The Routledge Handbook of Public Transport

Corinne Mulley, John D. Nelson, Stephen Ison

Public transport productivity and efficiency assessment

Publication details
Jonathan Cowie
Published online on: 13 May 2021

How to cite: Jonathan Cowie. 13 May 2021, Public transport productivity and efficiency assessment from: The Routledge Handbook of Public Transport Routledge
Accessed on: 14 Sep 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
12
PUBLIC TRANSPORT
PRODUCTIVITY AND
EFFICIENCY ASSESSMENT

Jonathan Cowie

Introduction

Over the years, many academic articles have been published assessing the ‘efficiency’ of public transport. What this indicates is a high level of interest in the subject area, underlined by the fact that a dedicated chapter on the subject is included in this Handbook. It does nevertheless raise a question seldom asked, namely why is there such a high level of interest in the subject? The main reason is normally associated with some kind of regulatory reform in the provision of public transport services, and as such, efficiency assessment is used as an analytical tool to examine an aspect of the outcome of such reforms. For example, rather than simply posing the question ‘how efficient is the German interstate bus industry and how has it evolved over time?’, the question could be more of the form ‘what impact have regulatory reforms had on the production efficiency of the German interstate bus industry?’ Hence the study of efficiency is being used to assess the impact of reform and, even more crucially, one aspect of that impact, specifically the productive efficiency of interstate bus services. As will be seen in the course of this chapter, however, efficiency assessment can also include elements regarding patronage levels. It basically depends on what ‘type’ of efficiency is being assessed; any such assessment, however, will still be from a production perspective. Hence the question could be ‘what impact have regulatory reforms had on the efficiency of the carriage of passengers in the German interstate bus industry?’ Efficiency assessment therefore has both a demand characteristic (patronage levels) and a supply perspective (the production of public transport services).

This chapter aims to outline the main issues and methods used to assess the productivity/efficiency of public transport from a practical perspective and should be viewed as complementary to a chapter on the subject co-authored in an earlier Routledge Handbook (Merkert & Cowie, 2018). As such, it will focus attention on econometric approaches, as these encompass underlying production theory and the implications within an empirical context. The study of efficiency is a vast area, however, and as a consequence, there is a limit as to what can be examined and outlined in a single chapter. Therefore, only the assessment of technical efficiency will be considered, namely the physical output to physical inputs ratio, more commonly known as total factor productivity, although all issues outlined provide the basis for other assessments of productive/cost efficiency. The chapter ends by considering some of the studies that have been carried out into the efficiency of public transport.
Efficiency and productivity – concepts and ideas

Consider the following equation:

\[
\frac{\sum_{i=1}^{M} Q_i}{\sum_{j=1}^{N} X_j} \quad [12.1]
\]

In Equation 12.1, the firm has \( M \) outputs \( Q \) and uses \( N \) inputs \( X \) to produce these. Hence, as an example, a bus company could have two outputs, timetabled and chartered services, and use three inputs to produce these, labour, vehicles and fuel. What Equation 12.1 actually shows is the total factor productivity index for this particular firm. Productivity is therefore the sum of the outputs divided by the sum of the inputs, simply referred to as the output to inputs ratio. Whilst this gives a single observation for a single year, productivity is normally viewed as a long-run concept; hence, the output to input ratio is monitored over time. Efficiency, on the other hand, is a relative assessment of that ratio at a specific point in time, with the reference point normally being against best practice. Both concepts are illustrated in Figure 12.1

‘Best practice’ could be defined as the theoretical maximum level of output which could be obtained with the inputs used in the production process, but the problem in reality is that best practice can only be assessed on what actually exists, not on the (unknown) theoretical maximum. As a consequence, efficiency measurement is the extent to which a service or operator performs in relation to the ‘best’ service/operator and, as such, is a relative rather than an absolute measure.

Practicalities, the choice of output

In terms of the practicalities of assessment, the first issue is to identify the ‘output’ of a public transport system. Public transport is a derived demand from the need for individuals to be somewhere else to undertake some other activity, for example, travel from home to work in order to earn an income. It also has a wider role to play both economically and in social welfare terms. Major cities would find it very difficult to sustain long-term economic growth without significant investment in the public transport network. In terms of social welfare, public transport is a participative requirement (Barr, 2020) that provides a minimum level of mobility for all members within society. What this leads to is the dual role of public transport as a carrier of
people and a provider of transport services. In terms of productivity, the former would best be reflected in an output measure that relates to usage/consumer demand, whilst the latter represents service provision/supply.

Regarding the main arguments in favour of demand-related measures, these are summarised by De Borger et al. (2002) as:

- Passengers or passenger kilometres partially capture the economic motive for providing services; therefore, it follows that a demand-related measure must be more relevant. This also assists in overcoming the fact that systems that carry fewer passengers produce more efficient services.
- Strong interdependency exists between demand and the spatial and quality attributes of supply and the appropriate specification of technology employed in the provision of public transport services. The basic argument, therefore, is that this is best ‘captured’ by a demand-based variable, as supply (even in a planned regulated market) will always ultimately be dependent upon demand. This has been particularly true in both the regulated London bus market and the British franchised rail market in recent years (see also Chapter 14).

Berechman (1993), on the other hand, gives the main arguments for the use of a supply-based measure:

- Inputs do not vary systematically with demand-related output measures so that they do not allow an adequate description of transit technology.
- Supply-related output indicators are under the control of operators to a greater extent than demand-related outputs. In unregulated environments, they represent direct decision variables for the transit companies, ones that would vary with the level of demand. In some ways, therefore, this provides a compromise over which output measure to use. Even where this is not the case, then as a result of government regulation of service levels, operators would still have some control via the negotiation process through which they will agree to the services to be provided, that is, the level of supply.
- Independent of the achievement of broader goals defined in terms of actual patronage levels, supplying public transport services in the least costly way may be considered a reasonable requirement for operators.

In many respects, the issue of which output measure to use will come down to the regulatory environment under which the public transport service is being provided (see also Chapters 1, 2 and 13). In heavily regulated environments with tight controls on service frequencies and fares, this aligns more with the idea of public transport as provider of a public service, and hence a supply-side measure would be used. This would be in contrast to a regulatory environment that was far more market based, such as is the case with the deregulated British bus market outside of London. In this case, operators are free to specify routes and fares, and the drive for profitability is achieved by carrying passengers as ‘efficiently’ as possible; hence, a demand-based measure would be used.

De Borger et al. (2002) also highlight that there is a general recognition of the heterogeneity of transport output in terms of temporal, spatial and quality characteristics. In other words, different operators operate under different conditions (geographical and regulatory), some of which are more favourable to the ‘efficient’ production of public transport services than others. Network coverage, for example, may be dense (good) or sparse (bad); regulatory differences may result in service differences with regard to quality as defined by speed, punctuality,
Productivity and efficiency assessment

frequency and travel links. Some argue that these characteristics should be an integral part of the technology description, that is, included in the efficiency assessment. This, however, assumes that all such factors can be accurately ‘measured’ and is hence highly positivistic in outlook. The counter to this is that there is no such variable or combination of variables that can ever reliably capture such differences between firms, and hence what is required is a far more pragmatic outlook and acceptance that such differences are unmeasurable. As a consequence, there is likely to be an optimal rather than a maximum level of efficiency that can be achieved due to such operating environmental differences, similar in many respects to the main principles that underpin benchmarking. One mechanism through which this whole issue can be partly addressed is to accept that, in most cases, what is of interest is how efficiency changes over time and not the absolute level of efficiency per se. Through such an approach, the impact of the operating environment on performance is considerably reduced, as, generally speaking, such external factors will change very slowly over time, hence allowing efficiency improvements to be isolated. As a result, any efficiency assessment should not be based on a single-year observation but rather undertaken over a period of time.

Practicalities, the choice of inputs

The choice of inputs by which to assess the level of efficiency/productivity in some respects is more straightforward than the choice of output, as these should reflect the main factors used in the production of public transport services, that is, the basic factors of production in the form of land and raw materials, labour and capital. In the empirical context, however, it is very difficult to obtain reliable data for land (a not-uncommon problem in economics); hence, the specified inputs generally reflect the main labour and capital input factors. On a practical note, De Borger et al. (2002), in a review of efficiency studies on public transit, found that in the 37 studies reviewed, there existed a wide range in the use of inputs. Whilst in some respects this reflects data availability, it also highlights that any estimation of the production process will be an approximation; hence, it will not be possible (or even required) to include every single factor used in the production of public transport services. The factors that are included should be the best approximation and, as such, all other factors approximated by their inclusion. This is underlined by the fact that inputs are complementary; hence, using as an example the basic labour and capital inputs, both are required to produce public transport services, and generally, more capital will require more labour (and more fuel, maintenance and administration), although a degree of substitution may exist at the margin.

The Malmquist productivity index

Whilst there exist a number of different indices that can be used to assess productivity, such as partial productivity measures like the Tornqvist and Fisher indices (Coelli et al., 2005), the Malmquist productivity index (Malmquist, 1953) is a key tool in the assessment of both productivity and efficiency over time. It tracks the output/inputs ratio to give total factor productivity, broken down into two further components, namely technical change and efficiency change. Technical change represents ‘real’ improvements in productivity, whilst efficiency change is the extent to which companies improve their productivity performance in relation to the ‘best’ firm. To put this in an economic context, technical change shifts the production frontier (which defines the maximum level of output that can be produced from a given level of inputs), whilst efficiency change moves the individual firms closer to/further from the production frontier. Both concepts are illustrated in Figure 12.2.
Figure 12.2 presents a very simplified example but one that allows a clear focus on the underlying concepts. In this assessment, there exist only two firms, firm A and firm B, and productivity is assessed across two time periods, s and t. The production function is assumed to exhibit constant returns to scale and hence is linear in form, but note that not all production functions need be linear. For both firms, the figure plots the production points for the combination of the inputs and the resultant level of output. As an example, if this were related to the production of train services, this would represent the combination of labour, rolling stock, infrastructure and energy and the resultant level of train services that the combination of these factor inputs provided (assuming a supply based measure of output). In the figure, a comparison of the relative production points for the two firms in the first time period s, as shown by points \( X_s^A \) and \( X_s^B \), demonstrates that for a given level of inputs, firm A produces a higher level of output than firm B. As such, in this simple two-firm example, firm A defines the production frontier as this represents ‘best practice’ and would be deemed ‘efficient’, or 100% efficient, to be exact. Through innovation, using the same physical level of inputs, firm A increases its output in the following time period t to production point \( X_t^A \), which, when compared to point \( X_t^B \), is once again a higher relative level, and hence in this two-firm example, this again represents ‘best practice’. Is this an efficiency improvement? The simple answer is no, as in time period s, firm A was already assessed as being 100% efficient; this cannot be improved. It would, however, again be assessed as being 100% efficient in time period t; hence, the efficiency improvement, or change, is zero, as in both time periods, firm A sits on the production frontier. The increase in output has arisen entirely due to technical change, and as stated previously, this represents ‘real’ gains in productivity. Firm A therefore would have a productivity improvement due to a large technical change and a zero efficiency change.

Contrast this with firm B. As can be seen in the figure, in the first time period s, firm B would be deemed inefficient, as its level of output, for the given level of inputs, is well below the standards being set by firm A. This is illustrated by its relative position to the production frontier in time period s (PF). Between the two time periods, however, it makes a major improvement and shifts its production position to point \( X_t^B \). Note that this is considerably nearer, in relative terms, to the standard being set by firm A. This improvement, however, is made up of a combination of technical and efficiency change. Hence, if firm B had only been able to keep pace with technological developments, it would have shifted its production position only as far as \( X_t^B' \). This, therefore, would be technical change (gain) relating to firm B. As can be seen,
Productivity and efficiency assessment

however, it has significantly improved beyond that point; hence, the net difference, in this very simplified example, would represent efficiency change.

Methods of assessing productivity/efficiency

In terms of practical assessment, in order to estimate a Malmquist productivity index, an estimation method is required that assesses both the efficiency and technical level of the public transport firm at specific points in time. Kerstens (1996) usefully categorises the main approaches used into a dichotomy of ‘parametric technology’ and ‘non-parametric technology’. The main difference between the two is the requirement in the former group of the specification of an underlying econometric model of production. This dichotomy is broadly reflected in the main methods that have appeared in the academic literature (although not the only ones used), with the main ones from each of these groups outlined in this section.

Data envelopment analysis

Data envelopment analysis (DEA) is a frontier-based approach, which builds upon the early work of Farrell (1957). Based on linear programming, the firm’s output-to-input ratio is maximised by applying a weight(s) to the output(s) and a different weight(s) to the input(s), such that when these same weights are applied to all other firms in the sample set, no one firm exceeds 100% efficiency. Charnes et al. (1978) were the originators of the approach, with the generic formal specification of any (basic) DEA problem, commonly known as the CCR model, given as

\[
\begin{align*}
\text{Max} & : \sum_{i=1}^{n} m_i y_{ik} \\
\text{Subject to:} & : \sum_{i=1}^{m} v_i x_{ik} = 1 & (12.2a) \\
& : \sum_{i=1}^{m} u_i y_{ik} - \sum_{i=1}^{m} v_i x_{ik} \leq 0 & (12.2b)
\end{align*}
\]

Where: \( j \) = decision making units, \( j = 1 \ldots k \ldots n \)
\( r \) = outputs, \( r = 1 \ldots t \)
\( i \) = inputs, \( i = 1 \ldots m \)
\( v_i, u_i \) = virtual input and output weights

Banker et al. (1984) proposed a further extension of the basic CCR model in order to incorporate variable returns to scale, which is of the form:

\[
\begin{align*}
\text{Max} & : \sum_{i=1}^{n} m_i y_{ik} - z_k \\
\text{Subject to:} & : \sum_{i=1}^{m} v_i x_{ik} = 1 & (12.3a) \\
& : \sum_{i=1}^{m} u_i y_{ik} - \sum_{i=1}^{m} v_i x_{ik} - z_k \leq 0 & (12.3b)
\end{align*}
\]

This is identical to the formulation given in Equations 12.2a to 12.2c but in very simple terms includes a scalar \( z_k \). This has the effect of adjusting for firms at different levels of production and thereby allows direct comparisons to be made; hence, any beneficial effects from variable returns to scale are neutralised.
Jonathan Cowie

Following Färe et al. (1994), in the context of the Malmquist productivity index for two time periods $s$ and $t$, four DEA ‘efficiency’ measures are required:

- $TE_s$, efficiency in time period $s$,
- $TE_t$, efficiency in time period $t$,
- $TE_{st}$, efficiency of period $s$ against the frontier in period $t$ and
- $TE_{ts}$, efficiency of period $t$ against the frontier in period $s$.

The first two enable the level of efficiency change to be assessed, which is a straightforward comparison of the relative distances of the firm’s production position to the production frontier in the two time periods; hence, efficiency change, $EC$, is given by:

$$EC_a = \frac{TE_s}{TE_t}$$  \[12.4a\]

The second two efficiency measures are an evaluation of how far the firm has shifted its production position in a cross-frontier comparison of the two relevant time periods, combined to provide the measure of technical change, $TC$. Following Coelli et al. (2005), this is calculated as:

$$TC_a = \left[\frac{TE_s \times TE_t}{TE_{st} \times TE_{ts}}\right]^{\frac{1}{2}}$$  \[12.4b\]

This actually represents the geometric mean (or the average) of the rate of technical change in the two adjacent time periods.

Finally, total factor productivity ($TFP$) is the product of the two:

$$TFP_a = EC_a \times TC_a$$  \[12.4c\]

Whilst this section has only presented the basic formulations used in DEA, one further development of the DEA approach that needs to be highlighted is the idea of bootstrapping. One of the key issues with the method is that it is highly deterministic, and thus there is a lack of statistical inference. As such, there is no way of validating if the specified inputs actually do have an effect on the level of output or how well the combination of the inputs ‘explain’ respective levels of output. In an efficiency context, therefore, a firm’s efficiency may be assessed on the basis of inputs that in reality have little or no impact on the level of output. As such, this would represent a major validity issue. ‘Bootstrapping’ the DEA, on the other hand, is an approach which draws samples from within the dataset against which firms are assessed, and hence what is produced is a distribution of efficiency scores for the individual firm. This can be used to introduce statistical inference, where a high level of variation in the efficiency score would clearly identify those issues highlighted. At the time of writing, the bootstrapping approach has not been used in the MPI context due to the computational difficulties that this presents but is a clear area for future development.

**Econometric approaches**

An econometric approach requires the estimation of the underlying production process; hence, a basic function of this relationship is estimated:

$$Q = f(X)$$  \[12.5\]
Productivity and efficiency assessment

Where the level of output is given by $Q$, and $X$ represents a vector of the inputs. This is the basic production function, and over the years, a number of different functional forms have been developed. One of the major differences between econometric approaches and DEA is that the former are grounded in economic theory. As a consequence, different specifications of the production function represent different assumptions regarding the production process, and whilst almost always overlooked, these can be key, as variations in the assumptions can cause considerable variation in how ‘efficient’ or ‘productive’ an individual firm may be estimated to be. Hence, to take a simple case of a two input model (labour and capital) and apply the simplest form of the production function, this would be the Cobb Douglas formulation:

$$Q = AL^aK^b$$

[12.6]

which, when specified in log form for estimation purposes, is given as:

$$\ln Q_a = a + b_L \ln L_a + b_K \ln K_a$$

[12.7]

And estimated as:

$$\ln \tilde{Q}_a = \tilde{a} + \tilde{b}_L \ln L_a + \tilde{b}_K \ln K_a + e_a$$

[12.8]

Hence the estimated value of the output $Q$ is a function of a scalar (estimated parameter $\tilde{a}$) and the two specified inputs Labour ($L$) and Capital ($K$). As the estimate will not be ‘perfect’, there will also be an error term ($e_a$) associated with each observation, which will simply be the estimated value subtracted from the actual. The effect of the inputs on the level of output is given by the $\tilde{b}$ terms in Equation 12.8; hence, the impact of labour is given by the estimated parameter $\tilde{b}_L$. There are, however, two issues with the use of the Cobb Douglas specification in efficiency/productivity assessment. If Equation 12.8 is examined more closely, then the output $Q$ is produced by a combination of the labour and capital inputs. As the equation is specified, however, a given level of output could be produced with a high level of labour and a low level of capital or a low level of labour and a high level of capital. This assumes, therefore, that the extent to which one factor input can be substituted for another will be at a fixed rate across the whole output range, an unrealistic assumption. Common sense (and economic theory) would deem that as one input factor is progressively substituted by the other, the productivity gains by doing so will reduce. This is known as elasticity of substitution, and production theory states that these vary across the production range. A second issue with the Cobb Douglas production function is that it does not account for varying returns to scale. Again, if Equation 12.8 is examined closely, across the full production range, the combined effect of labour and capital on the level of output remains fixed; hence, there are no variable scale effects. In the efficiency/productivity context, both of these limitations could be incorrectly attributed to inefficiency/poor productivity. What is required, therefore, is a function that relaxes these two assumptions, and in most cases, what is used is the translog form. For the two-input case, this is given by:

$$\ln \tilde{Q}_a = \tilde{a} + \tilde{b}_L \ln L_a + \tilde{b}_K \ln K_a + \tilde{c}_{L_L} \frac{1}{2}(\ln L_a^2) + \tilde{c}_{K_K} \frac{1}{2}(\ln K_a^2) + \tilde{c}_{L_K} \ln L_a \ln K_a + e_a$$

[12.9]

Whilst complex, this is an extension of the basic Cobb Douglas equation, with the inclusion of the individual squared terms allowing for the effect of the inputs to vary over the output range (hence allowing for variable returns to scale), and the inclusion of the cross product term
overcomes the problem of unitary elasticity of substitution; again, these vary over the output range. In the productivity context, time variables would be added to the equation; hence, the final form would be

\[
\ln Q_a = \tilde{\alpha} + \tilde{\beta}_L \ln L_a + \tilde{\beta}_K \ln K_a + \frac{1}{2} \tilde{\gamma}_{L} (\ln L^n)^{2} + \frac{1}{2} \tilde{\gamma}_{K} (\ln K^n)^{2} + \\
\tilde{\gamma}_{L} \ln L_a \ln K_a + \tilde{\theta}_t t + \tilde{\theta}_n t^2 + \tilde{\theta}_L t \ln L_a + \tilde{\theta}_K t \ln K_a + e_a
\]

[12.10]

Whilst probably appearing more complex than before, the addition of the time component allows the effect of time to vary over the period (by including \( t^2 \)) and also ‘allows’ the effect of the inputs on the outputs to vary over time (by inclusion of the time multiplied by the input factor variables).

In the context of the Malmquist productivity index, there is a need to derive measures of efficiency change (\( EC \)), technical change (\( TC \)) and total factor productivity change (\( TFP \)).

As with DEA (Equation 12.4a), efficiency change, \( EC \), is given by:

\[
EC_a = \frac{TE_s}{TE_t}
\]

[12.11a]

Again, this is a relative assessment of the respective production positions to the production frontier, although how efficiency in this case is estimated is considered subsequently. Technical change in many respects is far more straightforward than with DEA, as what has been estimated is a ‘model’ of production which includes a time dimension; hence, technical change is simply defined as the rate of change of output with respect to time and calculated by taking the partial differential of [12.10] with respect to time, hence:

\[
TC_a = \frac{\partial Q_a}{\partial t} = \tilde{\alpha}_t + \tilde{\beta}_L t + \tilde{\beta}_K \ln L_a + \tilde{\theta}_K \ln K_a
\]

[12.11b]

Given this is technical change between periods \( t \) and \( s \), rather than solely at a single point in time, a similar evaluation would be made at time period \( s \) and an average taken of the two.

**Production function to production frontier, corrected ordinary least squares**

The previous section outlined the basic components required to estimate efficiency and productivity using an econometric approach, although there still exist alternative measurements of ‘efficiency’, with the two main ones being corrected ordinary least squares (COLS) and stochastic frontier analysis (SFA), the differences between the two based on how the residual term \( e_i \) is used to estimate efficiency and the fact that SFA tends to be very heavy and exacting on data requirements.

Examining the COLS approach first, this is best illustrated by taking the simplest of the formulations given previously, that of the Cobb Douglas (12.8):

\[
\ln Q_a = \tilde{\alpha} + \tilde{\beta}_L \ln L_a + \tilde{\beta}_K \ln K_a + e_a
\]

[12.12]
For firm $i$ in time period $t$, therefore, given the combination of labour and capital inputs it uses, the level of output is estimated to be $\tilde{Q}_{it}$. In essence, this is the benchmark against which its efficiency is evaluated. Hence, the actual level of output is given by $Q_{it}$, with the residual in Equation 12.12 representing the difference between the actual and the estimated. Given that the parameter values $\tilde{b}_L$ and $\tilde{b}_K$ are estimated based on the whole dataset, then what these represent is the ‘average’ value of the influence of labour and capital on the level of output. If the firm is therefore using its inputs better than the ‘average’, then the model will underpredict its level of output, and hence the residual, $e_{it}$, will be positive. The reverse will be true in the case of underperforming companies. The residual therefore becomes the ‘gauge’ against which the firm is assessed in comparison to the whole dataset. It follows, therefore, that the ‘best’ firm must be the one with the largest positive residual value. This is the whole premise of the corrected ordinary least squares method to efficiency assessment.

Figure 12.3 illustrates how the residual term is used to move from a production function, which is a generic model of the relationship between the output and the inputs, to a COLS production frontier.

In Figure 12.3, the $x$ points represent individual firms production position, hence the aggregated combination of inputs and the level of output this produces. As may be expected, firms with a higher level of inputs generally produce a higher level of output, and as a consequence, all the points appear to lie around a straight line (although note one specified in log form). Estimation of this line is the basic production function (Equations 12.8 or 12.9, given the specification of the function), and what this represents is the formal relationship between the output and the inputs. As stated, this is derived from the whole dataset and hence represents the estimated level of output given the level of inputs for the ‘average’ firm across the full range of production. Firm $A$ is identified as ‘best practice’ as this has the largest positive residual (value of $e_{it}$). In other words, this is the firm that performs ‘best’ for a given level of inputs against the ‘average’ firm. By definition, therefore, it must be the most ‘efficient’. As a consequence, the whole function is shifted upwards by adding the largest residual (Max $e_{it}$) to the constant term ($\tilde{a}$ in Figure 12.3)}
Equations 12.8 and 12.9) to derive the production frontier. The frontier therefore represents the level of output that all firms in the dataset ‘should’ be producing if they were as productive as the most efficient. Hence taking Company B in Figure 12.3, it is producing at the level shown by point \( X_B \) but technically ‘should be’ producing at point \( X_B' \); hence, the difference in the two points represents inefficiency. As a formal measure, therefore, this would be calculated as the ratio \( I_BX_B/I_BX_B' \). Given the logarithmic nature of the function, this translates to:

\[
\text{Efficiency Company B} = \exp(e_{B'} - \max e_r)
\]

[12.13]

COLS-derived efficiencies are probably the most basic form of efficiency assessment using the econometric approach.

**Production function to production frontier, stochastic frontier approach**

One of the major criticisms with the COLS approach is that it assumes that the whole of the residual is due to inefficiency of the firm, whereas in reality, there are a number of other factors that could account for this. Stochastic frontier analysis overcomes this limitation by introducing a random element into the efficiency calculation; hence, the residual is split into two components, an (in)efficiency component and a random term, with a distribution (usually half normal, but others can be applied) assumed for the (in)efficiency component. Through an iterative approach, the function then attempts to fit the inefficiency distribution using the random element. This has often been described as ‘luck’, hence the explanation given that in some years the firm is ‘lucky’, whilst in others it is ‘unlucky’, or alternatively, that it experiences ‘random shocks’ (Brons et al., 2005). Both, however, are very poor explanations of the SFA approach. What the SFA approach recognises is that there will always be imperfections in the dataset, no matter how high quality the dataset is. This arises due to a multitude of factors, but in terms of public transport provision, these are usually caused by changes in the financial year or due to changes in operating areas. Other factors relate to the fact that whilst production (from which efficiency is derived) is a flow as such over time, this flow is assessed at specific points in time, and hence inconsistencies in the data can arise over that time period. To give a simple if trite example, a bus company gains a six-month contract to provide bus services between two rail stations whilst major work is undertaken on the rail line. For this, it leases the buses required and hires contract staff for the duration. Where this falls completely within a financial year, then very basic production statistics would show a significant increase in bus miles, but with the same level of staff and buses, and hence a massive improvement in efficiency. For all such similar major or minor occurrences, there is no rational way of adjustment. Hence, the SFA attempts to eradicate such factors through modelling an efficiency distribution (or inefficiency, to be more precise) where the random element is used to best fit the distribution.

**Which method to use?**

In practice, there is no single approach to the assessment of efficiency and no one method that has been proved to be superior to the others; all have been widely published in the academic literature. For example, of 40 studies in the review carried out by De Borger et al. (2002), 11 used SFA, 7 COLS and 15 DEA. In theoretical terms, the best approach to use is undoubtedly SFA, as this is particularly attractive due to the fact that it has both a strong theoretical foundation and does (in theory) account for data inadequacies. The reality, however, is that it has not
proven superior to either COLS or DEA in practical application and, as noted, is particularly heavy and very exacting on data requirements. Any researcher in the subject area therefore should be familiar with all approaches and indeed in their own (private) research use a range of methods to ensure consistency in results.

The application of efficiency assessment in the study of public transport

This section gives an overview of some of the research carried out into the efficiency of public transport. As stated previously, however, there have been a vast number of studies on the subject; hence, what is presented is only a small subset. The aim, however, is to give some insight into the type of research undertaken and the nature of the conclusions drawn from it.

Beginning with a review paper, Jarboui et al. (2012) find in the 24 studies included in the ‘core’ area of road-based public transport efficiency, most are directed to issues of general efficiency and returns to scale of the industry. Thus, as examples, Odeck (2008), in a study of 27 Norwegian bus operators using DEA, found the potential for efficiency gains in the sector to be on the order of between 4 and 11 percent, and overall industry efficiency had significantly increased following a high level of merger activity. Estimation of a Malmquist productivity index also suggested that both EC and TC were significantly higher for the merged companies. Von Hirschhausen and Cullmann (2010) used both deterministic and bootstrapped derived DEA estimates to examine the efficiency of 179 German bus operators over a 15-year period. What the authors found was very low annual industry average efficiency values, ranging from 40.1% to 46.0%. They also found evidence of increasing returns to scale for small and medium-sized operators, suggesting that the efficiency of the industry could be significantly improved through a process of merger and acquisition. Pina and Torres (2001) carried out an assessment of 15 public and private/public transport operators in Spain using a combination of DEA derived efficiencies, multiple regression and cluster analysis. Spain at that time had a combination of public and private provision, with the former by direct award and the latter through tendering. The authors, surprisingly, find higher levels of efficiency in publicly owned operators, which, whilst consistent with the first aspect of Vickers and Yarrow’s (1988) groundbreaking research that public ownership is not necessarily more inefficient, is completely inconsistent with the second that it is the level of competition that tends to drive efficiency improvements. All of the cited studies primarily give insights into the best way that the supply side of public transport industries should be organised and hence are primarily concerned with issues around big vs small operators, private vs public ownership, free market vs direct regulation and so on.

Other efficiency-based studies have given deeper insights into a wide range of issues connected with the provision of public transport. Nolan et al. (2002), for example, evaluated the effectiveness of the passing by the U.S. Congress of the Intermodal Surface Transportation Efficiency Act in 1991, which required operators to develop public transport systems that would be low polluting and safe and provide necessary public services. This was ‘converted’ to the empirical context through an assessment of social efficiency, defined in terms of use of non-diesel energy sources, safety incidents and route miles operated. Through employment of the DEA approach, two major components of inefficiency were identified, inefficient production technology and inefficiencies as a consequence of compliance with legally required social objectives. Twenty transit agencies were studied over six years, with the results showing that whilst a number of firms were ranked as technically efficient, very few placed similar emphasis on social efficiency. Closer examination actually shows decreases in the latter over the period reviewed and a very clear polarisation of firms, suggesting that the provisions of the Act were
Jonathan Cowie

not sufficiently robust to achieve the declared aims. In a similar vein, Wei et al. (2017), recognising the wider social role played by public transport, devised a framework in which both the operational efficiency and access equity could be evaluated. This combined DEA-based efficiency assessment with a geographic information system (GIS) analysis of bus network coverage of the extent to which each transit service provided a unique and necessary service coverage for disadvantaged populations. Operational efficiency scores from the DEA assessment and the GIS data were combined through a multiobjective optimisation model which evaluated the trade-offs between these two potentially competing goals and thus enabled an overall assessment of the performance of transit services to be made. The framework was applied to the 94 individual bus routes provided by the Utah Transit Authority. Hence, the ‘efficiency’ of public transport in this example did not only relate to one based on a measurable output (passengers carried) but also encompassed the wider social role played within society.

Finally, Cowie (2014) used efficiency assessment as the cornerstone of a study into the extent to which consumer sovereignty was present in the deregulated British bus industry. The theorem, originally proposed by Ludwig von Mises (1949), is a key requirement for an efficient free market. In the 97 bus markets analysed (and productive efficiency assessed), only in 28% of those studied was any evidence of consumer sovereignty found. The same author later extended the work and undertook a long-run productivity assessment of the British bus industry over a 21-year period (Cowie, 2018), which, with the addition of figures on profitability, concluded that the reforms introduced by the Transport Act 1985 had belatedly introduced a competitive market in the provision of bus services in Great Britain. Re-regulating the bus industry in Britain, therefore, without also taking measures to improve the underlying economics of bus provision would not address the issues of falling patronage and rising fares and costs.

To attempt to summarise all of this, as public transport tends to be heavily funded and specified by the public sector, a key issue with regard to performance is the extent to which what is provided represents value for money; hence, the evaluation of efficiency/productivity becomes central to any such assessment. What it also shows, however, is the diversity of the range of services provided by public transport and, as highlighted earlier in the chapter, its wider roles in support of social equity and (to a lesser extent in the studies highlighted) economic development. All can be encompassed in an efficiency framework and hence key issues identified with regards to the provision of public transport services in pursuit of these goals.

Closing comments

This chapter has outlined a very powerful toolkit through which the effectiveness of the supply side of public transport industries in the delivery of public transport services can be assessed. This should remain the essence of efficiency assessment; however, one concern is that what it can dissolve into is extreme positivism. It can be very strongly argued that the last two major advances in the approaches used to assess productivity/efficiency came with work on stochastic frontier analysis (Jondrow et al., 1982) and the bootstrapping of the DEA approach. In many senses, with the final development of these methods, the toolkit was completed. As such, efficiency assessment will never be perfect, but through a knowledgeable application, ecological validity will be high; that is, the results being produced will accurately reflect the reality being studied. There has been a tendency, however, in recent years to correct any environmental ‘inaccuracies’ by use of ever-more-sophisticated modelling approaches, and whilst this is a debatable point, there appears to be a clear shift to outright positivism and the blind belief that this can be accurately achieved. The problem, however, is what gets lost is that efficiency assessment is a means to an end, not an end in itself, and hence the danger is that the whole area
Productivity and efficiency assessment

purely becomes an exercise in mathematical modelling rather than an evaluation of policy and the extent to which public transport fulfils its key roles within society.

In terms of future directions, accurate efficiency analysis requires access to high-level data, but with public transport reforms continually moving in the direction of private sector involvement, this has led to various issues surrounding commercial sensitivity. As a consequence, whilst regulatory authorities require public transport operators to submit such data as part of the regulatory framework, this is then not accessible in the public domain due to commercial sensitivity. Hence, whilst the data may be used for regulatory purposes, further examination is prohibited due to data access restrictions. Whether such data is truly commercially sensitive may be open to some debate, and hence serious consideration should be given to making this more widely available.

As a final note, this chapter was written before the global outbreak of the COVID-19 pandemic and the lockdown regulations that ensued. At the time of writing, the COVID-19 pandemic of 2020 has had far-reaching and potentially long-lasting consequences for the provision and usage of public transport. With respect to this chapter, it will have a profound effect on the efficiency and productivity of such services. Nevertheless, it underlines the importance of time series data and the assessment of total factor productivity rather than simply ‘efficiency’ per se. The key point is that the pandemic has resulted in a significant drop in usage but one that has not been matched by a similar decline in provision, hence a dramatic decline in the basic output/inputs ratio. Furthermore, this situation may only recover slowly over time.

In terms of efficiency assessment, there are basically two ways this can be handled. The first is to recognise the importance of technical change and in particular adverse technical change, as what this represents is long- or medium-term structural decline. As such, the framework outlined in this chapter is sufficiently robust to incorporate these effects. In simple terms, a large decline in technical change might be expected but followed by periods of increasing technical change as some form of recovery in usage takes place over time. A second approach is to directly model the impact separately. Hence, for DEA-derived measures, this would be done in a second-stage regression, where the efficiency estimates are regressed on time, pre and post pandemic so as to ‘correct’ for this impact. In a similar manner, this could be directly incorporated in the model with an econometric approach through the use of dummy variables; hence, in Equation 12.8 for the Cobb Douglas function and 12.10 for the translog, both a step dummy (to account for the dramatic fall in patronage) and a slope dummy attached to time could be introduced to model the expected recovery as the consequences of the pandemic begin to ease over time.

Note

1 The remaining seven used a variant of DEA known as free disposal hull (FDH).

References


