form an Asymptotic Distribution of the estimator:

"An asymptotic distribution is a distribution that is used to approximate the true finite sample distribution of a random variable." (Greene, 2012, p 1124)

In order to determine the asymptotic distribution of an estimator it is necessary to establish what distribution the estimator converges to (as the sample size is increased). To do this, it is necessary to define convergence of a random variable to another random variable:

“ x_n converges in distribution to a random variable x with CDF[9] $F(x)$ if $\lim_{n\to\infty}|F_n(x_n)-F(x)|=0$ at all continuity points of F(x).” (Greene, 2012, p 1116)

And

“If x_n converges in distribution to x, where $F_n(x_n)$ is the CDF of x_n, then $F(x)$ is the limiting distribution of x_n. This is written $x_n \xrightarrow{d} x$.” (Greene, 2012, p 1116)

From the limiting distribution, the asymptotic distribution can be constructed. Central Limit Theorems provide limiting distributions for the estimators used in this thesis, namely least squares and maximum likelihood estimators. There are many types of central limit theorem and also supporting theorems and such exposition is too detailed for this thesis. Providing an appropriate central limit theorem can be utilised (which is the case with least squares and maximum likelihood estimators (Greene, 2012 p 492) subject to several technical requirements (Greene, 2012 p 489-492)), then for a consistent estimator (here denoted as a vector for generality)

[9] For completeness, the Cumulative Distribution Function (CDF) of a random variable X is defined as $F(x)=\Pr(X\le x)$.

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(0, \mathbf{V}\right) \qquad (3.1)$$

Then the asymptotic distribution of $\hat{\theta}_n$ is given as:

$$\hat{\boldsymbol{\theta}}_{\mathbf{n}} \overset{a}{\sim} N\left[\boldsymbol{\theta}, \frac{1}{n}\mathbf{V}\right] \qquad (3.2)$$

Note that (3.2) does not state that $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$ is distributed in this manner, but that it is approximately distributed. The distribution in (3.2) can be used as a basis of inference, such as a z or Wald statistic.

3.2.2. Estimation frameworks

In this sub-section, two general estimation frameworks are introduced, namely least squares and maximum likelihood. The properties of these estimators are considered. This is deliberately supposed to be high level and merely provides the *a priori* motivation for utilising these techniques in the thesis. As such it is brief. A full survey is provided in Greene (2012).

The econometric model considered in this thesis (bar some extensions for panel data) can be expressed in a general form as:

$$y_i = f\left(\mathbf{X_i}; \boldsymbol{\beta}\right) + \varepsilon_i \qquad \text{i=1,…,N} \qquad (3.3)$$

That is y_i, the dependent variable is modelled as a function $f(\cdot)$ of some other variables, $\mathbf{X}_i$, parameterised by a vector $\boldsymbol{\beta}$, and a stochastic error term, ε_i.

3.2.2.1.Least squares

Least squares techniques are very common in econometrics. The simplest variant, ordinary least squares (OLS), involves the minimisation of the sum of squared residuals. The popularity of OLS is mainly due to its desirable properties in the linear (in parameters) regression model when only relatively weak assumptions on the error terms are imposed (relative to the 'fully parametric' maximum likelihood type estimators which require 'full' distributional assumptions to be imposed, see 3.2.2.2 below). Least squares techniques applied to linear models yield not just consistent and unbiased estimates, but under the set of assumptions, attributed to Gauss-Markov, are minimum variance of possible unbiased estimators. Estimation can be conducted via closed form linear algebra expressions, which means computationally they are relatively simple.

The Gauss Markov conditions can be summarised as:

$G1\ f(\mathbf{X_i};\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$
$G2\ \ E(\varepsilon_i)=0$

$G3\ Var(\varepsilon_i^2)=\sigma^2$

$G4\ \ Cov(\varepsilon_i\varepsilon_j)=0\ \ i\neq j$

$G5\ \ E(\varepsilon_i\mathbf{X_i})=0$

Below, violations of the above assumptions are considered.

G2 is often violated in models involving inefficiency. In the presence of inefficiency, errors will have a positive expectation (in the case of a cost inefficiency). This is because the inefficiency error is defined to be non-negative (in the case of a cost inefficiency). However this does not present a barrier to applying least squares techniques. In particular least squares applied to such models (by simply ignoring the non-zero mean and proceeding as if it were zero) still yields unbiased and consistent estimates of all parameters except the constant. However as explain in sub-section 3.4.1, a suitable transformation can be applied *ex post* estimation to produce a consistent estimate of the constant term (in practice the constant is rarely of interest and so such a correction is not necessary, however it is necessary with respect to measuring inefficiency).

Violations of G3 and G4 mean that the error in the model suffers from hetroscedasticity and correlation across observations. This does not affect the unbiasedness or consistency properties of OLS, but OLS is no longer efficient. Instead there exists another least squares estimator, the Generalised Least Squares (GLS) estimator, which is more efficient, since the extra information contained in the pattern of hetroscedasticity (in the case of violation of G3) and/or correlation between observations (in the case of violation of G4) is exploited to better estimate the parameters. An example is the GLS estimator of the random effects model for panel data, which is often used in the semi-parametric models containing inefficiency. The GLS estimator is more efficient than OLS since it exploits the persistent correlation over time in the errors for each firm, while OLS ignores this.

Violation of G5 is more problematic. If it is the case that the regressors are correlated with the errors then OLS estimates of parameters will be biased. Intuitively this is because the implicit approach of the estimation process is to determine estimates of the parameters through attributing variation in the dependent variable to variation in a given explanatory variable. However, in this case, some variation in the error will also be attributed to variation in the explanatory variable; OLS can not distinguish between variations in **X** on y versus variation in the error on y. Instrumental variables can be used (Greene, 2012, chapter 8), however this is of limited relevance with respect to this thesis partly due to the difficulty in introducing instrumental variables into stochastic frontier analysis (the subject of sections 3.3-3.5). The only instrumental variables estimator used in this thesis is the fixed effects estimator used in Chapter 6. Here the difference between each regressor and the firm group mean of the regressor is used as an instrument for each regressor.

In non-linear models,a violation of assumption G1, least squares can still be applied, although estimation usually has to proceed via iterative techniques. The properties of these estimators are more difficult to establish and Greene (2012) sets out conditions for the estimator to be consistent (p. 227) and asymptotically normally distributed (p. 228).

In many economic models there may also be several equations that are linked through correlation in errors and cross equation restrictions i.e. the same parameters are contained in each equation. This is the case in the cost function and cost share equation estimated in Chapter 5. By Shepard's Lemma, the partial derivative with respect to (log) price of the (log) cost function is equal to the cost share of the input. As such the parameters which appear in the cost share equation(s) occur in the cost function as well. Further, given the economic relationship between to equations, the residual errors are likely to be correlated and so this should be exploited in estimation. Thus, while equation by equation OLS is unbiased and consistent, cross equation restrictions cannot be imposed and there are more efficient estimation methods. In particular the Feasible Generalised Least squares method initially developed by Zellner (1962) can be utilised and cross equation restrictions imposed. This is known as Seemingly Unrelated Regression (SUR). A non-linear version of this is used in the estimation of the cost model in Chapter 5.

In the remainder of this thesis, least squares techniques are referred to as 'semi-parametric' techniques, in the sense that the estimation techniques do not require a full set of distributional assumptions to be placed on the errors. Instead only assumptions on the first two moments (the mean and the variance covariance) have been made.

3.2.2.2.Maximum Likelihood

In contrast to least squares estimation techniques, maximum likelihood techniques require the distribution of the error terms to be fully specified. For example the classical linear regression model, to which OLS is best linear unbiased, can be estimated by maximum likelihood by assuming that the error is independently normally distributed with mean zero and homoscedastic variance. The method of maximum likelihood proceeds by choosing parameter values which maximise the probability that the observed sample was drawn from the distribution evaluated at the parameter values. This therefore requires the specification of a joint probability density function (PDF, joint over observations); hence the need for distributions to be fully specified. Because it is the parameter values which are subject to manipulation in the maximisation, the joint PDF (which is an expression relating to the data) is called a likelihood function (denoted L) which is the same function but gives an expression for the parameters conditioned on the data. For independent data (once conditioned on the regressors i.e. the errors are independent), this relation can be written:

$$L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} f\left(y_i \mid \mathbf{X_i}, \boldsymbol{\theta}\right) \tag{3.4}$$

Or in logarithms which makes the problem additive rather than multiplicative, aiding maximisation:

$$\log L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln f\left(y_i \mid \mathbf{X_i}, \boldsymbol{\theta}\right) \tag{3.5}$$

The benefit of this ‘fully parametric’ approach is that the techniques can exploit all the information contained within the assumed distribution to best fit the parameter values while semi-parametric approaches only exploit information up to a given number of moments (the first two in the case of OLS). The result is that maximum likelihood estimators are in general efficient, at least asymptotically, under the distributional assumptions. The trade-off is that they are less robust, in the sense that their attractive properties apply only if the errors are distributed as stated; semi-parametric methods permit more discretion as to the distribution of errors relative to fully parametric methods.

Like in the semi-parametric setting, in the fully parametric setting there are many ways to estimate the model parameters. The specific method of maximum likelihood is attractive for its large sample properties. Under regularity conditions reproduced in Greene (2012, p. 555), the properties of $\hat{\boldsymbol{\theta}}$, the ML estimator of $\boldsymbol{\theta}$ (the vector of parameters (which in turn would include $\boldsymbol{\beta}$ and other distributional parameters in (3.3))) are given below (Greene, 2012, p. 554, Theorem 14.1):

"*M1 Consistency:* $p\lim\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$

M2 Asymptotic normality: $\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left(\boldsymbol{\theta}, \{I(\boldsymbol{\theta})\}^{-1}\right)$*, where*

$$I(\boldsymbol{\theta}) = -E\left[\partial^2 \ln L \big/ \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\right]$$

M3 Asymptotic efficiency: $\hat{\boldsymbol{\theta}}$ *is asymptotically efficient [it achieves the Cramer Rao Lower bound – see Theorem C.2. in Greene, 2012]…*

…M4. Invariance: The maximum likelihood estimator of $\boldsymbol{\gamma} = c(\boldsymbol{\theta})$ *is* $c(\hat{\boldsymbol{\theta}})$ *if* $c(\boldsymbol{\theta})$ *is a continuous and continuously differentiable function.*" (notation amended to match this section of the thesis)

M1-3 are clearly desirable properties (see 3.2.1). M4 is very useful in terms of deriving the maximum likelihood estimator for combinations of parameters and their asymptotic sampling distribution.

Maximum likelihood methods are used throughout this thesis. They are essential to efficiency analysis in cross sectional data; without making full distributional assumptions it is impossible to recover a firm specific prediction of inefficiency. They are also used extensively in efficiency measurement using panel data (see section 3.5).

3.3. Efficiency methods

The previous section considered 'best' estimation of the model parameters. However efficiency is concerned with the 'gap' between the estimated frontier and firm's actual cost. This presents a challenge to econometric methods given that there are two other reasons for there being a gap between the estimated frontier and a firm's actual cost. Firstly any frontier will be estimated using a sample of data. This implies a sampling distribution for the model parameter estimates and thus the position of the frontier has a probabilistic distribution. Secondly, there is inherent noise in the cost data. This captures all factors outside of the model i.e. it is a recognition that the model is an abstraction from reality. Using least squares techniques for example, implies that the estimated frontier represents the conditional mean of cost. In practice the actual minimum cost for the firm may be higher or lower than this and part of the computed 'gap' may be noise rather than inefficiency.

The economic model for the cost frontier is given in (2.7) as:

$$\ln C = \ln\left(C\left(\mathbf{y},\mathbf{p};\boldsymbol{\beta}\right)\right) + u \qquad (3.6)$$

The parameter vector $\boldsymbol{\beta}$ is added to indicate that there will be a set of unknown parameters in the economic model. In addition, to keep the econometric discussion manageable it is assumed $\ln\left(C\left(\mathbf{y},\mathbf{p};\boldsymbol{\beta}\right)\right)$ is linear in parameters. Importantly it should be noted that data needs to be available on C, **y** and **p**. However u is unknown.

Furthermore, as shown in (2.9) it needs to be recovered from the analysis in order to compute cost efficiency.

However this is still only an economic model and not an econometric model. In order for this to become an econometric model and be amendable to estimation of the relevant parameters, an additional error term must be added. This is the 'noise' error term and represents the unexplained element of the economic model with respect to the real world data generating process. Ultimately it reflects the degree to which the economic model is an abstraction from reality.

Thus the econometric model is:

$$\ln C = \ln\left(C\left(\mathbf{y},\mathbf{p};\boldsymbol{\beta}\right)\right) + u + v \tag{3.7}$$

For the purpose of this thesis, the above model will be termed a stochastic frontier (SF) model. The model is a stochastic frontier because it simultaneously accounts for noise and inefficiency. This terminology is consistent with that in Coelli et al (2005), but is not strictly consistent with some other authors (e.g. Greene, 2008), particularly in the panel data context[10].

The remainder of this chapter considers how to estimate both the parameters in this

[10] Greene (2008) appears to define a stochastic frontier model in terms of the model estimation approach and whether the efficiency is measured in terms of absolute (with respect to a population distribution) – the case of a stochastic frontier – or relative to firms in sample. Importantly Greene defines the semi-parametric models of Schmidt ad Sickles (1984) and Cornwell, Schmidt and Sickles (1990) to not be stochastic frontier models, while using the definition in this thesis, they are. In practice this is a difference in terminology and not anything that precludes the use of either type of model in analysis of efficiency.

model and how to compute efficiency from the above model. This is a non-trivial problem and this review proceeds with two cases. Firstly, the model in the presence of cross sectional data and secondly, in the presence of panel data.

3.4. Cross sectional data analysis

Consider (3.7) using the cross sectional dimension only

$$\ln C_i = \ln\left(C\left(\mathbf{y}_i, \mathbf{p}_i; \boldsymbol{\beta}\right)\right) + u_i + v_i \quad i = 1,\ldots,N \tag{3.8}$$

Estimation of the frontier parameters, $\boldsymbol{\beta}$, can be undertaken through a variety of means provided the assumption that the error components are uncorrelated with regressors, that is $E[X_i\varepsilon_i] = 0$ where $\varepsilon_i = u_i + v_i$, holds. OLS provides consistent estimates of $\boldsymbol{\beta}$, with the exception of the constant term since the usual assumption of $E[\varepsilon_i] = 0$ does not hold.

By making explicit distributional assumptions on u_i and v_i, maximum likelihood estimation can be used and is the asymptotically efficient estimator, provided the distributional assumptions are correct. In practice maximum likelihood estimation tends to be adopted since the prediction of cost efficiency relies on exact distributional assumptions being made.

The distributional assumption commonly used for v_i is

$v_i \sim iidN\left(0, \sigma_v^2\right)$

And for u_i

$u_i \sim iid\left|N\left(0, \sigma_u^2\right)\right|$ or $u_i \sim iidN^+\left(\mu, \sigma_u^2\right)$ or $u_i \sim \text{iidexponential}\left(\sigma_u^2\right)$

In addition $\mathrm{v_i} \perp \mathrm{u_i}$ i.e. are independent of each other. With the above distributional assumptions, the models are known as the stochastic frontier normal-half normal, normal-truncated normal and normal-exponential models respectively.

Given these assumptions, there exists a closed form expression for the log likelihood of (3.5) (see Aigner et al (1977) for the normal-exponential and normal-half normal models and Stevenson (1980) for the normal-truncated normal model). As such the model parameters $\left(\boldsymbol{\beta}, \sigma_u^2, \sigma_v^2\right)$ can be estimated by maximum likelihood and therefore the estimates are consistent and asymptotically efficient.

There remains the issue of how to predict the level of firm efficiency from these models, important since it is likely to be the primary motivation for adopting stochastic frontier analysis.

There exist many possibilities for computation of overall industry wide efficiency (Jondrow et al, 1982). The aim is to measure the extent to which the "average" firm in population suffers from inefficiency. The first two possible methods work via estimates of the mean of $\mathrm{u_i}$. One option is to estimate the mean of the composite error, ε_i, by use of the residuals from the estimated regression of model, $\hat{\varepsilon}_i$, and to use this as the estimate of the mean of u_i. Intuitively $\hat{\varepsilon}_i$ can be averaged since the mean of v_i is zero and so any variation of the true mean of ε_i should be due to the u_i error component. The second is to estimate the mean of u_i by deriving the mean of ε_i analytically. In the truncated half normal case this is a function of the variance of u_i and so the estimated value of σ_u^2 is used in the computation. This or the measure in the first method can then be transformed into a measure of overall technical efficiency (see Aigner et al, 1977). A further option for calculating overall industry efficiency is to take the (cost weighted) average of the individual efficiency scores computed using the methods proposed in Jondrow et al (1982) (see below). It is not clear which measure is to be preferred and what the differences between the measures is likely to be.

However, while a measure of overall industry efficiency is useful as a summary measure for the industry, what early proponents of frontier models envisaged (such as Farrell (1957)) and indeed what industry regulators require, is a measure of inefficiency specific to a firm. It was five years from the basic SF model being proposed until Jondrow et al (1982) showed how to calculate a measure of firm specific inefficiency. The problem is non-trivial because by estimating parameters in the SF model, estimates are produced of the variances (and, if applicable, the means) of each of the component errors, v_i and u_i. However, the residual from the estimated model captures the sum of both v_i and u_i i.e. ε_i. Thus there is a problem as to how to decompose the residual given only the distributions of the error components are known[11]. The method that Jondrow et al (1982) propose proceeds by calculating the expected value of $\mathrm{u_i}$ given ε_{i} is known.

Jondrow et al (1982) propose the following point predictor of cost efficiency ($E\hat{F}F_i$):

$$E\hat{F}F_i = \exp[-E(u_i \mid \varepsilon_i)] \tag{3.9}$$

See Jondrow et al (1982) for the specific expressions for the normal-half normal and normal-exponential model of (3.9).

Battese and Coelli (1988) used the distribution of $\mathrm{u_i} \mid \varepsilon_{\mathrm{i}}$ to derive an alternative predictor of cost efficiency that is optimal in the sense of minimising mean square prediction error. This is an alternative to the estimate of cost efficiency express above.

Chapter 7 of this thesis is concerned with predicting inefficiency from cross sectional stochastic frontier models. As such the detailed discussion of the properties of the predictor in (3.9) is discussed in that chapter. The following extract however summarises the key issue:

“Point predictors for firm inefficiency are common in the literature and follow the

[11] This problem is further complicated by the fact that ε_{i} and the variance components of $\mathrm{v_i}$ and $\mathrm{u_i}$ are not known with certainty. Instead only estimates of these quantities are available. This is issue is returned to in Chapter 7 in terms of computing uncertainty around inefficiency predictions which incorporate uncertainty from parameter estimation.

methodology of Jondrow et al (1982). However in cross sectional models, these point predictors are known to be inconsistent [as the sample size increases] for the quantity of interest; namely the firm specific realisation of a random variable. The question then arises; how precise is the prediction of firm inefficiency? With this in mind, and the general desire of practitioners to understand uncertainty in their estimates, it is perhaps surprising that interval predictors are not commonly reported in the empirical literature." (This Thesis, p.202)

3.4.1. Alternative methods to incorporate cost efficiency

As described in section 3.3, the stochastic frontier model (3.7) is the true econometric counterpart to the economic model which includes firm inefficiency (3.6). However it is not without limitations particularly when estimated using cross sectional data (3.8). These include:

- The inconsistency of the prediction of cost efficiency for the quantity of interest (the actual level of firm inefficiency as opposed to the expectation of its distribution).
- The potentially arbitrary distributional assumptions on the error components

It is therefore considered reasonable to examine other approaches to predict the cost efficiency of firms.

There are two general sets of methods used to analyse the efficiency and productivity of decision making units (DMUs); those that utilise econometric estimation of parametric functions, termed parametric methods, and those that do not, termed non-parametric methods (Coelli et al, 2005). Subsequent to the release of this text book, a further class of 'semi-parametric' models have emerged which blur the distinction. However in this section the distinction is maintained. Semi-parametric models for panel data methods are reviewed in section 3.5[12]. Furthermore, 'efficiency methods' refer to methods to measure cost (or technical and allocative) efficiency rather than scale characteristics (also known as scale efficiency). Parametric and non-parametric methods are reviewed briefly below.

3.4.1.1.Parametric Methods

The parametric methods, alongside the stochastic frontier model, include restrictions of the stochastic frontier model in (3.5). Essentially, this involves assuming away one of the error components comprising ε_i; eliminating v_i produces a deterministic frontier whilst eliminating u_i ignores inefficiency.

Clearly, the latter model is simply the standard regression model and so it can be consistently estimated by OLS given the usual assumption that regressors are exogenous with respect to the error, $E[X_i v_i] = 0$. However no measure of firm inefficiency can be computed from this model; by assumption there is no inefficiency.

Given the failure of standard least squares techniques to account for firm

[12] The semi-parametric models in section 3.5 are only a sub-set of those in this wider literature. In particular in the semi-parametric models in section 3.5, the deterministic frontier is fully parametric while the error is not. However there are other models where the errors are fully parametric but the deterministic frontier is non-parametric e.g. Stochastic Non-smooth Envelopment of Data (StoNED) (Kuosmanen and Kortelainen, 2012).

inefficiency, one method to correct for this is to re-interpret the residual in the regression as firm inefficiency rather than statistical noise. Thus all errors are assumed to be positive and using the notation in (3.8), the following model is considered:

$$\ln C_i = \ln\left(C\left(\mathbf{y}_i, \mathbf{p}_i; \boldsymbol{\beta}\right)\right) + u_i \quad i = 1,\ldots, N \tag{3.10}$$

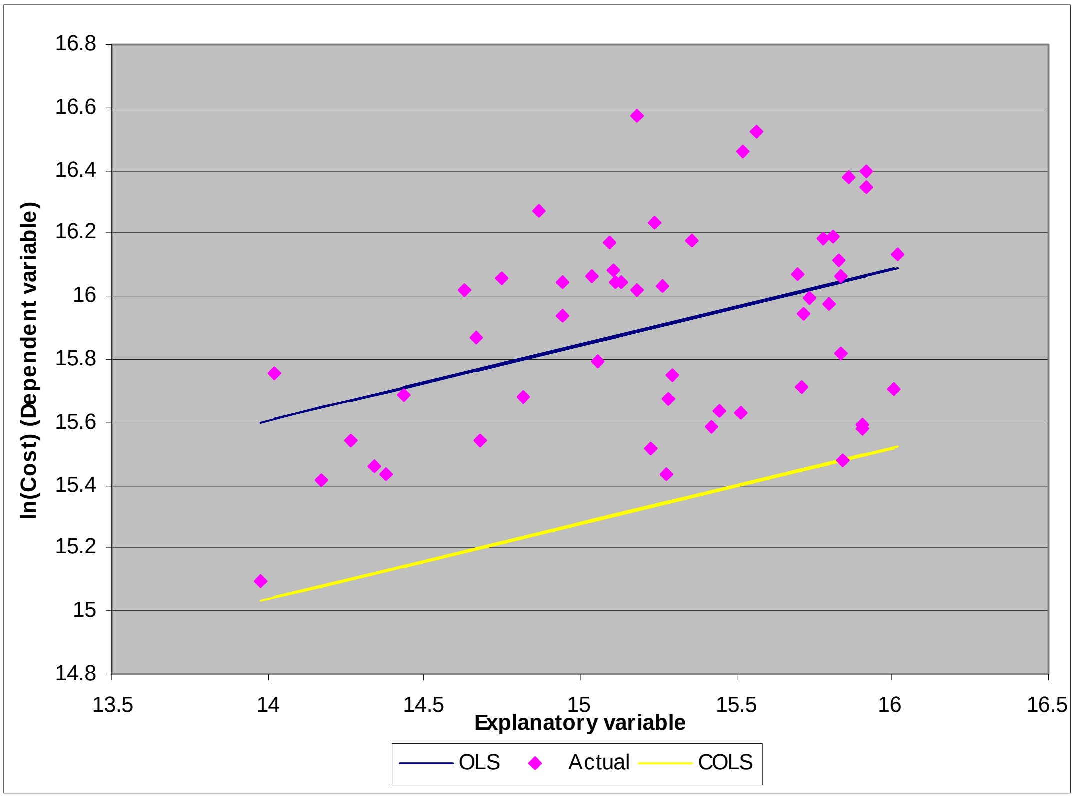
Estimation can be undertaken by ‘shifting’ the estimated OLS regression line down (in the case of a cost function). There are several techniques to do this. One such is Corrected Ordinary Least Squares (COLS). Shifting of the OLS line can be traced back to Winsten in his discussion contribution to Farrell (1957). Greene (1980) discusses the statistical properties of the estimators resulting from this shift.

COLS utilises the estimates of the slope coefficients from ordinary least squares but determines the intercept (β_1) coefficient as the ordinary least squares estimate plus the minimum residual in the case of a cost function. That is

$$\beta_1^{\mathrm{COLS}} = \beta_1^{\mathrm{OLS}} + \min\left(\hat{\varepsilon}_{\mathrm{i}}\right) \quad \forall \mathrm{i} \tag{3.11}$$

This is illustrated in Figure 3.1 which shows estimation of a cost function using OLS and then using COLS with respect to one output (all other things equal). The OLS line has been shifted down by the distance of the lowest residual. Thus the COLS line is now a frontier since all points are on or above the line.

Figure 3.1 The Corrected Ordinary Least Squares Method – Cost function



Source: Own analysis for illustration purposes

While this and other deterministic frontier techniques do improve on standard least squares methods by explicitly recognising firm inefficiency, they do so at the cost of assuming away any legitimate statistical noise. It should be noted that (3.10) is the original economic model and not an econometric model in the conventional sense since it does not allow for noise. Thus all deviations from the actual and fitted level of cost. are attributed to firm inefficiency as opposed to other sources such as measurement error and factors outside the firm's control not captured in the regression equation by the explanatory variables.

In addition, these techniques tend to be very susceptible to outlying values. An anomalous observation may appear very efficient relative to other firms, while, in fact, this value was subject to substantial measurement error. Thus other firms will have low efficiency scores due to the presence of the single outlier. In regulatory uses, regulators have often calculated efficiency scores relative to a frontier which has been shifted to the 75th percentile residual rather than the minimum (or maximum) (see Smith, Wheat and Nixon (2008) for an example).

3.4.1.2.Non-parametric methods

The non-parametric models are index number models and can be broken down into two streams:

- Total Factor Productivity (TFP) Measures which compute the ratio of inputs to outputs using an appropriate set of weights. This is used to calculate changes in TFP, but does not distinguish between the three drivers of TFP; efficiency change, technical change and scale effects; and
- Data Envelopment Analysis (DEA) techniques which use mathematical programming techniques to determine a frontier and calculate the distance of firms from the frontier allowing computation of the relative efficiency of firms.

Total Factor Productivity (TFP) Index Methods

Index number methods are concerned with measuring firm productivity and have been historically applied to aggregate time series data (Coelli et al, 2005 p. 6). An index number is defined as “a real number that measures changes in a *set of related* variables” (Coelli et al, 2005, p. 86, emphasis added). Essentially, index numbers are a method of summarising changes in a variety of variables via a method of aggregation justifiable either from theory or empirical observation. Productivity is the ratio of outputs to inputs. However, since a firm often produces several outputs and/or with several inputs, it is necessary to aggregate outputs and/or inputs using a set of appropriate weights. The weights that are used are likely to have a major effect on the results of any application. Index numbers are not considered further since the focus of this review is efficiency measurement. However, it should be noted that some index weighting systems have been shown to theoretically represent productivity growth given a set of assumptions regarding the underlying technology. See Coelli et al (2005, chapter 4) for a thorough treatment.

Data Envelopment Analysis

A method of accounting for firm inefficiency is using data envelopment analysis (DEA). This is a non-parametric technique in the sense that no parameters of a function are estimated. Instead mathematical programming is used to envelop the data and in particular determine which observations are on the frontier and which are not. Farrell (1957) was the first to propose the idea of enveloping data using a piecewise linear-hull i.e. a multi-dimension frontier, however it was not until Charnes, Cooper and Rhodes (CCR) (1978) proposed the first methodology termed DEA for the constant returns to scale case.

CCR proposed an input orientated method which can be represented as:

$$\max_{u,v}\left(\mathbf{u'q_i} / \mathbf{v'x_i}\right),$$

$$\text{st} \quad \mathbf{u}'\mathbf{q}_j \big/ \mathbf{v}'\mathbf{x}_j \le 1, \quad j = 1,2,...,N \tag{3.12}$$
$$\mathbf{u}, \mathbf{v} \ge \mathbf{0}$$

where $\mathbf{q}_i$ and $\mathbf{x}_i$ are Mx1 and Nx1 vectors of outputs and inputs respectively for the ith firm and $\mathbf{u}$ and $\mathbf{v}$ are corresponding Mx1 and Nx1 vectors of output and inputs weights. This formulation can be interpreted as the objective is to maximise the efficiency measure of firm i by choosing output and input weights subject to all efficiency measures for all the N firms calculated using the same output and input weights are less than 1 and all weights should be greater or equal to 1.

$(\mathbf{u}'\mathbf{q}_i / \mathbf{v}'\mathbf{x}_i)$ is interpreted as an efficiency measure (as opposed to simply a TFP measure) since this measure has to be between 0 and 1 and because it is calculated relative to other firms. That is, all firms are constrained to have efficiency measures between 0 and 1 with the same set of weights. The assumption of constant returns to scale means that firm i can be compared to all other firms even if the scale of production of some other firms is substantially different, since the ratio of efficient input to output is invariant with scale.

This problem has to be solved for each of the N firms (each will potentially have different weights). In order to get a unique solution, $\mathbf{v}'\mathbf{x}_i$ is set equal to 1.

DEA has a similar limitation as the COLS method since it does not explicitly allow for statistical noise in the data. Instead, a set of linear programming problems are solved for each firm and for each time period which determines whether, in relation to other firms, the firm is on the frontier. This forms a frontier which is piecewise linear.

A big advantage of DEA over parametric techniques is that there is no need to specify a functional form for the cost, production etc. relationship under consideration. Only a relatively weak assumption that the frontier can be represented by a piece-wise-linear convex hull is required. Models have been proposed which assume both constant and variable returns to scale. DEA is attractive to regulators because it allows each firm to have independent input and output weights which in turn mean the firm has a high chance of being efficient.

Finally, it should be noted that efficiency measures are computed rather than estimated in DEA, so that no standard errors and thus no confidence intervals can be constructed for them. Because of this, while measures of efficiency can be constructed, the likely reliability of such measures can not be determined[13].

3.5. Panel Stochastic Frontier Efficiency methods

In this section of the literature review, efficiency models which utilise panel data are considered. This is important given that Chapter 6 applies both fully parametric and semi parametric (to be defined below) methods to a multi-level dataset.

The econometric model considered is as in (3.7) but with amended subscripts to denote the two dimensions of panel data:

[13] It is noted that in recent years there has been considerable research to formulate a statistical foundation for DEA (see for example Daraio and Simar, 2007). This is very advanced material and beyond the scope of this review.

$$C_{it} = \alpha_0 + f(X_{it};\beta) + u_{it} + v_{it} \tag{3.13}$$

where i=1,…,N, t=1,…,T and X_{it} comprises k regressors.

While Schmidt and Sickles (1984) were not the first to develop panel data variants of models (for example Pitt and Lee (1981) proposed a time invariant efficiency model estimated by maximum likelihood methods), they were the first to articulate the principal benefits of panel data. These are threefold.

Firstly, the prediction of firm inefficiency is consistent, unlike in the cross sectional case (see 3.4 and Chapter 7). As the time period under consideration increases, so the probability that the prediction of firm inefficiency is an arbitrary distance from the true value tends to zero.

Secondly, it is possible to allow for correlation between explanatory factors and the inefficiency error component. This can be done by utilising a fixed effects approach (or some instrumental variables variant) (Schmidt and Sickles, 1984), or in the case of ML estimated models, incorporate these effects directly into the mean (or variance) of the inefficiency term (for example Haug and Liu, 1994 and Battese and Coelli, 1995).

Thirdly, there is a more general problem with the required distributional assumptions necessary in cross sectional analysis. Specific distributions have to be assumed for the two error components. Schmidt and Sickles (1984) make two points regarding this. First, they state that these assumptions have not been shown to be robust in terms of the sensitivity of results. Since their paper, there have been several studies which show, in general, that the results are not sensitive to the distributional assumptions. For example Greene (1990) found rank correlations between 0.747 and 0.980 when examining efficiency scores from the half-normal, truncated-normal, exponential and gamma distributions using a cross section of 123 US electric utilities.

Schmidt and Sickles' second point is perhaps more troubling, namely that skewness in the residuals is interpreted as inefficiency. All the distributional forms compared in the literature (normal-half normal, normal-truncated normal, normal-gamma) imply the same sign skewness. However there is no reason for suspecting that inefficiency would imply such skewness; all that is required is that the inefficiency distribution is one-sided. Indeed, other distributions can yield the opposite sign skewness (e.g. normal-doubly truncated normal (Qian and Sickles, 2007)) and further, the shape of these alternative distributions have some economic justification, for example, the doubly truncated normal mentioned above is motivated by an upper limit of inefficiency tolerated in a market (above this level the cost conditions force exit from the market). See Almanidis and Sickles (2010) for a further discussion.

Considering panel data allows either precise distributional assumptions to be eliminated altogether (for example, by adopting a panel data least squares approach) or their effect to diminish since an assumption is made regarding the behaviour of inefficiency over time as well as a specific distributional form. As Schmidt and Sickles (1984) articulate "essentially, evidence of inefficiency can be found in constancy over time as well as in skewness" (p. 367).

Time invariant models are first discussed and then time-varying models are considered.

3.5.1. Extension of the simple cross section model to time invariant models

Schmidt and Sickles proposed four time invariant formulations of the stochastic frontier model. In a time invariant model, inefficiency is assumed to vary across firms but is the same for all years for a given firm. The methods proposed (by Schmidt and Sickles, 1984) were the fixed effects model, generalised least squares random effects model, maximum likelihood formation of the random effects model (normal-half-normal model) and Hausman-Taylor instrumental variables (IV) estimation.

The basic time invariant model can be expressed (where the dependent variable is in logs) as:

$$C_{it} = \alpha_0 + f(X_{it};\beta) + u_i + v_{it} \tag{3.14}$$

where all variables are defined as in (3.13) except u_i which is time invariant but firm variant.

There are several possible distributional assumptions on u_i and v_{it} which yield different econometric models put forward in the literature.

3.5.1.1.Semi-parametric models

The first set of models are semi-parametric in the sense of the terminology of Sickles (2005) and the discussion in sub-section 3.2.2. In these specifications no specific distributions are imposed on either error terms, only assumptions on a set of moments of each error term and the regessors. Schmidt and Sickles (1984) outlined two possible specifications. First there is the fixed effects specification, where the u_i are assumed fixed parameters to be estimated.. The model can be written as:

$$C_{it} = \alpha_i + f\left(X_{it};\beta\right) + v_{it} \tag{3.15}$$

where $\alpha_i = \alpha_0 + u_i$

Once the model has been estimated the u_i can be recovered by assuming one firm in sample is totally efficient. In the case of a cost function (considered in the notation here), $\min(\hat{\alpha}_{\mathrm{i}}) = \hat{\alpha}_0$ i.e. all other firm intercepts have to be equal or greater to this, since they are either as or more inefficient than this firm.[14]

There are two important limitations of this model. Firstly, regressors cannot be time invariant. This is because it will be perfectly correlated with the fixed effect which will result in a singular regressor matrix. Secondly, because firm intercepts are fixed there is no potential to forecast hypothetical future firms. However, given the context

[14] Note that for a production function $\alpha_i = \alpha_0 - u_i$ and so $\max\left(\hat{\alpha}_{\mathrm{i}}\right) = \hat{\alpha}_0$.

of regulation, this seems less of a concern[15]. Therefore firm specific inefficiency can be calculated as:

$$\hat{\mathrm{u}}_{\mathrm{i}} = \hat{\alpha}_{\mathrm{i}} - \min(\hat{\alpha}_{\mathrm{i}}) \tag{3.16}$$

An alternative specification is to assume that u_i is a random effect. To keep this specification semi-parametric this model is estimated using generalized least squares just like any random effect panel data model (see Greene (2012) for the details of the method). The following is estimated:

$$C_{it} = f\left(X_{it};\beta\right) + \varepsilon_{it} \tag{3.17}$$

where $\varepsilon_{it} = v_{it} + \alpha_i$ and $\alpha_i = \alpha_0 + u_i$ i=1,…,N

Following estimation, firm specific estimates of α_i are recovered as the mean of the firm specific residuals:

$\hat{\alpha}_i = \sum_{t=1}^{T_i} \hat{\varepsilon}_{it}$ where T_i is the number of time periods observed for firm i and relative inefficiency is calculated as in (3.16). Efficiency can then be computed using the standard transformation for models with log dependent variables given in (2.9).

It is required to assume that the regressors are uncorrelated with both error

[15] In regulatory contexts, the firms from year to year tend to be the same and the primary interest is not to use the model to forecast performance of new firms

components[16] in order to yield unbiased and consistent parameter estimates. If this assumption holds, random effects yields more efficient parameter estimates than fixed effects. This approach can also accommodate time invariant regressors since coefficients are estimated using a weighting of the within estimator (fixed effects estimator) which sweeps out these regressors and the between estimator, which allows identification of these parameters separate to the time invariant firm effect. The assumption of regressors being uncorrelated with errors can be tested using the Hausman test (Hausman, 1978 and Hausman and Taylor, 1981).

Finally for the semi parametric models, there exists a set of estimators which are more efficient than fixed effects should some, but not all, regressors be correlated with the firm effects. These estimators were proposed by Hausman and Taylor (1981). Hausman and Taylor showed that under certain conditions, an estimator could be found which is more efficient than the fixed effects estimator and importantly could accommodate time invariant regressors. In general, provided the number of time varying regressors that are uncorrelated with the firm effects is at least as great as the number of time invariant regressors that are correlated with the firm effects, then a Hausman Taylor estimator exists. These estimators are called instrumental variables estimators as they use regressors which are uncorrelated with the firm effects as instruments for the regressors that are correlated with the firm effects.

[16] In the fixed effects there needs to be only no correlation between regressors and random noise as the inefficiency term is itself a regressor

3.5.1.2.Parametric Models

An alternative means of estimating the random effects model is to assume specific distributions on the two error components. The model is now given directly by (3.14). In this case the model is referred to as parametric and note that u_i represents absolute inefficiency rather than relative inefficiency. This is because here, u_i is modelled explicitly and is drawn from a distribution of possible u_i. This is unlike in the semi-parametric models where the firm effect, α_i is modelled explicitly (either drawn from a population distribution or assumed fixed) and then u_i is computed from this by setting the firm with least α_i equal to zero inefficiency (in the case of a cost model).

It is common to assume a normal distribution with zero mean for the v_{it} error term (Kumbhakar and Lovell, 2000). For the one sided term, u_i, distributions that have been applied include the half-normal (Pitt and Lee, 1981), truncated normal (Battese and Coelli, 1992), exponential (Econometric Software, 2010a) and gamma (Econometric Software, 2010a).

These models are generally estimated by maximum likelihood estimation although they can be estimated using Bayesian techniques[17] (see Greene, 2008 for a review). For the purpose of this review and thesis the focus is on estimation by maximum likelihood as this has received the most attention in the applied literature. The

[17] The Bayesian approach is not considered further as it has not been widely used in the applied literature. It is noted that this is a growing area of the theoretical literature in this field.

likelihood functions for the models are relatively easy to derive given the independence assumptions between error components and between the error components and the regressors. The likelihood function for the half normal model was first derived by Pitt and Lee (1981) and generalised to the unbalanced panel by Battese and Coelli (1988) who also extended it to the more general truncated normal model in which the half-normal is nested.

If the distributional assumptions are valid, then the parametric approach yields more efficient estimates of parameters and efficiency scores than the semi-parametric models because they utilise the extra information about the precise distribution of the error. If, however, the distributional assumption is invalid, then estimates are likely to be inconsistent and biased. See the earlier discussion (p. 69) regarding the sensitivity of model results to distributional assumptions.

3.5.2. Time varying inefficiency models

All the models considered in the above section assume that inefficiency is time invariant. As the length of the panel (the size of T) increases, this becomes an increasingly untenable assumption. This maybe for a number of reasons:

1) The average efficiency of all firms may change over time – all efficiency scores are scaled over time, all firm rankings stay the same;
2) Some firms ‘catch-up’ with others compared to which they were previously less efficient – efficiency scores change proportionally more or less for each firm but rankings stay constant;
3) Firms overtake each other in terms of relative performance – efficiency

scores vary differently for each firm and so rankings can change

In ascending order, these reasons require more flexible modelling approaches. The requirement for a model to address either 1, 2 or 3 will depend on the length of the panel since the longer a firm is observed the more likely it is to have radically changed performance.

3.5.2.1.Semi-parametric models

The essential problem for a semi-parametric model is that there can never be enough observations to identify T x N firm and time specific parameters and identify the k parameters involved in the deterministic frontier. As such some structure has to be imposed on the model to ensure identification (Kumbhakar and Lovell, 2000 p. 108). Two semi-parametric approaches are considered. Firstly, there are the methods proposed by Cornwell, Schmidt and Sickles (1990) which achieve identification by specifying a deterministic function of efficiency change for each firm. Secondly, consideration is given to the method of Lee and Schmidt (1993) which does not put such a ridged parametric function on the path of inefficiency over time but at the cost of defining the same path for all firms.

Cornwell, Schmidt and Sickles (1990) extended the fixed and random effects model to the case where in (3.15), α_i is replaced with some parametric function of time, with the following suggested:

$$\alpha_{it} = \Omega_{i1} + \Omega_{i2}t + \Omega_{i3}t^2 \tag{3.18}$$

It should be noted that this model accommodates all of the three reasons for adopting a time varying efficiency approach. Relative firm scores and rankings can change from year-to year and this variation is in a systematic (deterministic) way. This is because the Ω parameters vary from firm to firm. The problem is that there are now N x 3 firm specific parameters to estimate as well as the parameters required to fit the deterministic frontier. This presents two potential problems. First, there may not be sufficient observations to even identify all the parameters. For such a complex specification T must be greater or equal to 4 for the estimators to be computed. Second, even if estimation can proceed, the resulting parameter estimates may have large variances reflecting the large numbers of parameters required to be estimated.

This model is either estimated by fixed or random effects. For the fixed effects, the most simple case is to estimate the model directly by substituting the expression for α_{it} directly into (3.15). Efficiency is again a relative concept; now measured relative to the best performing firm in a given year. For each year, efficiency for firm i is computed as:

$$u_{it} = \alpha_{it} - \min_t(\alpha_{it}) \tag{3.19}$$

This presents an interesting possibility. While the path of α_{it} is smooth for each firm, the path of u_{it} may not be smooth over time because different firms may define the frontier at different points in time.

A further semi-parametric model was proposed by Lee and Schmidt (1993). This method differs from those imposed by Cornwell, Schmidt and Sickles (1990) because, unlike Cornwell, Schmidt and Sickles they do not impose a smooth function of efficiency change for firms. They trade this for the constraint of having to impose the same path of efficiency change on all firms. As such, this method can be seen as incorporating the reasons outlined in point 2) but not in point 3) above i.e., the method allows for all firms efficiency to change over time and for a degree of catch-up between firms. However the rankings of firms will not change over time. The general form of the model is the same as in (3.14), however u_i is substituted with:

$$u_{it} = \alpha(t) \cdot u_i \tag{3.20}$$

where $\alpha(t)$ is a set of time dummy variables and u_i is defined as before. Importantly, $\alpha(t)$ does not restrict the temporal pattern of inefficiency to any parametric form, but does apply the same path for all firms. It should be noted that there are T-1 parameters to estimate contained in $\alpha(t)$ rather than T since one parameter has to be fixed (usually set $\alpha(1)$=1) to identify u_i. Note also if $\alpha(t)$=1 for all t, then the model collapses to the time invariant panel model. Lee and Schmidt (1993) consider the u_i to be both fixed and random effects and the $\alpha(t)$ to be parameters. Notwithstanding, all these models are non-linear models which complicates estimation.

Once the model has been estimated, the inefficiency estimates are recovered as:

$$u_{it} = \left(\hat{\beta}_t \hat{u}_i\right) - \min_i \left\{\hat{\beta}_t \hat{u}_i\right\} \tag{3.21}$$

Thus firm efficiency in time t is relative to the best performing firm in time t.

It is important to note that because semi-parametric models yield estimates of relative efficiency, it is not possible to disentangle the effects of technical change (movement of the frontier) and movement of the firm which is most efficient, as by definition of relative efficiency, they are on the frontier. So, unlike in the case of the time invariant model, where the distinction between absolute and relative efficiency would seem academic, it matters here in terms of practical output. Indeed, the assumption of time invariant efficiency is sufficient to identify the technical change.

To illustrate this further consider the restricted case of (3.18) discussed in Kumbhakar and Lovell (2000, p. 109) where $\Omega_{i2} = \Omega_2$ and $\Omega_{i3} = \Omega_3$, that is there is one trend over time for all firms although each firm can have a different starting value at T=0, Ω_{1i}. In this case, the lack of identification between the most efficient firms absolute efficiency and technical change yields two extreme interpretations. Firstly, all firms improve (or worsen) efficiency over time as dictated by the relationship for α_{it}. In this case there is no technical change over the period by implication of the assumption. The second extreme interpretation is that inefficiency is time invariant and the time varying nature of performance is due to the movement of the frontier over time. Obviously there exists an infinite number of points between these extremes. Essentially further information is required in order to identify the two effects. In the case of the parametric model, this further information is given by specifying a population distribution for inefficiency.

3.5.2.2.Parametric models

Pitt and Lee (1981) identified two extremes of efficiency variation over time. The first is that there is no inefficiency variation across time i.e. inefficiency is time invariant. These models have been examined in detail in section 3.5.1. The other extreme is that inefficiency is independent across time. In this case a firm which was deemed very inefficient in the previous time period is no more likely to be very inefficient in the subsequent time period than any other firm; that is there is no correlation between efficiency scores across time. This may be unrealistic and a detriment to regulators (which want to use efficiency analysis to set efficiency targets for firms given past trends). However, it is noted that most empirical studies do report such models and also that some of the models proposed to deal with unobserved heterogeneity, treat inefficiency in such a way. Therefore this model (the pooled model) is discussed briefly in the following paragraphs.

This model is exactly the same as the cross section parametric models considered in the previous section (3.4), because the firm and time dimensions of the data are treated as one set of observations rather than explicit recognition of the panel nature of the data; the data could be thought of as a cross section with N x T observations. As such the model is often referred to as a pooled stochastic frontier model. The form of the model, assuming a normal-half normal composite error, is given below:

$$C_{it} = \alpha_0 + f(X_{it};\beta) + u_{it} + v_{it} \qquad (3.22)$$

where all variables are as defined in equation (3.14) except $u_{it} \sim N^{+}\left(0,\sigma_u^2\right)$ and is distributed independently of v_{it} and the regressors.

Being ultimately a large cross section model, this has the disadvantage of this class of models namely the lack of consistent prediction of the firm specific realisation of the inefficiency term. Only the conditional distribution of the inefficiency term can be derived and any summary measure of this distribution is not a consistent estimator of firm inefficiency as either N or T is increased (as effectively the cross section is just expanding).

Pitt and Lee (1981) also identified an intermediate case where the inefficiency scores for a firm are correlated over time. They proposed a model using a Seemingly Unrelated Regression (SUR) procedure (proposed by Zellner (1962) and discussed in sub-section 3.2.1); however this is not considered further here as it had the major flaw of not being able to yield firm specific estimates of inefficiency[18].

It was not until work by Kumbhakar (1990) and Battese and Coelli (1992), when parametric time varying models were proposed, that could importanty yield firm specific estimates of inefficiency. The general formulation of these models is:

[18] Indeed Pitt and Lee acknowledged this and proceeded to estimate the model in order to determine ‘how wrong’ the other two models were.

$$C_{it} = \alpha_0 + f(X_{it};\beta) + u_{it} + v_{it} \qquad (3.23)$$

where $u_{it} = f(t) \cdot u_i$ and all other variables and parameters are defined as in (3.14), including the independence of u_i with respect to v_{it} and the regressors. In the Kumbhakar specification:

$$\mathrm{f}(\mathrm{t}) = \left[1 + \exp\left\{\gamma \mathrm{t} + \delta \mathrm{t}^2\right\}\right]^{-1} \qquad (3.24)$$

and in the Battese and Coelli (1992) specification:

$$f(t) = \exp\{-\gamma(t - T)\} \text{ (a) or } \mathrm{f}(\mathrm{t}) = \left\{1 + \gamma(\mathrm{t} - \mathrm{T}) + \delta(\mathrm{t} - \mathrm{T})^2\right\} \text{ (b)} \qquad (3.25)$$

These models incorporate reasons 1 and 2 for time varying efficiency (given on page 71), but do not allow for firms to overtake/fall behind each other. This is because the same pattern of time variation is imposed on all firms. They do allow for catch-up between firms and the Kumbhakar and the (b) specification by Battese and Coelli (3.25) allow for a single turning point in the path of inefficiency for each firm over time. The Battese and Coelli models 'anchors' the u_{it} such that $\mathrm{u}_{\mathrm{iT}} = \mathrm{u}_{\mathrm{i}}$; i.e. the inefficiency in the last year is equal to the random draw from the distribution. The Kumbhakar model does not make such an anchor and so u_i does not have this interpretation in this model.

These models can be thought of as analogous to the Cornwell, Schmidt and Sickles

(1990) models with the constraint of $\Omega_{i2} = \Omega_2$ and $\Omega_{i3} = \Omega_3$[19]. The major difference is that now there is not ambiguity between whether this trend is the result of technical change or changes in inefficiency. Because inefficiency is now absolute, the impact of technical change can be modelled separately.

These models were further generalised by Cuesta (2000) and Orea and Kumbhakar (2004) to allow for firm specific time variation paths, incorporating reason 3 for generalising models to time varying inefficiency.

3.5.3. Accounting for unobserved heterogeneity

Time invariant unobserved heterogeneity is a term used in panel data modelling to describe systematic differences across firms which are constant over time but are not captured by the explanatory variables (regressors). There are many features of railways that are difficult to quantify, are constant over time and explain costs. Examples include the topography and climate in which the railway operates. These omitted factors form unobserved heterogeneity.

Unobserved heterogeneity can further be disaggregated to that which is correlated with the regressors and that which is not. Unobserved heterogeneity may be correlated with regressors if, for example, a variable such as proportion of track electrified is omitted in an infrastructure cost frontier but this is correlated with

[19] Or $\Omega_{i2} = \Omega_2$ and $\Omega_{i3} = 0$ in the case of the first Battese and Coelli specification. It is not strictly true since the parametric models adopt an exponential functional form, while the Cornwell, Schmidt and Sickles models adopt a linear form. However one is still a monotonic transformation of the other, so properties such as turning points remain.

passenger usage (as is often the case – greater used lines get electrified). The implication of ignoring this is that the parameter estimates are biased, which in turn distorts inefficiency predictions.

Alternatively, failure to account for unobserved heterogeneity which is not correlated with the regressors does not (at least in linear models) bias parameter estimates. However, it does mean that all systematic residual differences between firms are attributed to inefficiency rather than between the two factors. Thus, for efficiency analysis, accounting for both types of unobserved heterogeneity (i.e. time invariant inefficiency and other unobserved heterogeneity) is important.

In recent years, incorporating allowances for unobserved heterogeneity in stochastic frontier models has become a popular topic of research. There are four approaches to modelling unobserved heterogeneity in stochastic frontier models:

1) Ignore unobserved heterogeneity in the modelling. This is the case in all of the models considered in section 3.5 to this point. While simple to operationalize, this does imply that either there is no unobserved heterogeneity in the model other than inefficiency or that the effect of unobserved heterogeneity will show up in the parameter estimates and/or in the inefficiency error. This maybe especially problematic in the models that assume time invariant inefficiency, since this is likely to capture other time invariant unobserved heterogeneity.

2) Adopt a 'True' formulation. The terms 'True Fixed Effects' and 'True Random Effects' models originate in Greene (2005), but similar models can be traced back to work by Kumbhakar and associates in the 1990's

(Kumbhakar and Hjalmarsson, 1995, Kumbhakar and Heshmati, 1995, Kumbhakar, 1991 and Heshmati and Kumbhakar, 1994). The form of the models is as the simple pooled model (3.22) but with the addition of a time invariant firm effect:

$$C_{it} = \alpha_i + f(X_{it};\beta) + u_{it} + v_{it} \tag{3.26}$$

where α_i is either treated as a random variable, uncorrelated with regressors, or as a fixed effect. In the model by Greene, α_i captures time invariant unobserved heterogeneity. The term 'True' could be taken to imply an attractive property of the model, namely that it is correct in all cases. This is not the case and was articulated in the papers by Kumbhakar and associates in the early 1990s. In some papers (Kumbhakar, 1991 and Heshmati and Kumbhakar, 1994) α_i is also representative of time invariant unobserved heterogeneity but in other papers (Kumbhakar and Hjalmarsson, 1995 and Kumbhakar and Heshmati, 1995), α_i is taken to represent time invariant inefficiency. In reality α_i will represent some element of time invariant inefficiency and some element of other unobserved heterogeneity. This is to be expected, since the model in (3.26) has the time varying inefficiency component (u_{it}) that is independent over time, i.e. by assumption there is no element of inefficiency captured by this component which is persistent over time. In Chapter 6, the model (3.26) is applied to multi-level 'sub-company' data with the persistent and residual interpretation in line with Kumbhakar and Hjalmarsson (1995) and Kumbhakar and Heshmati (1995). The issue of the alternative interpretation, al la Greene (2005), is discussed in the chapter but also in the further work section of the conclusion.

3) Adopt the model proposed by Kumbhakar et al (2012). This comprises the following:

$$C_{it} = \alpha_i + f(X_{it};\beta) + u_{it} + w_i + v_{it} \tag{3.27}$$

where $u_{it} \sim \left|N\left(0,\sigma_u^2\right)\right|$, $w_i \sim \left|N\left(0,\sigma_w^2\right)\right|$, $v_{it} \sim N\left(0,\sigma_v^2\right)$ and $w_i \sim N\left(0,\sigma_w^2\right)$ and all error components uncorrelated with X_{it} and all error components independent from each other.

Kumbhakar et al, propose a three stage approach to estimate this model. This model appears to solve all problems, in the sense that time invariant inefficiency (w_i) is distinguished from other time invariant inefficiency (α_i). However, this model suffers from a number of arbitrary assumptions. Firstly it is only valid if the precise distributional assumptions are correct (as in the cross sectional discussion). Secondly, and perhaps more fundamental, inefficiency is viewed as comprising a time invariant component and a time independent component. This must be questioned in long panels, in the sense that inefficiency evolves over time rather than being persistent with random variation around such a level.

4) Use the Mundlak transformation (Mundlak, 1978) to account for unobserved heterogeneity which is correlated with regressors. This is explained in more detail in Chapter 6, but essentially requires the addition of firm group means to the model equation. This should eliminate (or at least reduce) any bias to parameter estimates resulting from unobserved heterogeneity. Importantly

this results in the correct capture of that part of variation in cost between firms that is uncorrelated with the regressors (which depending on whether approach 3 is adopted, is either interpreted as inefficiency or further decomposed into a mixture of inefficiency and unobserved heterogeneity).

3.6. Summary

This literature review has provided the econometric background for this thesis. It has covered both general econometric concepts, including properties of estimators and estimation frameworks, as well as commonly adopted methods to measure the efficiency of firms. The former material serves as background for Chapters 5 to 7, but is of particular value to Chapters 6 and 7. Chapter 6 considers methods to estimate various formulations of a dual level inefficiency model and Chapter 7 considers the statistical interpretation of intervals for firm specific efficiency predictions from stochastic frontier models. The thesis utilises both least squares estimation techniques (Chapters 5 and 6) and maximum likelihood techniques (Chapters 6 and 7).

The econometric literature dealing with efficiency measurement has matured considerably following the early work by Farrell (1957) and the seminal papers on stochastic frontier analysis in 1977. There are now a large number of econometric techniques to investigate and measure firm efficiency, especially when panel data is available. The key feature of this economic issue that warrants such a dedicated review of applicable econometric techniques is that the component of the model of interest, inefficiency, is commonly modelled as a random variable, which means that understanding and measuring the residuals of the model is key (Greene, 2008).

If panel data is available then inefficiency can be better disentangled from random noise. However, there is a trade-off between the need to impose restrictions on the model and to obtain relatively precise estimates of firm inefficiency, versus allowing the data to speak and determine the most appropriate form.

Crucially, there is also a need to appraise what exactly is being measured in terms of the inefficiency gap. Different models impose different assumptions which require different interpretations as to what the inefficiency error term is actually measuring. Clearly the (appropriate) incorporation of external influences on firms and their performance helps to illuminate this, but caution still needs to be taken in simply interpreting any net inefficiency as something within the firm’s control. Ultimately inefficiency is modelled as a residual and so is something unexplained by the model.

4. Review of Applications to Railways

4.1. Introduction

In this chapter, the relevant literature relating to the application of cost analysis to railways is briefly surveyed. Most of the work prior to the last two decades, has been applied to vertically integrated railways. This work is examined from the perspective of highlighting methodological issues rather than focusing on the specific results

Some work has been developed in the infrastructure element of railways, starting from the need to provide a marginal cost basis for charging for access to the infrastructure. However, there has over the last decade been a set of analyses focused on measuring the relative efficiencies of infrastructure managers. Both elements are reviewed.

Much less work has been undertaken regarding passenger train operations and the vast majority of this work has been applied to Britain where franchising of passenger train operations has been most comprehensively implemented within the EU and there exists a reasonable amount of data for analysis. This is developed further in Chapter 5.

4.2. Vertically integrated railways

In this section, the key features and results of cost and efficiency studies in vertically integrated railways are reviewed. This part of the review uses Oum et al (1999) as the basis for the studies considered but updates it to the present day. Partial and total productivity studies are ignored because they do not deal explicitly with inefficiency as defined in Chapter 2. Furthermore, the studies that have utilised programming techniques are not surveyed. While these studies are aimed at measuring inefficiency, their non-parametric nature means that there is little to be learned in terms of key features applicable to parametric studies which is not covered in the parametric literature. The reader is directed to Oum et al (1999) for a survey of these studies. In addition it is noted that the most recent literature on vertically integrated railways utilise linear programming to study the impact of railway reforms (Growitsch and Wetzel (2009), Asmild et al (2009), Cantos et al (2010)).

There have been a number of innovative methods used to study the performance of railways. Table 4.1 summarises the key studies and states what function was estimated and the key inputs and outputs used. In the remainder of this section, the key advances in the methodology associated with econometric analysis of railway performance are surveyed.

Table 4.1 Summary of the characteristics of parametric cost studies in railways

Study	Sample	Function estimated	Inputs or prices used	Outputs used
Andrikopoulos and Loizides (1998)	1969-1993. European rail companies.	Translog cost function. Returns to scale and productivity studied but not explicit allowance for inefficiency	Total cost. Includes capital costs (historic cost depreciation + interest).	Sum of passenger km and freight tonne km.
Christopolous, Loizides and Tsionas (2000)	1969-1992 European rail companies	Input specific technical inefficiency via Generalized McFadden cost function	Total cost. Includes capital costs (historic cost depreciation + interest).	Total train km.
Coelli and Perelman (1999)	1988-1983. European rail companies.	Deterministic production function	Number of employees. Rolling stock capacity.	Passenger km. Freight tonne km.

Study	Sample	Function estimated	Inputs or prices used	Outputs used
			Route kilometres.	
Coelli and Perelman (2000)	1988-1983. European rail companies.	Deterministic production and input and output distance functions	Number of employees. Rolling stock capacity. Route kilometres.	Passenger km. Freight tonne km.
Couto and Graham (2008)	1972-1999 27 European railway companies	Short-run variable cost function with first order cost shares to separate out technical and allocative inefficiency	Input prices for labour, service rendered by third parties, equipment (variable inputs) and measure of capital stock. Also some network characteristic variables.	Two models: 1) Passenger-km and freight tonne-km (final outputs) 2) passenger train-km and freight train-km (intermediate outputs)
Cowie and Riddington (1996)	1992. European rail companies.	Deterministic production functions	Number of employees. Capital (financial measure).	Passenger train km. Service provision index.
Deprins and Simar (1989)	1970-1983. Europe + Japan rail companies.	Deterministic Production Function	Number of employees. Number of coaches / wagons. Energy consumption. Route kilometres.	Total train km.
Farsi, Filippini and Greene (2005a)	50 railway companies in Switzerland 1985-1997	Various stochastic total cost frontier specifications examining the effect of controlling for time invariant characteristics	Input prices: Energy labour and capital	Passenger-km and freight ton-km
Gathon and Perelman (1992)	1961-1988. European rail companies.	Stochastic Factor Requirement function	Number of employees.	Passenger train km. Freight train km. Route km.
Gathon and Pestieau (1995)	1961-1988. European rail companies.	Stochastic Production function (also second stage regression)	Number of employees. Number of rolling stocks. Route kilometres.	Sum of passenger tonne km and freight tonne km.
Ivaldi and McCullough (2001)	25 US Class 1 Railroads 1978-1997	Translog Variable Cost Function	Prices: Indexes of labour, equipment, fuel and materials	Car miles of a) Bulk, b) high-value, c) general traffic and replacement of ties installed (infrastructure output) and also average length of Haul and length of road miles
Kumbhakar (1988a)	13 US Class 1 Railroads 1951-1975	Cobb-Douglas Stochastic distance function with demand system to separate out technical and allocative inefficiency	Quantities of labour, energy and capital	Passenger-km and Freight tonne-km

Study	Sample	Function estimated	Inputs or prices used	Outputs used
Kumbhakar (1988b)	42 US Class 1 Railroads 1951-1975	Stochastic distance function with demand system to separate out technical and allocative inefficiency	Quantities of labour, energy and capital	Passenger-km and Freight tonne-km
Kumbhakar, Orea, Rodriguez-Alvarez and Tsionas (2007)	1971-1994. Europe rail companies	Output and Input distance function in a latent class framework	Quantities of labour, energy and capital	Passenger-km and freight ton-km
Lan and Lin (2006)	39 international railways 1995-2002	Two distance functions one modelling technical efficiency the other modelling service effectiveness	Efficiency model: Number of passenger rolling units, number of employees Effectiveness model: Passenger train-km and freight train-km	Efficiency model: Passenger train-km and freight train-km Effectiveness model: Passenger-km and freight tonne-km
Loizides and Tsionas (2002)	1969-1992. Europe rail companies.	Short-run cost function (not frontier) with coefficients which vary by firm or year	Operating costs. Capital stock (financial measure).	Passenger km. Freight tonne km.
Parisio (1999)	8 European Railway companies for 1973-89	Short-run variable cost function with first order cost shares to separate out technical and allocative inefficiency	Input prices: Labour, energy, materials. Length of track is the measure of the fixed input.	Passenger-km and freight ton-km
Cantos and Villarroya (2000)	1970-1990. Europe rail companies	Stochastic cost	Variable cost (excludes capital cost).	Passenger train km. Freight train km.
Cantos and Villarroya (2001)	1970-1990. Europe rail companies.	Stochastic cost and revenue functions – Operating costs, revenue	Labour price, energy price, mater price – price of passenger and freight outputs (rev model)	Passenger km. Freight tonne km.
Tsionas and Christopolous (1999)	1969-1992. European rail companies.	Stochastic Production frontier with firm environmental variables as determinants of mean inefficiency	Number of employees. Energy consumption. Capital (financial measure).	Sum of passenger km and freight tonne km.

The multi-output nature of railways has motivated recent studies to use either the cost frontier or distance function. Distance functions are related to a multi-output generalisation of the production function (the transformation function) and yield estimates of technical inefficiency through considering feasible radial expansions (contractions) of outputs (inputs) with respect to the production set.

For cost frontier models, both variable cost frontiers (Parisio, 1999, Ivaldi and McCullough, 2001, Couto and Graham, 2008) and total cost frontiers (Cantos and Villarroya, 2000 and 2001, Farsi et al, 2005a) have been estimated; the difference depending on whether the infrastructure is deemed quasi-fixed or variable. This decision is partly determined by the robustness of the available capital stock level variable(s) versus the capital price variable. As noted by Parisio (1999), the cost function approach has not been too popular compared with the production/distance function approach. This has primarily been because of difficulties in developing data on input prices, particularly infrastructure capital. Instead, he estimates a variable cost function which requires data on the levels of capital and not their associated price.

4.2.1. Outputs Used

Network industries can be viewed as producing many different heterogeneous outputs. Transport networks in particular, given the non-storability of the product, the large number of origin and destination combinations and the many different trip purposes, produce a very large number of outputs. In the limit railways could be thought as producing individual travel opportunities, by time, space and purpose. Clearly, such a disaggregation of outputs is likely to be too extreme to undertake meaningful parametric analysis. As such, a more pragmatic approach has to be taken in specifying outputs.

Several common features of the output specification can be considered. First, two general classifications of outputs are common (Oum and Yu, 1994). One set are termed “available outputs” which are measures of the service that the railway produces (capital supplied) which are available to customers to consume. Examples include train-km, vehicle-km and seat-km. The second set are termed “revenue outputs” which are measures of consumed outputs. Examples include passenger-km and tonne-km of freight hauled. These two sets could also be thought of as intermediate versus final outputs of the railway system, although it must also be borne in mind that the demand for rail services is often a derived demand.

When choosing whether to use available or revenue outputs, it is important to consider what is required to be measured in the analysis and whether the implicit assumptions on what is under the firm’s control versus what is exogenous is reasonable. For example, using available outputs can be justified when considering the performance of a railway manager where the required outputs from the railway are heavily prescribed by a regulator or government. As such the railway manager does not have much discretion as to how many train-km, vehicle-km etc., can be run. This is instead set by the regulator. However if analysis of the effect of government policy is the aim of a study, then it is more appropriate to adopt revenue output measures as policy makers have discretion in the specification of railway services to best meet demand. Of course, available outputs might be used in this context, alongside revenue outputs as a measure of characteristics (quality) of the revenue output.

Any measured inefficiency from models reflects both inefficiency of the railway manager and of policy makers or regulators (Oum and Yu, 1994). Lan and Lin (2006) cite Fielding et al (1985) who define specific terms for these concepts. They define the degree of sub-optimal transformation of inputs into intermediate outputs as “technical inefficiency”, while they define degree of sub-optimal transformation of inputs into final outputs as “technical ineffectiveness”. They define a further concept, “service ineffectiveness”, as the degree of sub-optimal transformation of intermediate outputs into final outputs. They point out that it is the non-storability property of railway outputs which requires such distinctions. This thesis is concerned with the “technical inefficiency” concept, since the railway undertaking (at least in the short run) has to take its outputs as given. “Technical inefficiency” is bounded by quotation marks in order to distinguish the Fielding et al concepts from the definition of technical efficiency in production theory. In particular in this thesis, cost inefficiency is considered which includes allocative as well as technical inefficiency even though this applies to the transformation of inputs into intermediate outputs (and not final outputs). This is appropriate given the thesis is evaluating the cost characteristics and performance of different parts of a vertically separated industry.

The second general distinction that has been made is the need to distinguish between scale and density effects. Density effects comprise the effect on costs of increasing all outputs (in equal proportion) while holding network size constant. Scale effects comprise the effect on costs of increasing all outputs and network size in equal proportion. It is important to distinguish between density and scale effects since it is often argued that marginal costs in network transport industries are below average costs and this is a problem in terms of opening such markets to competition. However, it is not clear that the marginal cost of expanding the network to accommodate the marginal consumer (here marginal O-D pair) is less than average cost. There is however, stronger reason to suggest that the marginal cost of accommodating an additional consumer using the current network size through greater utilisation, is very small.

It is important to emphasise that network size is viewed in railways as a characteristic of railway outputs, since the size of the network affects the scope of travel opportunities available to users. This is in contrast to the use of network size as a proxy for the capital stock for which empirical estimation of related coefficients has yielded counter-intuitive signs (see the discussion about inputs below). Therefore, empirical evidence suggests that network size has a strong relation to the output of the railway rather than as a measure of the stock of capital of the railway.

As discussed in sub-section 2.2,1, Caves et al (1981 and 1984) outlined expressions for returns to scale and returns to density in cost functions. To recap, Caves et al showed returns to scale (RtS) and density (RtD) can be computed as follows:

$$\mathrm{RtS} = 1 \Big/ \left(\sum_{i=1}^{m-1} \varepsilon_{y_i} + \varepsilon_S \right) \tag{4.1}$$

$$\mathrm{RtD} = 1 \Big/ \sum_{i=1}^{m-1} \varepsilon_{y_i} \tag{4.2}$$

Where ε_{y_i} is the elasticity of cost with respect to the ith output (i=1,…,m-1) and ε_S

is the elasticity of cost with respect to the network size variable[20].

The need to distinguish between scale and density effects or the choice between revenue versus available outputs is only part of the wider issue of how to account for the heterogeneity of railway outputs, as introduced at the start of this section. One way to deal with the heterogeneity in outputs is to group outputs into m groups and include a further set of r variables which characterise the outputs

$$C(y_1,\ldots,y_m,q_1,\ldots,q_r,p_1,\ldots,p_n) \tag{4.3}$$

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 ambiguous as to what variables represent outputs versus output characteristics versus network size. By implication it also means that in practice, the distinction between the "technical inefficiency" and "technical ineffectiveness" of Fielding et al (1985), discussed earlier, is far from clear e.g. if train-km and passenger load factor enter the model.

The inclusion of characteristic variables in the cost function specification has

[20] For notational convenience and consistency with other equations which do not distinguish between the network size variable and other outputs, the network size variable is treated as the m'th output and so only the first m-1 output elasticities are used in the RtD equation (which excludes this output).

prompted new definitions of returns to 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 Zhang (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 RtS described above, however in Oum and Zhang (1997) these relate to changes in observed rather than unobserved variables. These ideas are applied to the analysis of TOC costs in Chapter 5 where several scale and density measures are proposed taking into account variations in output characteristics as well as ‘primary’ outputs.

While this formulation does simplify the problem to a traceable level, the resulting function may be very complicated, given the number of variables 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 is explained in more detail in Chapter 5, in which it is applied.

4.2.2. Input prices

The measures of the price of inputs should reflect the opportunity cost of a unit of those outputs. For example, the opportunity cost associated with one hour of labour is the wage rate. Less obvious is the price of capital. It should reflect the hourly rental of the capital. This is problematic to measure because of heterogeneity in capital (see below) but also due to the fact that capital tends to be owned rather than leased. Methods such as the perpetual inventory method (see Bishop and Thompson (1992)) have been developed to better capture a measure of capital price.

A further issue with the price of capital is the relationship between this and the network size which could be viewed as a measure of capital. In particular because of a positive coefficient on miles of railroad a negative marginal product of capital is suggested (Wilson, 1997). However it is clear that in a railway cost function network size is much more related to the scale of output of operation than a measure of the capital stock of the network.

In practice, there is a similar problem to defining input prices as in defining outputs, i.e. the problem of heterogeneity in inputs. For example, average salary is likely to be a poor measure of the labour price as workers may work a different number of hours across observations. Likewise there is the possibility of a different mix of workers across different observations. Thus one firm may face higher labour costs because it utilises more expensive but higher skilled labour. This is likely to distort coefficient estimates (and indeed estimates of inefficiency) due to endogeneity of explanatory variables. The usual way to remedy this is to disaggregate further the input prices in the model (such as wage rates per staff type), but this adds to the number of coefficients to be estimated and the data may simply not exist.

4.2.3. Functional form used

Early studies of railway costs used linear cost functions, i.e. cost is characterised by a fixed component plus a component proportional to output. This was adopted by US ICC rail road regulators between the 1930s and 1980s. There were several economists, such as Meyer (1958), Meyer et al (1959), Borts (1960), Meyer and

Kraft (1961), Friedlaender (1969) and Griliches (1972) also criticised the approach.

One major criticism was the incompatibility of the two part function with economic theory. In particular, the function lacked input prices which Nerlove clearly showed were an important component of any cost function which was representative of the underlying technology. However, this is not necessarily a problem with the linear functional form, but more with the variables included therein. Likewise, outputs can be better specified as discussed in the previous sub section.

The most important criticism of the linear functional form related to its simplicity. In particular, marginal costs were constant across all output levels and indeed any other domain of the cost function. This has significant implications. Firstly, all other things being equal, the functional form imposes increasing returns to density (or increasing returns to scale if there is no network size variable included). Secondly, returns to density necessarily decrease as usage increases, all other things being equal. While these may be desirable features of the industry, the function does not allow these to be tested.

Given the problems with the linear functional form, a natural replacement was the Cobb Douglas (CD) cost function which Nerlove had shown to be dual to the CD production function. An m output (each denoted y_i), n input (each price denoted p_j) CD cost function can be written

$$C = e^{\alpha}\left(\prod_{i=1}^{m} y_i^{\beta_i}\right)\left(\prod_{j=1}^{n} p_j^{\gamma_j}\right) \tag{4.4}$$

Importantly taking logarithms yields a model which is linear in parameters making estimation possible using linear techniques:

$$\ln C = \alpha + \sum_{i=1}^{m}\beta_i \ln y_i + \sum_{j=1}^{n}\gamma_j \ln p_j \tag{4.5}$$

For this function to be compatible with economic theory, the restriction $\sum_{i=1}^{n}\gamma_i = 1$ has to be imposed. Provided all input prices are accounted for in the cost function, a proportional rise in all input prices should increase cost by the same proportion i.e. cost should be linear homogenous in input prices.

However, like the linear functional form, the CD functional form imposes several restrictions on the underlying technology. For all inputs, the elasticity of cost is constant and thus the share of cost is always constant irrespective of the input prices. An alternative description of this property is that the elasticity of substitution of one factor for another is always unity. This is a very restrictive way to model how firms adjust inputs in response to factor price changes.

The model also implies restrictions on the relationships between costs and outputs. The elasticity of cost with respect to outputs is constant (over the whole domain of the function) which implies returns to scale and density is constant. This in turn implies a very restrictive path for marginal costs. In particular:

$$\frac{\partial C / \partial y_i}{C / y_i} = \frac{\partial \ln C}{\partial \ln y_i} = \beta_i$$
$$\rightarrow \partial C / \partial y_i = C / y_i \cdot \beta_i \tag{4.6}$$

i.e. marginal costs are proportional to average costs.

The restrictive nature of the CD and linear functional forms prompted a large amount of research into less restrictive functional forms. 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 and Christensen and Greene, 1976). This nests the CD as a special (restricted) case but 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[21]. The general form of the Translog cost function for m outputs and n inputs is represented as (ignoring time and cross sectional subscripts as applicable):

$$\ln C = \alpha_0 + \sum_{i=1}^{m} \beta_i \ln y_i + \sum_{i=1}^{n} \gamma_i \ln p_i + \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \beta_{ij} \ln y_i \ln y_j$$
$$+ \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \gamma_{ij} \ln p_i \ln p_j + \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{n} \delta_{ij} \ln y_i \ln p_j \tag{4.7}$$

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

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 CD (and the linear) is called first order flexible forms since they include only first order terms.

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 prices cannot be directly imposed in the Translog cost function. Instead, these restrictions have to be tested at each data point in 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 difficultly of choosing sufficiently flexible functional forms while maintaining the functional form has proper economic properties.

Finally, it should be noted 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 than single equation ordinary least squares.

4.2.4. Variable Cost function

In many regulatory settings, it is often not reasonable to assume that the firm can adjust the levels of all inputs. For example the size and configuration of railway infrastructure is often fixed. In these circumstances, the variable (short run) cost function is appropriate. It can be derived using duality from a production function under the assumption of cost minimisation, a level of the fixed input(s) and prices of the variable inputs. The resulting function for m outputs, n inputs (o of them fixed) is given as:

$$V\left(y_1,...,y_m,p_1,...,p_{n-o},z_1,...,z_o\right) \qquad (4.8)$$

where y and p are as before and z_i represents the level of the ith fixed input. The measure of variable cost in the function should only include the costs associated with the variable inputs and not those associated with the fixed input(s).

The issues raised in the discussions on output, functional form and inputs above are applicable to the variable cost function in addition to the total cost function but the measurement of RtS and RtD are subtly different. The reason is that there is a need to consider the effect of the fixed factor(s) when computing RtS and RtD. Caves et al (1981 and 1984) give the expressions as:

$$\mathrm{RtS} = \left(1 - \sum_{i=1}^{o} \varepsilon_{z_i}\right) \Big/ \left(\sum_{i=1}^{m-1} \varepsilon_{y_i} + \varepsilon_S\right) \tag{4.9}$$

$$RtD = \left(1 - \sum_{i=1}^{o} \varepsilon_{z_i}\right) \Big/ \sum_{i=1}^{m} \varepsilon_{y_i} \tag{4.10}$$

It is not entirely clear from the subsequent literature when (4.9) and (4.10) should be employed vis-à-vis (4.1) and (4.2). For example Wilson (1997) has two variables to capture the fixed factor. First length of rail road and second average speed rating. He defines RtS as (4.9) except ε_{z_i} only includes the variable cost elasticity with respect to length of rail road. This seems intuitive given the line speed measure is a characteristic of the track which may *a priori* not be expected to change with size of network. However RtD is given as (4.2) rather than (4.10) (Wilson, 1997 footnote 20), which seems odd given the definitions in Caves et al (1981). Caves et al (1984) compute RtD for their variable cost specification as (4.10) where ε_{z_i} is the cost elasticity with respect to capacity (defined as the sum of the annual service flows (measured in constant 1977 dollars) from flight equipment and ground property and equipment – footnote 19).

Clearly, either (4.1)/(4.2) and (4.9)/(4.10) could be valid measures of RtS and RtD in a variable cost function; ultimately the two sets of measures are aimed at answering subtly different questions. (4.1)/(4.2) are measuring how variable cost is impacted on by changing scale and density, while (4.9)/(4.10) are measuring how total cost are impacted on by changing scale and density. What is important in any analysis, is to clearly state to what costs RtS/RtD relate. To some extent, for this thesis, this point is mute, since it is often debatable, in a vertically separated industry such as railways, whether what is being estimated is a total or variable cost. In particular we can consider the cost function in Chapter 5 to be a total cost function, in the sense that cost comprises all costs under direct control of train operating companies. Similarly in Chapter 6, the sum of maintenance and renewals cost is all that is in control of the infrastructure manager, so again it is a total cost function. However from the perspective of the British railway each cost set is only a part of the wider system.

Related to this discussion is whether network length is viewed as a fixed factor in the variable cost function or simply an output to distinguish RtS from RtD (as it is described in the total cost function). Clearly, this decision affects the appropriate decision as to which expression for RtS and RtD to adopt. Lee and Baumel (1987) point out that a model with a fixed factor included alongside a capital price in a cost function violates the properties of both long run and short run cost functions. It is not clear which measure of RtS and RtD to adopt in practice.

4.2.5. Technical and allocative inefficiency

Early studies utilised deterministic frontier methodologies, such as COLS. These have now been superseded by the use of stochastic frontier methodologies. Furthermore, railway studies have often been used to highlight new theoretical developments. One such area has been in illustrating how cost inefficiency can be decomposed into technical and allocative inefficiency. This is either undertaken using a production frontier and factor demand system (Kumbhakar, 1988a, 1988b) or a system involving the cost frontier and factor share equations. Essentially to do this, data is required on both input quantities and prices. The statistics used to estimate the model are involved, since allocative efficiency appears in both the cost (production) frontier and in the cost share (input demand) equations. However, allocative inefficiency has a strictly positive effect on costs in the cost frontier (negative effect in the production frontier) but has a symmetric effect on the factor share equations. In order to identify the two sources of inefficiency, an intergratability condition has to be introduced which relates the two errors.

Even though some conditions have been proposed which reduce the amount of arbitrariness between this relationship (and thus in the decomposition), the modelling still requires other assumptions (such as assuming no noise in the share equations) which mean that this decomposition is far from clear cut. However, it is not the econometric complexity of this approach which is the model's main limitation, but that both input price and input quantity data has to be available for the decomposition. Furthermore, given that, outside the academic environment, frontier techniques are primarily used in railways for benchmarking purposes, it is unclear that the decomposition of cost inefficiency is sufficiently useful to warrant the collection of the data (more generally the issue of the usefulness of this approach in practice is noted in Greene (2008)).

4.3. Cost and efficiency studies for vertically separated railways

The literature in this area is very limited owing to only relatively recent reform of such railways. Importantly, only in the last two decades has both the data and indeed the research need developed such that industry and researchers have undertaken studies on specific vertically separate sections of the railway industry.

4.3.1. Infrastructure studies

4.3.1.1. Marginal Cost studies

While there has been a limited amount of work done on explicit efficiency studies for railway infrastructure operations, there has been a large amount of work to understand railway infrastructure maintenance and renewal cost from the perspective of quantifying the wear and tear by traffic on the infrastructure network. This is to inform marginal cost based pricing.

Beginning with research by Johansson and Nilsson (2002), there have been several studies that have estimated variable cost functions for infrastructure maintenance and for the sum of infrastructure maintenance and renewal costs. Studies have utilized either track section or regional data sometimes over a number of years.

Generally, there are two commonly defined (high level) cost categories relevant in determining infrastructure marginal wear and tear costs; maintenance cost and renewal cost. Maintenance generally contains expenditures on activities associated with day-to-day upkeep of the infrastructure, while renewals contains expenditures on activities on replacement of assets whose life is expired, on a like-for-like basis. Both cost elements contain substantial elements that are variable with traffic and so both should be analysed in econometric modelling. However, most studies in this area have considered maintenance expenditure only as the dependent variables and the limited number of studies that consider the sum of the two categories suffer from poor fit. This is because renewals expenditure tends to be lumpy (discrete in nature) and also depends on past, as well as current, traffic levels.

In terms of the choice of the sum of maintenance and renewal cost versus analysis of maintenance cost only, there is the obvious benefit of considering maintenance and renewal cost as the dependent variable, since this considers the majority of the infrastructure manager's activity that can (non-arbitrarily) be allocated to individual track sections or areas. It also avoids problems associated with different definitions of what exactly comprises maintenance versus renewal which can differ from zone to zone within an infrastructure manager and particularly from one infrastructure manager to another. However, there is less certainty that the cost functions for the maintenance and renewals combined has all of the appropriate variables within it due to the dynamic and lumpy nature of renewals expenditure. This could bias any efficiency estimates derived from the model. As such, a model for maintenance only expenditure is still a useful complement to a model with both cost categories as the dependent variable.

The type of outputs used are intermediate outputs. The primary reason for this choice is that the motivation for the costing exercise was to derive marginal costs with respect to intermediate outputs. This also corresponds to the type of output which the infrastructure manager perceives and so intermediate outputs are appropriate for measuring the efficiency (as opposed to effectiveness, Lan and Lin (2006)) of this decision making unit. The most popular output measure is gross tonne-km as, relative to commonly available alternatives, this seems to be most aligned to the true physical driver of damage and thus explains infrastructure costs. There may be *a priori* reason to believe that there is benefit from distinguishing between gross tonne-km of passenger and freight traffic is more cost reflective. While there has been some success in doing this (see Wheat et al (2009) for a synthesis of latest research) there is concern in the plausibility of the relative magnitudes of marginal costs for the two traffic types. In particular freight traffic seems to do less damage per gross tonne-km than passenger traffic by up to seven times (on average across the network) which seems implausible. Therefore, most studies have preferred to work with a single measure of output.

Much work has gone on into trying to better characterise the nature of the infrastructure. There seems to be three distinct measures of this input. First, measures of what the infrastructure actually is, i.e. its characteristics. Second, what the capability is of the infrastructure, given its composition, in terms of what quality of train service it can support. Third, there are measures that describe the condition of the infrastructure, although these are often interrelated with the second category. Table 4.2 gives examples of measures for each category through review of those used in several European studies. There a limited number of condition variables used in these studies. Potentially, the condition measures adopted by Kennedy and Smith (2004) (number of broken rails and infrastructure manager caused delays) could be useful to incorporate into these cost functions.

Table 4.2 Infrastructure variables used in previous railway infrastructure cost studies

Country	Great Britain	Sweden	Austria	France	Switzerland	Sweden	Finland
Study	Wheat and Smith (2008)	Andersson (2006)	Munduch et al (2002)	Gaudry and Quinet (2003)	Marti and N'schwander (2006)	Johansson and Nilsson (2002)	Johansson and Nilsson (2002)
Infrastructure characteristics	Track length Route length Length of switches	Track section distance Route length Tunnels Bridges Rail weight Rail gradient Rail cant Curvature Lubrication Joints Continuous welded rails Frost protection Switches Switch age Sleeper age Rail age Ballast age	Track section length Length of single-railed tunnels in meters Length of double-railed tunnels in meters Track radius Track gradient Length of the switches Station rails (as percentage of track length)	Number of track Apparatus Whether the track is electrified Route length Number of tracks, Automatic Traffic Control included or not	Track length Track distance (route length) Length of switches Length of Bridges Tunnels Level crossings Track Radius Track gradient Noise / fire protection Number of switches (by type) Shafts Platform edge	Track length Switches Bridges Tunnels	Track length Switches
Capability	Continuously welded rails Maximum line speed Maximum axle load	Rail weight Continuous welded rails Track quality class		Maximum line speed	Maximum line speed	Track quality index Secondary lines	Electrified Average speed
Condition	Rail age	Switch age Sleeper age Rail age Ballast age	Rail age		Rail age Sleepers age		

Source: Work carried out by Phil Wheat, ITS, University of Leeds. Reproduced from Link et al (2008)

Finally Table 4.3 presents RtD and RtS from a selection of studies in the literature that have examined infrastructure maintenance cost. The focus is on infrastructure maintenance cost (as opposed to maintenance and renewal together) given that this is the cost item used in the empirical example in Chapter 6. These studies have all found increasing RtD, with elasticities of cost with respect to traffic density of the order of 0.2-0.4 at the sample mean (Wheat et al (2009)).

Less clear is the evidence on RtS, with some studies finding large increasing RtS while other studies find only small increasing RtS. However, the usefulness of the RtS measure here has to be questioned, especially for studies that utilise observations by track sections (such studies are Johansson and Nilsson (2004), Tervonen and Idstrom (2004), Munduch et al (2002), Gaudry and Quinet (2003), Andersson (2006)). In particular the length of a track section has little to do with the organisation of maintenance and renewal activities, because typically maintenance/renewal teams are responsible for a number of track sections. Thus when analysing track section data, a more appropriate measure of RtS would relate to the overall track-km maintained/renewed by each operational crew which is likely to be greater than the track section-km and invariant across track sections within each operational area. In this review, no instances of any such variables being used within these cost functions has been found.

Table 4.3 Estimates of Returns to Scale and Density from infrastructure maintenance cost studies

Study	Country	Returns to Scale	Returns to Density
Johansson and Nilsson (2004)	Sweden	1.256	5.92
Johansson and Nilsson (2004)	Finland	1.575	5.99
Tervonen and Idstrom (2004)	Finland	1.325	5.74-7.51
Munduch et al (2002)	Austria	1.449-1.621	3.70
Gaudry and Quinet (2003)	France	Not reported	2.70
Andersson (2006)	Sweden	1.38	4.90
Wheat and Smith (2008)	Britain	2.074	4.18
Smith et. al. (2008)	International study	1.11	3.25
NERA (2000)	US	1.15	2.85

Source: Amended from Wheat and Smith (2008)

4.3.1.2.Efficiency studies

Published research in the academic literature on performance of railway infrastructure managers is also limited. As with the train operating company research, all (published) studies relate to the British infrastructure manager, but some do involve international comparisons with other infrastructure managers.

At the 2003 Access Charge Review, ORR commissioned LEK (LEK, 2003) to undertake internal benchmarking of Network Rail. This looked at potential efficiency savings for various expenditures categories based on comparisons across Network Rail's operating areas (seven in total). Some of the work involved statistical analysis but the analysis was far from a top down econometric efficiency study. Efficiency techniques employed were limited to OLS adjusted by either a COLS shift or lower quartile shift.

A more rigorous econometric study aimed at measuring disparity between the performances of individual geographical areas within the British infrastructure manager was undertaken by Kennedy and Smith (2004). This internal benchmarking study adopted both deterministic and stochastic input orientated distance function models and utilised relatively robust data sourced directly from the industry. They considered, in two separate models, maintenance only cost and the sum of maintenance and track renewal cost, as inputs, combined with delay minutes and broken rails as the two other inputs. The levels of these inputs were then assumed to be endogenously determined given a set of outputs (hence the input orientation of the distance function). The outputs were track-km and two traffic density variables –

freight tonne-km and passenger train-km both per track-km.

Their findings suggest that the infrastructure manager Railtrack (now replaced by Network Rail) made substantial improvements in efficiency from privatisation to 2000/01, but then their efficiency deteriorated post this period. There was a key event in October 2000 (the “Hatfield accident”) which for various reasons prompted a revision in the behaviour of the infrastructure manager and ultimately led to it going into administration and being replaced by Network Rail. In particular, they find that most of the earlier gains in efficiency were wiped out by the determination post Hatfield. They conclude that the substantial variation in efficiency between the geographical areas means that there were substantial opportunities to improve performance going forward.

However their stochastic frontier model can be deemed relatively crude from a methodological perspective (simple pooled model). In addition, there were only a limited number of variables included in the function which characterised the infrastructure. Along with the length of track, there were two variables measuring the state of the infrastructure manager’s assets (number of broken rails and caused delay minutes). As discussed below, recent research into cost functions of this industry have suggested several more variables which should be incorporated to better reflect the state of the infrastructure. The failure to incorporate more of these variables may explain why Railtrack was found to be improving performance up to the Hatfield accident in 2000 when it is a relatively well accepted view that Railtrack was not necessarily improving but instead under renewing and maintaining its assets (ORR, 2008). However, Kennedy and Smith did successfully demonstrate that suitable data existed within infrastructure and could be used to find evidence of inefficient practice.

Econometric efficiency analysis of Network Rail formed a very important part of the 2008 Periodic Review determinations. This comprised two pieces of analysis, both benchmarking studies utilising international comparators. The primary piece of analysis was utilised a data set collected by the UIC (International Union of Railways) and previously analysed for the Lasting Infrastructure Cost Benchmarking (LICB) project (UIC, 2008). This was data for a selection of railway infrastructure managers who were members of the UIC. The original LICB project was based on adjusted average cost calculations. Thus, unit cost were computed, but adjustments were made, based on the characteristics of railways (see Smith and Wheat (2010) for a review of the adjustment factors). However, the subsequent work sponsored by ORR undertook econometric efficiency analysis of the dataset (1993-2006) (Smith, Wheat and Nixon (2008) and Smith (2012)). The preferred model utilised a time varying inefficiency model which estimated firm specific paths of adjustment. The model found Network Rail to be 58% efficient. This analysis demonstrated that international comparisons of railway infrastructure managers could be made using econometric techniques. The modelling did suffer however from a limited number of explanatory variables (utilised variables were train-km (passenger and freight), route-km, proportion single track-km and proportion of track electrified – see below for a discussion of other variables that would ideally be within such models).

A supporting piece of econometric analysis was using a bespoke dataset collected by ORR comprising five infrastructure managers. This dataset included observations for regions within each infrastructure manager and, in some cases, data over time. At 2008 Periodic Review, this dataset was relatively new and so analysis was limited to verification of the inefficiency estimates from the main LICB data analysis (which were confirmed). Chapter 6 develops this analysis further. In particular, models are proposed which best exploit the multi-level structure of the data. It is also hoped that over time data can be obtained for many more variables than those currently available for analysis. This is because data is bespoke for this purpose rather than data such as the LICB data which was collected for a subtly different purpose.

4.3.2. Passenger train operations

There has been limited published work on the performance of passenger train operating companies (TOCs) in Britain. The papers have used a variety of methods including non-parametric DEA (Affuso et al., 2003 and 2002, 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, 2012a), production functions (Cowie, 2005) and distance functions (Affuso et al., 2003 and 2002). Clearly, the former methods can only consider cost or technical efficiency and produce no estimates regarding the actual cost structure.

The papers by Cowie consider three inputs: staff, rolling stock and network. This is deficient given that the network input is fixed and difficult to characterise. Cowie (2003) estimates a total cost function and uses access charge per route-km as the price of network. However the vast majority of the access charge is fixed and so this measure over estimates the marginal charge for access. Given the regulatory regime, network access can easily be thought of as a pass-through with respect to franchised TOCs since TOCs are compensated directly with respect to changes in access charges as a condition of the franchise contract. Thus a cost function which considers TOC cost less access charges as the dependent variable seems most appropriate, rather than the cost function estimated by Cowie. This is the approach taken in Chapter 5.

In Cowie (2002b and 2005) route-km are used as the input for network, however as Cowie (2009) acknowledges this is a poor proxy for the true network input. Cowie (2009) replaces this with the cost of access for each TOC as measured by the charges paid to Network Rail. Given the arbitrariness of the allocation of the fixed charge to individual operators, the usefulness of this measure has to be questioned. Also, post 2002, the infrastructure manager was not fully funded by TOC access payments. Instead, the Network Grant (direct payment from Government) was introduced alongside access charges. This further distorts any ‘price’ for network access post 2000 (affecting the Cowie (2009) study).

Affuso et al (2003) do not include any network inputs into their distance function, but this simple exclusion does not seem optimal since the network may affect the transformation function. It is considered that a better way to deal with this is to estimate a variable cost function with infrastructure held fixed.

Finally on data, there are issues with the consistency of data from year to year and TOC to TOC given the data sources used in the studies referred to above. In particular, all the studies which estimated cost functions utilise data in TOC accounts to determine the cost of network access. However, investigation of the accounts as part of the background to the research in Chapter 5, revealed inconsistencies across TOCs and over time as to what elements of access charges are itemised under the heading. Also, the series on train-km derived from National Rail Trends (ORR, 2012) seems to have unexplained step changes over time for some TOCs indicating that this series maybe unreliable. In Chapter 5, access charge data and train-km and vehicle-km data are sourced direct from Network Rail which ensures the quality of the data.

Turning to the results on TOC performance, all the studies report improved performance over the period from privatisation to the period 2000/01. A consistent finding is that this improvement in performance, as measured by a Malmquist total factor productivity measure, has tended to be driven by positive technical change with only a small improvement in average technical efficiency over the period. Thus, while the best performing TOCs seemed to be improving up to 2000/01, there was little evidence that all firms were converging i.e. that franchising was successfully driving out poor performance.

Cowie (2009) and Smith and Wheat (2012a) are the only studies to have considered the period following the Hatfield accident in October 2000. Cowie's study covered the years 1996/97-2003/04, while Smith and Wheat extended the sample to 2005/06. Cowie found that, following Hatfield, there was a deterioration in TFP and this was across all TOCs i.e. was found to be as a result of negative technical change growth rather than a deterioration in technical efficiency of a sub-set of firms (see Figures 2 and 3 in Cowie (2009)). In fact Cowie finds that average technical efficiency improves over the post Hatfield period. This suggests that, even with the distribution of some firms moving to renegotiated contracts, franchising had still begun to proliferate best practice across the industry. This finding has to be moderated however by the finding that, overall, TFP was not found to be substantially different at the end of the period than at the first year following privatisation. Smith and Wheat (2012a) found also found that technical change was, in the early years of their sample, beneficial in terms of lowering costs, however following the Hatfield accident, not only was there a statistically significant upward shift in costs, but the direction of technical change shifted, such that costs began to increase over time. These observations are the same with respect to overall TFP in the Smith and Wheat (2012a) model.

For the parametric studies, it should be possible to derive returns to scale and density results from the models. In Cowie (2005) and Affusso et al (2003) these properties of the models are not discussed in the text. Furthermore, the fact that the data does not appear to be normalised at the sample mean, coupled with the adoption of Translog functional forms, means that the results in the papers cannot be used to derive these results. Of the non-parametric research, Merkert et al (2009) did estimate a variable RtS model and found that British and Swedish TOCs were below minimum efficient scale, while the large German operators were above.

Only Cowie (2002b) and Smith and Wheat (2012a) provide an explicit discussion of the returns to scale properties of the models. Cowie defines returns to scale simply in relation to his single output train-km (there are of course different possible measures of RtS in this context such as returns to network size, train-km and train length). His results seem to suggest decreasing returns to scale at low train-km, but then increasing RtS at higher train-km.

Smith and Wheat (2012a) put forward a model which yields estimates of the extent of both returns to scale and returns to density, where the primary usage output is train-km rather than train-hours. They found constant RtS and increasing RtD. One limitation of the Smith and Wheat (2012a) 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 returns to scale and density.

Overall, the received studies on passenger train operations have concentrated on technical change, cost efficiency and overall TFP trends. The motivation for concentrating on these issues, were, firstly studies focus on Britain and, secondly, at the time the railway in Britain suffered from a substantial cost shock which resulted in several franchises getting into financial difficulty.

However, rail passenger franchising in Britain is now more mature and government policy is towards larger franchises. Further, as detailed in sub-section 1.1, recent European Commission policy changes mean that it is likely that competitive tendering of passenger railway operations will become more wide spread. Thus research can inform the most cost effective way to design tenders and also provide analysis as to whether larger franchises are preferred on cost grounds. However, this is an under researched area. There is a need to account for output heterogeneity in the cost function.

4.4. Summary

In this chapter, the railway specific context of cost research has been outlined. The early material has discussed the difficulty in characterising output of railways. In reality there are many ‘final’ (user) outputs with train services being provided between different origins and destinations, departing at different times and taking different amounts of travel time; all of which implies that some summary measure(s) are needed for feasible empirical work. The concepts of returns to scale versus returns to density are important, as are more general measures of the characteristics of output, but this is still a limited distinction and as such there have been attempts to measure traffic by type and/or include characteristics of output to enrich the description of the underlying technology. Better characterising output is the motivation for the hedonic cost function approach in Chapter 5 for train operations, and this has been shown to be an under researched area.

The studies on railway infrastructure have shown good progress in developing cost functions to describe the cost structure within a railway, for the purpose of marginal wear and tear cost estimation. However, when the exercise is amended to compare the efficiency performance of railways, then it becomes clear that there is a difficulty in gathering sufficient data to conduct meaningful analysis. This motivates the research in Chapter 6 on models to exploit dual-level panel data.

5. Passenger Train Operating Company cost analysis[22]

5.1. Introduction

A key outstanding research issue regarding passenger train operations comprises the optimal size and composition of tender contracts[23]. This is motivated from two perspectives.

Firstly, there is a clear policy motivation given the movement towards competitive tendering in many countries around the world. For example, in Europe, successive reforms have seen infrastructure separated from operations to a greater or lesser degree and, though not required yet by legislation, many countries (in particular Britain, Sweden and Germany) have introduced competitive tendering or franchising of passenger rail services (see for example Alexandersson and Hulten (2007) and Brenck and Peter (2007) for a review of the Swedish and German experience respectively). In recent legislation proposals, the European Commission has set out further steps to encourage competitive tendering in domestic passenger train operations (European Commission, 2013). Competitive tendering in rail has also been used outside Europe. Examples include: Melbourne, Australia; Latin America; and some North American commuter services (Smith and Wheat, 2012a).

Thus optimal specification of tenders is of importance in transportation operations

[22] This chapter is based on Wheat and Smith (2013). The content is essentially the same, with minor extra material relating to more detail on the model.

[23] The chapter, and indeed the thesis as a whole, uses the terms ‘tender’ and ‘franchise’ interchangeably. Franchising is the term used for passenger rail operations contracts in Great Britain, however competitive tendering is the more common international terminology.

and this, coupled with the competition for the market that tendering introduces, will result in services being delivered at least cost. This is in contrast to railway infrastructure, where there is often a monopoly infrastructure manager. Thus more direct RPI-X regulation requiring explicit efficiency analysis maybe required.

In the British context, which is the focus of the empirical analysis in this chapter, a current policy question is whether to remap existing TOCs into fewer, larger TOCs. This has two impacts. First, individual tenders are likely to increase in size and geographical coverage with an associated increase in services per tender. Second, due to the removal of overlap between tenders, utilisation per TOC (train hours per route-km operated) are also likely to increase (increase in density). The former effect is measured by defining returns to scale (RtS) as the cost effect of increasing the size (route-km and stations operated) and usage (primarily train hours) by the same proportion. The latter effect is measured by defining returns to density (RtD) as the cost effect of increasing usage, holding network size constant (measured by route-km and stations operated).

Secondly, in addition to pressing policy issues discussed above, the existing research literature reviewed in 4.3.2 clearly points to more research being needed to understand how train operating costs are affected, not just by the size of operation, but also by the intensity of usage of a given network and the degree of heterogeneity in service types. More generally, the research is motivated by noting that the conventional result in transportation economics is that increasing the density of utilisation of infrastructure will lower average costs (per train-km) (Hensher and Brewer (2000), Button (2010)). This may be expected when the costs associated with infrastructure are considered (e.g. Wheat and Smith (2008), Smith and Wheat (2012a) and Andersson et al (2012)). However, scale and/or density effects are also likely to be apparent in situations where industries are structured on an operation only basis, as in the case where passenger rail services are subject to competitive tendering, for example in Europe.

For train operations, while fixed costs are not as abundant as in railway infrastructure, there is a degree of overall management cost which is invariant to the overall scale of the operation. There are also likely to be improvements in the utilisation of assets (e.g. rolling stock diagramming) for larger and more intensively (densely) used networks / operations which ultimately may reduce unit costs.

It should be noted that these two definitions (RtS and RtD) refer to the effect on train operations costs only and not anything to do with infrastructure costs. RtS and RtD are distinguished since there are two conceptual ways for a train operator to grow[24]. Firstly, a train operator can become geographically larger i.e. operating to and from more points. This is captured by the RtS concept. Secondly, a train operator can grow by running more train hours over a fixed network. This is captured by the RtD concept.

A priori it is reasonable to expect RtD to be larger than RtS given there is likely to be more scope to reduce unit costs by producing more train hours on a fixed network (by, say, better diagramming of existing routes) rather than expanding the network

[24] See Caves et. al. (1981) and Caves et. al. (1984) for use of the terms returns to scale (RtS) and returns to density (RtD) in empirical applications.

(where there may simply be more routes and places to serve).

The key contributions of this chapter are to provide new insights into the structure RtS and RtD in passenger railway operations and also to consider whether heterogeneity in services provided by TOCs affects the estimates of RtS and RtD. This second innovation is important to address the current policy question in Britain since several of the merged franchises now produce services across the three generic service types (intercity, regional and London South Eastern commuting) whilst previously they provide only services from predominantly one service type.

Fundamentally, the research question is, conditional on finding RtS and RtD, can these still be exploited if the services provided by merging franchises are very different? For example, there may be difficulties in sharing rolling stock across different service types. To do this, a hedonic cost function approach is adopted which allows incorporation of measures of the characteristics of outputs. Importantly, this allows incorporation of measures of TOC heterogeneity which are central to evaluate the cost effect of merging heterogeneous TOCs.

The structure of this chapter is as follows. The past literature on RtS and RtD studies has been reviewed in section 4.3.2. and as such this is not repeated. Section 5.2 outlines the methodological approach and in particular the motivation of the hedonic cost function and Section 5.3 identifies the data and demonstrates the improvements in data available for this study relative to previous studies. Section 5.4 discusses the empirical findings relating to overall scale and density returns and the impacts of influence on costs of heterogeneity in outputs. It also presents, for illustration, predicted cost changes for three re-mappings and discusses the reasons for the each cost change. Section 5.5 concludes.

5.2. Methodology

In this section the general economic device and estimation method that is used in the analysis is outlined. The following section populates the model with the choice of variables. A cost function derived from the behavioural assumption of cost minimisation is represented as

$$C_{it} = C(\mathbf{y}_{it}, \mathbf{p}_{it}; \boldsymbol{\beta}) \quad \text{i=1,...,N} \quad \text{t=1,...,T} \tag{5.1}$$

where C_{it} is the cost of firm i in year t, $\mathbf{y}_{\mathrm{it}}$ and $\mathbf{p}_{\mathrm{it}}$ are L and M dimension vectors of outputs and prices of inputs respectively again for firm i in year t. Firms provide a great deal of different train service outputs. For example, TOCs provide train services with different stopping patterns and running speeds. Thus, one approach is to consider this an issue of returns to scope. The amount of each numerous output however cannot be specified due to a number of reasons. Firstly, the data does not exist on outputs at such a level of disaggregation. Secondly, if data did exist then the model would have a large number of parameters such that partial analysis would be imprecise. Thirdly, the Translog cost function cannot accommodate zero levels of outputs very satisfactorily. Instead, the hedonic cost function approach first used by Spady and Friedlaender (1978) is used, which provides a parsimonious method of incorporating output characteristics (termed output quality in their paper) to characterise heterogeneity in outputs. This provides a means of incorporating measures of heterogeneity of output both across and within firms. The former is important for consideration of the cost effect of merging TOCs. As discussed in Jara Diaz (1982), failure to account for output characteristics can result in incorrect policy recommendations in relation to optimal firm size.

Using the notation of Spady and Friedlaender (1978), replace the lth element of $\mathbf{y}_{\mathrm{it}}$, y_{lit}, with ψ_{lit} where

$$\psi_{lit}\left(y_{lit},\mathbf{q_{lit}}\right)=y_{lit}\cdot\phi\left(q_{1lit},...,q_{Blit}\right) \tag{5.2}$$

Where y_{lit} is now the lth "physical output" and q_{blit} is the bth quality characteristic of the lth physical output. ψ_{lit} is assumed homogenous of degree one in the physical output. This implies that a doubling of y_{lit} results in a doubling of ψ_{lit}; this is required for identification of the function within the wider cost function and sets y_{lit} to be the numeriere of ψ_{lit}. $\phi_l\ \ \forall l$ is considered to be Cobb Douglas as in Bitzan and Wilson (2007) (as opposed to Translog as in Spady and Friedlaender's formulation) given the large number of quality variables in this formulation.

Spady and Friedlaender (1978) discuss the implicit restrictions associated with adopting the hedonic formulation. They term the function "quality separable" since the impact of the quality variables on the associated primary output is independent of prices (and also of the level of other primary outputs). Ultimately this restriction is the price of adopting the hedonic function, but it makes the model far more manageable in terms of parameters to be estimated (34 parameters for the hedonic formulation, but the unrestricted Translog would require estimation of circa 140 parameters; there are only 243 observations). Given the Cobb Douglas form for ϕ_l in (5.2), an eloquent way to describe the implication of the “quality separable” restriction is that the elasticity of cost with respect to the quality variable is proportional to the elasticity of cost with respect to the primary output.

A Translog cost function (in $\boldsymbol{\psi}_{\mathbf{it}}$, $\mathbf{p}_{\text{it}}$ and, given that the model utilises panel data, a cost non-neutral technology trend, t) is used

$$\ln(C_{it}) = \begin{cases} \alpha + \sum_{l=1}^{L} \beta_l \ln(\psi_{lit}) + \sum_{m=1}^{M} \delta_m \ln(P_{mit}) + \gamma_T t + \frac{1}{2}\sum_{l=1}^{L}\sum_{b=1}^{L} \beta_{lb} (\ln(\psi_{lit}))(\ln(\psi_{bit})) \\ + \frac{1}{2}\sum_{m=1}^{M}\sum_{c=1}^{M} \delta_{mc} (\ln(P_{mit}))(\ln(P_{cit})) + \sum_{l=1}^{L}\sum_{m=1}^{M} \kappa_{lm} (\ln(\psi_{lit}))(\ln(P_{mit})) \\ + \sum_{l=1}^{L} \lambda_{Tl} t \ln(\psi_{lit}) + \sum_{m=1}^{M} \varphi_{Tm} t \ln(P_{mit}) + \gamma_{TT} t^2 \end{cases} \qquad (5.3)$$

Shephard’s Lemma is applied to (5.3) to yield the cost share equations:

$$\frac{\partial \ln(C_{it})}{\partial \ln(P_{mit})} = S_m = \delta_m + 2\delta_{mm} \ln(P_{mit}) + \sum_{l=1}^{L} \kappa_{lm} \ln(\psi_{lit}) + \varphi_{Tm} t \quad \text{m=1,...,M} \qquad (5.4)$$

The model parameters are estimated as a system of the cost function and the factor shares, to aid both the precision of estimates and also to ensure that the estimated cost shares are as close as possible to the true cost shares (which by (5.4) is a requirement of economic theory). In addition to the cost shares, economic theory associated with the existence of a dual cost function provides a set of useful restrictions to aid estimation. Firstly, symmetry of input demand with respect to price requires $\delta_{mc} = \delta_{cm}$ and also there is symmetry in the cross derivatives of outputs, $\beta_{lb} = \beta_{bl}$. Secondly, the cost function must be linear homogenous of degree 1 in prices. This requires:

$$\left\{\begin{array}{ll} \sum_{m=1}^{M}\delta_{m} = 1 & \\ \sum_{c=1}^{M}\delta_{mc} = 0 & m = 1,...,M \\ \sum_{m=1}^{M}\kappa_{lm} = 0 & l = 1,...,L \\ \sum_{m=1}^{M}\varphi_{Tm} = 0 & \end{array}\right\} \quad (5.5)$$

A convenient way of imposing (5.5) on (5.3) and (5.4) is to divide input prices and cost by one of the input prices (see Heathfield and Wibe (1987) for this derivation).

Given there are parameters implicit in ψ_{lit}, estimation is undertaken using non-linear Seeming Unrelated Regression. To avoid the errors in the cost shares summing to zero for each observation, one of the cost shares has to be dropped. The cost share for the Mth input is removed from estimation (i.e. the input whose price is used to divide cost and all other prices by).

Cost efficiency is not modelled explicitly. Partly this is a pragmatic approach; including cost inefficiency in the cost equation potentially requires an intricate econometric framework to link the allocative inefficiency component in the cost shares to the cost inefficiency term in the cost equation (Kumbhakar and Lovell, 2000). Also such modelling is not without assumptions in its own right which maybe tenuous. For example, the maximum likelihood approach by Kumbhakar (1997) allows for allocative inefficiency to enter the model appropriately but with the expense of assuming away any noise error in the cost share equations. However, perhaps the most important reason is that the measurement of firm technical or cost inefficiency is a separate question, and here the focus is on scale and density efficiency. Furthermore, estimation of the model produces consistent estimates for the model parameters except the constant term in the cost equation even in the presence of inefficiency (due to the assumption commonly made in the efficiency literature that inefficiency is uncorrelated with regressors). Since the constant term is not of direct interest no correction is made to this.[25]

Finally on efficiency, as noted in Spady and Friedlaender (1978, p. 162) and applicable to both the hedonic cost function or other functions where output characteristic variables are used, there is the potential problem of endogeniety of the characteristic variables. Some variables may be under the control of the firm and thus are endogenous. Those variables that are under the firm's control depend on the regulatory conditions affecting the market. There are statistical techniques that can produce consistent estimates of parameters in models with endogenous regressors (instrumental variables techniques). Spady and Friedlaender caution however that any estimated function would still be an ambiguous description of the technology of the firm because such a function would be determined by both the supply and demand characteristics of the market. This same point applies when input markets are not perfectly competitive and so input prices can be influenced by firms. The

[25] This could however be done (and cost inefficiency measured) by applying a two stage estimation approach. MLE would be used to estimate the noise and inefficiency variance components using the residuals from the cost equation estimated using the method in this chapter (Kumbhakar and Lovell, 2000).

implication is that inefficiency may be confounded within the parameter estimates rather than solely in the residuals. It should be noted that it is the inclusion of output characteristics potentially under the firm’s control that creates the problem, not the hedonic formulation per se. Thus this is a limitation of any efficiency and indeed cost study. Given the focus away from measuring efficiency in this chapter and given the limited opportunities for TOCs to influence the demand side of the market, then the analysis proceeds as if all regressors (including output characteristics) are exogenous.

Therefore, after imposing symmetry and linear homogeneity of degree one in input prices on (5.3) and (5.4), the system of M equations to be estimated is:

$$\left\{ \begin{array}{l} \ln\left(\frac{C_{it}}{P_{Mit}}\right) = \left\{ \begin{array}{l} \alpha + \sum_{l=1}^{L}\beta_l \ln(\psi_{lit}) + \sum_{m=1}^{M-1}\delta_m \ln\left(\frac{P_{mit}}{P_{Mit}}\right) + \gamma_T t \\ + \frac{1}{2}\sum_{l=1}^{L}\sum_{b=1}^{L}\beta_{lb}(\ln(\psi_{lit}))(\ln(\psi_{bit})) \\ + \frac{1}{2}\sum_{m=1}^{M-1}\sum_{c=1}^{M-1}\delta_{mc}\left(\ln\left(\frac{P_{mit}}{P_{Mit}}\right)\right)\left(\ln\left(\frac{P_{cit}}{P_{Mit}}\right)\right) \\ + \sum_{l=1}^{L}\sum_{m=1}^{M-1}\kappa_{lm}(\ln(\psi_{lit}))\left(\ln\left(\frac{P_{mit}}{P_{Mit}}\right)\right) \\ + \sum_{l=1}^{L}\lambda_{Tl}\, t \ln(\psi_{lit}) + \sum_{m=1}^{M-1}\varphi_{Tm}\, t\ln\left(\frac{P_{mit}}{P_{Mit}}\right) + \gamma_{TT}t^2 \end{array}\right. \\ S_m = \delta_m + 2.\delta_{mm}\ln\left(\frac{P_{mit}}{P_{Mit}}\right) + \sum_{l=1}^{L}\kappa_{lm}\ln(\psi_{lit}) + \varphi_{Tm}t \quad \mathrm{m} = 1,\ldots,(\mathrm{M}-1) \end{array} \right\} \qquad (5.6)$$

In addition to the symmetry and linear homogeneity in prices, the cost function has to be concave in input prices. This cannot easily be imposed on the Translog function form since the restrictions are a function of the data. Instead, the matrix of second derivatives of input prices is computed at each data point to verify if it is negative definite; a necessary and sufficient condition for concavity in prices. The matrix for the Translog function is given in Diewert and Wales (1987) as:

$$\nabla^2_{PP} C(\boldsymbol{\psi}, \mathbf{P}, t) = \begin{bmatrix} \delta_{11} + S_1(S_1 - 1) & \delta_{12} + S_1 S_2 & \cdots & \delta_{1M} + S_1 S_M \\ \delta_{12} + S_1 S_2 & \delta_{22} + S_2(S_2 - 1) & & \vdots \\ \vdots & & \ddots & \vdots \\ \delta_{1M} + S_M S_1 & \cdots & \cdots & \delta_{MM} + S_M(S_M - 1) \end{bmatrix} \tag{5.7}$$

A further condition that is not imposed, but checked post estimation, is that the factor demand own-price elasticities are negative for all inputs. The Allen-Uzawa own-price elasticities and partial elasticities of substitution are given as:

$$\sigma_{mm} = \left(\delta_{mm} + S_m(S_m - 1)\right)/S_m^2 \tag{5.8}$$

and

$$\sigma_{mc} = \left(\delta_{mc} + S_c S_m\right)/S_c S_m \tag{5.9}$$

respectively. If σ_{mc}<0, the two inputs are complements, if σ_{mc}>0 then they are substitutes.

5.3. Data

A panel data set of 28 TOCs over 11 years (2000 to 2010[26]) is used. The panel is unbalanced with a total of 243 observations. The unbalanced nature of the panel

[26] Quoted years are for year end to 31st March e.g. 2000 is April 1999 to March 2000.

reflects the re-franchising and, importantly, re-mapping of franchises over time.

TOC cost is defined as total reported cost less access charge payments to Network Rail (the railway infrastructure manager). This definition follows from Smith and Wheat (2012a). Netting off access charge payments is important as they are (indirectly) merely transfer payments from Government to the infrastructure manager and are not reflective of the cost of network access for a given TOC (at least in a given year). Importantly, TOCs are compensated for changes in the access charge payments over time by the construction of the franchise contracts[27]. It is therefore important to note that netting off access charge transfer payments to Network Rail does not mean that a variable cost function is estimated. It should be noted that it is considered that what is estimated is a total cost function since this cost represents the total cost under the control of the franchisee (for the duration of the franchise).

The cost data is sourced from the TOC's publicly posted accounts, while access charge payments are sourced direct from Network Rail. These are the best sources of these data given that the TOC accounts do not report access charges in a consistent manner across all TOCs.[28]

Regarding the explanatory variables, Table 5.1 summarises the data. There are three

[27] It should also be noted that since 2001/02 Network Rail received some of its funding directly from central government via the Network Grant. As such the sum of access charges over all TOCs does not reflect the full cost of infrastructure provision for years beyond 2002. This is another reason that access charges do not reflect the opportunity cost of network access.

[28] In particular it is obvious that some TOCs are itemising in their accounts only variable access charges rather than the sum of variable and (generally the much larger) fixed charge.

primary outputs; route-km, train-hours and number of stations operated. It is considered that TOCs produce train services (train hours) and operating stations. In addition, route-km is included to distinguish between geographical size and intensity of operations. Thus it is analogous to the use of route-km in integrated railway studies to distinguish between scale and density effects (Caves et al, 1985). Conceivably, route-km could have been included as a characteristic of the primary train hours output. However, adopting this approach would have imposed, *a priori*, a more restrictive relation between scale and density effects; the hedonic function adopted imposes proportionality between the cost elasticity with respect to the primary output and the cost elasticity with respect to the quality variable. Given the focus of this study towards optimal size/utilisation of TOCs, it was deemed that the more flexible approach should be adopted.

With respect to other studies, there are a number of improvements to note with regard to the specification of outputs in this study. Firstly, both stations operated and train operations measures are included. Station operation is an important activity for some TOCs but less so for others and as such should not be ignored. Only Smith and Wheat (2012a) considered stations within analysis. Secondly, train hours data are used in this study. This, along with distance measures (incorporated via average speed measures) and train length measures are the key drivers of costs since these measures include both time based and distance based cost drivers. To the author's knowledge, no previous railway cost study, either of vertically integrated or separated railways, has taken account of train hours, length and speed in the model.

A key element of this study is to consider the cost implications of merging TOCs which produce outputs with different characteristics. Thus it is important to account for output characteristics which measure the extent of heterogeneity within a given TOC as well as across TOCs. In addition to including the average characteristics of TOC output (train length, speed and passenger load factor), therefore, two further sets of measures to account for diversity in TOC service provision are included. The first is the proportion of train-km that correspond to each of three service groups (intercity, London South Eastern (LSE) commuting and the remainder regional). q_{42} and q_{52} pick up systematic cost differences, over and above that captured by the other output characteristics, from TOCs providing intercity and LSE commuting services respectively (the proportion for regional services is dropped to prevent perfect collinearity). For example, it can be expected that intercity TOCs will, all other things being equal, be more expensive due to such factors such as the need to provide higher quality rolling stock and better on train services. As well as including these terms, interactions between the service group proportions are included. The majority of TOCs provide only one service group, thus the interaction variables are only non-zero for a select set of TOCs, the majority of which were formed from re-mappings of TOCs that provided a single service type but in the same geographical area, and have subsequently been merged into one. Thus the coefficients on these interaction variables would provide an indication of any cost increasing (or decreasing) impact of TOCs providing heterogeneous service mixes, over and above any change in other service level characteristics.

The second is the number of generic rolling stock types operated by a TOC is to be included in the model. These are taken from the rolling stock classifications within the Department for Transport’s Network Modelling Framework model (data supplied direct from DfT’s Rail Analysis Division). Essentially they classify rolling stock into speed bands and traction source (electric or diesel) and whether they are multiple units or loco-hauled. The more rolling stock types that are operated, the more likely there is heterogeneity in service provided within a TOC.

It should be noted that when it comes to evaluating franchise re-mappings, it will not just be the rolling stock type and franchise service type proportion heterogeneity that affect the cost change. The other average heterogeneity characteristic variables will be different. Thus it is difficult to assess the impact of changes in heterogeneity by looking at the signs on the service type and rolling stock type variables in isolation. This is returned to in the results section at 5.4.

Two input prices are defined; payroll staff costs and non-payroll costs. Payroll staff costs include all labour costs from staff which are directly employed by the TOC. A natural price measure is staff cost divided by staff numbers. The divisor for non-payroll costs is less clear. Firstly, once access charge payments are removed, the publically available accounts do not allow for costs to be consistently broken up any further than staff and non-payroll costs. Non-payroll costs includes rolling stock capital lease, rolling stock non-capital lease and other outsourced maintenance costs and energy costs. The only divisor that is available is number of rolling stock units and this is adopted in the price calculation. This is a limitation of the data but it is believed that this is the best solution. *Ex post* estimation, concavity in input prices is checked and this is fulfilled at all data points which gives some reassurance that the input prices data are not having perverse effects. Perhaps the most important implication of the definition of input prices is that it would be expected that there is a reasonable degree of substitutability between the two inputs at the margin since functions such as train maintenance can be outsourced and thus staff activity can be taken off the payroll.

Two TOCs do not operate any stations (Cross Country and Gatwick Express). This presents a difficulty for the model given that the logarithm of zero is not defined. There are several small changes to the function which can be made. One option is to input the variable in levels or via a Box-Cox transformation instead of the logarithm. Another is to introduce a small positive shift in all of the data (to avoid zeros). However, inspection of the data reveals that most TOCs operate many stations while two operate no stations. This would suggest that it is more appropriate to model this extreme of the sample differently to the rest. Clearly, it is unlikely to be feasible to produce an entirely separate model for those TOCs that operate no stations due to the small number of observations for this group. Instead, those TOCs with no stations are modelled as a cost function comprising only two outputs and the two input prices. Furthermore, the coefficients with respect to the route-km (and the interactions with other variables) are allowed to be different for those TOCs that do operate stations. As a sensitivity exercise, an attempt was made to estimate the model with different coefficients associated with the train hours variable but this model failed to converge.

Table 5.1 variables used

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 (ORR, 2012 and past volumes)
$\psi_2 = y_2 q_{12}^{\phi_{12}} q_{22}^{\phi_{22}} q_{32}^{\phi_{32}} e^{\phi_{42}q_{42}} e^{\phi_{52}q_{52}} e^{\phi_{62}q_{62}} e^{\phi_{72}q_{72}} e^{\phi_{82}q_{82}} e^{\phi_{92}q_{92}}$			
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)

Given the variable definitions in Table 5.1 the system can be estimated based on

(5.6) as:

$$\left\{\begin{array}{l} \ln\left(\frac{C_{it}}{P_{2it}}\right)=\left\{\begin{array}{l} \alpha+\sum_{l=1}^{3}\beta_l\ln(\psi_{lit})+\delta_1\ln\left(\frac{P_{1it}}{P_{2it}}\right)+\gamma_T t+\frac{1}{2}\sum_{l=1}^{3}\sum_{b=1}^{3}\beta_{lb}(\ln(\psi_{lit}))(\ln(\psi_{bit})) \\ +\delta_{11}\left(\ln\left(\frac{P_{1it}}{P_{2it}}\right)\right)^2+\sum_{l=1}^{3}\kappa_{l1}(\ln(\psi_{lit}))\left(\ln\left(\frac{P_{1it}}{P_{2it}}\right)\right) \\ +\sum_{l=1}^{3}\lambda_{Tl}t\ln(\psi_{lit})+\varphi_{T1}t\ln\left(\frac{P_{1it}}{P_{2it}}\right)+\gamma_{TT}t^2 \end{array}\right. \\ \qquad S_1=\delta_1+2.\delta_{11}\ln\left(\frac{P_{1it}}{P_{2it}}\right)+\sum_{l=1}^{3}\kappa_{lm}\ln(\psi_{lit})+\varphi_{Tm}t \end{array}\right\} \quad (5.10)$$

for those TOCs that operate stations and:

$$\left\{\begin{array}{l} \ln\left(\frac{C_{it}}{P_{2it}}\right)=\left\{\begin{array}{l} \alpha+\beta'_1\ln(\psi_{1it})+\beta_2\ln(\psi_{2it})+\delta_1\ln\left(\frac{P_{1it}}{P_{2it}}\right)+\gamma_T t \\ +\frac{1}{2}\sum_{l=1}^{2}\sum_{b=1}^{2}\beta'_{lb}(\ln(\psi_{lit}))(\ln(\psi_{bit}))+\delta_{11}\left(\ln\left(\frac{P_{1it}}{P_{2it}}\right)\right)^2 \\ +\kappa'_{11}(\ln(\psi_{1it}))\left(\ln\left(\frac{P_{1it}}{P_{2it}}\right)\right)+\kappa_{21}(\ln(\psi_{2it}))\left(\ln\left(\frac{P_{1it}}{P_{2it}}\right)\right) \\ +\lambda'_{T1}t\ln(\psi_{1it})+\lambda_{T2}t\ln(\psi_{2it})+\varphi_{T1}t\ln\left(\frac{P_{1it}}{P_{2it}}\right)+\gamma_{TT}t^2 \end{array}\right. \\ \qquad S_1=\delta_1+2.\delta_{11}\ln\left(\frac{P_{1it}}{P_{2it}}\right)+\kappa'_{11}(\ln(\psi_{1it}))+\kappa_{21}(\ln(\psi_{2it}))+\varphi_{Tm}t \end{array}\right\} \quad (5.11)$$

for those TOCs that do not operate stations. Parameters followed by ’ in (5.11) indicate those parameters which are allowed to vary relative to those in (5.10).

5.4. Results

This section is divided into four sub-sections. In the first, the suitability of the estimated model in terms of being consistent with economic theory and whether the model is suitably parsimonious is considered. The second focuses on the scale and density properties of the model. In the third, sub-section the impact of heterogeneity of output on costs and scale and density is considered. The final section shows how these three factors (scale, density and heterogeneity) affect the expected cost changes for two specific mergers in this dataset and also for one hypothetical, but currently highly topical, potential merger.

5.4.1. Consistency with economic theory

It is first appropriate to consider the suitability of the estimated model in terms of their consistency with economic theory. The parameter estimates are shown in Table 5.2. The R^2 measure of fit for the cost function equation and the cost share equation are 0.928 and 0.489 respectively. The higher R^2 for the cost function primarily reflects the fact that the dependent variable is in logarithms while it is in levels in the cost share equation. The fitted cost shares are all between zero and one and the Hessian has been evaluated at each data point and found to be negative definite for all observations; thus the function is concave in input prices over the relevant range.

Table 5.2 Parameter Estimates

Parameter	Estimate	P-val		Parameter	Estimate	P-val	
Main parameters				*Hedonic output (ψ_2) parameters*			
α	7.729	0.001	***	ϕ_1	0.701	0.000	***
β_1	-1.831	0.000	***	ϕ_2	0.856	0.000	***
β_2	-0.464	0.256		ϕ_3	0.059	0.609	
β_3	0.592	0.076	*	ϕ_4	0.425	0.031	**
δ_1	1.048	0.000	***	ϕ_5	0.309	0.005	***
γ_T	0.039	0.420		ϕ_6	-1.520	0.002	***
β_{11}	0.100	0.003	***	ϕ_7	-0.157	0.763	
β_{22}	0.048	0.048	**	ϕ_8	-0.463	0.631	
β_{33}	0.109	0.000	***	ϕ_9	0.021	0.139	
β_{12}	0.078	0.045	**				
β_{13}	-0.189	0.000	***	*No-stations model free parameters*			
β_{23}	0.010	0.819					
δ_{11}	0.080	0.000	***	β_1'	-1.170	0.011	**
κ_{11}	-0.058	0.000	***	β_{11}'	0.035	0.323	
κ_{12}	0.067	0.000	***	β_{13}''	0.050	0.335	
κ_{13}	0.004	0.545		κ_{11}'	-0.046	0.000	***
λ_{T1}	0.002	0.663		λ_{T1}'	0.005	0.278	
λ_{T2}	-0.008	0.119					
λ_{T3}	0.002	0.545		R^2			
φ_{T1}	-0.006	0.000	***	Cost Function		0.928	
γ_{TT}	-0.001	0.539		Share Equation		0.489	

***, **, * Statistically significant from zero at the 1%, 5% and 10% levels respectively

The Allen-Uzawa own-price elasticities and partial elasticities of substitution (given in (5.8) and (5.9)) have also been computed. The mean estimated own-price elasticities are -0.297 and -1.345 for other expenditures and staff price respectively, which are both negative and so in line with expectations. The own-price elasticities are negative for all observations. The cross elasticity is 0.632 which is positive and thus indicates the two inputs are substitutes and this is the case when the elasticity is evaluated for each observation. This may reflect the degree to which some labour activity can be taken in-house (therefore appear on payroll costs) versus by out-sourcing (appearing under non-payroll costs). This is likely to be the case for non-capital rolling stock expenditure activities where maintenance can be performed in-house or by a third party or ROSCO. More generally, at the margin it is reasonable that there are some substitution possibilities between staff and rolling stock (capital) (choosing rolling stock that requires less staffing costs).

Thus it appears that the estimated function does represent a cost function consistent with economic theory at least in terms of sensible cost shares, substitution elasticities and concavity in input prices (other restrictions such as homogeneity of degree one in input prices and symmetry are guaranteed by imposition). As such, there is confidence that the estimated cost function can be used to infer the properties of the underlying technology.

Finally, before reporting on the scale and density properties of the model, several restrictions on the Translog can be tested both with a view of obtaining a more parsimonious function and to test economic hypotheses about the underlying technology. Of interest are:

- Homotheticity – the cost function is homothetic if it can be written as the product of a function in outputs and a function in input prices (and, since the study uses panel data, time) i.e. $C(\boldsymbol{\psi}, \mathbf{P}, t) = f(\boldsymbol{\psi}).g(\mathbf{P}).h(t)$. Thus it requires that $\kappa_{1l} = 0$, $\lambda_{Tl} = 0$ l=1,2,3, $\kappa'_{12} = 0$, $\lambda'_{T1} = 0$ and $\varphi_{T1} = 0$ - 9 restrictions.

- Homogeneity – This refers to homogeneity in outputs. It is a special case of homotheticity in the sense that it implies unchanging returns to scale i.e. constant output elasticity i.e. $f(\psi) = \psi_1^{\beta_1}\psi_2^{\beta_2}\psi_3^{\beta_3}$. It requires $\kappa_{1l} = 0$, $\lambda_{Tl} = 0$, $\beta_{lb} = 0$ l=1,2,3 b=1,2,3, $\kappa'_{12} = 0$, $\lambda'_{T1} = 0$, $\varphi_{T1} = 0$ and $\beta'_{l2} = 0$ l=1,2 - 17 restrictions.
- Unitary Elasticity of Substitution – This implies that $\sigma_{12} = 1$ in (5.8). This requires $\delta_{12} = 0$ which given the restrictions imposed by linear homogeneity of degree one in input prices implies $\delta_{11} = 0$ - 1 restriction
- Homogeneity and Unitary Elasticity of Substitution – This is the Cobb-Douglas restrictions (if additionally Homogeneity in the time trend is imposed) – 19 restrictions (additional $\lambda_{TT} = 0$)
- No hedonic characteristics – This requires $\phi_i = 0$ i=1,..,9. If this is supported the model reduces to one which is linear in parameters – 9 restrictions.

All hypotheses are rejected as reported in Table 5.3. This shows that the flexible specification is required to describe the underlying technology. Thus the model in Table 5.2 is retained as the preferred model. Now the findings on returns to scale and density are considered.

Table 5.3 Results of specification tests

	Homotheticity	Homogeneity	Unitary Elasticity	Cobb-Douglas	Hedonic
Number of Restrictions	9	17	1	19	9
Test statistic - Chi-sq	142.24	371.11	360.63	660.79	114.48
p - val	0.0000	0.0000	0.0000	0.0000	0.0000

5.4.2. Returns to Scale and Density

As described in the introduction section returns to scale (RtS) and returns to of density (RtD) are defined specifically for train operations. RtS measures how costs change when a TOC grows in terms of geographical size. RtD measures how costs change when a TOC grows by running more services (measured by train-hours) on a fixed network. Applying these definitions to the model in (5.10) the expressions are:

$$RtS_{it} = \frac{1}{\left(\frac{\partial \ln C_{it}}{\partial \ln \psi_{1it}} + \frac{\partial \ln C_{it}}{\partial \ln \psi_{2it}} + \frac{\partial \ln C_{it}}{\partial \ln \psi_{3it}} \right)} \tag{5.12}$$

and

$$RtD_{it} = \frac{1}{\left(\frac{\partial \ln C_{it}}{\partial \ln \psi_{2it}} \right)} \tag{5.13}$$

The definition of RtD and RtS adopted is in relation to the hedonic output. Given the normalisation of train hours within the hedonic function, the findings on RtD and RtS with respect to ψ_2 can interchangeably be described in terms of variation in train hours (holding stations operated and network length and other things, including output characteristics, equal).

The rejection of the null hypothesis of homogeneity in outputs indicates that RtS and RtD will be non-constant and vary with the levels of the hedonic outputs, time and the level of prices. Figure 5.1 plots RtS and RtD for all observations against train hours. 45% and 94% of observations exhibit increasing RtS and RtD respectively.

The definitions of RtS and RtD are that there are increasing returns if the estimate is greater than unity, constant returns if the estimate is unity and decreasing returns if the estimate is less than unity. RtS and RtD evaluated at the sample mean of the data are 0.891 and 1.209 respectively. Constant RtS is rejected in favour of decreasing RtS at the 1% level (p-val=0.0055) and RtD is rejected in favour of increasing RtD at at the 1% level (p-val=0.0034). Thus from these statistics it does seem that British TOCs exhibit increasing RtD but decreasing RtS.

This is an economically plausible finding. TOCs are likely to be able to lower unit costs by running more services on a fixed network. For example, by better diagramming of rolling stock and staff they can reduce wasted time. Thus it is likely rolling stock can be used more intensively in a given time period which ultimately spreads any fixed lease charges over more units of output (train hours). Ultimately inputs into the production process suffer from indivisibilities and these can be more productively combined at higher usage levels[29].

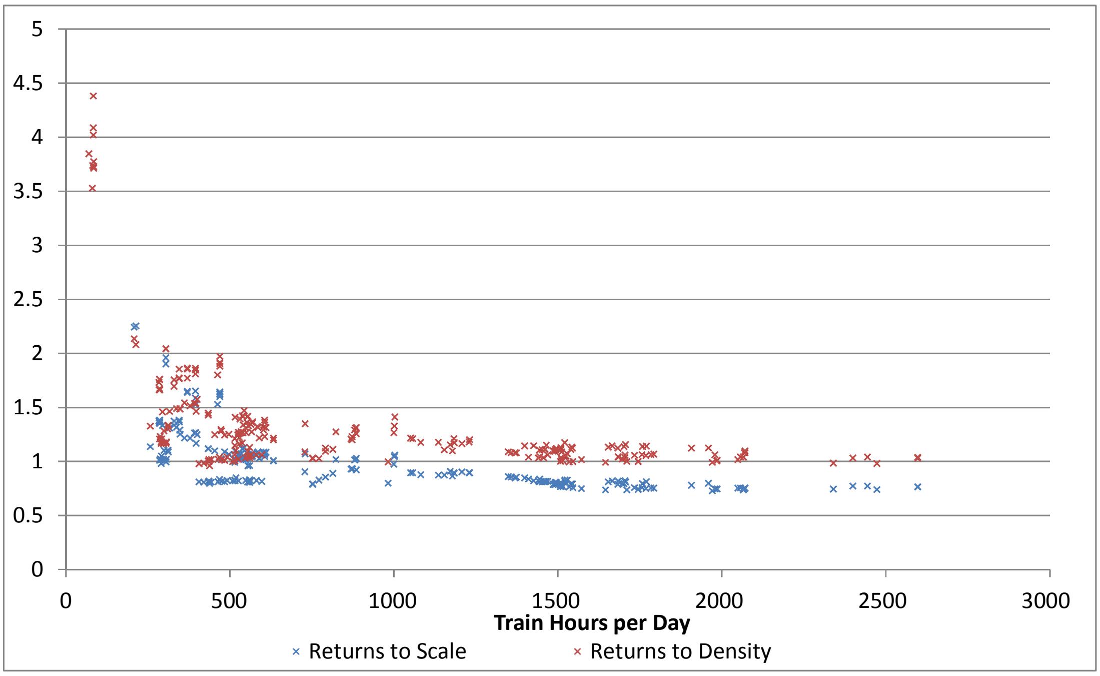
However, TOCs may struggle to make unit cost savings or even prevent unit costs increasing when the size of the network served increases, holding utilisation (train hours per route-km) constant. This can arise since (to some extent) indivisibilities in inputs are route specific rather than network specific. For example, it can be envisaged that the utilisation benefits of running more trains between point A and B will be greater than utilisation benefits from running a set of services from A to B and then adding a new service from two unrelated points C and D. The latter

[29] Importantly indivisibility of inputs is a RtD issue rather than a cost efficiency issue since the explanation relates to the characteristic of the production technology rather than the extent to which minimum cost conditional on a level of output is achieved.

scenario (for the same total train hours) is likely to require more rolling stock units and more staff hours than the former since there are two rather than one operational routes. To provide a less abstract (but extreme) example, the addition of a branch line to an existing network would not be expected to exploit higher utilisation of rolling stock since it is (almost) an independent operation to the rest of the network.

RtS is actually be found to be decreasing for some observations i.e. unit costs increasing as scale increases. To explain, an appeal is made to the theory of the firm which considers that there is an optimal scale of a firm and that at some output level it gets very difficult to coordinate inputs, and thus unit costs start to rise (the firm is larger than the minimum efficient scale point). Note that the same pattern of variation in RtD is found, that is there exists a minimum efficient density level, but no TOC (yet) operates at a high enough density to attain it.

Figure 5.1 Estimated Returns to Scale and Density against Train Hours for the sample



The RtS and RtD findings are now broken down by TOC types – intercity, commuting (into London - LSE), regional and mixed TOCs

Figure 5.2 provides a plot which considers RtD against train density for different TOC types holding all other characteristics[30] at the TOC type sample mean. The density range of only the middle 80% of the distribution observed for each TOC type is used, as this avoids showing RtD estimates from the model which are clearly out of sample and not realistic e.g. intercity TOC services always operate at low densities due to the long distance nature of the services and so are only plotted over this range.

Overall, holding characteristics at the sample mean and over the middle 80% of the distribution, Figure 5.2 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 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 in response to future increases in passenger demand. This is important given the strong upward trend in passenger demand since rail privatisation in Britain and also noting that the trend seems to be continuing, even during the recession at the end of the sample period

[30] In this sub-section ‘characteristics’ refers to all other variables in the cost function and not just the output characteristic variables in ψ_{2it}.

(ORR, 2012). It is also relevant for recent policy in Britain following Sir Roy McNulty's Rail Value for Money study, since unit cost reductions of around 25% are targeted for the TOCs, and according to the results this chapter (increasing RtD), part of this unit cost reduction will occur naturally as train hours increase on a fixed network (though other savings will also be needed and the ability to grow volumes will be constrained to some extent by capacity and also by demand). In the wider EU context, the European Commission has aggressive targets for rail passenger usage and market share which will increase passenger train density and therefore should reduce unit costs (assuming that train-km can be expanded without the need for investment in infrastructure). The results show that the LSE service type has substantial scope for unit cost savings from increasing usage and this also holds for many regional TOCs given the large spread of usage levels across this group. However there is less scope for unit cost savings (and possibly a risk of decreasing RtD from large increases in usage) for intercity TOCs and regional TOCs at the high usage end of the spectrum.

Figure 5.2 Returns to density for different TOC types holding other variables constant

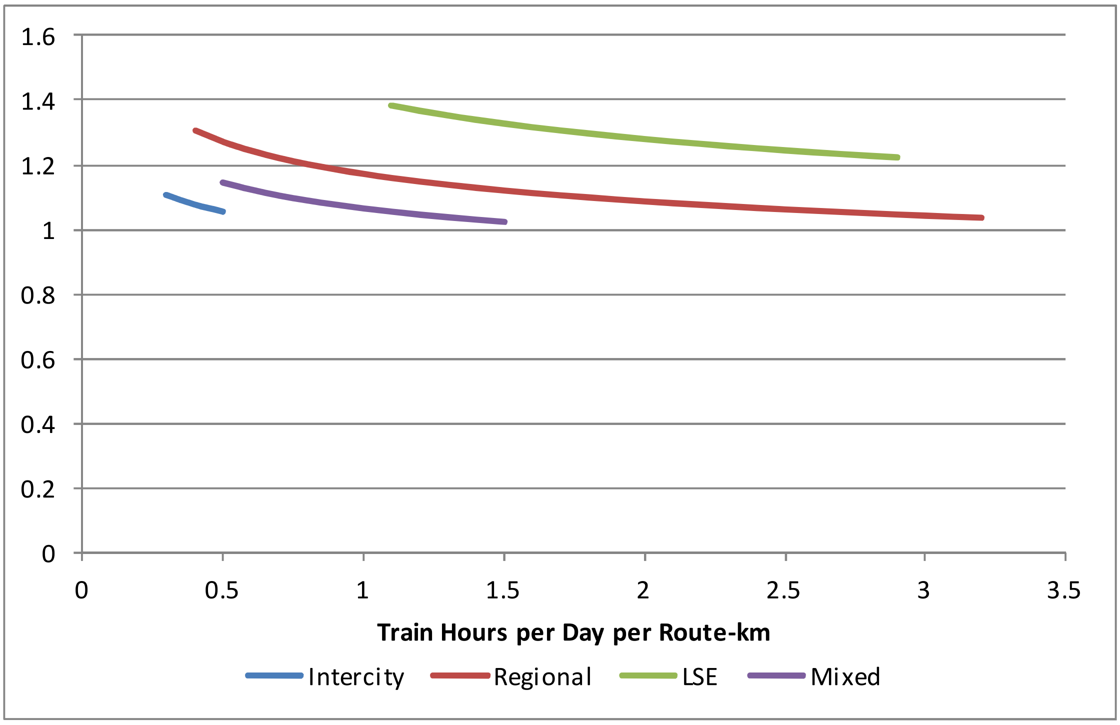


Figure 5.3 Returns to scale for different TOC types holding other variables constant

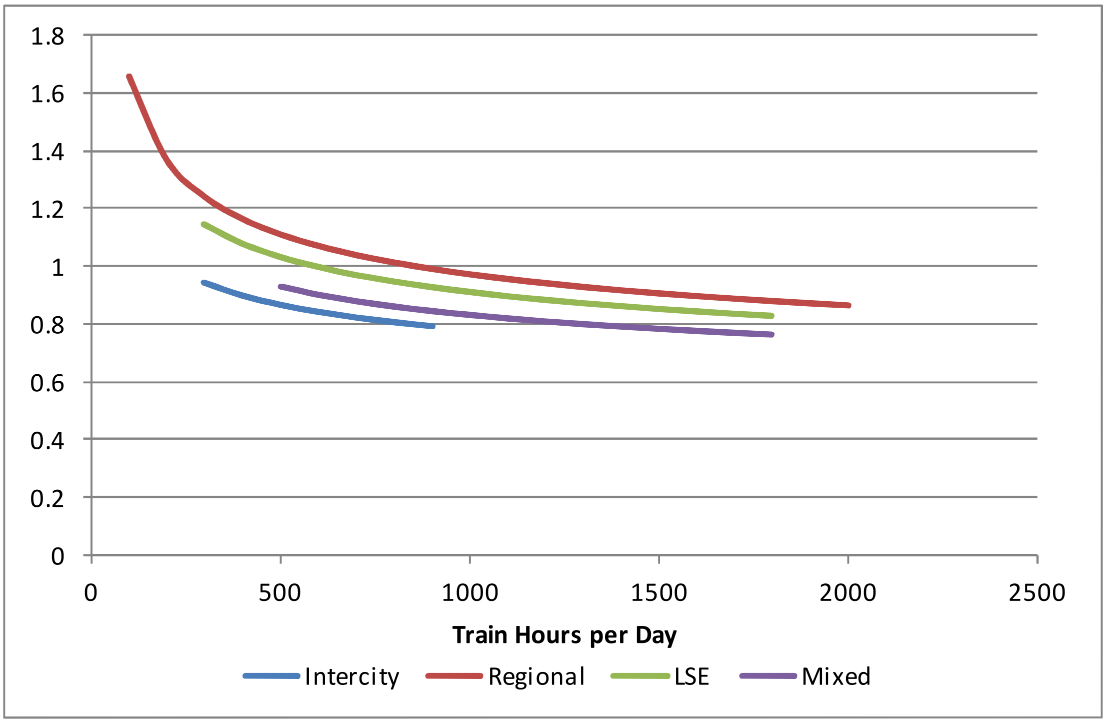


Figure 5.3 provides a similar plot for RtS. This shows that for all of the middle 80% of the train hours distribution, intercity (and mixed TOCs) exhibit decreasing RtS.

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. Thus 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. The chief aim of these mergers was to capture the benefits of sharing of staff and rolling stock between services and to reduce the number of operators running out of London stations, so as to improve timetabling and real time control of use of infrastructure (platforms etc). This has tended to result in larger franchises e.g. Great Western re-mapping (an example considered in 5.4.4), which implies an increase in the size of TOCs which, given the findings on RtS, is likely to increase rather than reduce unit costs. However, there are a number of other factors that change through re-mapping TOCs relevant to the model, notably possible reduction in overlap of franchises (which increases the density of operation) and a move to a mixture of the type of services provided. The model shows that TOCs tend to have increasing RtD which acts to reduce unit costs following TOC mergers. As discussed above, there are also important heterogeneity factors to take into account. Which effect will dominate in a given situation is an interesting research question. Next the findings regarding heterogeneity are discussed, followed by consideration of the cost implications for mergers, via a set of real world examples.

Finally in considering the policy implications of our findings on RtD and RtS, it must be remembered that the analysis concerns the costs of passenger train operations only. Just because unit costs can be reduced by running more train hours or by franchise remapping does not mean that this is the best course of action; best from the perspective of either minimizing whole system cost or maximizing welfare. There may be demand side constraints such that running extra train services may not yield a sufficiently large increase in passenger usage to justify the extra cost. There may also be a reduction in competition between franchises if franchise overlap is reduced, which may result in a net disbenefit. Finally running extra train services may have negative externalities to other services due to infrastructure congestion and other infrastructure costs. Thus this analysis should be used alongside analyses of other aspects of the railway system to evaluate the merits or demerits of specific interventions. Note that when merging/remapping TOCs is considered in sub-section 5.4.3, then these issues of congestion and demand side constraints are less important given we are simply rearranging the provision of existing services.

5.4.3. Implications of heterogeneity

This section considers the impact of TOC heterogeneity on costs; the other variables populating the hedonic cost function i.e. the q_{j2} j=1,..,9 variables and related coefficients in Table 5.1. The elasticity of cost with respect to average train length, train speed and passenger load factor are proportional to the elasticity with respect to train hours, with the coefficient on the characteristic acting as the proportionality constant:

$$\frac{\partial \ln C_{it}}{\partial \ln q_{j2it}} = \phi_{j2} \frac{\partial \ln C_{it}}{\partial \ln \psi_{2it}} \text{ j=1,2,3} \qquad (5.14)$$

All ϕ_{j2} j=1,2,3 coefficients are less than unity indicating that the cost elasticities with respect to these characteristics are lower than for train hours. This is intuitive. Generally from an operations perspective, it is cheaper to add vehicles to existing trains (q_{12}) rather than run more train services (e.g. there is still only one driver). Likewise the passenger load factor coefficient (q_{32}) is very low which indicates the very low marginal cost of carrying extra passengers once the number of train hours and train length are controlled for. The train speed coefficient (q_{22}) implies that running trains a greater distance, holding train hours constant, increases costs less than increasing train hours and distance together. This result is due to both staff and most of vehicle costs being time based rather than distance based, all other things being equal.

In terms of implications for RtD and RtS, given the findings of decreasing RtD and RtS with the size of ψ_{2it}, a TOC operating the same train hours can be expected to have greater RtD and RtS if it operates shorter trains, slower trains and/or has a lower passenger load factor. This follows from the fact that the level of the hedonic output, ψ_{2it} is found to be an increasing function of q_{12}, q_{22} and q_{32}. Furthermore, these findings are intuitive.

Turning to the findings specifically on the effect of TOCs providing a mixture of service types, which is given by the coefficients on the interaction proportion variables and number of generic rolling stock types operated i.e. q_{j2it} j=4,...,9. To explain the findings, it is useful to consider some stylised examples. Table 5.4 presents the growth in the hedonic output ψ_2 from the base case of a wholly regional TOC. Table 5.4 firstly considers the impact of mixing service types and then considers the additional impact of a TOC operating more rolling stock types which is likely when TOCs provide more service types (highlighted grey). Importantly, it shows that while mixed TOCs are more expensive than regional TOCs, they are not more expensive than exclusively intercity or LSE TOCs, all other things being equal. Adding in the effect of increasing rolling stock types increases the growth rate in the hedonic output further relative to a wholly regional TOC, however mixed TOCs still are less costly than pure intercity and LSE TOCs.

Thus Table 5.4 would seem to indicate that allowing TOCs to produce mixed services is beneficial. However, it should be noted that heterogeneity and changes in heterogeneity are captured in the model via a complex set of variables (including train speed, train length and passenger load factor) as well as the TOC type dummies/number of rolling stocks etc. All these characteristics will change following a franchise re-mapping (and not just the TOC type dummies/ rolling stock variable). Thus the overall effect is a complex interaction of all heterogeneity characteristics, density, scale and input prices. As such when specific re-mappings which result in mixed TOCs are considered, the overall heterogeneity effect may actually be cost increasing (as is indeed the case in the Greater Western example consider in the next sub-section).

Table 5.4 Heterogeneity findings – Growth in hedonic output (ψ_2) relative to a regional only TOC

TOC Type Composition			Increase in rolling stock			
Regional	LSE	Intercity	types	Growth rate	p-val	
100%	0%	0%	0	0.0%	N/A	
0%	100%	0%	0	36.2%	0.000	***
0%	0%	100%	0	52.9%	0.000	***
33%	33%	33%	0	0.7%	0.588	
50%	50%	0%	0	3.9%	0.563	
0%	50%	50%	0	-1.3%	1.603	
50%	0%	50%	0	18.9%	0.000	***
33%	33%	33%	6	14.5%	0.000	***
50%	50%	0%	3	10.8%	0.157	
0%	50%	50%	3	5.2%	0.002	***
50%	0%	50%	3	26.8%	0.000	***

Notes: a) The growth rate is constructed as the percentage increase in ψ_2 resulting from a change in the composition of the TOC relative to the base case (a 100% regional TOC). Formally, Growth rate = $\left(e^{\phi_{42}q_{42}} e^{\phi_{52}q_{52}} e^{\phi_{62}q_{62}} e^{\phi_{72}q_{72}} e^{\phi_{82}q_{82}} e^{\phi_{92}q_{92}}\right)-1$.
b) The computation is indifferent to the number of rolling stock types in the base case
c) The impact of combining rolling stock types is included by implicitly assuming each TOC type operates three unique rolling stock types.

5.4.4. The impact of franchise re-mapping

In this sub-section, the estimated model is used to predict the cost change from re-mapping franchises[31]. The franchise re-mapping in recent years has, in most cases, the following implications:

- In general there has been a rationalisation to larger franchises. Thus there will be scale effects, which given the finding of decreasing RtS for large TOCs

[31] Note simply comparing the sum of costs for the pre-re-mapped TOCs with those from the post-re-mapped TOCs is not valid because there is output, input price and technical change growth between the time periods that they are observed in the dataset. Also, the last year and first year of data are often cost data with the most measurement error given the required adjustments to align costs to match a standard financial years (when in fact re-mappings occur within years). Thus the model is used to predict the cost change.

could increase unit costs.

- Irrespective of whether the re-mapped TOC(s) are larger, the move to integrating TOCs of various service types results in a removal of franchise overlap which implies that the sum of the route-km for all the re-mapped TOC(s) will be less than the sum of the route-km for the previous TOCs. This implies that for a given usage level (train hours), density of usage increases. Thus, there will be density effects which, given the finding of increasing returns to density, implies a decrease in unit costs.

- The re-mapped franchises now provide more than one service type, as opposed to the previous TOCs which, in most part, operated only one service type. Thus the TOCs formed from re-mapping will have TOC heterogeneity measures (length of train, average speed etc.) which are weighted averages of the previous TOCs. This will not necessarily be cost neutral given the flexible form that the quality variables enter into the model (there are non-constant elasticity effects in the model). The new TOCs will also have non-zero values for some of the TOC service type heterogeneity interaction terms i.e. there will be effects from the TOC providing a mixed service. Furthermore, they may be operating different numbers of rolling stock types (see Table 5.4).

The extent to which mergers can deliver cost savings through exploiting increasing RtD depends on the relative heterogeneity characteristics before and after re-mapping. This effect is qualified by providing the evaluated ψ_2 divided by route-km for the TOC, which is termed the 'heterogeneity adjusted density' measure. It is this that determines the extent to which a

TOC can exploit any increasing RtD since RtD is defined with respect to the hedonic output. It should be noted that it is the proportional change in this measure from the before to after re-mapping situation which gives the extent to which density is changing; the absolute number is meaningless (it is a function of the units of the data). If the proportional change in heterogeneity adjusted density is greater than the proportional change in train hours density then heterogeneity is reinforcing the returns to density (and scale) effects. This is because the density measure that is actually driving RtD/RtS is increasing more than the naive measure of density (train hours density). Similarly, if the reverse is true heterogeneity is dampening the RtD (and RtS) effects.

Clearly, *a priori* for a given merger, there are conflicting effects; with increasing density generally reducing costs, increasing scale of operations increases costs and the impact of changes in heterogeneity being ambiguous. Two real world mergers are considered and also a hypothetical merger, which is quite topical at present, due to the policy aspiration of several northern English regions to expand and become franchisor of the enlarged Northern franchise. The characteristics of each merger are described in Table 5.5, alongside the predicted cost changes. The following observations can be made:

- Greater Western merger – This is found to increase costs. This is for two reasons. Firstly, there is an exhaustion of RtS i.e. the new franchise is simply too large. Secondly there is a large fall in the impact of heterogeneity on ψ_2. The result is that while train hours density increases by 57%, heterogeneity adjusted train density increases by only 12%. This implies that the Greater

Western TOC is unable to exploit increasing RtD as much as would be expected based on the large increase in train density, thus there is only a weak off-setting cost reduction effect from density relative the cost increasing scale effect (the impact of heterogeneity is to dampen any density effect). In any case, the new franchise has even exhausted RtD savings (being estimated to exhibit roughly constant RtD)

- London Eastern re-mapping – This is found to decrease costs. Importantly both the new franchises have increasing RtD and one TOC still has increasing RtS (the other has constant returns to scale). Thus it is concluded that these TOCs are not operating at output levels above the minimum efficient scale points.

- New Northern franchise – This results in a small increase in costs. This seems to be due to the decreasing RtS faced by both the Northern and New Northern TOCs and constant RtS of Transpennine Express. Furthermore, it is predicted by the model that the New Northern franchise will have exhausted RtD. Overall the effect of heterogeneity changes is approximately neutral from one mapping to the other.

Table 5.5 The predicted cost impacts of franchise re-mapping

	Pre-remapping TOCs								
Year of remapping	Name	TOC Type	Route-km	Train-hours	Train Hours Density	Hetrogeniety Adjusted Density	RtS	RtD	Predicted Cost
2006/07	Great Western	Intercity	1368	598	0.437	165.1	0.815	1.061	278
	Great Western Link	LSE	581	550	0.947	108.1	1.023	1.267	138
	Wessex	Regional	1394	529	0.380	26.2	1.021	1.218	92
	Total		*3343*	*1677*	*0.502*	*97.3*			*508*
2004/05	Anglia	Regional	669	312	0.467	69.2	1.087	1.336	47
	Great Eastern	LSE	235	555	2.362	404.8	1.041	1.417	95
	WAGN	LSE	414	886	2.139	300.8	0.923	1.256	167
	Total		*1318*	*1753*	*1.330*	*201.8*			*308*
2010/11 - hypothetical	Northern	Regional	2746	2597	0.946	48.5	0.744	0.984	389
	Transpennine Express	Regional	1251	633	0.506	40.4	1.023	1.170	137
	Total		*3996*	*3230*	*0.808*	*46.0*			*527*

	Post-remapping TOCs								
Year of remapping	Name	TOC Type	Route-km	Train-hours	Train Hours Density	Hetrogeniety Adjusted Density	RtS	RtD	Predicted Cost
2006/07	Greater Western	Mixed	2129	1677	0.788	109	0.737	0.992	554
	Total		*2129*	*1677*	*0.788*	*109*			*554*
2004/05	ONE	Mixed	1001	1028	1.027	142	0.786	1.091	170
	Great Northern	LSE	275	725	2.637	383	1.116	1.430	120
	Total		*1276*	*1753*	*1.374*	*194*			*290*
2010/11 - hypothetical	New Northern	Regional	3019	3230	1.070	62	0.724	0.990	579
			3019	*3230*	*1.070*	*62*			*579*

	Percentage change in Characteristics (+ indicates increase)						Cost Change	
Year of remapping	Name	Route-km	Train-hours	Train Hours Density	Hetrogeniety Adjusted Density		£'000	Percent
2006/07	Greater Western	-36%	0%	57%	12%		**45.6**	**9%**
2004/05	ONE/Great Northern	-3%	0%	3%	-4%		**-17.9**	**-6%**
2010/11 - hypothetical	New Northern	-24%	0%	32%	34%		**52.6**	**10%**

Notes: 1) Method for calculating metrics for Post-remapping TOCs: Route-km: taken from actual values in subsequent years; Train-hours: sum of pre-remapping TOCs allocated to post-remapping TOCs through proportion split between post-remapping TOCs in a subsequent year; Predicted cost - in addition to the aforementioned variables, assumptions needed to be made regarding the level of other variables in the function i) input prices - averages of input prices for pre-mapping TOCs ii) levels of other variables in the hedonic output function - taken from actual data for post-remapping TOCs in the subsequent year iii) number of stations operated is taken from subsequent year data for post-remapping TOCs.

2) The New Northern TOC is hypothetical: Measures are calculated as in Note 1) with the exception: i) route-km this is given as the Northern route-km plus the additional route length of Transpennine Express of the North West route to Glasgow ii) number of stations operated is the sum of the stations operated by the two merging TOCs.

5.5. Summary

In this chapter, a hedonic Translog cost function for TOCs in Great Britain has been estimated. The model includes three hedonic outputs: route-km, stations operated and train hours. The use of train hours (over train-km) is a data innovation in itself but, in addition, the train hours hedonic output has a number of characteristics also included within the hedonic function, which characterise TOC heterogeneity. Thus the model is rich in its characterisation of firms' technology.

This richness allows establishment of a deeper understanding of the variation in returns to scale and density in the industry. In particular, different scale and density effects can be distinguished depending on the output characteristics of the TOC and not just the usual overall output level and input price level as in a simple (non-hedonic) Translog cost function. This is important since there has recently been a move towards re-mapping franchises to larger, more heterogeneous franchises, which requires such a rich model to determine whether this increases or reduces costs.

The analysis indicates over 50% of TOCs in the sample operate under decreasing RtS. Furthermore, returns to scale fall with the size of operation, which is consistent with a u-shaped average cost curve. The implication of these findings is that the current mappings of TOCs in Britain are such that operations are above their optimal size given that most TOCs operate with decreasing RtS i.e. on the upward part of the average cost curve. Thus there is an argument for more, smaller TOCs.

However, it is found that there are increasing RtD, i.e. unit cost savings from running more trains on a fixed network. This has two implications. Firstly, increasing capacity (train hours) to meet increasing passenger demand should reduce unit costs. Secondly, there is scope to reduce unit costs by removing franchise overlap; this effect therefore working in the opposite direction to the scale effect (as the density finding suggest that TOC mergers with reduce unit costs). There can however be impacts of changes in the output mix (heterogeneity of services) which prevent TOCs from exploiting any RtD even though train hours per route-km increases.

Three examples of the cost changes from mergers are presented which demonstrate the importance of changes in scale of TOCs which generally increases unit costs, changes in the density of operation of TOCs which generally reduces unit costs and changes in the output heterogeneity which affect the extent to which any RtD can be exploited. It is found that two mergers increase costs while one considered reduces cost. Overall, the analysis shows that it is important to model the intricate relationship between cost and scale, density and heterogeneity explicitly, rather rely on simple heuristics (e.g. large TOCs increase unit costs, but denser TOCs reduce unit costs) as it is the interactions between these three factors which results in the direction and magnitude of a cost change. In particular, changes in heterogeneity characteristics played a substantial role in the Great Western re-mapping since these changes prevented exploitation of the returns to density, which implied an overall cost increase due to relatively substantial decreasing RtS. Since franchise mergers also reduce on rail competition which maybe undesirable (Jones, 2000), the supposed cost savings from exploiting RtD are important in supporting the case for mergers. It is therefore illuminating that this study suggests that these returns may not be realised in all cases.

Though the empirical example is focused on the British TOCs and offers some important insights in respect of rail policy in Britain, it also has wider implications. The findings suggest that previous estimates of scale and density properties in railways may have been biased, to the extent that they did not adequately model the interaction between scale/density and heterogeneity of services. In terms of regulatory policy, in interpreting evidence on scale and density returns in railways, the model suggests that policy makers need to take service heterogeneity into account. Failure to do so may mean that policy decisions are made on the basis of supposed scale/density returns that cannot be realised in practice. Modelling railway operations is complex and thus to address specific policy questions (such as the cost implications of mergers) a rich model, such as that developed in this chapter, is required.

6. Infrastructure cost analysis[32]

6.1. Introduction

One clear finding from the railways infrastructure review (sub-section 4.3.1) was that railway infrastructure suffers from increasing returns to density i.e. there are average cost savings from a single infrastructure manager at least in a given geographical area i.e. between the same origin and destination points. Thus the cost characteristics and the organisation in Britain is such that there is a single, monopoly infrastructure manager. With monopoly comes price setting power and as such there is a need to provide economic regulation in an attempt to correct the market failure. Economic regulation requires information on the (efficient) cost structure of the industry and thus some form of benchmarking is required to establish this.

In this chapter, the task of benchmarking a single infrastructure manager in a country with other international comparators is considered. Relying on time series analysis for a given infrastructure manager is difficult since efficiency and technical change cannot be distinguished separately. Furthermore, even assuming away any technical change factors, any time series approach would rely on a firm having been efficient in the past to gauge the current performance of the firm relative to best practice. Thus some cross sectional comparisons need to be made and these may or may not be augmented with repeat observations over time (forming a panel data set). Past

[32] This chapter is based on Smith and Wheat (2012b). There is additional discussion of potential modelling innovations, covering the use of the Mundlak (1978) transformation in 6.4.3 and Closed Skew Normal distributions as a basis for estimating the dual-level model in 6.4.4.

econometric analysis has adopted one of two approaches to develop cross sectional comparators.

Firstly, internal benchmarking can be undertaken. This considers data on business units within the single company. Thus cross sectional observations are provided by considering the integrated firm as comprising several smaller units. This was the approach undertaken by Kennedy and Smith (2004) for seven zones comprising the British railway network (operated by the single infrastructure manager, Network Rail). This is likely to provide some indication as to the likely variability of the performance of the units within a firm but will ignore any persistent under performance of all of the business units that can arise from inefficiency common to all units and with respect to joint costs not considered within the analysis. Regarding data, costs can be reported using consistent definitions across the units given all the units considered are from same firm, providing the researcher with reassurance that data is comparable and thus inefficiency measurement based on residuals is robust. This is with the proviso that the units are meaningful from an organisational perspective and that joint costs are not arbitrarily allocated to units. Failure of either of these conditions will result in misleading comparisons.

The second approach is to collect data on a number of infrastructure managers across different countries. Smith (2012) and Smith, Wheat and Smith (2010) discuss the use of this approach by benchmarking Network Rail against European comparator railways using data from the UIC. This approach also underpinned the efficiency determination for Network Rail in the 2008 Periodic Review (see ORR (2008)). In general, this method has the advantage of considering performance benchmarks outside of the firm of interest; thus comparing the firm to best practice elsewhere. A key constraint is data, both in terms of collecting data on the same variables across countries, but more subtly, ensuring that the definitions of variables, particularly cost definitions are the same. Obviously, failure to collect enough variables will result in many exogenous cost drivers being excluded from the analysis and, given efficiency analysis majors on what variation in costs are left over after controlling for observed factors, lead to misleading efficiency conclusions. Nonetheless, the same problem will arise if variables are defined in an inconsistent way from firm to firm. The need for consistent data is likely to constrain the number of firms that can be considered. The 2008 Periodic Review work is one of the few examples of this approach undertaken in practice; partly a result that the dataset was already collected over a number of years for a separate purpose which made analysis feasible.

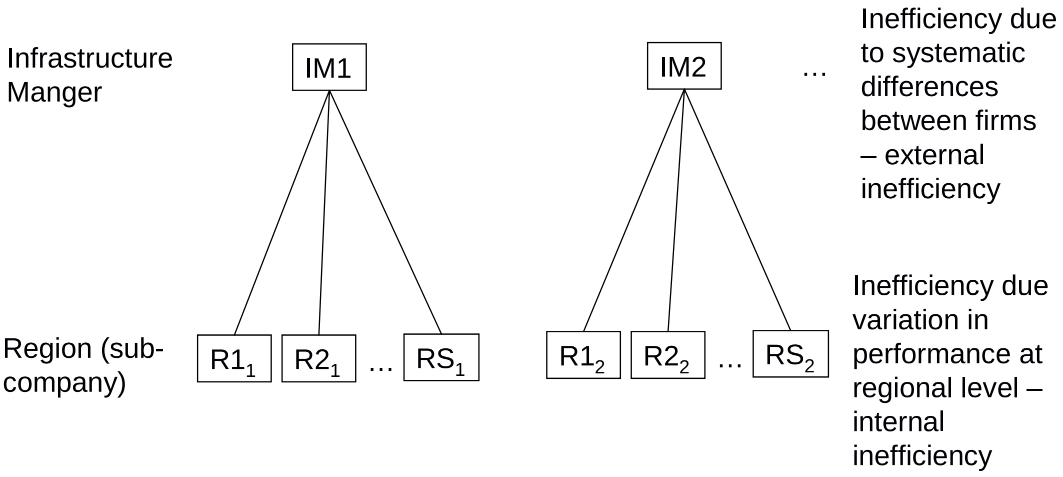
Clearly either approach is likely to have limitations. The subject of this chapter is to consider a combination of the two approaches outlined above. It considers how best to exploit regional data for a number of countries simultaneously.

The structure of the chapter is as follows. Section 6.2 discusses the unique features of the dataset. Section 6.3 develops a stochastic frontier model to suite the data structure. Section 6.4 outlines econometric estimation of the proposed models in section 6.3. Section 6.5 applies the models to a data set of European and North American railway infrastructure managers.

6.2. Sub-company data structure

Figure 6.1 shows the structure of the sub-company data that this chapter is focusing on. The data is collected for business units (here in called “sub-companies”) for a number of railway infrastructure managers (IMs). This requires the following notation. Denote infrastructure managers by index i, with i=1,...,N, and the sub-companies comprising each IM by index s, with s=1,...,S(i). Thus the notation is flexible enough to accommodate different numbers of sub-companies per IM which is important given that there is no *a priori* reason for the same number of sub-companies to be observed per IM. For cases where the sub-companies are observed over a number of time periods a further index t is required, t=1,...,T(i); again the index accommodates different numbers of time periods per IM.

Figure 6.1 Sub-company data structure



Infrastructure Manger
IM1
IM2
…
Inefficiency due to systematic differences between firms – external inefficiency
Region (sub-company)
$R1_1$
$R2_1$
…
RS_1
$R1_2$
$R2_2$
…
RS_2
Inefficiency due variation in performance at regional level – internal inefficiency

Thus the dataset is tiered with multiple regional observations per IM. It is important to realise, however, that the data itself is only at one geographic level for each firm; data is collected by sub-company and not at the IM level. The ‘dual-level’ refers to the tiered modelling of inefficiency discussed below.

There are two possible advantages of utilising data comprising observations on business units across a number of companies. The first is that using data on sub-companies rather than restricting analysis to a single observation per IM implies more observations, which is clearly an important consideration given sample sizes are often small for regulatory benchmarking purposes. The second is that, by comparing the regulated IM to other IMs, benchmarking will be relative to external, rather than simply internal, best practice.

A more subtle benefit to the data structure is regarding the conceptualisation and modelling of inefficiency. Utilising this data structure explicitly allows us to consider inefficiency that may arise due to both overall IM management policies (head office) and how these are implemented across the sub-companies. In particular, econometric methods can be used to distinguish between inefficiency that differs between sub-companies within a given IM and any inefficiency which is persistent across all the sub-companies in a given IM. This model is labelled as the ‘Dual-Level Inefficiency Model’. This model distinguishes between inefficiency due to systematic differences between IMs (external inefficiency) and variation in the performance between sub-companies in a given IM (internal inefficiency). A more detailed discussion of the economic interpretation of this model is provided in section 6.3.2.

6.3. Sub-company model of inefficiency

In this section, a stochastic frontier cost[33] model which allows for both persistent, firm-specific and sub-company level inefficiency effects (external and internal inefficiency respectively) is developed. Interesting special cases are also outlined, which are subsequently used as (nested) comparator models in the empirical illustration in Section 6.5.

6.3.1. Dual-level inefficiency model

The model consists of a cost frontier which has been transformed by taking logarithms:

$$\ln \mathrm{C}_{\mathrm{its}} = \alpha + \mathrm{f}\,(\mathbf{X_{its}};\boldsymbol{\beta}) + \mathrm{u}_{\mathrm{its}} + \mathrm{v}_{\mathrm{its}} \qquad \text{i=1,…,N, t=1,…,T(i), s=1,…,S(i)} \qquad (6.1)$$

where C_{its} is the cost for sub company unit s in firm i in time period t, α is a constant, $\mathbf{X_{its}}$ is a k dimension vector of outputs and input prices (and covariants if applicable), $\boldsymbol{\beta}$ is the conformable vector of parameters, $\mathrm{v}_{\mathrm{its}}$ is a random variable representing statistical noise and $\mathrm{u}_{\mathrm{its}}$ is a variable representing inefficiency. $\mathrm{v}_{\mathrm{its}}$ is assumed to be distributed independently from the regressors and $\mathrm{u}_{\mathrm{its}}$. The inefficiency term(s), while initially multiplicative, are additive following the logarithm transformation and thus inefficiency is a Farrell (1957) type radial

[33] As widely noted in the literature, the model can easily be translated into a production function by reversing the sign on $\mathrm{u}_{\mathrm{its}}$.

measure.

In order to consider inefficiency effects at two levels within the firm, u_{its} is decomposed into:

$$\mathrm{u_{its}} = \mu_{\mathrm{it}} + \tau_{\mathrm{its}} \text{ with } \mu_{\mathrm{it}} \perp \mathrm{v_{its}},\ \mu_{\mathrm{it}} \sim \mathrm{iid} \text{ and } \tau_{\mathrm{its}} \sim \mathrm{iid} \qquad (6.2)$$

In this formulation $\mathrm{u_{its}}$ is split into two components: μ_{it}, which is the persistent element of inefficiency that applies across all sub-companies within the same firm; and τ_{its}, which is the residual component that varies randomly across all sub-companies. Both inefficiency terms may either be fixed over time or vary in some way.

In order to explain the economic interpretation of the model, and its position within the literature, the t subscripts from the model are dropped to allow focus on the sub-company structure of the data. Following this, the different assumptions that may be made concerning the variation of inefficiency over time are considered. Re-writing (6.1) and (6.2) without the time subscripts yields:

$$\ln \mathrm{C_{is}} = \alpha + \mathrm{f}(\mathbf{X_{is}};\boldsymbol{\beta}) + \mathrm{u_{is}} + \mathrm{v_{is}} \qquad \text{i=1,…,N, , s=1,…,S(i)} \qquad (6.3)$$

where $\mathrm{u_{is}} = \mu_{\mathrm{i}} + \tau_{\mathrm{is}}$ with $\mu_{\mathrm{i}} \perp \mathrm{v_{is}}$, $\mu_{\mathrm{i}} \sim \mathrm{iid}$ and $\tau_{\mathrm{is}} \sim \mathrm{iid}$.

This formulation is analogous to that presented in Kumbhakar and Hjalmarsson

(1995) and Kumbhakar and Heshmati (1995). In their formulation, applicable to standard panel data (i and t subscripts only), μ_i represents the persistent (over time) element of inefficiency, and τ_{it} is the residual component of inefficiency[34] (both of which are one-sided). Here the same distinction between persistent and residual inefficiency is made, but this time over sub-companies comprising a firm, rather than time.

Since u_{is} is the inefficiency of each sub-company in the sample (comprising a persistent and random element), a further step is required to produce an overall measure of firm inefficiency. Overall inefficiency for an individual firm is computed therefore as the sum of the persistent element and a weighted average of the random component for each of the sub-companies within the firm:

$$\bar{u}_i = \mu_i + \frac{\sum_{\forall s} C_{is} \cdot \tau_{is}}{\sum_{\forall s} C_{is}} \qquad (6.4)$$

6.3.2. Economic Interpretation

The model in (6.3) can be interpreted in terms of separating out at what geographic level inefficiency within an IM varies; either at the IM level or at the sub-company level. μ_i represents the component of inefficiency which is persistent across all sub-companies within an IM, thus yields a measure of performance that varies across

[34] The terms persistent and residual inefficiency are adopted since they are terminology in Kumbhakar and Hjalmarsson (1995).

firms but not across sub-companies within an IM. This is termed external inefficiency. τ_{it} gives an additional sub-company performance measure; i.e. the inefficiency of a given sub-company additional to the inefficiency for all sub-companies within the IM. This is termed internal inefficiency.

Care should be taken not to confuse the above interpretation with one which states where the inefficiency resides in terms of which level of management (head office versus regional) is ‘responsible’ for the inefficiency. Indeed, in presenting this work at conference[35] and other industry/regulatory meetings, the $\mu_i \perp v_{is}$ assumption has been challenged as unrealistic. The criticism arises since it is perceived that μ_i measures the inefficiency that can be attributed to central management failings, while τ_{it} measures the inefficiency that can be attributed to regional management failings. However, some comments have suggested that if central management have a high degree of incompetence then they may be less rigorous in their appointment of regional managers which may in turn lead to on average worse performance of the managers in the IM. Thus, the argument follows that there is dependence between the two error components. However this argument only follows if the model is viewed as yielding measures that can be attributed to each level of management. All the model actually yields is which element of inefficiency varies by sub-company versus that which is persistent across all sub-companies in a given IM. If the model is to be viewed as yielding measures as to where efficiency resides, then care should be taken to be clear that any inefficiency arising from appointment of incompetent regional managers (the example above) is the ‘fault’ of the higher level managers; it

[35] Workshop on Efficiency and Productivity Analysis (EWEPA) June 2009, Pisa, Italy.

appears in the high level inefficiency score.

In addition to interpreting the allocation of inefficiency between the two error components, a discussion is required as to the validity of interpreting μ_i as inefficiency as opposed to other sub-company invariant unobserved factors. Such factors may include climate effects e.g., it may be expected the prevailing weather conditions in different countries will affect the costs of providing railway infrastructure. If these are not specifically included as regressors in the model and are to some extent sub-company invariant then it can be expected that these would influence μ_i. Thus it is not clear that μ_i represents inefficiency vis-à-vis sub-company invariant unobserved heterogeneity. There is no definitive practical solution to this dilemma; the preferable mitigating method is to try and incorporate as many cost driving explanatory variables into the analysis as possible. This is not an issue confined to the exploitation of sub-company data; in the standard panel literature the μ_i term has also been interpreted as a measure of time invariant unobserved heterogeneity (see Greene, 2005, Kumbhakar, 1991, Heshmati and Kumbhakar, 1994). The use of the Mundlak (1978) transformation/decomposition of the fixed effects model is discussed in section 6.4.3 as a possible way to separate out unobserved heterogeneity from inefficiency. Nethertheless, this decomposition requires the assumption that unobserved heterogeneity is correlated with the regressors and inefficiency is not.

Finally, as noted in the introduction, the use of sub-company data has benefits for performance analysis and more generally cost analysis beyond the ability to measure dual level performance. It can substantially increase the number of observations for analysis which addresses a common problem in economic regulation (small N). Furthermore, aggregation bias can arise in an estimated cost function if data is aggregated at a level that is not equivalent to the level at which operational decisions are actually made (Theil, 1954). For example, analysing infrastructure of railways using national data may lead to misleading estimates of returns to network size if the railway is in fact organised into zones. A more useful concept would be to look at the returns relating to network zone size. Much depends on what the analyst is trying to understand in the first place, but overall sub-company data provides a much richer dataset to investigate much more subtle distinctions regarding returns of size and density. Clearly, if aggregation bias is present in the deterministic frontier, then this will lead to systematic under or over prediction in inefficiency (omitted variable bias).

Both simulation and analytic evidence to confirm the above point has been provided by Brorsen and Kim (2013) for the stochastic frontier model. In particular they consider the case where managerial autonomy is at the disaggregated level, but costs are modelled at the aggregate level. The empirical example used to inform their simulation is schools (disaggregate level) and school districts (aggregate level). It is shown that if costs are modelled at the aggregate level then returns to scale estimates are biased downwards (towards decreasing returns) and inefficiency (measured by the magnitude of the inefficiency variance) is also biased downwards (under estimated).

6.3.3. Sub-company inefficiency invariance model

One interesting special case that is nested within the model outlined in (6.3) is the sub-company inefficiency invariance model. Where it is reasonable to assume that $\tau_{is} = 0 \; \forall \; i,s$, that is, all inefficiency is persistent across sub-companies in a firm, and thus there is no additional inefficiency variation between sub-companies comprising a firm, then the model can be written:

$$\ln C_{is} = \alpha + f(\mathbf{X_{is}};\boldsymbol{\beta}) + \mu_i + v_{is} \qquad i=1,\ldots,N,\ s=1,\ldots,S(i) \qquad (6.5)$$

In this case the model has reduced to a more conventional model, analogous to the time invariant inefficiency models of Pitt and Lee (1981) or Schmidt and Sickles (1984) but with inefficiency invariance in sub-companies comprising a firm rather than across time.

It should be noted that one of the weaknesses of the time invariant model in the standard panel inefficiency model literature is that it may not be appropriate to assume that inefficiency is invariant over time, particularly when panels are long (and it is exactly when panels are long that the benefits of the panel approach to inefficiency estimation are fully felt). Whilst the assumption of sub-company inefficiency invariance may likewise be challenged – in fact, the presence of sub-company effects is the motivation behind the dual-level efficiency model – this assumption may be a reasonable approximation in some circumstances (when there is little sub-company autonomy). Furthermore, the assumption does not necessarily become more implausible as the number of sub-company units is increased (as is the case for long panels). Importantly, since this model is nested within the dual-level inefficiency model, the absence of sub-company inefficiency variation can be tested.

6.3.4. The pooled model

The restriction $\mu_i = 0 \ \ \forall i$ yields a simple pooled model in which the inefficiency of each sub-company ($u_{is} = \tau_{is}$) is assumed to be identically and independently distributed across all sub-company units irrespective of which firm they belong to. In this case the central management in each firm plays no role at all from an inefficiency perspective. Since this model is nested within the dual-level inefficiency model, the absence of a persistent, firm-specific inefficiency component can be statistically tested.

6.3.5. Assumptions about inefficiency variation over time

The empirical application comprises data both at sub-company level and over time. The focus of the sections above has been on the sub-company dimension of the data structure. Therefore, the dual-level inefficiency model outlined in (6.1) and (6.2) makes a simple assumption concerning the variation in inefficiency over time ($\mu_{it} \sim \text{iid}$ and $\tau_{its} \sim \text{iid}$). The pooled model likewise makes a simple assumption regarding the variation in inefficiency over time ($u_{its} = \tau_{its} \sim \text{iid}$). In the sub-company invariance inefficiency model (6.5), where $\tau_{its} = 0$, firm inefficiency (μ_i) is assumed to be invariant over both sub-company and over time.

It is possible to make alternative assumptions about the behaviour of both the μ_{it} and τ_{its} inefficiency terms over time. These include independence and invariance over time as noted above, but could be extended to allow varying inefficiency over time via a deterministic scaling model (presented in the most general forms in Kumbhakar and Lovell (2000), Orea and Kumbhakar (2004)). Section 6.4 shows how to estimate such paths for the case of the sub-company invariance model. Importantly, sub-company data structure potentially provides a powerful way to estimate firm specific paths of inefficiency over time, since there can be many observations per firm relative to the number of time periods. This is in contrast to the use of panel data where the number of observations per firm is equal to the number of time periods to which they are observed.

6.4. Estimation

6.4.1. Dual-level inefficiency level model

The estimation framework draws on the approach by Kumbhakar and Hjalmarsson (1995) and Kumbhakar and Heshmati (1995). Consider (6.1) re-written as:

$$lnC_{its} = \alpha_{it} + f(x_{its};\beta) + \tau_{its} + v_{its} \qquad (6.6)$$

where $\alpha_{it} = \alpha + \mu_{it}$.

At this stage, no distributional assumptions on the two inefficiency error components have been made, except that they are distributed independently of the random noise term v_{its} and independently of each other. Now additional assumptions are made to facilitate estimation. Firstly, assumptions need to be made as to whether μ_{it} or correspondingly α_{it} are correlated with the regressors. If so, α_{it} is considered to be a fixed effect. If not, then α_{it} is considered to be a random effect. This assumption is considered in more detail in section 6.4.2 and an alternative formulation to decompose the fixed effect is outlined. Secondly, the assumption that τ_{its} is uncorrelated with the regressors and τ_{its} is a random effect is made. Treating τ_{its} as a random effect is a necessary assumption for the case of T=1 since fixed effects can not be identified from the regressors for T=1.

This model could be estimated in several ways. The first two methods use maximum likelihood to estimate the model in one stage. These are variants of the 'True' fixed and random effects models proposed by Greene (2005). In both cases $\tau_{its} \sim \text{iid}\left|N\left(0,\sigma_{\tau}^{2}\right)\right|$ and $v_{its} \sim \text{iidN}\left(0,\sigma_{v}^{2}\right)$, however it is possible to relax the assumption of homoscedasticity and zero mean of the (untruncated) distributions (Greene 2005). The formulation is the same as the original formulation of the pooled stochastic frontier model proposed by Aigner et al (1977), but with effects by firm per time period[36].

[36] Note that by effects by firm per time period it is not meant that the model has two way effects in firm and time. Instead there is one set of effects, with one effect for each year and firm. This is very general and could be replaced with an assumption that the persistent inefficiency of sub-companies in a firm is also time invariant, in which case $\alpha_{it} = \alpha_i = \alpha + \mu_i$. This is the assumption that is used in the empirical application. A further assumption could be that $\alpha_{it} = \alpha_{i1} + \alpha_{i2}t + \alpha_{i3}t^2$, i.e. that the persistent inefficiency follows a Cornwell et. al. (1990) type variation over time.

In the True fixed effects case, α_{it} is treated as a fixed effect and maximum likelihood is used to estimate the model. This case allows α_{it} to be correlated with the regressors. A potential disadvantage of this estimation approach is that, because of the presence of fixed effects, estimates of all parameters in the model (not just the fixed effects) may be inconsistent and biased. This is known as the incidental parameters problem (Neyman and Scott, 1948 and Lancaster, 2000). Greene (2005) provides Monte Carlo evidence that the bias does not appear to be substantial when T=5, which is encouraging given the short nature of panels typically available for performance analysis studies.

Estimation of this model by maximum simulated likelihood yields estimates of $\alpha_{it}, \boldsymbol{\beta}, \sigma_v^2, \sigma_\tau^2$. Ignoring for now the fact that the α_{it}'s are estimates and not population values, following Schmidt and Sickles (1984),

$$\min(\alpha_{it}) \xrightarrow[p]{} \alpha \quad T \to \infty \quad N \to \infty \tag{6.7}$$ [37]

As such a consistent estimator of μ_{it} is given by

$$\hat{\mu}_{it} = \alpha_{it} - \min(\alpha_{it}) \xrightarrow[p]{} \mu_{it} \quad T \to \infty \quad N \to \infty \tag{6.8}$$ [38]

For finite N and T, this method of recovery of μ_{it} results in a measure of relative

[37] This is a correction to the published paper. Consistency in this case requires both N and T to expand (given the effect varies by time and firm).

[38] The use of "min" is also a correction from the published paper. This also effects equations (6.9) and (6.14).

inefficiency (relative to the best performing firm/time observation). However this estimator cannot be constructed because the α_{it}'s have to be estimated. Thus the feasible estimator of μ_{it} is:

$$\hat{\mu}_{it} = \hat{\alpha}_{it} - \min(\hat{\alpha}_{it}) \tag{6.9}$$

The conditional expectation predictor proposed by Jondrow et al (1982) can be used to calculate a point predictor for the residual component of inefficiency, τ_{its}:

$$E[\tau_{its} \mid \varepsilon_{its}] = \rho_{its*} + \sigma_* \frac{\phi(\rho_{its*}/\sigma_*)}{1-\Phi(\rho_{its*}/\sigma_*)} \tag{6.10}$$

where $\rho_{its*} = \sigma_\tau^2 \varepsilon_{its} / (\sigma_\tau^2 + \sigma_v^2)$, $\sigma_*^2 = \sigma_\tau^2 \sigma_v^2 / (\sigma_\tau^2 + \sigma_v^2)$ and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal PDF and CDF respectively. To operationalise this, σ_τ^2 and σ_v^2 are replaced with their corresponding estimates and ε_{its} with

$$\hat{\varepsilon}_{its} = \ln C_{its} - \hat{\alpha}_{it} + f(\mathbf{X_{its}}; \hat{\boldsymbol{\beta}}) \tag{6.11}$$

In the true random effects case, α_{it} is treated as random and assumed independent of the regressors. Estimation proceeds by the method of simulated maximum likelihood, rather than simple maximum likelihoods because simulation is used to integrate out the random effect α_{it} from the likelihood function. Unlike the formulation in Greene (2005), a normal distribution cannot be assumed for this effect, since this variable is truncated from below at α. Instead, α_{it} is assumed to comprise:

$$\alpha_{it} = \alpha + \mu_{it} \quad \mu_{it} \sim \text{iid}\left|N\left(0, \sigma_{\mu}^{2}\right)\right| \tag{6.12}$$

The model now comprises the usual composite error term as proposed by Aigner et al (1977) distributed independently by each sub-company and by time, but also a random parameter, the constant term, which varies independently by firm and by time period. Estimation of this model by maximum simulated likelihood yields estimates of $\alpha, \boldsymbol{\beta}, \sigma_{v}^{2}, \sigma_{\tau}^{2}, \sigma_{\mu}^{2}$. Firm and time specific predictions of α_{it}, denoted $\hat{\alpha}_{it}$, are predicted as the expectation of α_{it} conditional on the data and the estimated parameters as given in equation 32 in Greene (2005). This is a consistent predictor as $S \to \infty$[39] (Train 2009, p. 269). This is approximated during the simulation of the likelihood function in estimation. μ_{it} is then predicted as:

$$\hat{\hat{\mu}}_{it} = \hat{\alpha}_{it} - \hat{\alpha} \tag{6.13}$$

Importantly, it should be noted that the prediction of μ_{it} is a prediction of absolute persistent inefficiency as opposed to the relative measures which are produced by the other estimation methods discussed in this chapter. This is because α is estimated through the maximum simulated likelihood process since it is the truncation point and mean of the underlying normal distribution of α_{it}. A prediction of the residual

[39] This is a correction to the published paper. In Train's case the effect varies by firm (individual, N). Here it varies by firm and time, so consistency requires expansion of the S dimension.

component of sub-company inefficiency, τ_{its}, is the same as for the True fixed effects case.

An alternative estimation framework is the multistage approach outlined in Kumbhakar and Hjalmarsson (1995) and Kumbhakar and Heshmati (1995). In this approach, the model is first estimated by either within or generalised least squares estimation, depending on whether the α_{it}s are treated as fixed or random effects respectively. Following this estimation, the residuals, $\hat{\varepsilon}_{\text{its}}$, are computed and these are used to compute the fixed or random effects, $\hat{\alpha}_{\text{it}}$ (as outlined in Kumbhakar and Hjalmarsson (1995) and Kumbhakar and Heshmati (1995)). An estimate of μ_{it} is then recovered as

$$\hat{\hat{\mu}}_{\text{it}} = \hat{\alpha}_{\text{it}} - \min\left(\hat{\alpha}_{\text{it}}\right) \tag{6.14}$$

The second stage comprises the use of conditional maximum likelihood estimation[40] to estimate the parameters of the specified distributions of τ_{its} and v_{its}. Kumbhakar and Hjalmarsson (1995) and Kumbhakar and Heshmati (1995) utilise a half normal and normal distribution for the two error components. Adopting these distributions for τ_{its} and v_{its} the conditional log likelihood function (for each observation) is[41]:

[40] Conditional on the (consistent) estimates in the first stage.

[41] Note that the sign on $\varpi_{\text{its}}\lambda/\sigma$ is reversed vis-à-vis Kumbhakar and Heshmati (1995) since a cost frontier is being estimated.

$$\ell_{\mathrm{ist}}\left(\sigma,\lambda\,|\,\boldsymbol{\beta},\alpha,\varpi_{\mathrm{its}}\right)=\mathrm{constant}-\ln\sigma+\ln\Phi\left(\varpi_{\mathrm{its}}\lambda/\sigma\right)-\frac{1}{2}\left(\varpi_{\mathrm{its}}/\sigma\right)^{2} \tag{6.15}$$

Where $\varpi_{\mathrm{its}}=\tau_{\mathrm{its}}+\mathrm{v}_{\mathrm{its}}$, $\lambda=\sigma_{\tau}/\sigma_{\mathrm{v}}$ and $\sigma=\sqrt{\sigma_{\tau}^{2}+\sigma_{\mathrm{v}}^{2}}$. To operationalise the model, ϖ_{its} is replaced with the consistent estimates given in the earlier stages by $\hat{\varpi}_{\mathrm{its}}=\hat{\varepsilon}_{\mathrm{its}}-\hat{\hat{\mu}}_{\mathrm{it}}$.

Summing over all observations and maximising with respect to σ and λ yields consistent estimates of the parameters of the distributions of τ_{its} and v_{its}. Following this, the Jondrow et al (1982) estimator can be applied as above to yield a prediction of τ_{its}.

Importantly in the first stage, no distributions have been specified for any error components. As such, the main parameter estimates, $\boldsymbol{\beta}$, are consistently estimated even if the resulting distributional assumptions in the second stage prove incorrect. Also, if the α_{it} ‘s are treated as fixed effects, the multistage approach has the advantage that this model does not suffer from the incidental parameters problem in the first stage, since in this first stage the incidental parameters are swept out by the within transformation. Thus, it is possible to introduce correlation between the firm persistent inefficiency component and the regressors without the potential inconsistency resulting from the incidental parameters problem. Of course, the incidental parameters problem may be present for the second stage parameters, since the estimated fixed effects implicitly enter the conditional likelihood function. The inevitable trade-off against this robustness is a loss of estimation efficiency relative to specifying (correctly) a full maximum likelihood function (such as using the approach by Greene (2005) above).

Since the remaining error components ($\tau_{its} + v_{its}$) are assumed not to be correlated with the regressors and μ_{it} both estimation methods are consistent[42].

6.4.2. Sub-company inefficiency invariance model

When sub-company inefficiency invariance is imposed, as in (6.5), then a host of ‘standard’ panel data models can be applied to model inefficiency (at the firm level only by assumption). As explained in section 6.3, the choice of model to apply depends on the assumption regarding inefficiency variation over time.

The case of both time invariant inefficiency and independence over time is an extension of the Pitt and Lee model with slightly different subscripts. As such the reader is referred to Pitt and Lee’s (1981) paper for details of the likelihood function. Likewise for the time varying models these are trivial extensions of the general time varying presented in Kumbhakar and Lovell (2000) and Orea and Kumbhakar (2004). The likelihood function for the model for standard panel data is presented in Kumbhakar and Lovell (2000) and this requires only trivial sub-script amendments to form the required likelihood functions for the variants of the model discussed in 3.5. It is assumed that the distribution of the inefficiency term is:

[42] Provided, in the GLS case, the regressors and μ_{it} are uncorrelated as discussed earlier.

$\mu_{it} \sim N^{+}\left(\pi_i, \sigma_{\mu}^{2}\right)$ when independence over time is assumed for inefficiency and $\mu_{it} = g_i(\boldsymbol{\delta}'\mathbf{Z_{it}}) \cdot \mu_i$ with $\mu_i \sim N^{+}\left(\pi_i, \sigma_{\mu}^{2}\right)$ when dependence of inefficiency over time is allowed for.

For all of these models, except the model which assumes independence over time of inefficiency, a prediction of firm inefficiency is given by the conditional expectation of the inefficiency component and is amended from Greene (2008) and given below:

$$E[\mu_{it} \mid \varepsilon_i] = g_i(\cdot)E[\mu_i \mid \varepsilon_i] = g_i(\cdot)\left[\tilde{\mu}_i + \tilde{\sigma}_i\left(\frac{\phi(\tilde{\mu}_i/\tilde{\sigma}_i)}{\Phi(\tilde{\mu}_i/\tilde{\sigma}_i)}\right)\right] \tag{6.16}$$

Where $\varepsilon_i = \varepsilon_{i11}, \ldots, \varepsilon_{i1S(i)}, \varepsilon_{i2S(i)}, \ldots, \varepsilon_{iT(i)S(i)}$, $\tilde{\mu}_i = \dfrac{(1-\gamma)\pi_i - \gamma\sum_{t=1}^{T(i)}\sum_{s=1}^{S(i)} g_i(\cdot)(d\varepsilon_{its})}{(1-\gamma) + \gamma\sum_{t=1}^{T(i)}\sum_{s=1}^{S(i)}(g_i(\cdot))^2}$,

$\tilde{\sigma}_i^2 = \dfrac{\gamma(1-\gamma)\sigma^2}{(1-\gamma) + \gamma\sum_{t=1}^{T(i)}\sum_{s=1}^{S(i)}(g_i(\cdot))^2}$, $\gamma = \sigma_{\mu}^2 / \sigma^2$, $\sigma^2 = \sigma_{\mu}^2 + \sigma_v^2$,

$$d = \begin{Bmatrix} 1 \text{if production function} \\ -1 \text{if cost fucntion} \end{Bmatrix}$$

The equivalent estimator for the case of independence across time is a trivial adaptation of the estimator presented in Battese and Coelli (1988) (summation over s rather than over t) and so it is not presented here.

6.4.3. An alternative approach for accounting for correlation between μ_{it} and the regressors

In the dual-level inefficiency model, the interpretation of μ_{it} is that it represents firm inefficiency. Clearly, it could also capture some unobserved sub-company invariant factors. A further issue arises when μ_{it} is found to be correlated with the regressors. As Farsi et al (2005b, footnote 16 p. 2131) argue, it is not clear why cost inefficiency would be correlated with the regressors, since cost inefficiency here does not take into account such things as scale inefficiency and conceptually is a measure of the incompetence of management to manage resources given the outputs and input prices that the firm faces. Thus, if regressors are correlated with μ_{it}, then it can be argued that such correlation arises due to other unobserved sub-company invariant factors apart from inefficiency.

Assuming that unobserved heterogeneity is correlated with regressors, but inefficiency is not, a natural question to ask is whether unobserved heterogeneity can be separated out from inefficiency. One approach would simply be to impose the no correlation assumption and adopt a random effects approach for μ_{it}. This is unsatisfactory since it simply ignores the correlation, thus random effect estimation (1st stage) of the model parameters will be biased. An alternative approach is to partition the fixed effect into two parts:

$$\alpha_{it} = \boldsymbol{x'}_{\boldsymbol{its}}\boldsymbol{\rho} + \delta_{its} \qquad \delta_{its} \sim iid\left(0, \sigma_{\delta}^{2}\right) \tag{6.17}$$

The first term captures the correlation between α_i and the regressors and represents the impact on the fixed effects from unobserved heterogeneity. The second term is a random variable which is uncorrelated with the regressors and (once transformed by the Schmidt and Sickles operation (6.14)) is assumed to represent inefficiency which is persistent across all sub-company units within a given firm. This decomposition was first proposed by Mundlak (1978) to highlight the relationship between fixed effects and random effects estimation. The relationship in (6.17) becomes highly useful when it is averaged over sub-company units:

$$\alpha_{it} = \bar{\mathbf{x}}'_{\mathbf{it}}\boldsymbol{\rho} + \delta_{it} \qquad \delta_{it} \sim iid\left(0, \sigma_{\delta}^{2}\right) \tag{6.18}$$

(6.18) has become known as the Mundlak transformation (Farsi et al, 2005a and 2005b). Now substitute (6.18) into (6.6):

$$lnC_{its} = \bar{\mathbf{x}}'_{\mathbf{it}}\boldsymbol{\rho} + \delta_{it} + f(\mathbf{x}_{\mathbf{its}}; \boldsymbol{\beta}) + \tau_{its} + v_{its} \tag{6.19}$$

where $\delta_{it} \sim iid\left(0, \sigma_{\delta}^{2}\right)$, $\tau_{its} \sim iid|N(0, \sigma_{\tau}^{2})|$, $v_{its} \sim iidN(0, \sigma_{v}^{2})$ and $\delta_{it} \perp \tau_{its} \perp v_{its}$

Thus the model simply has k additional regressors (relative to the fixed effects model), with the inefficiency components being random variables. The model can be estimated in three ways. First, GLS random effects can be applied to (6.19), and then ML estimation can be applied to the residuals from this regression (as is the case for the multi- stage random effects formulation described in sub-section 6.4.1).

Second, a three stage approach can be adopted. Firstly, a fixed effects regression is undertaken, then, secondly, the fixed effects are regressed on the group means of the regressors. The residuals of this decomposition then form the predictor of δ_{it}. Through application of the transformation in (6.14), sub-company invariant (external) inefficiency can be predicted. Thirdly, the residuals from the original fixed effects regression can be used (in the same way as detailed for the multi-stage approach in sub-section 6.4.1) to predict sub-company varying (internal) inefficiency via a further ML estimation.

Third and the final way to estimate (6.19), is to estimate (6.19) via the single stage 'True Random Effects' procedure as detailed in sub-section 6.4.1.

Importantly (and the fundamental result of Mundlak's paper), the estimates of the $\boldsymbol{\beta}$ using the multistage procedure are identical to the fixed effects approach. As such they are consistent (unlike running random effects ignoring the correlation). The implication is that the sub-company varying (internal) inefficiency from this model (estimated via the multi-stage method only) is the same as from a fixed effects formulation without the Mundlak decomposition. The single stage estimation method is non-linear and it is not necessarily the case that the fixed effects estimates will coincide with the estimates from (6.19) (Farsi et al, 2005b, footnote 6, p. 2128).

The benefit of this approach is the orthogonal decomposition of the fixed effects; the value of which is dependant on the researcher's view as to whether inefficiency is correlated with the regressors or not. If they are willing to assume it is not correlated with the regressors, which Farsi et al (2005b) argue is reasonable, then it would seem that the Mundlak approach is a useful way to 'strip out' an element of unobserved heterogeneity from the prediction of sub-company invariant (external) inefficiency.

Finally, note that if $\boldsymbol{\rho} = \mathbf{0}$ then the model in (6.19) becomes the random effects formulation of (6.6). As such the joint test of the null hypothesis of $\boldsymbol{\rho} = \mathbf{0}$ is a test of random effects against fixed effects i.e. an alternative test statistic of the Hausman test (Greene, 2012, p. 421). This can be tested using a Wald test and is known as a (Wu) variable addition test.

An important caveat to this approach is to note that for (6.17) to be an identified relationship, there has to be sufficient unique values of $\bar{x}_{it}$ to identify $\boldsymbol{\rho}$ and σ_{δ}^{2}. This is equivalent to there being at least k+1 fixed effects (in the equivalent fixed effects dummy variable model) and this may not be a trivial requirement when using sub-company data since there may be data on only a small number of firms and/or a wish to impose time invariant inefficiency effects. This requirement is indeed an issue in the empirical application (section 6.5) where there are only 5 firms, but 5 regressors and time invariant inefficiency. Thus there are only 5 unique values of each group mean, which is not sufficient to identify 6 parameters.

One option to implement the approach by excluding one group mean. It was decided not to take this approach forward into the empirical example. The main concern was that it would be arbitrary which group mean was dropped and this still would mean there are only just enough unique observations to identify the extra parameters. Further, when such an approach was implemented in statistical software (LIMDEP v9 (Econometric Software, 2010)), the standard errors did not seem plausible (relative to the estimation output from the fixed effects model) and, further, the test statistic for the Wu variable addition test was highly volatile to the random effects estimation routine applied.[43]

The issue of small N relative to k, is one limitation of the sub-company data structure. One of the key reasons put forward in section 6.1 for using sub-company data is that N is often small (however it should be noted that there are other reasons). The Mundlak transformation is a useful device which would appear to better disentangle inefficiency from unobserved heterogeneity. Thus, further work to understand the robustness of the transformation when N is small (or not substantially larger) relative to k would be of value, however this is out of scope for this thesis.

6.4.4. An alternative means to estimate the random effects dual-level inefficiency model

It should be noted that the paper referred to in sub-section 6.2 by Brorsen and Kim (2013), does consider, as an extension to their work, a dual-level inefficiency model (see the Appendix of their paper)[44]. Their model is estimated using the results relating to Closed Skew Normal (CSN) distributions. Essentially, such distributions can be used to provide an analytic expression for the true random effects type models explained in sub-section 6.4.1. See Brorsen and Kim (2013) for a survey of such distributions.

[43] This was the case as well when Eviews v6 (Quantitative Micro Software, 2007) was used for comparison.

[44] It should be noted that the Smith and Wheat (2012b) paper, on which this chapter is based, was published before the Brorsen and Kim paper.

A CSN distribution can, in theory, be maximized directly with respect to parameters, rather than appealing to the simulation methods outlined in Greene (2005) and sub-section 6.4.1. In practice however, such maximization is difficult in the case multiple regressors. Brorsen and Kim (2013) state,

"*We demonstrate estimating such a model on a very simple example. Because of the difficulties in estimating the closed skew normal distribution, the focus of our paper is not about estimating such models. Our paper is trying to determine what would happen if such a model was the true model and a stochastic frontier model was estimated with the aggregated data*." p. 27.

Brorsen and Kim also state that the simulation approach by Greene (2005) (one of the methods proposed in this chapter) is a more feasible way to estimate the model. The use of CSN distributions is relatively new to stochastic frontier modelling and, given the difficulties in estimation, are not taken forward into the empirical example section of this chapter.

6.5. Application to international railway infrastructure comparisons

6.5.1. Context

The dual-level inefficiency model is now applied to data on five railway infrastructure managers, comprising firms from North America alongside European national infrastructure managers (IMs). A railway infrastructure manager is responsible for the management (maintenance and renewal) of the railway infrastructure (permanent way, structures, lineside equipment and stations and depots). An infrastructure manager is different conceptually from a train operator who actually runs the train services. In the case of Britain, the infrastructure manager is institutionally separate from train operating companies. For the other companies, the IM also runs the train services but, importantly, separate accounts are available for the IM side and also the structure of the companies is such that the two functions can be considered divorced in terms of business organisation.

This analysis builds on a previous study conducted for the British Office of Rail Regulation (ORR) as part of the 2008 Periodic Review of the British infrastructure manager's efficiency performance[45]. In that work, which was exploratory in nature, and based on a smaller sample than is now available, the authors estimated the simplest, single-level efficiency versions of the models presented in this Chapter (namely the pooled and sub-company invariant models; see sub-section 6.3.3 and 6.3.4 respectively).

Each IM in the sample is divided into a number of regions. The number of regions per IM (S(i) using the terminology in section 6.3) ranges from 3 to 18. The difference in the number of regions per IM reflects both the availability of data (in respect of the number of years available for each firm) and also, importantly, the organisational structure of the IM. Thus the definition of regions for each IM is such that it is expected that there exists some management autonomy at the regional level as well as at the firm head office level. Hence, there is a need at least to consider a

[45] See Smith et. al. (2008) and ORR (2008) for details of the work undertaken. Note that the list of railway companies considered are slightly different in the analysis for this Chapter than in the Periodic Review analysis.

dual-level inefficiency model.

As noted in section 6.3, it is beneficial for both efficiency performance analysis and more generally cost analysis to analyse data at a level of geographical aggregation that corresponds to how firms organize their activities. This allows both for any dual-level inefficiency to be captured, but also allows for the true scale and density properties of the cost frontier to be established. Thus while the range of regions per IM may appear large, this is partly due to the overall size differences of the IMs considered. Furthermore, assurances have been received from the participating IMs that these breakdowns have degrees of autonomy, thus making it appropriate to analyse efficiency at this level.

For some IMs the dataset is supplemented by having repeat observations over time (T(i) ranges from 1 to 5). The panel covers the period 2002 to 2007, though is unbalanced in time as noted. Overall there are a total of 89 observations on the five IMs. As discussed in section 6.3.5, an assumption about how inefficiency behaves over time is required in this case. Given the unbalanced nature of the observations over time and the generally small number of time periods for most IMs, a time invariant model is adopted. Thus both the firm and sub-company inefficiency components are time invariant in the model.

The data structure enables the investigation of efficiency variation between rail systems in different countries, whilst also looking at inefficiency at the sub-company level within each country. The use of sub-company data also expands the sample size substantially without the need to collect a long panel. The utilisation of sub-company data can thus be seen as interesting and important in an international benchmarking context where cross-sections may well be small and panels short.

For data confidentiality reasons, data is anonymised and it is not possible to publically reveal which firms were involved in the study and what the corresponding efficiency score was for each firm. The firms are however identified in a confidential Appendix for the benefit of the examiners (subsequently removed from this published version).

6.5.2. Data

The data is summarised in Table 6.1. The dependent variable is maintenance cost, comprising all elements of railway infrastructure maintenance (e.g. permanent way, structures and signalling). Note that in railway accounts, maintenance is distinct from renewals activity, where renewals expenditure is the like-for-like replacement of assets following life expiration, and maintenance expenditure is the day to day upkeep of the assets to keep them in safe and operable condition. Whilst there could be definitional differences between countries which affect this variable (as is the case in any international study) as part of the data collection process, considerable efforts were made to harmonise definitions across countries which adds to confidence. Country specific cost data is converted into US dollars using purchasing power parity (PPP) exchange rates and data is also converted to 2006 constant prices.

The explanatory variables comprise tonne density, defined as gross tonne-km per track-km (TTKD) and track-km (Track) for outputs in order to account for scale and density effects. In addition, the proportion of track length that is electrified (ProElect) is included as a proxy for the quality of the infrastructure. Price indices for capital between countries are not available, but the PPP exchange rate adjustment should account for some of the differences across countries. Wage rate data is used for each of the IMs. However, it should be noted these are company-wide rather than sub-company specific and that in some cases the data is based on all staff employed by the railway, not just infrastructure maintenance. Thus the Wage variable is relatively crude and as such the sensitivity of the results to its inclusion is discussed in the results. The data is normalised to the sample mean which implies the coefficients on the first order variables represent elasticities at the sample mean[46].

Table 6.1 Summary of data used in the study (un-normalised data)

Variable	Mean	Standard Deviation	Min	Max
Maintenance Cost	43,801,077	28,162,452	9,103,240	114,210,161
Tonne Density (Tonne-km / Track-km) (TTKD)	8,059,323	6,157,594	1,077,481	21,808,976
Track-km (Track)	928	588	252	2,988
Proportion of track-km electrified (ProElect)	0.65	0.41	0.00	1.00
Average staff cost per staff member (Wage)	57,408	9,473	39,791	84,378

Note: costs are in 2006 US $

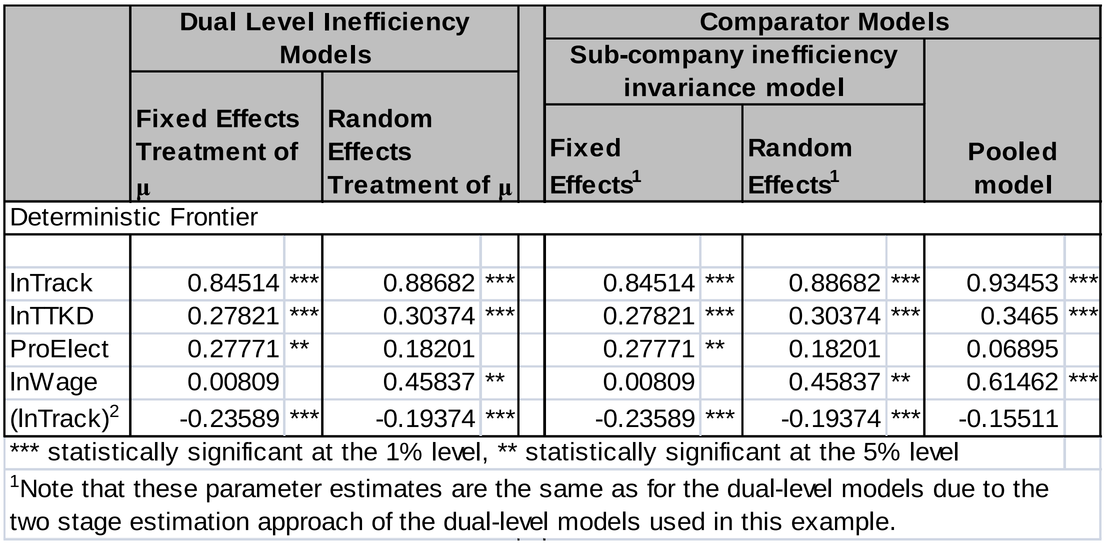
6.5.3. Results

Table 6.2 presents the parameter estimates from the dual-level efficiency models, estimated by assuming the α_i are fixed and random effects in turn (LIMDEP v9 is used to operationalise the multistage fixed and random effects estimation approaches (details of the code are available on request)). The parameter estimates for the two special (nested) cases of the dual-level model as discussed in section 6.3 are also

[46] Note ProElect is not normalised to the sample mean.

presented. First, the sub-company inefficiency invariance model (fixed and random effects cases), which corresponds to the fixed/random effects models used as the first stage in the dual-level model. Second, the special case where inefficiency is only sub-company varying (no persistent, firm-specific effects) is shown, which is referred to as the pooled model in line with the terminology used in section 6.3[47].

Table 6.2 Parameter estimates for dual-level Inefficiency models and comparator models



	Dual Level Inefficiency Models				Comparator Models					
					Sub-company inefficiency invariance model				Pooled model	
	Fixed Effects Treatment of μ		Random Effects Treatment of μ		Fixed Effects[1]		Random Effects[1]			
Deterministic Frontier										
lnTrack	0.84514	***	0.88682	***	0.84514	***	0.88682	***	0.93453	***
lnTTKD	0.27821	***	0.30374	***	0.27821	***	0.30374	***	0.3465	***
ProElect	0.27771	**	0.18201		0.27771	**	0.18201		0.06895	
lnWage	0.00809		0.45837	**	0.00809		0.45837	**	0.61462	***
$(\ln Track)^2$	-0.23589	***	-0.19374	***	-0.23589	***	-0.19374	***	-0.15511	

*** statistically significant at the 1% level, ** statistically significant at the 5% level

[1]Note that these parameter estimates are the same as for the dual-level models due to the two stage estimation approach of the dual-level models used in this example.

The functional form was chosen by first estimating a Translog and then testing down. The vast majority of second order terms had very low t statistics and in addition to the squared track term, only an interaction term between wage and track was significant at any reasonable significance level. However inclusion of this term yielded a model with implausible negative wage elasticities for many observations within the sample. For this reason, this term was dropped. Importantly, the joint restriction that all of the omitted second order terms (including the wage/track

[47] Note that while the terminology "pooled model" accurately describes the pooled nature of the data over sub-companies, it should be noted that time invariance is assumed. As such the model is actually an analogue to the time invariant model first proposed by Pitt and Lee (1981).

interaction) were equal to zero could not be rejected at any reasonable significance level (e.g. Wald test in random effects model treatment gave a statistic value of 12.12 and an associated p value of 0.19804 (9 degrees of freedom)). As such it is concluded that the specification is both a useful and intuitive economic model of the underlying cost characteristics while its parsimony is supported by the data.

Turning to the choice of fixed versus random effects, it should be firstly noted as discussed in section 6.3, this refers to the persistent, firm-specific effect in the model (α_i). The Hausman test gives a p value of 0.0861 which indicates a preference for random effects at the 5 per cent significance level. Nonetheless, the fixed effects results are still reported for comparative purposes to compare with the random effects results, given that evidence of correlation between effects and regressors is found at the 10% level.

A key feature of any cost frontier is how costs change with output. As explained in the literature review of applications to railways (Chapter 4), it is usual to distinguish between returns to scale (RtS) and returns to density (RtD). The inverse of the sum of the coefficients on lnTrack and lnTrack^2 give the estimates of RtS and because the data is normalised to the sample mean, the inverse of the coefficient on lnTrack gives RtS at the sample mean. The inverse of the coefficient on lnTTKD give the estimate of RtD in this model. For the preferred random effects treatment of 1.13 (at the sample mean) and 3.29 for RtS and RtD respectively. The two separate restrictions of constant RtS and RtD (=1) can be rejected for each at the 5% and 1% levels respectively (p values of 0.0156 and 0.0000 respectively). Thus the cost frontier indicates moderate RtS and large RtD. RtD being greater than RtS is intuitive given the fixed costs of providing railway infrastructure (for reference see the results and supporting commentaries in the references in Table 6.3).

Table 6.3 compares the estimates of RtS and RtD from this study to those studies considered in Chapter 4 for infrastructure managers. In terms of RtS, the estimate is towards the lower end of the received literature (although all studies indicate increasing RtS point estimates). The estimate of RtD is consistent with the received literature and indicates that 30% of maintenance cost is deemed variable with usage. This is in line with the recommendations from the CATRIN project which also recommended 30% of maintenance cost to be variable with traffic (Wheat et al, 2009) for the purpose of setting access charges based on marginal wear and tear cost.

Table 6.3 Estimates of Returns to Scale and Density from infrastructure maintenance cost studies

Study	Country	Returns to Scale	Returns to Density
This chapter	**International study**	**1.13**	**3.29**
Johansson and Nilsson (2004)	Sweden	1.256	5.92
Johansson and Nilsson (2004)	Finland	1.575	5.99
Tervonen and Idstrom (2004)	Finland	1.325	5.74-7.51
Munduch et al (2002)	Austria	1.449-1.621	3.70
Gaudry and Quinet (2003)	France	Not reported	2.70
Andersson (2006)	Sweden	1.38	4.90
Wheat and Smith (2008)	Britain	2.074	4.18
Smith et. al. (2008)	International study	1.11	3.25
NERA (2000)	US	1.15	2.85

Source: Amended from Wheat and Smith (2008)

The *a priori* sign of ProElect is ambiguous given the extent to which the variable is a proxy for track quality (i.e. higher quality track might be expected to have lower maintenance costs). On the other hand, electrification means that there are more assets to maintain, makes access to the infrastructure more complex and may also be associated with higher speed services which increases cost. Thus the positive coefficient on ProElect (only significant in the fixed effects model) is neither in line nor at odds with prior expectations. The literal interpretation of the coefficient in the random effects model, given that ProElect is a proportion variable, is that electrifying the network (from 0% to 100% track-km electrified) increases maintenance costs by exp(0.18201)-1=20%.

The coefficient on the wage variable is statistically significant in the random effects model. It is believed that the wage coefficient is insignificant in the fixed effects model since this variable is invariant for each IM at a given point in time. Thus it is likely there is some correlation between this and the fixed effects. However, in both models, the null hypothesis that the coefficient is different from the average labour cost share (65%)[48] fails to be rejected even at the 10% level. Dropping the wage variable does not seem to affect the estimates of the deterministic cost frontier.

Overall, the results show that the parameter estimates are in line with expectations and previous evidence, thus giving confidence in the resulting efficiency findings, which are now discussed.

Firstly, the statistical significance for each of the inefficiency components within the model is considered[49]. The persistent, firm-specific inefficiency effects are modelled

[48] Owing to lack of data, this is an estimate based solely on Network Rail data.

[49] As noted in sub section 6.5.1, given the unbalanced nature of the observations over time and the generally small number of time periods for most IMs, both the firm-specific and sub-company inefficiency components are time invariant in the model.

as either fixed or random effects, the latter being estimated by generalised least squares in the two-stage approach that is adopted here. As such LR tests are not undertaken for whether the variance parameters are zero as these are not estimated in this estimation framework. Instead an F test is used to test the joint significance of the fixed effects and an LM test for the appropriateness of a model without effects. The F test has a value of 5.34 which yields a p value of 0.00073. The conclusion of the test is that the fixed effects are jointly statistically significant.

For the LM test, the Moulton/Randolph standardised form (SLM, Moulton and Randolph (1989)) is adopted which is appropriate for unbalanced panels and is a one sided test (the variance of the random effect can only be non-negative). Thus it can be expected that the test should have greater power than the more standard Breusch and Pagan (1980) test. The value of the SLM statistic is 4.59 and is distributed standard normal under the null of zero random effect variance. Thus a model with no effects can be rejected at any reasonable significance level. Thus all of the tests provide evidence of significant persistent firm-specific effects. These are then transformed into persistent efficiency scores via a Schmidt and Sickles (1984) transformation as described in section 6.4.

Turning now to the statistical significance of the sub-company varying inefficiency term, in the two stage approach adopted for this example, this term is estimated by maximum likelihood. As such an LR tests is undertaken with respect to the significance of the variance parameter of the inefficiency distribution. For the dual-level random effects model, the LR statistic is 18.15 and for the dual-level fixed effects model the LR statistic is 33.23. As described in Coelli et al (2005), this statistic has a non-standard mixed chi square distribution (1 degree of freedom). The large statistic values mean that in both cases the null hypothesis of zero variance is rejected at any reasonable significance level. As such it is concluded that the data set exhibits dual-level inefficiency.

6.5.3.1. Efficiency scores from the dual-level models

Table 6.4 shows overall firm efficiency scores for each infrastructure manager. It also decomposes the efficiency scores into the two components; persistent and sub-company varying. As explained in section 6.3, these two components can be interpreted as the degree of external and internal inefficiency respectively. In this example, the average persistent efficiency scores for the dual-level models are 0.849, 0.835 and 0.840 (random and fixed effects formulations respectively), and 0.851 and 0.690 for the sub-company varying component (random and fixed effects formulations respectively). Thus the random effects formulation points to roughly equal external and internal components, while the fixed effects formulations point to more internal than external inefficiency. As discussed earlier, at the 5% significance level the random effects results are preferred due to the result of the Hausman test. Overall firm efficiency is the product of the two components, and is higher, on average, for the random effects dual level model (0.724) than for the fixed effects alternatives (0.564). The overall efficiency scores for the preferred random effects dual level model are within plausible ranges while the fixed effects scores appear slightly low.

Table 6.4 Summary of efficiency results

	Dual-Level Inefficiency Models			Comparator Models		
				Sub-company inefficiency invariance model		
Firm	Fixed Effects Treatment of μ	Random Effects Treatment of μ		Fixed Effects	Random Effects	Pooled model
Persistent Efficiency Score - External Efficiency						
1	1.000	1.000		1.000	1.000	1.000
2	0.770	0.880		0.770	0.880	1.000
3	0.925	0.840		0.925	0.840	1.000
4	0.617	0.687		0.617	0.687	1.000
5	0.862	0.839		0.862	0.839	1.000
Average	*0.835*	*0.849*		*0.835*	*0.849*	*1.000*
Sub-company Varying Efficiency Score - Internal Efficiency						
1	0.621	0.881		1.000	1.000	0.916
2	0.734	0.857		1.000	1.000	0.879
3	0.593	0.819		1.000	1.000	0.779
4	0.853	0.849		1.000	1.000	0.761
5	0.649	0.850		1.000	1.000	0.830
Average	*0.690*	*0.851*		*1.000*	*1.000*	*0.833*
Overall Efficiency Score						
1	0.621	0.881		1.000	1.000	0.916
2	0.565	0.754		0.770	0.880	0.879
3	0.549	0.688		0.925	0.840	0.779
4	0.527	0.583		0.617	0.687	0.761
5	0.560	0.713		0.862	0.839	0.830
Average	*0.564*	*0.724*		*0.835*	*0.849*	*0.833*

6.5.3.2.Comparator models

The comparator models are the pooled model and sub-company inefficiency invariance model. The former assumes that there is no persistent inefficiency within firms, and the inefficiency of each sub-company is assumed to be identically and independently distributed across all sub-company units irrespective of the firm to which they belong. The second comparator model comprises persistent, firm-specific effects only, representing the case where there is no variation in efficiency performance between sub-company units within the same firm (the model parameters for these models are simply those for the dual-level models reported in

Table 6.2). It should be noted that the average overall firm efficiency is considerably lower using the dual-level model as compared to all three of the comparator models. This is because the comparator models are constrained models and only consider one source of inefficiency. As discussed above, both restrictions are rejected for this dataset so the dual-level models are preferred.

In summary, this empirical example has demonstrated the possibility of separating firm inefficiency into a persistent and a sub-company varying component. Thus the dual-level model can be applied successfully to real data.

It also shows that the failure to account for the dual-level nature of inefficiency, for example, by estimating one of the three, simpler comparator models, may cause overall firm inefficiency to be systematically under predicted. This is an additional issue to the influence of aggregation bias on inefficiency predictions discussed in sub-section 6.3.2 (since in this empirical example, the correctly aggregated scale measure is included in the deterministic frontier). Here the under prediction emerges since some residual error in the comparator models (be it sub-company varying (in the case of the sub-company invariance model) or sub-company invariant (in the case if the pooled model)) is not being characterised as inefficiency and will instead be capture by the random noise term.

6.6. Summary

This chapter has outlined econometric techniques to analyse data on geographical regions of multiple companies. This is termed ‘sub-company’ data. This is data which both exists and is desirable for performance analysis of horizontally integrated elements of vertically separated railways. This is commonly the infrastructure managers and at a simple level can be used to overcome the problem of few external comparator observations.

Importantly, as well as providing extra observations to estimate cost frontiers, the data structure permits estimation of a dual level inefficiency model which separates sub-company persistent inefficiency from sub-company varying inefficiency. This is a useful decomposition in itself, since it gives an indication as to whether inefficiency predominantly varies within firm or between firms which is useful in terms of identifying where efforts should be made to eliminate inefficiency. In addition to benefits with respect to the capturing of inefficiency variation, since management decisions are often made (to some extent) at sub-company levels within firms, it is likely that failure to disaggregate data to this level will lead to bias in returns to scale results. This is because the inappropriate aggregate measure of scale is used in an aggregate firm analysis. Further this will also bias inefficiency predictions as these are transforms of the residuals.

The techniques have been applied to railway infrastructure managers. This has shown that there is statistically significant dual-level inefficiency. The empirical example also indicates that failure to take account of the dual level inefficiency variation may result in under estimation of inefficiency. Average efficiency is 0.724 for the preferred model, comprising 0.849 and 0.851 for the persistent and sub-company varying components respectively.

Appendix 6 – Identification of infrastructure managers included in the empirical example (Confidential)

[This has been removed from the published thesis for reasons of commercial confidentially]

7. Uncertainty in efficiency analysis[50]

7.1. Introduction

It is common in efficiency analysis to present point predictors for firm efficiency. This was the approach in the preceding chapter. However, less common is the computation of interval predictors for firm efficiency (Simar and Wilson, 2010). As highlighted in section 3.4 of the econometric literature review and developed further in this chapter, there are many sources of uncertainty associated with predictors of firm inefficiency. Thus it is reasonable to suppose that under some circumstances such uncertainty is non-trivial and it would be wise for practitioners to present interval predictions alongside point predictions.

This chapter develops interval predictors for firm efficiency for cross sectional stochastic frontier models. That is the models are fully parametric in the sense that full distributional assumptions are applied to all error components. At the outset it is important to note how this fits into the analysis elsewhere in the thesis. Clearly Chapter 5 does not consider efficiency measurement. Chapter 6 does consider efficiency measurement, but it should be noted that the models developed are using a multi-dimensional data set (analogous to panel data). As such the discussion in this chapter is not strictly applicable, although the techniques can easily be applied to panel data settings.

[50] This chapter is based on an early draft of a paper with William Greene and Andrew Smith (Wheat, Greene and Smith, 2013) but differs from the paper in that i) there is an extended introduction, ii) it provides more clarification on the nature of prediction intervals for firm specific inefficiency estimates in section 6.2 and iii) it provides more interpretation of the quantitative results empirical example.

7.2. Methodological Background

There are a number of means to evaluate the appropriateness of a stochastic frontier model. Firstly, the econometric approach provides estimates of the frontier. These can be compared with *a priori* expectations and should they match or at least not be inconsistent with expectations, give reassurance that the frontier is appropriate, which is important given that the frontier is what efficiency is measured against. Secondly, the rankings and summary statistics (mean, maximum and minimum) of the firm efficiency scores can be compared with the researchers and/or industry's *a priori* expectations as to performance in the sector.

The above methods clearly fit into the category of comparing statistical results to underlying economic/industry expectations. Complementing these diagnostics are various statistical inference procedures that can help the researcher understand the uncertainty associated with estimates. The obvious candidates are hypothesis tests on the model parameters such as those defining the frontier and the variance parameters on the inefficiency distribution. These provide useful information as to whether the frontier parameters are statistically significant and whether there is statistically significant levels of inefficiency in the sample as a whole.

A further set of statistical techniques relate to understanding uncertainty in firm specific estimates of inefficiency. Many empirical studies (and the study in chapter 5) have simply reported point estimates for firm inefficiency. There exists a wider set of techniques which have been applied to firm specific inefficiency predictions but there is a complication to applying the usual techniques since firm inefficiency is modelled as a realisation of a random variable and not a parameter to be estimated. Understanding the implications of this complication is the subject of this chapter, both in terms of appropriate interpretation of the techniques that have been proposed in the literature and also extensions to the techniques in order for the techniques to truly represent prediction intervals for firm inefficiency, which is the interpretation put forward in this chapter.

To understand better the features of the stochastic frontier model which make intervals for firm inefficiency interesting from a statistical point of view, it should be noted that stochastic frontier models have the attractive property that the unobserved residual component of the model is comprised of both noise and inefficiency. Thus the model can discriminate between observed factors (regressors), noise and inefficiency. Point predictors for firm inefficiency are common in the literature and follow the methodology of Jondrow et al (1982). However in cross sectional models, these point predictors are known to be inconsistent for the quantity of interest; namely the firm specific realisation of a random variable. The question then arises; how precise is the prediction of firm inefficiency? With this in mind, and the general desire of practitioners to understand uncertainty in their estimates, it is perhaps surprising that interval predictors are not commonly reported in the empirical literature.

While a body of literature exists on such intervals, overall it is not clear as to what the properties and limitations are with respect to each innovation. The purpose of this chapter is to clarify, and in places develop, the existing literature on the subject that has grown over the last two decades. The literature is decomposed into two themes. As a starting point, the case where the parameters in the model are known (as opposed to having to be estimated) is considered. The starting point are the intervals proposed by Horrace and Schmidt (1996) (HS intervals). Importantly an interpretation is offered that the correct way to view HS intervals are prediction intervals for firm inefficiency rather than confidence intervals. This point has previously been made either explicitly in the case of Simar and Wilson (2010) and eluded to in both Coelli et al (2005) and Greene (2008). For the purpose of this chapter, it is considered that this distinction is beyond simple semantics. Viewing the HS intervals as prediction intervals does help evaluate some of the further claims made in the literature. For example the chapter shows that the intervals are not confidence (or otherwise) intervals for $E[u_i|\varepsilon_i]$ since $var[u_i|\varepsilon_i]$ is not the variance of $E[u_i|\varepsilon_i]$. An important further implication is that the HS intervals cannot be used for hypothesis testing as implied by Bera and Sharma (1999). It is also shown that the HS intervals do not represent minimum width intervals for u_i and as such are not optimal interval predictors. How to calculate minimum width intervals is then discussed and how such intervals either include or exclude zero as a lower bound depending on where the probability mass of the distribution of $u_i|\varepsilon_i$ resides. This is important since it has useful implications for practitioners and policy makers. In the empirical example, the difference between the two-sided HS intervals and minimum width intervals is illustrated using data from Chapter 5.

The more realistic case where model parameters are estimated is then considered. Thus how, through taking into account additional uncertainty due to estimation of parameters, an interval which is truly analogous to a prediction interval can be developed is outlined. Simar and Wilson (2010) have outlined a method using bootstrapping, but in this chapter, a method which samples from the asymptotic distribution of the parameters is proposed. Irrespective of the method taken to compute the parameter uncertain prediction intervals, it is considered that the reporting of this interval is crucial for understanding uncertainty surrounding firm specific estimates of inefficiency, particularly where sample sizes are small when uncertainty due to parameter estimation may be high.

The remainder of the chapter is structured as follows. Section 7.3 formalises the model under consideration, outlines the HS interval computation, brings together the literature and discusses what are the key features of the received intervals under the assumption that the model parameters are known. Section 7.4 discusses approaches to incorporating parameter uncertainty, which is required for the intervals to be truly analogous to prediction intervals for u_i. Within this, an alternative method based on numerical distributional sampling methods rather than bootstrapping is presented. Section 7.5 presents an empirical example where the proposed intervals are compared with those from the received literature and section 7.6 concludes.

7.3. The case of known parameters

In this section, interval estimation for firm specific inefficiency estimates is considered, assuming that the model parameters are known. Importantly this assumption means that intervals can be derived analytically.

The stochastic frontier model was first proposed simultaneously by Aigner et al (1977) and Meeusen and van den Broeck (1977). The simple cross sectional normal-half-normal formulation for a cost function can be represented as:

$$y_i = f(\boldsymbol{x_i};\boldsymbol{\beta}) + \varepsilon_i,\ \varepsilon_i = u_i + v_i \ \ u_i \sim \left|N\left(0,\sigma_u^2\right)\right|,\ v_i \sim N\left(0,\sigma_v^2\right) \ i=1,\ldots,N \tag{7.1}$$

where y_i is the dependent variable (cost), $\mathbf{x_i}$ is a vector of regressors, parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u^2, \sigma_v^2)$, u_i is the one sided cost inefficiency random variable and v_i is the symmetric noise variable. In order for u_i to yield a Farrell (1957) type radial measure of efficiency, it is assumed that the dependent variable is in logarithms. Estimation usually proceeds via maximum likelihood which is a consistent and efficient estimator given the distributional assumptions.

Jondrow et al (1982) provided the expression for the conditional density of u_i given the realised value of the composite error term ε_i, $(u_i|\varepsilon_i)$ under different unconditional distributional assumptions for u_i. For the model in (7.1), the density is

$$(u_i|\varepsilon_i) \sim N^+(\mu_{i*}, \sigma_*^2)$$

$$\mu_{i*} = \frac{\varepsilon_i \sigma_u^2}{(\sigma_u^2+\sigma_v^2)} \tag{7.2}$$

$$\sigma_*^2 = \frac{\sigma_v^2 \sigma_u^2}{(\sigma_u^2+\sigma_v^2)}$$

Jondrow et al suggest the mean or mode of this distribution as a point estimator of firm inefficiency. Waldman (1984) provides theoretical results for cross sectional models that show that the Jondrow et al (1982) conditional expectation estimator is superior to the best linear estimator in terms of correlation with the underlying inefficiency error and mean-squared error. Intuitively, this is because the Jondrow et al estimator exploits the distributional information available while the best linear estimator does not. As such, there is some theoretical justification for some point estimators of inefficiency over others. However, in cross sectional models, and indeed pooled panel models, the variance of the distribution in (7.2) does not tend to zero as the sample size increases[51]. Thus point estimators of firm inefficiency are inconsistent, even though $E[\widehat{u_\imath|\varepsilon_\imath}] \underset{p}{\to} E[u_i|\varepsilon_i]$ as the sample size increases. The limitation of the Jondrow et al estimator is not that the sample estimator is inconsistent for its population equivalent but that the population equivalent is not the statistic of interest to the researcher. $E[u_i|\varepsilon_i]$ is just one point within the distribution of $u_i|\varepsilon_i$; understanding the spread of this distribution should also be of interest. As such interval prediction for u_i is important.

[51] Notable exceptions are point estimates of firm inefficiency from the class of time invariant panel models (Pitt and Lee, 1981) and the class of deterministically time varying models (Batesse and Coelli, 1992 and Cuesta, 2000) which yield consistent estimates as $T \to \infty$. For the purpose of this chapter, however, attention is restricted to cross sectional models (or equivalently pooled panel models).

Horrace and Schmidt (1996) provided expressions for interval brackets of the conditional distribution in (7.2) and propose two-sided intervals of the conditional distribution of firm inefficiency as a measure of the uncertainty surrounding firm specific inefficiency. The α-percentile is given by (see Simar and Wilson (2010) and Horrace and Schmidt (1996) for derivation):

$$\rho_{\alpha} = \mu_{i*} + \sigma_{*}\Phi^{-1}\left[1 - (1 - \alpha)\Phi\left(\frac{\mu_{i*}}{\sigma_{*}}\right)\right] \qquad (7.3)$$

where $\Phi(\cdot)$ is the standard normal CDF and $\Phi^{-1}[\cdot]$ is the standard normal quantile function.

Horrace and Schmidt (1996) consider cross sectional models as well as time invariant panel models, while Hjalmarsson et al (1996) extended the Horrace and Schmidt (1996) analysis to produce intervals for time varying inefficiency models. Bera and Sharma (1999) derived the same intervals. Simar and Wilson (2010) note that Taube (1988) proposed similar intervals to Horrace and Schmidt with the expectation that they were one-sided intervals.

The first observation, and one from which the rest of the results follow, is that the intervals given in (7.3) do not correspond to the usual definition confidence intervals for firm inefficiency. As discussed in Simar and Wilson (2010, footnote 9), but also alluded to in Coelli et al (2005) and Greene (2008), Horrace and Schmidt (1996) incorrectly used the terminology ‘confidence intervals’ when in fact they are prediction intervals for the random variable u_i (and not a parameter), using the information available in the realized composite error term. Importantly, a prediction interval does not collapse in width as $N \to \infty$, which is clearly the case here. The useful distinction emphasized here is that a prediction interval is a statement about likely values of a random variable that is the object of forecast.

Following from the above, and also pointed out in Simar and Wilson (2010), the intervals in (7.3) are therefore not confidence intervals for an estimator of $E[u_i|\varepsilon_i]$. Horrace and Schmidt (1996) did not make this claim but there seems to have been some misinterpretation in the literature (Simar and Wilson, 2010), particularly since this is the point estimator reported commonly in the literature. The HS intervals are brackets for the distribution of $u_i|\varepsilon_i$ only. The variance of the estimator of $E[u_i|\varepsilon_i]$ is not the variance of $u_i|\varepsilon_i$. Given the conditioning on ε_i (and σ_u^2 and σ_v^2), the estimator of $E[u_i|\varepsilon_i]$ is computed with certainty, since all of the parameters comprising $f(u_i|\varepsilon_i)$ are (assumed at this point to be) known – there is no sampling distribution.

A further point, and one not discussed in the literature, given that the intervals are prediction intervals, it is not possible to use them to conduct hypothesis testing on specific values of firm specific inefficiency (random variable). Of particular interest is the null hypothesis of $u_i = 0$ (no inefficiency). This is highlighted since Bera and Sharma (1999) seem to contradict this by providing ‘critical values’ for such an hypothesis test. However, these ‘critical values’ are simply the one sided percentiles of the conditional distribution (corresponding to one minus the significance level). They suggest that the computed $E[u_i|\varepsilon_i]/(var[u_i|\varepsilon_i])^{0.5}$ should be compared to the critical values. However, this is not hypothesis testing for u_i and not even hypothesis testing for $E[u_i|\varepsilon_i]$ given $var\big(\widehat{E[u_i|\varepsilon_i]}\big) = 0 \neq var(u_i|\varepsilon_i)$. To clarify, classical hypothesis testing proceeds by using the distribution of a sample statistic to determine if a null hypothesis regarding an unknown population parameter can be rejected at a given level of statistical significance. Conditioned on a given realisation of ε_i, there is no sampling involved in determining the distribution $f(u_i|\varepsilon_i)$ and thus no sample statistic. Furthermore, there is no unknown parameter(s); u_i is a random variable and all the other parameters comprising the distribution are assumed known. Finally, the probability that a given firm can have zero inefficiency (or any specific value for that matter) is itself zero; this is by construction of the model. To ask whether it is likely that a firm's inefficiency would be 'near' would be best given by an interval estimator, but this is not inference in its familiar sense.

It is suggested that a better candidate for inference regarding firm specific inefficiency estimates might be the set of techniques outlined in Horrace (2005) and Flores-Lagunes, Horrace, and Schnier (2007) which provide inference on the rankings of independent truncated normal distributions. These techniques yield a probability of firms being efficient (relative to other firms in sample). It should be noted, however, that these techniques do not take into account parameter uncertainty and so at present are only valid when the model parameters are known.

7.3.1. Construction of minimum width intervals

The intervals proposed by Horrace and Schmidt and others are central two sided intervals. Because the distribution of $(u_i|\varepsilon_i)$ is asymmetric, these will not be minimum width and so such an interval does not represent an efficient interval predictor for u_i. Greene (2012) discusses how to compute minimum width intervals for distribution functions which are asymmetric or otherwise. He considers the solution to the Lagrangian problem of minimising the width of the prediction interval subject to the interval containing the desired probability mass. Formally, the problem is

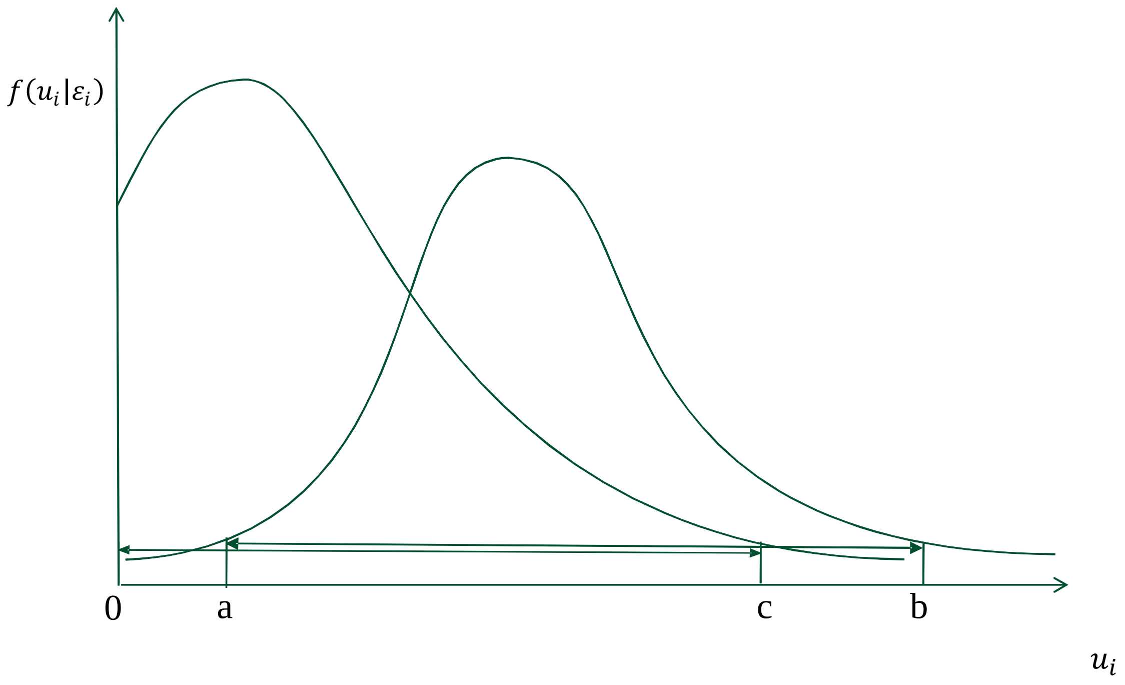
$$min(L,U): U - L + \lambda\big(F(L) + \big(1 - F(U)\big) - \alpha\big) \tag{7.4}$$

where L and U are the lower and upper bounds of the prediction interval respectively, $F(\cdot)$ is the cumulative distribution function of the random variable under consideration (here $(u_i|\varepsilon_i)$), α is the desired significance level and λ is the Lagrange multiplier.

The first order conditions yield two possible solutions (L^*, U^*) depending on the exact shape of the distribution. If $\lambda \neq 0$, (L^*, U^*) are such that $f(L^*) = f(U^*)$ where $f(.)$ is the probability density function. Alternatively, if $\lambda = 0$, the solution will be $L^* = 0$ and U^* such that $1 - F(U^*) = \alpha$; for the case of a negatively skewed distribution truncated at zero. This is the interval corresponding to the one sided interval proposed by Taube (1988). The two solutions are illustrated in Figure 7.1. If the probability mass of the distribution is sufficiently away from zero, then there exists non-zero (L^*, U^*) such that $f(L^*) = f(U^*)$ (represented by a and b respectively in Figure 7.1). Thus this is a minimum width prediction interval for u_i which does not include the zero boundary of the distribution. However if the probability mass is sufficiently close to zero, then no (L^*, U^*) exists such that $f(L^*) = f(U^*)$ and in this case the minimum width interval includes the zero boundary of the distribution, the upper bound being the one sided interval represented by c in Figure 7.1. Note that the minimum width interval corresponds to the highest posterior density (HPD) interval in Bayesian inference.

It is worth emphasising the practical importance of the above result. Under some circumstances the minimum width prediction interval for u_i does not include $u_i = 0$ (no inefficiency) while in other circumstances it does. While this result should not be confused with “inference” (see the discussion above), it is envisaged that adopting such a prediction interval will convey useful information to policy makers/regulators regarding whether a firm suffers from inefficiency or not. Thus the minimum width interval provides useful information regarding whether firm specific inefficiency is likely to be close to zero or not.

Figure 7.1 Possible minimum prediction intervals for u_i



Greene (2012) considers that the easiest way to determine the minimum width interval in a practical setting is via a grid search. Given analytic expressions for the percentiles of the distribution are known (given in (7.3)), this should not be an onerous task. This approach is used in the empirical example.

However, for the normal/truncated-normal, normal/exponential stochastic frontier model it should be noted that $(u_i|\varepsilon_i)$ is distributed truncated normal (see Jondrow et al, 1982). In these cases it is possible to derive minimum width predictive intervals analytically. For a random variable $(u_i|\varepsilon_i)\sim N^+(\mu_{i*},\sigma_*^2)$ the lower and upper bounds of the minimum width predictive intervals are given as:

$$U^* = \mu_{i*} + \sigma_* \Phi^{-1}\left[\left(1-\frac{\alpha}{2}\right)\left(1-\Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right)\right] \tag{7.5}$$

$$L^* = \mu_{i*} + \sigma_* \Phi^{-1}\left[\left(\frac{\alpha}{2}\right)\left(1-\Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right)\right] \tag{7.6}$$

Provided $L^*, U^* \geq 0$, otherwise

$$L^* = 0 \tag{7.7}$$

$$U^* = \mu_{i*} + \sigma_* \Phi^{-1}\left[1-\alpha.\Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right] \tag{7.8}$$

The derivation is given in the Appendix 7A.

It should be noted that one property of minimum width intervals, in contrast to central two-sided or one sided intervals, are that the percentiles that provide the minimum width bounds for one distribution are not necessarily those that provide minimum width bounds for a monotone transformation of the original distribution. In particular, the minimum width bounds for efficiency are not necessarily simple transformations of the minimum width bounds for inefficiency. This analysis has not attempted to provide analytic intervals for efficiency and it is recommended that these are determined numerically via a grid search over the intervals of this distribution. This is relatively easy to undertake, since the α percentile of the distribution of the efficiency distribution is just $e^{(-\rho_\alpha)}$ where ρ_α is as defined in (7.3).

7.4. Incorporating parameter uncertainty into prediction intervals for $\boldsymbol{u_i}$

The received prediction intervals are computed using population parameters that are in reality unknown. Therefore, in finite samples, they are likely to be too narrow. This section is concerned with techniques to incorporate such additional uncertainty.

Before the chapter proceeds to discuss methods regarding how to incorporate parameter uncertainty within the prediction intervals, a brief comment is needed on the underlying distribution of $E[u_i|\varepsilon_i]$. For clarity, it is the sampling properties of a function of $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma_u^2}, \widehat{\sigma_v^2}\right)$ that is the estimator of $E[u_i|\varepsilon_i]$ that is being considered here. The analogous case would be an estimator of $E[y|\boldsymbol{x}]$ in a linear regression context. While $E[u_i|\varepsilon_i]$ is indeed a random variable with a mean and variance (Wang and Schmidt, 2009), only if parameter estimation uncertainty is taken into account is there uncertainty regarding the *estimator* of $E[u_i|\varepsilon_i]$.[52] Simar and Wilson (2010) show how to compute confidence intervals for an estimator of $E[u_i|\varepsilon_i]$ assuming that model parameters are estimated, using the bootstrapped distribution of the model parameters. The resulting intervals are confidence intervals for $E[u_i|\varepsilon_i]$ in the strict statistical sense since $E[u_i|\varepsilon_i]$ is an unknown parameter when ε_i is known (realized), and the sampling distribution of the estimator of it, $E\widehat{[u_\iota|\varepsilon_\iota]}$ is used to provide an interval estimate with the usual resampling properties. In particular, since the population parameters are consistently estimated, the estimator $E\widehat{[u_\iota|\varepsilon_\iota]}$ will converge in probability to the function evaluated at the realized ε_i.

[52] Professor Greene should be credited with this remark.

While such a confidence interval does help to clarify the received literature, it is not what is sought here, because this interval ignores the decompositional uncertainty surrounding u_i (the uncertainty in distinguishing u_i from v_i given ε_i; the exact uncertainty that the distribution of $u_i|\varepsilon_i$ captures). A prediction interval for u_i, itself, will capture both sources of uncertainty and thus give a true reflection of likely upper and lower boundaries of inefficiency for a given firm.

One further interesting digression, not discussed elsewhere in the literature but relevant to the discussion about hypothesis testing in section 7.2, is whether computing a standard error for the estimator of $var[u_i|\varepsilon_i]$ could yield a hypothesis test for the null of no inefficiency for a given firm. In particular, if $var[u_i|\varepsilon_i] = 0$, then the distribution of $u_i|\varepsilon_i$ collapses to zero. A standard error could be computed using bootstrapping or a simulation approach similar to that proposed in 7.4.2. Bera and Sharma (1999) have analysed the expression for $var[u_i|\varepsilon_i]$ and provided useful results. Importantly $var[u_i|\varepsilon_i]$ can only equal zero if σ_u^2=0. Thus any test would reduce to a test for whether the model as a whole exhibits inefficiency rather than the desired test at the firm level.

To date, the only analytical work to incorporate parameter uncertainty with firm inefficiency estimates has been outside the maximum likelihood (ML) stochastic frontier framework. The multiple and marginal comparisons of best approaches (Horrace and Schmidt (1996, 2000) and Kim and Schmidt (2008)) incorporate such variability but these apply to the fixed effects model. Amsler et al (2010) have computed confidence intervals taking into account parameter uncertainty for deterministic frontier models, i.e. models with a single error term rather than a composite error term, under a variety of distributional assumptions. They state however that expressions for "[exact prediction intervals in finite samples are] not possible in stochastic frontier models or in non-parametric models like DEA" (p.5)

Kumbhakar and Löthgren (1998) show by Monte Carlo simulation that applying the HS intervals to cases with parameter uncertainty does not provide the required coverage (e.g. 95% of the realisations of u_i) by some margin for small sample sizes. This is as expected since the intervals need to be wider than that indicated by the Horrace and Schmidt expressions to cover the additional parameter uncertainty. It should be noted that Kumbhakar and Löthgren's model only has one regressor parameter to estimate. The impact of parameter uncertainty may be expected to be greater for models with more regressors.

7.4.1. Simar and Wilson's bagging approach

Two approaches to incorporate parameter uncertainty within prediction intervals for firm inefficiency are considered. Simar and Wilson (2010) outline a method of bootstrapping, known as bootstrap aggregating, or *bagging*, to compute such intervals. Drawing on the notation in their paper, the approach proceeds via the following steps:

Algorithm 1

[1] Maximize the log-likelihood function for the model in (7.1) to obtain ML estimates $\widehat{\boldsymbol{\theta}}$. Recover estimates $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ from elements of $\widehat{\boldsymbol{\theta}}$.

[2] Set b=1.

[3] For $i = 1, \dots, N$ draw $v_i^* \sim N(0, \hat{\sigma}_v^2)$ and $u_i^* \sim N^+(0, \hat{\sigma}_u^2)$ and compute $y_{ib}^* = g(x_i|\widehat{\boldsymbol{\beta}})e^{v_i^* + u_i^*}$.

[4] Using the pseudo-data $\mathfrak{I}_{b,n}^* = \{\mathbf{x_i}, y_{ib}^*\}_{i=1}^{n}$, compute bootstrap estimates $\widehat{\boldsymbol{\theta}_b^*} = argmax_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathfrak{I}_{b,n}^*)$.

[5] Draw u_{ibk}^*, $k = 1, \dots, K$, from $f(u_i|\widehat{\boldsymbol{\theta}}_{\mathrm{b}}^*, \mathbf{x_i}, y_i)$ given in (7.2), for each $i = 1, \dots, N$.

[6] Increment b by one

[7] Repeat steps [3]-[6] B times, yielding a set $\mathfrak{U} = \{u_{ibk}^*\}_{b=1,k=1}^{B,K}$ of BK values for each $i = 1, \dots, N$.

[8] Form the set $\mathfrak{U}_i^* = \{u_{ibk}^* | u_{ibk}^* \in \mathfrak{U}_i, u_{ibk}^* \neq 0\}$ for each $i = 1, \dots, N$.

[9] Order set $\mathfrak{U}_i^*$ and compute the percentiles of interest.

Prediction intervals for each observation i are then computed by ranking the elements of set $\mathfrak{U}i^*$ and choosing the appropriate percentiles as the lower and upper bounds. Essentially, like all bootstrapping procedures, the validity of this approach relies on the distribution $(\widehat{\boldsymbol{\theta}^*} - \widehat{\boldsymbol{\theta}})$ approximating the distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$.

Step [5] in their procedure is relatively computationally undemanding (it is simply drawing from a truncated distribution a total of BK times following the iterations). However, step [3] requires B re-estimations of the model. While the values of B and K used by Simar and Wilson are not stated in the paper, their model is very simple (a constant and one other regressor). When a more complex model is estimated, it is reasonable to assume that this approach may involve substantial computing time.

7.4.2. An alternative asymptotic approach

As an alternative, this study proposes to draw from the multi-variate distribution $\hat{\boldsymbol{\theta}} \sim \mathrm{N}\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Omega}}\right)$ in place of steps [3] and [4]. This is not bootstrapping and appeals to the asymptotic properties of $\widehat{\boldsymbol{\theta}}$ as an approximation of the small sample distribution of $\widehat{\boldsymbol{\theta}}$. Such an approach can be traced back to Krinsky and Robb (1986) who used the approach to approximate the distribution of estimated substitution elasticities in a Translog cost model. The method is commonly used in the literature as a means of approximating the distribution of relatively complicated functions of estimated parameters as an alternative to the delta method (e.g. willingness to pay in mixture discrete choice models (Train, 2009) – see Greene (2012 p 649-651) for a discussion on application). This approach is termed the multi-variate normal approach. More formally, by asymptotic theory:

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathrm{N}\left(\boldsymbol{\theta}, \boldsymbol{\Omega}\right) \qquad (7.9)$$

where $\boldsymbol{\Omega}$ is the asymptotic variance covariance matrix.

To operationalise this distribution there is a need to substitute $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$, which is the ML estimate of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\Omega}}$ for $\boldsymbol{\Omega}$ which is the estimated asymptotic variance covariance matrix.

It is necessary to sample from the truncated normal distribution for different draws of $\widehat{\boldsymbol{\theta}}$ to determine empirically the shape of the distribution for $\hat{u}_i \mid \hat{\varepsilon}_i$, i.e.:

$$f\left(\hat{u}_i \mid \boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{x_i}, y_i\right) = \int_{\hat{\boldsymbol{\theta}}} f\left(\hat{u}_i \mid \hat{\boldsymbol{\theta}}, \mathbf{x_i}, y_i\right) d\hat{\boldsymbol{\theta}} \tag{7.10}$$

In sum, the following alternative approach is proposed:

Algorithm 2

[1] Maximize the log-likelihood function for the model in (7.1) to obtain ML estimates $\widehat{\boldsymbol{\theta}}$. Recover $\hat{\boldsymbol{\Omega}}$, the estimator of the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$**.**

[2] Set b=1.

[3] Draw $\widehat{\boldsymbol{\theta}}_{\mathbf{b}}^{*}$ from $\mathrm{N}\big(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}\big)$.

[4] For $\mathrm{i} = 1, \dots, \mathrm{N}$ draw u_{ibk}^{*}, $k = 1, \dots, K$, from $f\big(u_i \big| \widehat{\boldsymbol{\theta}}_{\mathrm{b}}^{*}, \mathbf{x_i}, y_i\big)$ given in (7.2), for each $i = 1, \dots, N$.

[5] Increment b by one

[6] Repeat steps [3]-[5] B times, yielding a set $\mathfrak{U} = \{u_{ibk}^{*}\}_{b=1,k=1}^{B,K}$ of BK values for each $i = 1, \dots, N$.

[7] Form the set $\mathfrak{U}_i^{*} = \{u_{ibk}^{*} | u_{ibk}^{*} \in \mathfrak{U}_i, u_{ibk}^{*} \neq 0\}$ for each $i = 1, \dots, N$.

[8] Order set $\mathfrak{U}_i^*$ and compute the percentiles of interest.

Thus, apart from the change in steps [3] and [4], the bagging and multi-variate normal approaches are the same. The multi-variate normal approach has the advantage of not requiring re-estimation of the model which saves computing time. However it is not clear to what extent the $\widehat{\boldsymbol{\theta}}$ do follow the distribution in (7.9) in finite samples, both in terms of the multi-variate normal shape and the extent to which the estimated covariance matrix $\widehat{\boldsymbol{\Omega}}$ does approximate the true covariance matrix. This is analogous to the extent to which the distribution $\left(\widehat{\boldsymbol{\theta}^*} - \widehat{\boldsymbol{\theta}}\right)$ approximates the distribution of $\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ in the bagging approach in finite samples. Thus, both approaches are approximations to the true distribution in finite samples but, due to the asymptotic results, it is likely that by adopting either approach the intervals will better reflect the true intervals over simply ignoring parameter uncertainty.[53]

[53] An issue that arises in comparing the 'bagging' and asymptotic approaches is the possibility that the point estimate of the variance of the inefficiency term is zero - this is the 'wrong skewness' problem. In this instance, the asymptotic approach produces a zero width interval, by construction, but the bagging approach may still produce a nonzero width interval (Simar and Wilson, 2010). This issue has attracted some attention in recent discussions of stochastic frontier modelling. The possibility is noted, but, an attempt to confront this substantive issue is not made in this chapter (the analysis assumes a nonzero estimate of the variance). This question is for further research. [Note that Professor Greene should be credited with this remark.]

7.5. Empirical Example

In this section, the various prediction intervals referred to in the chapter using a dataset for a stochastic frontier model reported within the transport economics literature, are computed. This model considers the cost and efficiency of train operating companies (TOCs) in Great Britain between 1996 and 2006 (see Smith and Wheat, 2012a). This work is a forerunner to the work reported in Chapter 5 but the emphasis of the Smith and Wheat work was to examine changes in cost efficiency from different contract types.

An actual 'real world' dataset and model specification is used, rather than a simulated or a more stylised empirical example. This provides a realistic number of parameters that have to be estimated – the model has 28 regressors – when compared to models which may be estimated in practice. Thus it is envisaged that there may be a realistic degree of parameter uncertainty.

The model parameter estimates and a description of the variables used are given in Appendix 7B. A pooled model is estimated, when in fact the preferred Smith and Wheat (2012a) model was a panel data model (based on the formulation by Cuesta (2000)). Only the pooled model is considered since the purpose of its inclusion in this chapter is to provide an illustrative example. Analysing a panel dataset as a pooled dataset is, for this purpose, essentially the same as analysing a dedicated cross sectional dataset; cross sectional data being the focus of this chapter. Given potential temporal correlation in errors the corrected covariance matrix presented in

Alvarez et al (2006) is utilised, although this makes very little difference to this analysis's findings on interval widths.

The following discussion follows the delineation used between section 7.3 and 7.4; namely, to first consider the difference between central two sided intervals and minimum width prediction intervals for firm efficiency. The extension to incorporate parameter uncertainty is then considered. The estimation of the pooled stochastic cost frontier model and the bagging runs of the stochastic cost frontier model were undertaken using LIMDEP v9 (Econometric Software Inc., 2010b). The computation of the central and minimum width intervals under both the assumption of known parameters and estimated parameters were undertaken using the matrix programming language suite [R] (R Development Core Team, 2010). Codes for both packages are available on request.

7.5.1. Findings on the width of intervals parameter known case

Table 7.1 summarises the results for the upper and lower bounds of the central and minimum width intervals for 95th, 90th and 85th percentage prediction levels for firm efficiency, by displaying the observations which comprise each decile ordered by lower bound.

The minimum width intervals are at least as narrow as the central intervals. On average, the minimum width intervals are 5.9%, 6.2% and 6.4%. narrower than the central intervals for the 95th, 90th and 85th percentile prediction intervals respectively[54]. For some observations, in particular those with small $E[u_i|\varepsilon_i]$, the minimum width intervals are up to 17% narrower than the corresponding central interval (90th percentile prediction interval).

Several further points help illuminate the reasons for these findings. Firstly, as described in section 7.3, the upper bound for efficiency for the minimum width intervals is one for those observations with probability mass of $u_i|\varepsilon_i$ close to zero. Given the homoscedastic variance assumption on u_i, this is equivalent to those observations with small $E[u_i|\varepsilon_i]$ having an upper predictive bound for efficiency of one. This conveys useful information to practitioners and policy makers regarding the likely inefficiency of a given observation. By definition, central intervals cannot make this distinction.

Secondly, as the predictive interval significance level increases, so the number of observations with upper bounds of one decreases and the intervals becomes narrower. 75%, 62% and 54% of observations have minimum width upper boundaries of one for efficiency for the 95%, 90% and 85% predictive intervals respectively.

Thirdly, for observations with large $E[u_i|\varepsilon_i]$, the minimum width interval approaches the central interval. This is because the distribution of $u_i|\varepsilon_i$ tends to a symmetric normal distribution as $E[u_i|\varepsilon_i]$ increases. Thus adopting minimum width intervals over central intervals gives most gains in terms of shrinkage of intervals for

[54] Given efficiency is often expressed as a percentage it is worth clarifying that the percentage reductions given above are percentages of the interval width rather than absolute percentage points.

those observations with little inefficiency. As discussed above, this can be a substantial savings for these observations.

Fourthly, as discussed in Bera and Sharma (1999, p.205) for central intervals, it is found that the minimum width intervals appear at first to be monotonically increasing with $E[u_i|\varepsilon_i]$, however for the larger $E[u_i|\varepsilon_i]$, the intervals start to decrease in width. Bera and Sharma (1999) attribute this behaviour to the exponential transformation associated with efficiency in the multiplicative SF models (they show a strictly monotonically increasing relationship between inefficiency and width). Given that for large $E[u_i|\varepsilon_i]$, the minimum width intervals approach the central intervals, it is not surprising that the same results are found here as in Bera and Sharma.

Table 7.1 Deciles of the upper and lower predictive interval bounds for firm efficiency – the case of known parameters – ordered by minimum width lower bound

	95% Predictive Interval							90% Predictive Interval							85% Predictive Interval						
	Minimum Width			Central (HS)			Diff.	Minimum Width			Central (HS)			Diff.	Minimum Width			Central (HS)			Diff.
Decile	UB	LB	Width	UB	LB	Width	%	UB	LB	Width	UB	LB	Width	%	UB	LB	Width	UB	LB	Width	%
0%	0.773	0.597	0.176	0.779	0.603	0.176	0.2%	0.756	0.609	0.148	0.763	0.615	0.148	0.2%	0.747	0.618	0.129	0.753	0.624	0.129	0.2%
10%	0.971	0.759	0.212	0.967	0.756	0.212	0.1%	0.951	0.771	0.180	0.951	0.771	0.180	0.2%	0.939	0.781	0.158	0.940	0.782	0.158	0.1%
20%	0.993	0.790	0.203	0.983	0.778	0.205	0.8%	0.977	0.801	0.176	0.971	0.794	0.176	0.3%	0.965	0.809	0.156	0.961	0.805	0.156	0.0%
30%	1.000	0.809	0.191	0.988	0.793	0.196	2.4%	0.990	0.822	0.169	0.979	0.809	0.171	1.2%	0.980	0.829	0.150	0.972	0.820	0.152	1.1%
40%	1.000	0.821	0.179	0.992	0.805	0.187	4.4%	1.000	0.840	0.160	0.985	0.821	0.164	2.6%	0.992	0.847	0.144	0.979	0.832	0.147	1.7%
50%	1.000	0.835	0.165	0.994	0.819	0.176	6.2%	1.000	0.855	0.145	0.989	0.835	0.154	5.6%	1.000	0.868	0.132	0.984	0.846	0.138	4.4%
60%	1.000	0.850	0.150	0.996	0.834	0.163	7.8%	1.000	0.869	0.131	0.992	0.850	0.142	8.2%	1.000	0.882	0.118	0.989	0.861	0.128	8.0%
70%	1.000	0.866	0.134	0.997	0.849	0.148	9.3%	1.000	0.885	0.115	0.995	0.866	0.129	10.4%	1.000	0.897	0.103	0.992	0.877	0.115	11.0%
80%	1.000	0.875	0.125	0.998	0.859	0.138	10.1%	1.000	0.894	0.106	0.996	0.875	0.120	11.7%	1.000	0.906	0.094	0.993	0.886	0.107	12.6%
90%	1.000	0.894	0.106	0.998	0.878	0.120	11.6%	1.000	0.911	0.089	0.997	0.894	0.103	13.9%	1.000	0.923	0.077	0.995	0.904	0.092	15.5%
100%	1.000	0.924	0.076	0.999	0.911	0.088	14.2%	1.000	0.939	0.061	0.998	0.924	0.074	17.3%	1.000	0.948	0.052	0.997	0.933	0.065	19.5%

UB – Upper Bound, LB – Lower Bound.

Table 7.2 Deciles of the upper and lower 90% predictive interval bounds for firm efficiency – incorporating parameter uncertainty– ordered by minimum width lower bound

	Minimum Width												Central Intervals											
	No Parameter Uncertainty (1)			Asymptotic Approach (2)			Bagging Approach (3)			Percentage Differences			No Parameter Uncertainty (1)			Asymptotic Approach (2)			Bagging Approach (3)			Differences		
Decile	UB	LB	Width	UB	LB	Width	UB	LB	Width	(2)/(1)	(3)/(1)	(3)/(2)	UB	LB	Width	UB	LB	Width	UB	LB	Width	(2)/(1)	(3)/(1)	(3)/(2)
0%	0.756	0.609	0.148	0.784	0.589	0.195	0.839	0.541	0.298	32.3%	101.8%	52.5%	0.763	0.615	0.148	0.792	0.595	0.197	0.885	0.564	0.320	33.2%	117.0%	63.0%
10%	0.951	0.771	0.180	0.961	0.765	0.196	0.971	0.753	0.219	8.8%	21.5%	11.6%	0.951	0.771	0.180	0.960	0.764	0.195	0.972	0.752	0.220	8.4%	22.2%	12.7%
20%	0.977	0.801	0.176	0.979	0.797	0.182	0.987	0.789	0.198	3.7%	12.5%	8.5%	0.970	0.794	0.176	0.974	0.791	0.183	0.981	0.781	0.200	4.1%	13.4%	8.9%
30%	0.990	0.822	0.169	0.992	0.817	0.175	1.000	0.817	0.183	4.0%	8.6%	4.4%	0.979	0.809	0.170	0.981	0.803	0.179	0.985	0.798	0.187	4.8%	9.9%	4.8%
40%	1.000	0.840	0.160	1.000	0.834	0.166	1.000	0.832	0.168	4.0%	5.5%	1.5%	0.985	0.821	0.164	0.986	0.815	0.172	0.988	0.812	0.176	4.8%	7.5%	2.6%
50%	1.000	0.855	0.145	1.000	0.849	0.151	1.000	0.852	0.148	3.7%	1.7%	-2.0%	0.989	0.835	0.154	0.990	0.828	0.161	0.991	0.832	0.159	4.9%	3.0%	-1.8%
60%	1.000	0.869	0.131	1.000	0.865	0.135	1.000	0.870	0.130	3.6%	-0.3%	-3.8%	0.992	0.850	0.142	0.993	0.844	0.149	0.994	0.849	0.145	4.8%	1.9%	-2.7%
70%	1.000	0.885	0.115	1.000	0.880	0.120	1.000	0.890	0.110	3.8%	-4.6%	-8.1%	0.995	0.866	0.129	0.995	0.859	0.135	0.996	0.871	0.124	5.1%	-3.5%	-8.2%
80%	1.000	0.894	0.106	1.000	0.889	0.111	1.000	0.901	0.099	4.3%	-6.2%	-10.1%	0.996	0.876	0.120	0.996	0.869	0.126	0.996	0.883	0.114	5.3%	-5.2%	-10.0%
90%	1.000	0.911	0.089	1.000	0.907	0.093	1.000	0.922	0.078	5.0%	-12.5%	-16.7%	0.997	0.894	0.103	0.997	0.888	0.109	0.998	0.906	0.092	5.9%	-11.1%	-16.1%
100%	1.000	0.939	0.061	1.000	0.936	0.064	1.000	0.951	0.049	5.3%	-19.2%	-23.2%	0.998	0.925	0.074	0.998	0.920	0.079	0.999	0.938	0.061	6.8%	-16.9%	-22.2%

7.5.2. Findings on incorporating parameter uncertainty

Table 7.2 gives the deciles for the 90th percentile prediction intervals for observation specific efficiency for three cases; assuming the parameters are known (as in sub-section 7.5.1), approximating the distribution of the parameters as multi-variate normal (Krinsky and Robb, 1986) and alternatively using the bagging approach of Simar and Wilson (2010). The discussion is restricted to the 90th percentile prediction intervals as similar observations can be made for other percentiles. It is important to note that the purpose of implementing both bagging and the multi-variate normal approach is to illustrate the feasibility of each approach and to discuss the increase in interval widths relative to the shrinkage from adopting minimum width. Any differences should not be taken as evidence that one method is superior to another. Such conclusions can only be drawn from a robust simulation study where the coverage of intervals can be assessed. This is left for further research.

For the multi-variate normal method, the minimum width prediction intervals are always larger than the corresponding intervals assuming no parameter uncertainty. On average they are 4.7% wider. For the bagging method, the intervals are 4.8% wider. Overall, it would seem that incorporating parameter uncertainty increases the interval width by slightly less compared to the reduction in width by adopting minimum width as opposed to central intervals. It should be recalled that the empirical example was chosen because there are many parameters to be estimated within the frontier. Thus, *a priori* while it may be expected that there is a relatively large contribution from parameter uncertainty for this example, in practice the added uncertainty is still less than the gain from adopting minimum width intervals.

Furthermore, incorporating parameter uncertainty increases the width of intervals (even in proportional terms) for those observations with the greatest $E[u_i|\varepsilon_i]$; the opposite relationship to that found between the width saving from adopting minimum width over central intervals in the parameter known case. The relationship appears to not be strictly monotonic, which is to be expected given that the additional uncertainty is arising from several parameters which are multiplied by regressors, taking different values for each observation. Even with this caveat, overall observations which benefit the most from adopting minimum width over central prediction intervals (in terms of shrinkage of intervals), i.e. those observations with least inefficiency, are the observations whose intervals increase proportionally least when additionally allowing for parameter uncertainty. For example, even when parameter uncertainty is taken into account, adopting minimum width intervals can still reduced predictive intervals by 13.5% for the most efficient observations (90% predictive interval) in this example.

One difference between the bagging approach and the multi-variate normal approach is that where $E[u_i|\varepsilon_i]$ is relatively small, the bagging prediction intervals are narrower than the intervals computed assuming parameters are known. This is a counter intuitive result, given that incorporating parameter uncertainty should introduce more uncertainty and thus widen the distribution. At the other end of the efficiency scale, the bagging intervals are considerably wider than both the intervals where parameter uncertainty is not taken into account and those from the multi-variate normal approach. Thus while the bagging and multi-variate normal approaches yield similar increases in widths in terms of average interval width gain, the extremes are very different, with the bagging approach yielding counter intuitive results at the extremes. This is left for future research to determine if the results hold in more applications.

7.5.3. Specific illustrations of intervals for the analysis of TOCs

In order to highlight the applicability of the techniques in this chapter to vertically separated railways, Table 7.3 presents 90 and 95 per cent prediction intervals for the TOCs in the last year of sample. The Jondrow et al (1982) point estimator (of efficiency) is provided for comparison. The rank of TOCs implied by each boundary is also presented. The estimated average efficiency score for this year of the model is 90%. This is similar to the average for the model as a whole (91%).

The following observations can be made:

- Irrespective of which interval type (HS versus minimum width), the width of prediction intervals is large. For example, the average width for 90% Horrace and Schmidt prediction interval is 14 percentage points. Clearly there is a need to report interval predictors as well as the Jondrow et al (1982) efficiency point predictor.
- The minimum width intervals distinguish between those TOCs which have a one-sided interval with an upper bound at 100% from those that have two sided intervals. As such this provides a criterion to distinguish between those firms that exhibit strong evidence of inefficiency versus those that do not. It is found that 14 and 15 out of 23 TOCs (for the 90 and 95 per cent intervals respectively) have upper bounds for the minimum width intervals which

include 100%. As such there is some evidence that these firms are efficient (although this is not inference in the conventional sense).

- Contrasting the above findings to those that can be gleaned from the two sided HS intervals, by construction the upper bound must be less than 100%. While it is clear that many upper bounds are close to 100%, it is not clear whether there is a sensible means to judge how close is close enough to make meaningful statements about which TOCs are likely to be efficient versus inefficient. For example, at the 90% prediction level, ONE has a two sided upper bound of 98.3% but is still consistent with a minimum width interval that includes 100%.
- With the exception of the rankings implied by the upper bounds of the minimum width intervals; the rankings for each TOC seem consistent, irrespective of which measure is used[55]. This is as expected given the results of Bera and Sharma (1999) regarding the monotonic relationship between the (inefficiency) point estimate and the width of intervals.
- The reason for the inconsistency of the ranking from the upper bound of minimum width intervals is that there are multiple TOCs ranked first given the construction of the intervals. An implication of this result may be that, if relative rankings are of interest, then a useful measure for policy/regulatory purposes which summaries whether a firm is inefficient or not and, if it is, how it compares to other TOCs, is to use the ranking of the minimum width upper bound. Thus, 14 TOCs occupy the top rank with the remainder having a position assigned. It should be noted that this is not inference in the conventional sense. It would be interesting to compare this approach to that

[55] Small differences are likely to be due to sampling error.

yielded from the work by Horrace (2005) and Flores-Lagunes, Horrace, and Schnier (2007), although neither technique is ideal given the latter ignores parameter uncertainty.

Table 7.3 Efficiency predictions for TOCs in the final year of the dataset (2006)

TOC name	Jondrow et al (1982) efficiency predictor		90% Prediction Interval								95% Prediction Interval							
			HS Interval				Minimum Width with Parameter Uncertainty				HS Interval				Minimum Width with Parameter Uncertainty			
			LB		UB		LB		UB		LB		UB		LB		UB	
Arriva Trains Wales	88.5%	*16*	80.0%	*15*	96.7%	*16*	79.6%	*16*	98.2%	*16*	78.1%	*16*	98.4%	*16*	78.6%	*15*	99.7%	*16*
Central Trains	87.2%	*17*	78.3%	*17*	95.9%	*18*	77.6%	*17*	97.4%	*18*	76.7%	*18*	97.6%	*18*	76.3%	*18*	99.0%	*18*
Chiltern	97.2%	*1*	93.0%	*1*	99.7%	*1*	93.7%	*1*	100.0%	*1*	91.1%	*1*	99.9%	*1*	92.1%	*1*	100.0%	*1*
Southern	95.9%	*5*	90.2%	*5*	99.6%	*3*	91.0%	*5*	100.0%	*1*	88.0%	*5*	99.9%	*5*	89.1%	*5*	100.0%	*1*
South Eastern	87.1%	*19*	78.2%	*19*	96.0%	*17*	77.5%	*18*	97.5%	*17*	76.6%	*19*	97.6%	*19*	76.2%	*19*	99.1%	*17*
Cross Country	94.1%	*8*	86.7%	*8*	99.2%	*8*	87.6%	*8*	100.0%	*1*	84.9%	*8*	99.7%	*8*	85.6%	*8*	100.0%	*1*
Gatwick Express	96.0%	*4*	90.4%	*4*	99.6%	*3*	91.2%	*4*	100.0%	*1*	88.3%	*4*	99.9%	*4*	89.3%	*4*	100.0%	*1*
GNER	96.1%	*3*	90.7%	*3*	99.6%	*3*	91.5%	*3*	100.0%	*1*	88.6%	*3*	99.9%	*3*	89.7%	*3*	100.0%	*1*
Great Western	92.7%	*10*	85.0%	*10*	98.8%	*10*	85.9%	*10*	100.0%	*1*	82.8%	*10*	99.6%	*10*	83.9%	*10*	100.0%	*1*
c2c	88.8%	*15*	79.8%	*16*	97.2%	*15*	79.9%	*15*	99.0%	*15*	78.4%	*15*	98.5%	*15*	78.5%	*16*	100.0%	*1*
Merseyrail	87.2%	*17*	78.3%	*17*	95.9%	*18*	77.5%	*19*	97.3%	*19*	76.7%	*17*	97.6%	*17*	76.3%	*17*	99.0%	*19*
Midland Main Line	95.4%	*6*	89.3%	*6*	99.5%	*6*	90.2%	*6*	100.0%	*1*	87.0%	*6*	99.8%	*6*	88.3%	*6*	100.0%	*1*
Northern	90.2%	*14*	80.1%	*14*	98.2%	*14*	81.2%	*14*	100.0%	*1*	79.9%	*14*	99.0%	*14*	78.7%	*14*	100.0%	*1*
Scotrail	95.3%	*7*	88.8%	*7*	99.4%	*7*	89.7%	*7*	100.0%	*1*	86.9%	*7*	99.8%	*6*	87.7%	*7*	100.0%	*1*
Silverlink	91.6%	*12*	83.5%	*12*	98.4%	*11*	84.4%	*12*	100.0%	*1*	81.4%	*12*	99.4%	*12*	82.3%	*12*	100.0%	*1*
South West Trains	83.8%	*21*	75.3%	*21*	93.1%	*21*	73.8%	*21*	94.0%	*21*	73.6%	*21*	94.9%	*21*	72.5%	*21*	96.3%	*21*
Thameslink	77.6%	*22*	69.3%	*22*	86.8%	*22*	67.6%	*22*	87.5%	*22*	68.1%	*22*	88.0%	*22*	66.0%	*22*	89.8%	*22*
Thames Trains First Great Western Link	91.8%	*11*	83.9%	*11*	98.4%	*11*	84.8%	*11*	100.0%	*1*	81.7%	*11*	99.4%	*11*	82.8%	*11*	100.0%	*1*
WAGN	96.7%	*2*	91.9%	*2*	99.7%	*1*	92.6%	*2*	100.0%	*1*	89.9%	*2*	99.9%	*2*	90.9%	*2*	100.0%	*1*
Wessex	85.8%	*20*	77.1%	*20*	94.8%	*20*	76.0%	*20*	95.9%	*20*	75.4%	*20*	96.6%	*20*	74.7%	*20*	97.9%	*20*
West Coast	93.8%	*9*	86.4%	*9*	99.1%	*9*	87.3%	*9*	100.0%	*1*	84.5%	*9*	99.7%	*9*	85.3%	*9*	100.0%	*1*
Transpennine Express	68.7%	*23*	60.7%	*23*	78.1%	*23*	58.9%	*23*	78.6%	*23*	60.3%	*23*	77.9%	*23*	57.5%	*23*	81.1%	*23*
One	91.4%	*13*	83.4%	*13*	98.3%	*13*	84.3%	*13*	100.0%	*1*	81.3%	*13*	99.3%	*13*	82.2%	*13*	100.0%	*1*

Rank is given in underlined italics

7.6. Summary

The prediction of firm specific inefficiency is the primary objective of any analysis which utilises stochastic frontier techniques (or, at the very least, one of the most important).[56] Interval prediction is notable given that the available point estimates, for use with cross sectional data, first proposed by Jondrow et al (1982) are inconsistent for the underlying quantity of interest; namely the realisation of a random variable representing inefficiency. Ultimately, there is likely to be substantial uncertainty surrounding predictions of firm specific inefficiency and this needs to be quantified.

In this chapter, the prediction intervals proposed by Horrace and Schmidt (1996) for firm specific inefficiency estimates in cross sectional stochastic frontier models have been considered. The chapter makes two contributions to the literature. Firstly, prediction intervals assuming the model parameters are known (as assumed in the original Horrace and Schmidt paper) have been examined. It is noted that the existing intervals are not efficient predictive intervals in the sense that they do not correspond to the minimum width prediction intervals. How to compute these is explained and it is pointed out that as well as being statistically efficient, the minimum width interval provides policy makers with useful information in terms of whether the firm prediction interval spans zero inefficiency or not. This in turn can be used by policy makers as evidence as to whether a given firm has scope for

[56] This is a logical statement, given that under the assumptions of the majority of stochastic frontier models, OLS estimation provides consistent and unbiased estimation. Therefore the motivation for using specific stochastic frontier estimation techniques must come from the desire to measure inefficiency, rather than to correct or provide more robust slopes relative to OLS.

efficiency improvement or not. In contrast, by construction, central intervals cannot make this distinction since there will always be some probability mass below the lower bound.

Secondly, the extensions required to the intervals to incorporate parameter uncertainty, given that model parameters have to be estimated, are considered. An alternative to the bagging procedure presented in Simar and Wilson (2010) which utilises an approximation to the asymptotic distribution of the estimated parameters (following Krinsky and Robb (1986)) is introduced, partly for its computational simplicity. Nonetheless, either is a candidate method given that both methods implicitly make distributional approximations. It is considered to be a further research issue as to which performs best in different circumstances, and the analysis leaves investigation to such methods as a rigorous simulation experiment and/or further empirical applications.

The empirical illustration utilises data from a published railway operations study. It is found that the minimum width prediction intervals are approximately 6-6.5% narrower than the equivalent central two sided width interval reported in the literature (parameters assumed known). Furthermore, adopting the minimum width interval reduces the width of predictive intervals most (vis-à-vis central intervals) for those observations with little inefficiency since the distribution for inefficiency for these observations is highly asymmetric. When additionally uncertainty associated with parameter estimation is added, it is found that the minimum width predictive intervals expand by about 4.5%. Thus, overall, even after incorporating parameter uncertainty, it is found that the minimum width predictive intervals are narrower than the two sided parameter known intervals discussed in the literature.

Appendix 7A – Derivation of minimum width predictive intervals for the truncated normal distributions

Consider $(u_i|\varepsilon_i) \sim N^+(\mu_{i*}, \sigma_*^2)$ (7.11)

There are two solutions to the Lagrangian problem. Either:

$f(L^*) = f(U^*)$ exists such that $\int_{L^*}^{U^*} f(u_i|\varepsilon_i)\, du_i = (1-\alpha)$ and $L^*, U^* \geq 0$ (7.12)

Otherwise

$L^* = 0$ and U^* such that $\int_0^{U^*} f(u_i|\varepsilon_i)\, du_i = 1-\alpha$ (7.13)

U^* for the case in (7.13) is given by Horrace and Schmidt (1996) and reproduced in (7.3) as

$$U^* = \mu_{i*} + \sigma_* \Phi^{-1}\left[1 - \big(1 - (1-\alpha)\big)\Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right]$$

$$U^* = \mu_{i*} + \sigma_* \Phi^{-1}\left[1 - \alpha.\Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right] \quad (7.14)$$

Now consider (7.12).

Define, X~N(μ_{i*}, σ_*^2)

Then

$$\int_{L^*}^{U^*} f(u_i|\varepsilon_i)\, du_i = 1 - \alpha \leftrightarrow \int_{L^*}^{U^*} f(X)\, dX = (1 - \alpha) \int_0^{\infty} f(X)\, dX$$

$$\int_{L^*}^{U^*} f(X)\, dX = (1 - \alpha) \left(1 - \Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right) \tag{7.15}$$

Given the symmetry of $f(X)$, for $f(L^*) = f(U^*)$,

$$\int_{-\infty}^{U^*} f(X)\, dX = \left(1 - \frac{\alpha}{2}\right)\left(1 - \Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right) \tag{7.16}$$

$$\int_{-\infty}^{L^*} f(X)\, dX = \left(\frac{\alpha}{2}\right)\left(1 - \Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right) \tag{7.17}$$

Yielding

$$U^* = \mu_{i*} + \sigma_* \Phi^{-1}\left[\left(1 - \frac{\alpha}{2}\right)\left(1 - \Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right)\right] \tag{7.18}$$

$$L^* = \mu_{i*} + \sigma_* \Phi^{-1}\left[\left(\frac{\alpha}{2}\right)\left(1 - \Phi\left(\frac{\mu_{i*}}{\sigma_*}\right)\right)\right] \tag{7.19}$$

Intuitively, L^* and U^* in (7.12) are the boundaries of the central interval of the untruncated normal distribution with mean μ_{i*} and variance σ_*^2, since the normal distribution is symmetric. However, they do not correspond to the usual $\frac{\alpha}{2}$ and $\left(1 - \frac{\alpha}{2}\right)$ percentiles of the normal distribution since the actual distribution is truncated and thus a correction is necessary for the untruncated distribution to integrate to unity.

Appendix 7B – Parameter estimates for the empirical example

Table 7.4 gives the output for the preferred model in Smith and Wheat (2012a) re-estimated for a normal-half normal pooled model (See Smith and Wheat (2012a) for more details on the model formulation and interpretation). Overall, it is considered that the model parameter estimates are broadly in line with those from the Smith and Wheat model, which was a panel data model, but here the data is analysed as a pooled model. Importantly, the conclusions regarding constant returns to scale are the same as that found in Smith and Wheat, although returns to train density at the sample mean are no longer found (see Smith and Wheat (2012a) for details of computation in this context). The average point efficiency scores ($exp(-E[u_i|\varepsilon_i])$) are 0.90 for the panel model and 0.91 for the pooled model, although the correlation between the scores is only 0.6 which is not surprising given the added structure imposed to efficiency variation in the panel model. Overall, while a pooled model is not the preferred model for modelling TOC costs, it is considered to be a reasonably credible alternative for the illustrative purpose of this chapter.

Table 7.4 Model coefficient estimates

Variable	Description	Coefficient	
Dependent variable			
LCOST	Operating cost of train operating companies (1996/97 to 2005/06) - 238 observations		
Explanatory variables			
ONE	Constant	5.39046	***
ROUTE	ln(route length)	0.819282	***
TDEN	ln(traffic density)=ln(train-km/route-km)	1.02807	***
STAT1	Number of stations operated	0.210935	***
TIME	Time trend	-0.09417	***
INP	ln(Wage)	0.401388	***
TLEN	ln(average length of train)=ln(vehicle-km/train-km)	0.270768	***
LFAC	ln(average passenger load factor)	0.148172	**
LNAGE	ln(average age of rollingstock)	0.056336	**
TDEN2	TDEN^2	0.103867	***
STAT12	STAT1^2	-0.02384	**
TIME2	TIME^2	0.007183	***
TLEN2	TLEN^2	0.277243	***
DENSTAT1	TDEN*STAT1	-0.00343	
TDENLEN	TDEN*TLEN	-0.20026	***
STAT1LN	STAT1*TLEN	0.08822	**
ONWARDS2	Dummy variable: = 1 for observations in year 2000 onwards	0.170486	***
_1_YEAR_	Dummy variable: = 1 iff franchise in last year	-0.03871	
MANBF	Dummy variable: = 1 for years before a franchise was placed on to a management contract if it was subsequently placed on to such contract	0.052184	
MANAF	Dummy variable: = 1 for years after a franchise was placed on to a management contract if it was placed on to such contract	0.247369	*
RENBF	Dummy variable: = 1 for years before a franchise was placed on to a renegociated contract if it was subsequently placed on to such contract	0.143143	**
RENAF	Dummy variable: = 1 for years after a franchise was placed on to a renegociated contract if it was placed on to such contract	0.341777	**
INTERCIT	Dummy variable: = 1 iff franchise is classed as an intercity operator	0.475924	***
LSE	Dummy variable: = 1 iff franchise is classed as a London South Eastern operator	-0.02452	
MANBFT	MANBF*TIME	-0.02055	
RENBFT	RENBF*TIME	-0.03906	**
MANAFT	MANAF*TIME	-0.01966	
RENAFT	RENAF*TIME	-0.03312	
Cost frontier error component parameters			
Lambda		1.52756	***
Sigma		0.14271	***
***,**,* Statistically significant at the 1%, 5%, 10% level respectively			

8. Conclusion

8.1. Introduction

In this final chapter, the thesis is concluded. The structure is that section 8.2 provides a summary of the thesis context. Section 8.3 describes how the research aims and objectives have been achieved through a detailed examination of the conclusions from each research chapter. Section 8.4 provides a wider, more holistic view of the research in terms of cross chapter conclusions. Finally, section 8.5 ends the thesis by considering further research opportunities.

8.2. Summary of thesis context

Vertical separation refers to separation in management and administration of different aspects of the production process. For the purpose of this thesis, vertically separated railways refer to railways where passenger and freight operations (the running of trains) are separate to the infrastructure. Such separation is becoming more important in railways, both in Europe and the wider world, with market reforms aimed at opening up operations to competition and infrastructure managers to greater cost (reduction) pressures.

Research has been directed to examine several pressing needs within the sector, identified as:

- The need to understand the returns to scale and density properties of passenger train operations – this is important given the move to competitive tendering of railway operations in terms of specifying cost minimising tender specifications.

- the need to exploit multi-layered datasets to better predict the efficiency of infrastructure managers – this is of particular importance given the limited number of comparators available to make an assessment and it is of particular relevance to the railway in Britain given the recent rises in infrastructure costs and the need to assess the scope for cost reduction.
- The need to understand the uncertainty in efficiency predictions – this is of significance given the maturing of the rail regulatory sector in Britain and the perceived elimination of the ‘easy wins’ for efficiency improvements post privatisation.

8.3. Reconciliation against aims and objectives

The aim of this thesis is to apply appropriate econometric techniques to better analyse the cost structure of vertically separated railways - specifically the infrastructure management and passenger train operations activities - to inform regulatory bodies and policy makers. The research chapters have all contributed to this aim.

Six specific objectives were set and addressed through three research chapters (Chapters 5 to 7). Below, for each of the research chapters, the contribution to the relevant objectives are outlined. This is a prelude to discussion of more general conclusions in section 8.4.

8.3.1. Chapter 5 Passenger Train Operating Company cost analysis

Objectives addressed:

- To explore the use of a hedonic cost function approach to incorporate measures of output heterogeneity in the analysis of train operating companies (TOCs) costs
- To provide new empirical evidence as to the cost implications of redrawing franchise boundaries, crucially drawing on the scale and density properties of the estimated model and how these vary with heterogeneity of the TOC's output

In Chapter 5 a hedonic Translog cost function for TOCs in Great Britain has been estimated. The model includes three hedonic outputs: route-km, stations operated and train hours. The model is rich in output characteristic variables which is unique in the railway operations literature. The approach in Chapter 5 is pragmatic; given so many measures of heterogeneity of output available, a feasible manner to operationalize a Translog type functional form, in order to keep the number of parameters manageable, is to use a hedonic approach.

The estimated cost function conforms with the economic restrictions required for the cost function to represent the underlying technology. The use of train hours (over train-km) is a data innovation in itself but, in addition, the train hours hedonic output has a number of characteristics also included within the hedonic function, which characterize TOC heterogeneity. Thus the model is rich in its characterisation of firms' technology.

This richness allows establishment of a deep understanding of the variation in returns to scale and density in the industry. In particular, different scale and density effects can be distinguished, depending on the output characteristics of the TOC, not just the usual overall output level and input price level as in a simple (non-hedonic) Translog cost function. This has importance since there has recently been a move towards re-mapping franchises to larger, more heterogeneous franchises which requires a rich model to determine whether this increases or reduces costs.

The analysis indicates over 50% of TOCs in the sample operate under decreasing RtS. Furthermore, returns to scale fall with the size of operation, which is consistent with a u-shaped average cost curve. The implication of these findings is that the current mappings of TOCs in Britain are such that operations are above their optimal size given that most TOCs operate with decreasing RtS i.e. on the upward part of the average cost curve. Thus there is an argument for more, smaller TOCs.

It is also found however that there are increasing RtD, i.e. unit cost savings from running more trains on a fixed network. This has two implications. Firstly, increasing capacity (train hours) to meet increasing passenger demand should reduce unit costs. Secondly, there is scope to reduce unit costs by removing franchise overlap; this effect therefore working in the opposite direction to the scale effect (as the density finding suggests that TOC mergers will reduce unit costs). Nethertheless, there can be impacts of changes in the output mix (heterogeneity of services) which prevent TOCs from exploiting any RtD even though train hours per route-km increases. Of the three example mergers considered, it is found that two mergers actually increase cost (Greater Western and New Northern) and one reduces cost (ONE/Greater Northern).

The findings suggest that previous estimates of scale and density properties in railways may have been biased to the extent that they did not adequately model the interaction between scale/density and heterogeneity of services. The model in Chapter 5 contains both measures on inter and intra TOC heterogeneity which permits control for both average differences in output heterogeneity between TOCs and also the extent to which output differs within each TOC. This in turn leads to a complex interplay between heterogeneity and more ‘standard’ concepts of returns to density and scale. In terms of regulatory policy, in interpreting evidence on scale and density returns in railways, the model suggests that policy makers need to take service heterogeneity into account. Failure to do so may mean that policy decisions are made on the basis of supposed scale/density returns that cannot be realised in practice. Modelling railway operations is complex and thus to address specific policy questions (such as the cost implications of mergers) a rich model, such as that developed in Chapter 5, is required.

8.3.2. Chapter 6 Infrastructure cost analysis

Objectives addressed:

- To explore via econometric analysis the exploitation of a multi-layered panel dataset to predict the inefficiency level of infrastructure managers
- To provide new empirical evidence on the potential efficiency saving of the infrastructure managers in sample.

Chapter 6 has outlined econometric techniques to analyse data on geographical regions of multiple companies. This is termed ‘sub-company’ data. This is data which both exists and is desirable for performance analysis of horizontally integrated elements of vertically separated railways. This is common for infrastructure managers and, at a simple level, can be used to overcome the problem of few external comparator observations.

Importantly, as well as providing extra observations to estimate cost frontiers, the data structure permits estimation of a dual-level inefficiency model which separates sub-company persistent inefficiency from sub-company varying inefficiency. This is a useful decomposition in itself, since it gives an indication as to whether inefficiency predominantly varies within firm or between firms, as well as having utility in terms of identifying where efforts should be made to eliminate inefficiency. Furthermore, it is often more sensible from an economic perspective to model infrastructure costs using sub-company data. This is because aggregation bias can be avoided particularly with reference to measurement of returns to scale.

Several candidate estimation techniques have been proposed and their properties discussed. A selection of techniques have been applied to the analysis of 5 railway infrastructure managers. Given the small number of firms and the short panel length, it is only by having data at the sub-company level that estimation of a cost model is feasible from the perspective of having sufficient data for reasonable precision in estimation. It is shown that the parameter estimates are in line with the received literature. With respect to inefficiency, there is statistically significant dual-level inefficiency. The empirical example also indicates that failure to take account of the dual-level inefficiency variation may result in under estimation of inefficiency.

The average efficiency score is 0.724 for the preferred model, comprising 0.849 and 0.851 for the persistent and sub-company varying components respectively. Thus there exists a large amount of cost saving potential for infrastructure managers to assimilate best practice from other infrastructure managers, but also to consistently apply their own best practice across sub-companies within their organisation.

8.3.3. Chapter 7 Uncertainty in efficiency analysis

Objectives addressed:

- To explore the most appropriate predictor of firm efficiency from parametric stochastic frontier models covering point and interval predictors
- To provide a new method to incorporate the effect of parameter uncertainty into predictors of firm efficiency and illustrate these concepts via application to TOCs in Britain.

In Chapter 7, the appropriate way to predict inefficiency for specific firms, as opposed to industry average inefficiency, from stochastic frontier models which utilise cross sectional data is considered. The majority of stochastic frontier studies simply report a point predictor for firm inefficiency, following the method of Jondrow et al (1982). However, it is well established that such a predictor is not consistent for the quantity of interest, a specific realisation of a random variable. Thus in cross sectional models, firm specific inefficiency is likely to be predicted with a large amount of uncertainty, irrespective of sample size. Such an observation is borne out in numerous empirical applications, including Horrace and Schmidt (1996), Bera and Sharma (1999) and the empirical example in Chapter 7.

With the above in mind, interval prediction of firm inefficiency is of value as it quantifies uncertainty. A literature exists on such intervals and the first part of Chapter 7 clarifies the literature in terms of what uncertainty is captured and what uncertainty is not captured within the received literature, starting with the intervals proposed by Horrace and Schmidt (1996).

In terms of methodological advancements, the chapter makes two contributions to the received literature. Firstly, it is noted that the existing intervals are not efficient predictive intervals in the sense that they do not correspond to the minimum width prediction intervals. How to compute these is explained and it is pointed out that, as well as being statistically efficient, the minimum width interval provides policy makers with useful information in terms of whether the firm inefficiency prediction interval includes zero inefficiency or not (equivalently whether the prediction interval for firm efficiency includes unity). This in turn can be used by policy makers as evidence as to whether a given firm has scope for efficiency improvement or not. In contrast, by construction, central intervals cannot make this distinction since there will always be some probability mass either side of the boundaries.

Secondly, what extensions are required to the intervals to incorporate parameter uncertainty given that model parameters have to be estimated are considered. An alternative to the bagging procedure (Simar and Wilson, 2010) which utilises an approximation to the asymptotic distribution of the estimated parameters (following Krinsky and Robb, 1986) is introduced. This is partly introduced for its computational simplicity; however either is a candidate method given that both methods implicitly make distributional approximations.

The empirical example highlights that prediction uncertainty is not trivial in cross sectional models. For example, the average width for 90% Horrace and Schmidt prediction interval is 14 percentage points for TOCs in the last year of the sample. The innovations proposed in Chapter 7 do not, on average, change the interval width substantially. This is because the use of minimum width contracts the intervals, while additionally accounting for uncertainty in parameter estimates results in expanded interval. However, the innovations proposed do result in changes in the width, upper bounds and lower bounds for individual observations. In particular, those observations with small residuals have intervals for efficiency with the upper bound at unity and width considerably narrower (of the order of 13.5% narrower). For those observations with the largest residuals, the intervals tend to be wider than the received intervals in the literature. The reason is that accounting for parameter uncertainty increases the width of intervals while for those observations with large residuals, the minimum width interval corresponds to the central interval. Thus there are substantial increases in width from incorporating uncertainty in estimating parameters but no (or very little) contraction from adopting minimum width.

For the TOCs considered in the empirical example, it is found that 14 and 15 out of 23 TOCs in the final year of sample (for the 90 and 95 per cent intervals respectively) have upper bounds for the minimum width intervals which include 100%. As such there is some evidence that these firms are efficient.

8.4. Overall Conclusions

The three research Chapters 5, 6 and 7 have necessarily focused on addressing specific pressing issues in vertically separated railway cost analysis. Taken together, the material does provide several insights into key challenges in vertically separated railway cost analysis and these are explored below:

8.4.1. Data

Any cost analysis, be it primarily to understand characteristics of the cost function or measure inefficiency requires high quality data. This in turn has many facets. Clearly, there needs to be enough observations with sufficient variation to facilitate precise estimation of parameters. Panel data is obviously desirable since it increases both the number of observations (relative to a single cross sectional of the same data), but also allows for a richer analysis of the data, since both between and within group variation in the data can be exploited. Also, the effect of omitted time invariant factors can be controlled. Panel data was available in a reasonable dimension for the analysis of train operating company costs.

However such data is not always easily to come by – as in the infrastructure case – and may suffer from limited variation over time in the case of industries with very long asset lives, such as railway infrastructure. Expanding the sample size is one of the key motivations for the models exploiting multi-layer ‘sub-company’ datasets but, more subtly, there is the advantage that there is likely to be more variation in the dataset relative to a standard panel, with the total number of observations held constant, for the reason of lack of time variation in panel data in this context.

For both efficiency analysis and more general cost analysis, Chapter 6 clearly demonstrates that there is a need to model costs at the level that management autonomy resides. Failure to do so can result in misleading predictions of efficiency as it mismatches returns to scale properties of the cost function with efficiency. This point has also been made by Brorsen and Kim (2013) who used data on schools and school districts to show that if the model was estimated using data at district level then returns to scale are found to be decreasing rather than finding that these schools are inefficient. Ultimately the aggregation bias is resulting in correlation between errors and regressors, since true measures of scale/density (at the disaggregate level) are not included in the model.

Chapter 6 also showed that in the presence if disaggregate, sub-company, data, failure to take into account dual-level inefficiency results in under prediction of inefficiency. Thus, such data, while desirable, should be analysed through a dual-level inefficiency model.

8.4.2. Accounting for heterogeneity

All chapters illustrate the need to explicitly account for heterogeneity of railways in any cost analysis. In Chapter 5, it was clearly shown the importance of output heterogeneity in accurately determining the RtS and RtD characteristics of passenger train operations franchises. Importantly the policy implications can be quite different when only high level RtS and RtD results are available. In practice, RtD results, for example, may not be able to be realised in practice (such as for two out of three of the merger examples considered in Chapter 5).

In Chapter 6, the results of Mundlak (1978) were exploited to demonstrate the importance of modelling explicit relationships between unobserved heterogeneity (heterogeneity which is not directly measured by explanatory variables) and the regressors, in obtaining accurate predictions of cost inefficiency. Failure to take this into account may result in bias parameter estimates (if a random effects estimation approach is implemented) or lead to attribution of sub-company invariant unobserved heterogeneity (or time invariant in the case of a standard panel model) to sub-company invariant cost inefficiency (if a fixed effects estimation approach is implemented).

Finally, Chapter 7 provides a treatment of prediction intervals for firm inefficiency. Of crucial importance for valid intervals to be constructed, is the need for the residual from any regression to tend in probability to the realisation of the error term for a given observation ($\hat{\varepsilon}_i \xrightarrow[p]{} \varepsilon_i$). This only occurs if the frontier function is specified correctly i.e. if all variables that explain costs are included in the model i.e. heterogeneity needs to be accounted for.

8.4.3. Accounting for uncertainty in estimation, prediction and interpretation

Chapter 7 is concerned with quantifying uncertainty with respect to efficiency predictions in cross sectional stochastic frontier models. What is clear from Chapter 7 is that ignoring uncertainty in efficiency predictions, through simply reporting the Jondrow et al (1982) predictor gives a misleading impression of the accuracy of a cross sectional stochastic frontier model.

It is not just Chapter 7 which considers uncertainty. All the research chapters contribute to a better understanding of uncertainty. Chapter 5 is concerned with uncertainty in the estimation of the cost model parameters. In particular, the use of the hedonic cost function is a pragmatic approach to deal with the availability of so many output characteristics. Adopting a full Translog cost function would require the estimation of over 140 parameters as opposed to the 35 in the adopted functional form.

The issue of uncertainty is a characteristic throughout Chapter 6 since the chapter is concerned with the interpretation of various stratifications (averages) of residuals from models. What elements of the dual model represent inefficiency versus other unobserved time or sub-company invariant factors affecting cost? To some extent, the chapter simply assumes away other factors, with the exception of the attempt to utilise the Mundlak transformation to decompose sub-company invariant unobserved factors into invariant inefficiency and unobserved heterogeneity. However, throughout the chapter, it is acknowledged that there are potentially other interpretations of the error components and this is discussed further in sub section 8.5.1 below.

8.5. Future research needs

In this final section, the scope for further research is examined. Four areas can be identified and each is taken in turn below.

8.5.1. A stochastic hedonic cost frontier

Chapter 5 demonstrated the value of the hedonic cost function for incorporating characteristics of output. This led to a richer understanding of the cost characteristics of the industry, and, importantly, led to different policy conclusions in terms of optimal franchise sizes than relying on simpler measures of RtS and RtD. However Chapter 5 does not incorporate allowance for cost inefficiency in the analysis.

Incorporating cost inefficiency is desirable both in the context of Chapter 5 and more generally. It should be noted that following the McNulty (2011) review, ORR is expected to provide oversight on TOCs costs as well as for the IM, Network Rail. Thus it is likely that cost benchmarking will be relevant to this aim. More generally, the hedonic cost function is a useful device to incorporate characteristics of output into cost functions. Such incorporation is also desirable from an efficiency measurement perspective. Ultimately, inefficiency should capture the extent to which actual costs are above best practice minimum cost. Thus, controlling for output characteristics which are outside the control of management will enhance the measurement of cost efficiency.

One difficulty with considering cost inefficiency in the model in Chapter 5 is that the cost share equations would contain the impact of allocative inefficiency. As explained in Chapter 5, this implies a intricate error structure between the cost share equations and the cost function (see p. 125 of this thesis for further discussion).

An approach to estimating such a model could be to estimate the cost frontier as a single equation and ignore the cost share equations. This could be undertaken using maximum likelihood estimation. However, it should be noted that this would require bespoke computer code, as, to the author's knowledge, no commercially available software, such as LIMDEP (Econometric Software, 2010a), FRONTIER (Coelli, 1996) and STATA (STATA Corp., 2013) allows for non-linear (in parameters) functional forms in the context of stochastic frontier models. Even if this could be programmed, it is likely that the estimator would be less efficient relative to one which exploited the cost share relationship.

One pragmatic approach would be to use the parameter estimates from the SUR estimation and undertake further ML regression of the residuals to decompose them into noise and cost inefficiency. This is a multi-stage procedure similar to those considered in Chapter 6 (see footnote 25, p. 126).

8.5.2. Unobserved heterogeneity in the dual level inefficiency model

Chapter 6 considered the multi-level dataset comprising sub-company observation on several companies. In Chapter 6, apart from the noise error term and the group means in the model with the Mundlak decomposition, all other error components were interpreted as representative of inefficiency at either the sub-company level or the more aggregate firm level. As discussed in section 3.5.3 and in Chapter 6, there are alternative interpretations on the error components that they are actually representative of other sub-company invariant and/or time invariant unobserved factors, not just inefficiency.

The recent paper by Kumbhakar et al (2012) provides a useful starting point, in terms of decomposing a time or sub-company invariant effect into two components. The Kumbhakar et al (2012) model was given in (3.23) and reproduced below

$$C_{it} = \alpha_i + f(X_{it};\beta) + u_{it} + w_i + v_{it} \qquad (8.1)$$

where $u_{it} \sim \left|N\left(0,\sigma_u^2\right)\right|$, $w_i \sim \left|N\left(0,\sigma_w^2\right)\right|$, $v_{it} \sim N\left(0,\sigma_v^2\right)$ and $\alpha_i \sim N\left(0,\sigma_w^2\right)$ and all error components uncorrelated with X_{it} and all error components independent from each other.

A clear analogy for sub-company data using the notation in Chapter 6 is:

$$C_{it} = \alpha_i + f(X_{its};\beta) + u_{is} + w_i + v_{its} \qquad (8.2)$$

where $u_{is} \sim \left|N\left(0, \sigma_u^2\right)\right|$, $w_i \sim \left|N\left(0, \sigma_w^2\right)\right|$, $v_{its} \sim N\left(0, \sigma_v^2\right)$ and $\alpha_i \sim N\left(0, \sigma_w^2\right)$ and all error components uncorrelated with X_{its} and all error components independent from each other.

Thus this model distinguishes between unobserved heterogeneity which is time and sub-company invariant, inefficiency which is sub-company and time invariant and inefficiency which is time invariant (but differs across sub-companies). Thus this is the Kumbhakar et al, type extension of the model used in the empirical example in Chapter 6 (with time invariant inefficiency).

There are potentially other formulations which could be investigated to exploit the multi-level nature of the data (such as additionally controlling for sub-company varying but time invariant unobserved heterogeneity). The feasibility of estimation and sensitivity to distributional assumptions should be examined. However, it is important to remember that ultimately inefficiency is modelled as a residual and so is something unexplained by the model. As such, in the absence of including actual measures of heterogeneity directly into the cost function, there will always be some arbitrary distinction between (unobserved) inefficiency and unobserved heterogeneity, irrespective of the sophistication of the modelling approach. The approach in Chapter 5 to incorporating heterogeneity within cost functions still has substantial merit.

8.5.3. Accounting for uncertainty in panel data stochastic frontier models

Chapter 7 applied methods of approximating prediction intervals for firm specific inefficiency in stochastic frontier models using cross sectional data. Further research would apply the techniques to fully parametric panel data models. Note that the distributions required both the bagging and multi-variate normal approaches (the equivalent distribution to that in (7.2)) are known for panel data models, hence such an application should be possible to implement. Note, however, that the semi-parametric models (e.g. the time invariant models of Schmidt and Sickles (1984)) require different techniques such as Multiple Comparisons with the Best (Horrace and Schmidt, 1996 and Horrace and Schmidt, 2000).

There is an interesting research question which emerges from such an extension. What is the trade-off between parameter uncertainty and decomposition uncertainty with respect to panel data model formulation? In particular, panel data models such as Pitt and Lee (1981), Battese and Coelli (1992) and Cuesta (2000) have the property that decompositional uncertainty diminishes over time (as a result of exploiting persistency in inefficiency over time in the data). However, parametric functions are needed in order to allow for variation in inefficiency over time. For full flexibility in variation over time, this requires firm specific parameters to be estimated, which in small samples will add to parameter uncertainty. Thus there emerges an interesting research topic - the extent to which inefficiency is predicted more precisely in panel models vis-à-vis pooling the data.

8.5.4. Uncertainty in regulatory application

More generally, there is an issue as to how regulators of infrastructure companies should use top down efficiency analysis. In particular, to determine the ‘X’ in the RPI-X regulatory approach, regulators tend to weight the results of a number of studies. This can often involve implicitly weighting the results of studies which use different datasets (and methods), but also models which adopt different methods/formulations using the same dataset. For example, ORR’s efficiency determination for Network Rail in the 2008 Periodic Review which considered a mixture of top down and bottom up analysis as well as examining a number of models estimated on two data sets (one of which is the dataset used in Chapter 6) (ORR, 2008).

Chapter 7 is concerned with quantifying uncertainty in the prediction of firm specific inefficiency. Several questions emerge. To what extent does this quantification capture the same uncertainty as that of (somehow) synthesising the results of many models estimated using the same or an alternative dataset? How should regulators weigh the evidence at their disposal? Are there any particular issues with the fact that ‘robustness’ often refers to the robustness of the prediction of the inefficiency of one firm (the regulated firm) out of many firms, when most statistical tests of a model relate to how costs are described over all firms? Clearly a model which is robust for all firms is desirable, but what if a model gives odd predictions for the regulated firm (which of course is possible due to ‘luck’ in sampling)? These are exciting and highly relevant future research questions.

References

Affuso, L., A. Angeriz, and M.G. Pollitt (2002): 'Measuring the Efficiency of Britain's Privatised Train Operating Companies', Regulation Initiative Discussion Paper Series, no: 48, London Business School.

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).

Aigner, D.J., Lovell, C.A.K, and Schmidt, P. (1977), 'Formulation and Estimation of Stochastic Frontier Production Function Models', *Journal of Econometrics*, 6 (1), 21-37.

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.

Almanidis, P. and Sickles, R. (2010). 'The Skewness Problem in Stochastic Frontier Models: Fact or Fiction?', in Keilegom, I and Wilson, P (Eds*) Exploring Research Frontiers in Contemporary Statistics and Econometrics:* A Festschrift in Honor of Leopold Simar, Springer Publishing, New York.

Alvarez, A., C. Amsler, L. Orea, and P. Schmidt (2006). ‘Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics’, *Journal of Productivity Analysis*, 25, 201-212.

Amsler C., Leonard, M. and Schmidt, P. (2010). ‘Estimation and Inference in Parametric Deterministic Frontier Models’, Working Paper.

Andersson, M. (2006). ‘Marginal railway infrastructure cost estimates in the presence of unobserved effects’, *Case study 1.2D I Annex to Deliverable D 3 Marginal cost case studies for road and rail transport, Information Requirements for Monitoring Implementation of Social Marginal Cost Pricing*, EU Sixth Framework Project GRACE (Generalisation of Research on Accounts and Cost Estimation).

Andersson, M, Smith, A., Wikberg, A. and Wheat, P. (2012). ‘Estimating the marginal cost of railway track renewals using corner solution models’, *Transportation Research Part A*, 46 (6), 954-964.

Andrikopoulos, A. and Loizides, J, (1998). ‘Cost Structure and Productivity Growth in European Railway Systems’, *Applied Economics*, 30, 1625-1639.

Arrow, K., Chenery, H., Minhas, B. and Solow, R. (1961). ‘Capital-Labor Substitution and Economic Efficiency’, *Review of Economics and Statistics,* 45, 225-247.

Asmild, M, Holvad, T, Hougaard, JL and Kronborg, D. (2008). Railway Reforms:

Do They Influence Operating Efficiency? Department of Economics Discussion Papers, No. 08_05. Copenhagen: University of Copenhagen.

Battese, G.E. and Coelli, T.J. (1988). ‘Prediction of Firm-Level Technical Efficiencies With a Generalised Frontier Production Function and Panel Data’, *Journal of Econometrics,* 38, 387-399.

Battese, G.E. and Coelli, T.J. (1992). ‘Frontier Production Functions and the Efficiencies of Indian Farms Using Panel Data from ICRISAT’s Village Level Studies’, *Journal of Quantitative Economics*, 5, 327-348.

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.

Beesley, M., Littlechild, S. (1988). ‘The regulation of privatized monopolies in the United Kingdom’. *RAND Journal of Economics* 20, 454–472.

Bera, A. K. and Sharma, S. C. (1999). “Estimating Production Uncertainty in Stochastic Frontier Production Frontier Models”, *Journal of Productivity Analysis* 12, 187-210.

Bishop, M. and Thompson, D. (1992). ‘Regulatory Reform and Productivity Growth in the UK’s Public Utilities’, *Applied Economics*, 24, 1181-1190.

Bitzan, J. D. And Wilson, W.W. (2007). ‘A hedonic Cost Function Approach to Estimating Railroad Costs’. *Research in Transportation Economics*, 20 (1), 69-95.

Borts, G. (1960), ‘The Estimation of Rail Cost Functions.’ *Econometrica*, 28, 108-131.

Brenck, H. and Peter, M. (2007). ‘Experience with Competitive Tendering in Germany’ in European Conference of Ministers of Transport *Competitive Tendering for Rail Services,* Paris.

Breusch, T. S. and Pagan, A. R. (1980). ‘The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics’, *Review of Economic Studies* 47,. 239-253.

Brorsen, B.W. and Kim, T. (2013). ‘Data aggregation in stochastic frontier models: the closed skew normal distribution’, *Journal of Productivity Analysis*, 39, 21-37.

Button, K. (2010). *Transport Economics*. 3rd Edition, Edward Elgar Publishing Ltd.

Cantos, P., Pastor, J.M. and Serrano, L., (2010). ‘Vertical ,and Horizontal Separation in the European Railway Sector and its effects on productivity’, *Journal of Transport Economics and Policy*, 22 (2) 139-160.

Cantos, P. and Villarroya, J. (2000), ‘Efficiency, Technical Change and Productivity in the European Rail Sector: A Stochastic Frontier Approach’, *International Journal* *of Transport Economics*, 27(1), 55-76.

Cantos, P. and Villarroya, J (2001). ‘Regulation and Efficiency: The Case of European Railways’, *Transportation Research Part A*, 35 (5), 459-472.

Caves, D.W., Christensen, L.R. and Swanson, J.A. (1981). ‘Productivity Growth, Scale Economies, and Capacity Utilisation in U.S. Railroads, 1955-74, *American Economic Review*, 71 (5). 994-1002.

Caves, D.W., Christensen, L.R. and Tretheway, M.W. (1984). ‘Economies of Density versus Economies of Scale: Why Trunk and Local Service Airline Costs Differ’, *The RAND Journal of Economics*, 15 (4) 471-489.

Caves, D.W., Christensen, L.R., Tretheway, M.W. and Windle, R.J. (1985), ‘Network Effects and the Measurement of Returns to Scale and Density for U.S. Railroads’, in Daughety, A.F., ed., *Analytical Studies in Transport Economics*, Cambridge, Cambridge University Press, pp. 97-120.

Chambers, R. G. (1988). *Applied production analysis.* Cambridge University Press, UK.

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

Christensen, L.R., and Greene, W.H. (1976). ‘Economies of scale in U.S. electric power generation’, *Journal of Political Economy* 84, 1, 655-676.

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

Christopoulos, D.K., Loizides, J., and Tsionas, E.G. (2000). ‘Measuring Input-Specific Technical Inefficiency In European Railways: A Panel Data Approach’, *International Journal of Transport Economics*, 27 (2), 147-171.

Coelli, T. J. (1996). “A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation.” *CEPA Working Paper 96/07*. University of New England, Australia.

Coelli, T. and Perelman, S. (1999). ‘A Comparison of Parametric and non-Parametric Distance Functions: With Application to European Railways’, *European Journal of Operational Research*, 117, 326-339.

Coelli, T. and Perelman, S. (2000). ‘Technical Efficiency of European Railways: A Distance Function Approach’, *Applied Economics*, 32, 1967-1976.

Coelli, T. J., Rao, D. S. P., O’Donnel, C. J. and Battese, G. E (2005). *An Introduction to Efficiency and Productivity Analysis*. Second Edition, Springer.

Cornwell, C., Schmidt, P. and Sickles, R.C. (1990). ‘Production Frontiers With Cross-Sectional And Time-Series Variation in Efficiency Levels’, *Journal of Econometrics*, 46, 185-200.

Couto, A. and Graham, D.J. (2008). ‘The contributions of technical and allocative efficiency to the economic performance of European railways’, *Portuguese Economic Journal*, 7, 125-153.

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.

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

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.

Daraio, C. and Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis: methodology and applications.* Springer, New York.

Debreu, G. (1951). “The coefficient of resource utilization,” *Econometrica* 19:3 (July), 273-292.

Department for Transport (2012). *Reforming our Railways: Putting the Customer First.* Command Paper. Printed in the UK by The Stationery Office Limited
on behalf of the Controller of Her Majesty’s Stationery Office

Deprins, D. and Simar, L. (1989). ‘Estimating Technical Inefficiencies With Correction For Environmental Conditions’, *Annals of Public and Cooperative Economics*, 81-102.

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

Diewert, W. E and Wales, T. J. (1987). ‘Flexible Functional Forms and Global

Curvature Conditions' *Econometrica* 55 (1), 43-68.

Econometric Software Inc. (2010a). LIMDEP v9. http://www.limdep.com, Plainview, NY.

Econometric Software Inc. (2010b). *LIMDEP, User's Manual*, http://www.limdep.com, Plainview, NY.

European Commission (2013). The Fourth Railway Package – Completing the Single European Railway Area to Foster European Competitiveness and Growth. COM (2013) 25 Final.

Farrell, M.J. (1957). 'The Measurement of Productive Efficiency', *Journal of the Royal Statistical Society*, Series A (general), 120 (3), 253-290.

Farsi, M., Filippini, M. and Greene, W.H. (2005a). 'Efficiency Measurement in Network Industries: Application to the Swiss Railway Companies', *Journal of Regulatory Economics*, 28 (1), 69-90.

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

Fielding, G. J., T. T. Babitsky, and M. E. Brenner (1985). 'Performance Evaluation for Bus Transit', *Transportation Research*, 19A, 73–82.

Flores-Lagunes, A. Horrace, W. C. and Schnier, K. E. (2007). ‘Identifying technically efficient fishing vessels: A non-empty, minimal subset approach’, *Journal of Applied Econometrics* 22 729-745.

Friedlaender, A. F. (1969). *The Dilemma of Freight Transport Regulation.* Brookings.

Fuss, M. and McFadden, D. (1979). *Production economics: a dual approach to theory and applications,* Edited by Melvyn Fuss and Daniel McFadden, North Holldan: Amsterdam.

Gathon, H.J. and Perelman, S. (1992). ‘Measuring Technical Efficiency in European Railways: a Panel Data Approach’, *The Journal of Productivity Analysis*, 3, 135-151.

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

Gaudry, M. and Quinet, E. (2003), *Rail Track Wear-and-Tear Costs by Traffic Class in France*, Universite de Montreal, Publication AJD-66.

Greene, W. H. (1980). ‘Maximum Likelihood Estimation of Econometric Frontier Functions’, *Journal of Econometrics* 13:1 (May), 27-56.

Greene, W.H. (1990). ‘A Gamma-distributed Stochastic Frontier Model’, *Journal of Econometrics*, 46, 141-164.

Greene, W.H. (2003). *Econometric Analysis.* 5th Edition, Pearson Prentice Hall.

Greene, W.H. (2005).’Reconsidering Heterogeneity in Panel data Estimators of the Stochastic Frontier Model,’ *Journal of Econometrics* 126, 269-303.

Greene, W.H. (2008), ‘The Econometric Approach to Efficiency Analysis’, in Fried, H.O., Lovell, C.A.K. and Schmidt, S.S., eds., *The Measurement of Productive Efficiency Growth*, 2nd Ed. New York, Oxford University Press.

Greene, W. H. (2012), *Econometric Analysis*, 7th Edition, Prentice Hall, New York.

Griliches, Z. (1972), ‘Cost allocation in railroad regulation’ *Bell Journal of Economics and Management Science*, Spring, 26-41.

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.

Haug, C.J. and Liu, J. (1994). ‘Estimation of a Non-Neutral Stochastic Frontier Production Function’, *Journal of Productivity Analysis,* 5, 171-180

Hausman, J. A. and Taylor, W. (1981). ‘Panel data and Unobservable Individual

Effects,’ *Econometrica*, 49, 1377-1398.

Hausman, J. A. (1978). “Specification Tests in Econometrics,” *Econometrica*, 46, 1251-1272.

Hensher, D. and Brewer, A (2000). *Transport: An Economics and Management Perspective.* Oxford University Press.

Heshmati, A. and S. Kumbhakar, (1994). ‘Farm Heterogeneity and Technical Efficiency: Some Results from Swedish Dairy Farms’, *Journal of Productivity Analysis*, 5, 45-61.

Heathfield, D. F. and Wibe, S. (1987). *An Introduction to Cost and Production Functions.* Macmillian Education Ltd, Hampshire UK.

Hicks, J. R. (1935). ‘The theory of monopoly: A survey’, *Econometrica* 3:1 (January), 1-20.

Hjalmarsson, L., Kumbhakar, S.C., and Heshmati, A. (1996). ‘DEA, DFA and SF: A comparison’, *Journal of Productivity Analysis*, 7, 303-327.

Horrace, W. C. (2005). ‘On ranking and selection from independent truncated normal distributions’, *Journal of Econometrics*, 126, 335-354.

Horrace, W.C. and Schmidt, P. (1996). ‘Confidence Statements for Efficiency

Estimates from Stochastic Frontier Models', *Journal of Productivity Analysis*, 7:2/3, 257-282.

Horrace, W.C. and Schmidt, P. (2000). 'Multiple comparisons with the best, with Economic Applications', *Journal of Applied Econometrics*, 15, 1-26.

Ivaldi, M. and McCullough, G.J (2001). 'Density and integration effects on class I U.S. freight railroads', *Journal of Regulatory Economics*, 19 (2), 161-182.

Jara-Diaz, S. R (1982). 'The estimation of transport cost functions: a methodological review', *Transport Reviews*, 2 (3), 257-278.

Johansson, P. and Nilsson, J. E. (2002), *An Economic Analysis of Track Maintenance Costs*, Deliverable 10 Annex A3 of UNITE (UNIfication of accounts and marginal costs for Transport Efficiency), Funded by EU 5th Framework RTD Programme. ITS, University of Leeds, Leeds. Online: http://www.its.leeds.ac.uk/projects/unite/.

Johansson, P. and Nilsson, J. (2004). 'An economic analysis of track maintenance costs', *Transport Policy*, 11(3), 277-286.

Jondrow, J., Lovell, C.A.K., Materov, I.S. and Schmidt, P. (1982). 'On Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model', *Journal of Econometrics*, 19, 233-238.

Jones, I. (2000). ‘Developments in Transport Policy. The Evolution of Policy Towards On-Rail Competition in Great Britain’. *Journal of Transport Economics and Policy*, 34 (3), 371-384.

Jupe, R. and Crompton, G. (2006). ‘A deficient performance: The regulation of the train operating companies in Britain’s privatised railway system’, *Critical Perspectives on Accounting* 17 (2006) 1035–1065.

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*, 38 (2), 157-190.

Kim, Y. And Schmidt, P. (2008). ‘Marginal comparisons with the Best and the Efficiency Measurement Problem’, *Journal of Business & Economic Statistics.* 26 (2) 253-260.

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

Krinsky, I., Robb, A., (1986). ‘On approximating the statistical properties of elasticities’. *Review of Economics and Statistics* 68 (4), 715-719.

Kuosmanen, T. and Kortelainen, M. (2012). ‘Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints', *Journal of Productivity Analysis,* August 2012, 38 (1), 11-28

Kumbhakar, S.C. (1988a). 'Estimation of Input-specific technical and Allocative Inefficiency in Stochastic Frontier Models', *Oxford Economic Papers*. 40 (3), 535-549.

Kumbhakar, S.C. (1988b). 'On the Estimation of Technical and Allocative Inefficiency Using Frontier Functions: The Case of U.S. Class I Railroads', *International Economic Review*. 29 (4), 727-43.

Kumbhakar, S.C. (1990). 'Production Frontiers, Panel Data, and Time-Varying Technical Inefficiency', *Journal of Econometrics*, 46, 201-211.

Kumbhakar, S., (1991). 'Estimation of Technical Inefficiency in Panel Data Models with Firm- and Time Specific Effects', *Economics Letters*, 36, 43-48.

Kumbhakar, S.C. (1997). 'Modelling allocative inefficiency in a Translog cost function and cost share equations: an exact relationship', *Journal of Econometrics* 76 351–356

Kumbhakar, S., and A. Heshmati, (1995). 'Efficiency Measurement in Swedish Dairy Farms 1976-1988 Using Rotating Panel Data', *American Journal of Agricultural Economics*, 77, 660-674

Kumbhakar, S., and L. Hjalmarsson, (1995). ‘Labor Use Efficiency in Swedish Social Insurance Offices’, *Journal of Applied Econometrics*, 10, 33-47

Kumbhakar, S.C., Lien, G. and Hardaker, J.B. (2012). ‘Technical efficiency in competing panel data models: a study of Norwegian grain farming’, *Journal of Productivity Analysis*. Online first.

Kumbhakar, S. C. and Löthgren, M. (1998). ‘A Monte Carlo Analysis of Technical Inefficiency Predictors’, Working Paper Series in Economics and Finance, No. 229, Stockholm School of Economics.

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

Kumbhakar, S.C., Orea, L., Rodriguez-Alvarez, A. and Tsionas, E.G. (2007). ‘Do we estimate an input or an output distance function? An application of the mixture approach to European railways’, *Journal of Productivity Analysis,* 27, 87-100.

Lan, L.W. and Lin, E.T.J. (2006). ‘Performance Measurement for Railway Transport: Stochastic Distance Functions with Inefficiency and Ineffectiveness Effects’ *Journal of Transport Economics and Policy,* 40 (3), 383–408.

Lancaster, T. (2000). ‘The Incidental Parameters Problem Since 1948’, *Journal of Econometrics*, 95, 391-414.

Lee, T. and Baumel, C. P. (1987). 'The Cost Structure of the U.S. Railroad Industry Under Deregulation.,' *Journal of the Transportation Research Forum* 27 (1) 245-253.

Lee, Y.H. and Schmidt, P. (1993). 'A Production Frontier Model with Flexible Temporal Variation in Technical Efficiency', in Fried, H.O., Lovell, C.A.K. and Schmidt, S.S., eds., *The Measurement of Productive Efficiency*, New York, Oxford University Press.

Lalive, R. and Schmutzler, A. (2008). 'Entry in Liberalized Railway Markets: The German Experience', *Review of Network Economics*, De Gruyter, 7 (1), 3.

LEK (2003). *Regional Benchmarking: Report to Network Rail, ORR and SRA,* London.

Link, H., Stuhlemmer, A. (DIW Berlin), Haraldsson, M. (VTI), Abrantes, P., Wheat, P., Iwnicki, S., Nash, C., Smith, A., CATRIN (Cost Allocation of TRansport INfrastructure cost), Deliverable D 1, Cost allocation Practices in the European Transport Sector. Funded by Sixth Framework Programme. VTI, Stockholm, March 2008.

Loizides, J. and Tsionas, E.G. (2002). 'Productivity Growth in European Railways: a New Approach', *Transportation Research Part A*, 36 (7), 633-644.

Marti, M. and Neuenschwander, R. (2006). *Case study 1.2E: Track Maintenance*

Costs in Switzerland, Annex to GRACE (Generalisation of Research on Accounts and Cost Estimation) Deliverable D3: Marginal Cost Case Studies for Road and Rail Transport. Funded by 6th Framework RTD Programme. Ecoplan, Berne, Switzerland.

McNulty, R. (2011). *Realising the potential of GB Rail: Report of the Rail Value for Money Study. Summary Report.* Report for the Office of Rail Regulation and Department for Transport, Great Britain.

Meeusen, W. and van Den Broeck, J. (1977). 'Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error', *International Economic Review*, 18 (2), 435-444.

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*.

Meyer, J. R. (1958). 'Some Methodological Aspects of Statistical Costings as illustrated by the Determination of Rail Passenger Costs', *American Economic Review* 48 (2) 209-222.

Meyer, J. R. et al (1959). *The Economics of Competition in the Transportation Industries.* Harvard University Press.

Meyer, J. R. and Craft, G. (1961). 'The Evaluation of Statistical Costing Techniques as Applied in the Transportation Industry', *American Economic Review* 51 (2) 313-334.

Moulton, B. R. and Randolph, W. C. (1989). 'Alternative tests of the error components model', *Econometrica*, 57, 685-693.

Mundlak, Y. (1978). 'On the pooling of time series and cross section data', *Econometrica*, 64, 69–85.

Munduch, G., Pfister, A., Sögner, L. and Stiassny, A. (2002), *Estimating Marginal Costs for the Austrian Railway System*, Working Paper 78, Department of Economics, Vienna University of Economics and B.A., Vienna, Austria.

Nash, C.A. and Nilsson J.E. (2009). 'Competitive tendering of rail services – a comparison of Britain and Sweden', *paper presented to the international conference on competition and ownership in land passenger transport*, Delft, September 2009.

NERA (2000). *Review of Overseas Railway Efficiency: A Draft Final Report for the Office of the Rail Regulator*, London.

Nerlove, M. (1963). 'Returns to Scale in Electricity Supply', In *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld.* Stanford University.

Neyman, J. and E. Scott. (1948). 'Consistent Estimates Based on Partially Consistent

Observations', *Econometrica*, 16: 1-32.

Office of Rail Regulation. (2008). *Periodic review of Network Rail's outputs and funding for 2009-2014***.** London.

Office of Rail Regulation (2012). *National Rail Trends*. Available at http://www.rail-reg.gov.uk/server/show/nav.2026 accessed 29/11/2012.

Orea, L. and Kumbhakar, S.C. (2004). 'Efficiency measurement using a latent class stochastic frontier model', *Empirical Economics*, 29, 169-183

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

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*, 28, 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.

OXERA (2009). *Recommendations on how to model efficiency for future price reviews*. Available at http://www.rail-reg.gov.uk/server/show/nav.2499, accessed 05/11/2013.

Parisio, L. (1999). ‘A comparative analysis of European railroads efficiency: a cost frontier approach’, *Applied Economics*, 31, 815-823.

Passenger Demand Forecasting Council (2005). *Passenger Demand Forecasting Handbook (PDFH).* Version 5.1. Published April 2013.

Pitt, M.M. and Lee, L.F. (1981). ‘Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry’, *Journal of Development Economics*, 9, 43-64.

Qian, J. and Sickles, R. C. (2007). *Stochastic Frontiers with Bounded Inefficiency.* Working Paper, Rice University.

Quantitative Micro Software (2007). *Eviews v.6.* http://www.eviews.com/home.html.

R Development Core Team (2010). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

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

Sickles, R. C. (2005). ‘Panel estimators and the identification of firm-specific efficiency levels in parametric, semiparametric and nonparametric settings’, *Journal of Econometrics,* 126, 305-334.

Simar, L. And Wilson, P. W. (2010). ‘Inferences from Cross-Sectional, Stochastic Frontier Models’, *Econometric Reviews*, 29 (1), 62-98.

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, 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. and Wheat, P. (2010). *Sensitivity analysis on the UIC harmonisation factors.* Report for RailConsult. April 2010

Smith A.S.J and Wheat P. (2012a). ‘Evaluating alternative policy responses to franchise failure: Evidence from the passenger rail sector in Britain’, *Journal of*

Transport Economics and Policy, 46 (1), 25-49.

Smith A.S.J and Wheat P. (2012b). ‘Estimation of Cost Inefficiency in Panel Data Models with Firm Specific and Sub-Company Specific Effects’, *Journal of Productivity Analysis*, 37 (1), 27-40.

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 Nixon, H. (2008). *International Benchmarking of Network Rail’s Maintenance and Renewal Costs*, joint ITS, University of Leeds and ORR report written as part of PR2008, June 2008. Presentation available at www.rail-reg.gov.uk.

Smith, A.S.J., Wheat, P.E. and Smith, G. (2010). ‘The role of international benchmarking in developing rail infrastructure efficiency estimates’, *Utilities Policy,* 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.

STATA Corp. (2013). STATA v 13. http://www.stata.com/.

Stevenson, R. E. (1980), ‘Likelihood Functions for Generalized Stochastic Frontier

Estimation', *Journal of Econometrics* 13 (1) (May), 57-66.

Studenmund, A. H. (2011). *Using Econometrics: A Practical Guide.* 6th Edition. Pearson Eductation Inc. Boston, MA US.

Taube, R. (1988). *Möglichkeiten der Effizienzmess ung von öffentlichen* Verwaltungen. Berlin: Duncker & Humbolt GmbH.

Tervonen, J. and Idstrom, T. (2004). *Marginal Rail Infrastructure Costs in Finland 1997-2002,* Report by the Finnish Rail Administration. Available at www.rhk.fi [Accessed 20/07/2005].

Theil, H. (1954). *Linear Aggregation of Economic Relations,* North Holland Publishing Company, Amsterdam.

Train, K. E. (2009). *Discrete Choice Methods with Simulation,* 2nd Ed, Cambridge University Press, New York.

Tsionas, E.G. and Christopoulos, D.K. (1999). 'Determinants of Technical Inefficiency in European Railways: Simultaneous Estimation of Firm-Specific and Time-Varying Inefficiency', *KONJUNKTURPOLITIK*, 45, 240-256.

UIC (International Union of Railways) (2008). Lasting Infrastructure Cost Benchmarking (LICB). December 2008, available at http://www.uic.org/IMG/pdf/li08C_sum_en.pdf [Accessed 20/12/2013].

Vickers, J. and Yarrow, G. (1988). *Privatization: An Economic Analysis.* MIT Press.

Waldman, D. M (1984). ‘Properties of technical efficiency estimators in the stochastic frontier model’, *Journal of Econometrics,* 25, 353-364.

Wang, W. S. and Schmidt, P. (2009). ‘On the distribution of estimated technical efficiency in stochastic frontier models’, *Journal of Productivity Analysis*, 148, 36-45.

Wardman, M. (2006). ‘Demand for Rail Travel and the Effects of External Factors’, *Transportation Research. Part E,* 42 (3), 129-148.

Wheat, P., Greene, W. and Smith, A. (2013). ‘Understanding prediction intervals for firm specific inefficiency scores from parametric stochastic frontier models', Journal of Productivity Analysis. In print.

Wheat, P. and Smith, A. (2008). ‘Assessing the Marginal Infrastructure Maintenance Wear and Tear Costs for Britain’s Railway Network’, *Journal of Transport Economics and Policy,* 42 (2), 189-224.

Wheat, P. and Smith, A. (2013). ‘Do the usual results of railway returns to scale and density hold in the case of heterogeneity in outputs: A hedonic cost function approach’. Accepted for publication in Journal of Transport Economics and Policy.

Wheat, P. Smith, A. and Nash, C. (2007). *Rail Research UK Project B4: System Level Cost Framework for the Assessment of Sub-System Trade-Offs: Final Report*. Available from the authors by request.

Wheat, P., Smith, A.S.J. and Nash, C.A. (2009). CATRIN (Cost Allocation of TRansport INfrastructure cost), Deliverable 8 - Rail Cost Allocation for Europe. Funded by Sixth Framework Programme. Coordinated by VTI, Stockholm.

Wilson, W. W. (1997). ‘Cost Savings and Productivity in the Railroad Industry’, *Journal of Regulatory Economics*, 11, 21-40.

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