TRANSPORT REVIEWS, 2002, VOL. 22, NO. 1, 1–38

Public transit performance: what does one learn from frontier studies?

BRUNO DE BORGER†, KRISTIAAN KERSTENS‡ and ÁLVARO COSTA§

†Department of Economics, University of Antwerp (UFSIA), Prinsstraat 13, B-2000 Antwerp, Belgium

‡LABORES, Université Catholique de Lille, BP 109, 60 Boulevard Vauban, F-59016 Lille Cédex, France

§Faculdade de Engenharia da Universidade do Porto, Rua dos Bragas, 4099 Porto Codex, Portugal

(*Received 6 May 2000; revised 12 September 2000; accepted 29 August 2001*)

This paper provides a comprehensive survey of the literature on production and cost frontiers for public transit operators, and it evaluates the contributions of frontier analysis to the understanding of the performance of the public transport sector. The authors first succinctly contrast best practice (or frontier) and average practice specifications of technology. They also review relevant performance indicators and the methods to measure them. Next, the existing frontier studies measuring urban transit performance are systematically summarized and critically assessed. It is shown that the organization of the market, contract design, the degree and nature of the regulatory regime, and the characteristics of the network being served are all important determinants of inefficiency. However, although the frontier literature has substantially contributed to the knowledge of urban transit technologies and the determinants of performance, it is found that many important issues remain unresolved.

1. Introduction

Despite a declining trend in transit demand in most industrial economies, urban transit remains an important transport mode. Companies under different ownership regimes provide urban passenger services in a highly regulated environment, making use of a diversity of vehicles (bus, tramway, metro, light rail, etc.). Government intervention in the sector is widespread and has traditionally been justified by reference to a series of market failures. In the past two decades, however, concerns about possible regulatory failures led to a reassessment of the role of the state in the organization of the sector (for reviews, see Glaister 1990, Berechman 1993). The relative merits of public and private provision and the implications of different pricing policies and regulatory regimes for the performance of urban transit firms became important issues. In this respect, a crucial question is which type of operating and regulatory environment is best suited to stimulate productivity growth and

*Author for correspondence; e-mail: bruno.deborger@ufsia.ac.be

Transport Reviews ISSN 0144-1647 print/ISSN 1464-5327 online © 2002 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/01441640010020313

efficiency in the industry. Not surprisingly, efficiency has played a prominent role in political and academic arguments guiding recent privatization and deregulation policies (e.g. Mackie *et al.* 1995).

It is only quite recently that appropriate methodologies have been developed that allow a careful study of the efficiency of urban transit firms. Indeed, the early analyses of the supply of urban transit services relied on estimating average practice technologies by simply estimating functions through the middle of the data. This traditional analysis is of little help when looking for excellence in production and ignores the inherent frontier nature of cost and production functions. Conceptually based on the seminal contribution by Farrell (1957), from the early 1980s on various frontier estimation techniques have been developed to determine 'best practice' behaviour in an industry (e.g. Lovell 1993). The frontier methodologies allow one to distinguish between efficient and inefficient production, and to estimate the degree of inefficiency by considering observed best practice standards in the industry as a benchmark. Moreover, the estimated frontiers allow separation of productivity changes over time from changes in efficiency. Finally, it has been recognized that frontier estimates of technologies may well imply different production characteristics (e.g. scale and scope economies) than average practice functions.

Not surprisingly, frontier methods have found their way to the transport sector, and studies on the productivity and efficiency of almost all transport modes are now available in the literature. For instance, Viton (1997) considered technical efficiency of US multimode bus transit, Good *et al.* (1993) compared the performance of European and US airlines, Norwegian ferries were analysed in Førsund (1992), while De Borger (1993) studied the behaviour of the Belgian railroads, etc. The primary purpose of this paper is to provide a comprehensive survey of the literature on production and cost frontiers for the public transit sector, and critically to assess the contributions of this literature to one's understanding of performance of this sector. What are the main determinants of efficiency differences between transit firms? What determines productivity growth in the sector? What is the role of ownership in determining efficiency? To what extent do environmental factors and network characteristics affect performance? What are the effects of different regulatory regimes on performance? This survey provides information on these questions and on a number of related issues.

The survey presented below hopes to fill a small gap in the available literature. Indeed, whereas an overview of frontier studies on railroads has recently appeared (Oum *et al.* 1999), a comprehensive survey of frontier methodologies and empirical results for public transit is not yet available. Although some results are reviewed in Berechman (1993) and De Borger and Kerstens (2000), neither study serves this purpose. The former obviously does not contain information on recent frontier estimates; the latter is more selective and condensed. In addition, neither study specifically focuses on frontier analyses.

The material is organized as follows. Section 2 introduces efficiency and productivity measurement using cost and production frontiers. Section 3 discusses the appropriate specification of inputs and outputs in the available studies, and some general characteristics of the typical urban transit technology specifications are summarized. Section 4 systematically and critically assesses the contributions of the frontier studies in evaluating urban transit performance. Finally, Section 5 concludes.

2. Efficiency and productivity measurement: concepts and methodologies

The purpose of performance measurement is to compare behaviour of organizations over time, across space, or both. Furthermore, benchmarking comparisons can be made within a sector or across sectors, comparisons can be limited to the national level or may have an international character, etc. Of course, a crucial preliminary question is to specify the goals of the organizations being evaluated. In principle, public sector activities (such as public transit) may serve a series of objectives, making the evaluation of their performance a difficult exercise. Indeed, from a welfare economic viewpoint, the public sector serves four main goals: efficiency, equity, financial balance and macroeconomic stabilization (Marchand *et al.* 1984, Rees 1984). However, despite the existence of multiple objectives, the focus in many empirical studies in the transport industry is on issues of productivity and (mainly technical) efficiency. There are at least two reasons for this phenomenon. First, a transparent framework for productivity and efficiency measurement has been developed, unlike for the other objectives. Second, it has been forcefully argued that, independent of the other objectives, a first and indispensable demand for all public sector activities is to operate technically efficient (Marchand *et al.* 1984, Pestieau and Tulkens 1993). In the remainder of this section, we, therefore, review the most relevant efficiency and productivity notions and explain how to make them operational.

2.1. *Efficiency and productivity*

Farrell (1957) introduced the idea of best-practice frontiers and provided the first measurement scheme for efficiency. Recent contributions have defined more elaborate taxonomies of efficiency concepts (Färe *et al.* 1994). In a static context, one distinguishes between technical, scale, structural and allocative efficiency.

First, technical efficiency (TE) is defined as production on the boundary of the production possibility set. This set summarizes all technological possibilities of transforming inputs into outputs open to the organization. A producer is technically inefficient if production occurs in the interior of this production possibility set. Second, scale efficiency (SCE) relates to a possible divergence between actual and ideal production size. The ideal configuration coincides with the long run competitive equilibrium, where production is characterized by constant returns to scale. A producer is scale efficient if its choice of inputs and outputs is situated on a constant returns to scale frontier; it is scale inefficient otherwise. Third, structural efficiency (STE) is closely related to the definition of technical efficiency. A technically efficient producer is structurally efficient if production occurs in the uncongested or 'economic' region of production. Structural inefficiency occurs when production experiences congestion. In that case, some of the inputs have negative marginal products. Otherwise stated, the producer could benefit from actually reducing the congesting input factors. Common examples of congestion include agriculture (too much rain spoils crops) or today's city traffic. Finally, allocative (or price) efficiency (AE) requires the specification of a behavioural goal and is defined by a point on the boundary of the production possibility set that satisfies this objective, given certain constraints on prices and quantities. Most often organizations are thought to be minimizing costs. In that case, a technically efficient producer is allocatively inefficient if there is a divergence between observed and optimal costs.

These static efficiency concepts, except structural efficiency, are illustrated in figure 1. (Graphically, structural inefficiency would imply that the input sets bend backward at the extremes.) Figure 1 presents two input sets, describing all combinations of inputs able to produce a given output level, and their boundaries. The input set $L(y)^{CRS}$ is characterized by constant returns to scale, embodying the ideal of a long run competitive equilibrium, while the input set $L(y)^{VRS}$ allows instead for variable returns to scale.

Efficiency is traditionally measured equiproportionally (Farrell 1957). The radial efficiency measure in the inputs varies between 0 and 1, with efficient production on the boundary represented by unity. It has a cost interpretation and indicates the proportional reductions in inputs that leave the output level unaffected. For instance, an efficiency measure of 0.80 means the organization could produce the same outputs with only 80% of its current inputs, making a 20% cost reduction feasible. (Similarly, output-oriented efficiency measures have a revenue interpretation.)

This radial measurement of static efficiency concepts is illustrated on figure 1 for observation a in the interior of input set $L(y)^{VRS}$. First, this observation implies technical inefficiency, since it uses more of both inputs to produce exactly the same output vector as, for example, observation a′ on the boundary of the input set $L(y)^{VRS}$. The degree of technical efficiency (TE) is represented by the ratio of distances Oa′/Oa and is measured relative to the variable returns to scale technology $L(y)^{VRS}$. If both inputs are reduced according to the scalar Oa′/Oa, then the resulting input vector a′ on the boundary of input set $L(y)^{VRS}$ is technically efficient. Second, observation a′ is scale inefficient (SCE) because it needs more inputs to deliver the same output level as e.g. observation a″ on the boundary of the constant returns to scale technology $L(y)^{CRS}$. SCE is defined by the ratio Oa″/Oa′, i.e. by comparing short ($L(y)^{VRS}$) and long ($L(y)^{CRS}$) run technologies. This ratio indicates the lowest

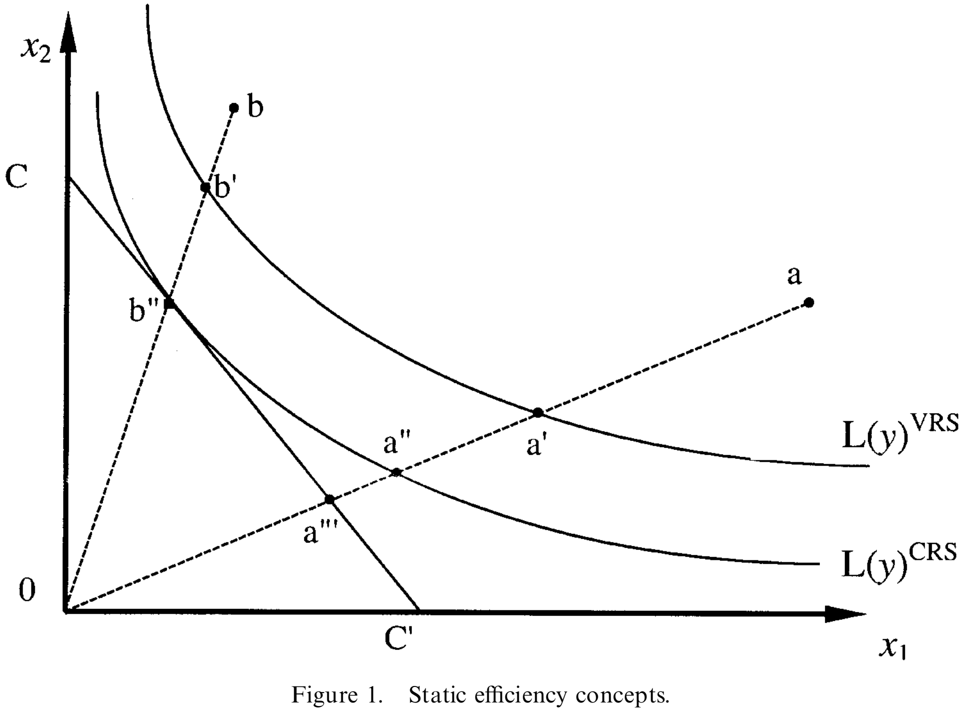

Figure 1. Static efficiency concepts.

possible input vector a″, situated on the boundary of the long run constant returns to scale technology $(L(y)^{CRS})$, able to produce the same output as the technically efficient combination a′. Finally, observation a″ is still not allocatively efficient. If the producer aims to minimize costs, then the given input prices define an isocost line CC′. For these prices, total costs are minimized at the intersection of the input set $L(y)^{CRS}$ and this line CC′ (point b′). Observation a″ is, therefore, not allocatively efficient, since it requires a higher budget to produce the same outputs. Allocative efficiency (AE), captured by the ratio Oa‴/Oa″, indicates the cost reduction resulting from reallocating inputs from a″ to a‴. Of course, the inputs of a‴ cannot yield output y on the boundary of $L(y)^{CRS}$ but, at the same cost, input vector b″ is available that does produce this output level.

In a dynamic perspective it is important to allow for the possibility of technological innovation. Technical or productivity change is traditionally conceived as a move of the production possibilities set over time due to product and process innovations. When the frontier has shifted over time due to productivity change, then it is essential to know whether production units have been able to catch up with these developments. Dynamically, measuring productivity changes and technical efficiency are, therefore, related: the former measures the shift in the production possibilities, whereas the latter indicates the extent to which organizations maintain their position relative to an eventual shifting frontier.

This basic idea is illustrated on figure 2: two production functions are shown at two different points in time. The shift of the production function is caused by technological innovation that allows producing the same outputs with less inputs (or more outputs with the same inputs). Organizations may or may not adjust to these new production possibilities. For instance, firm b is positioned on the frontier both

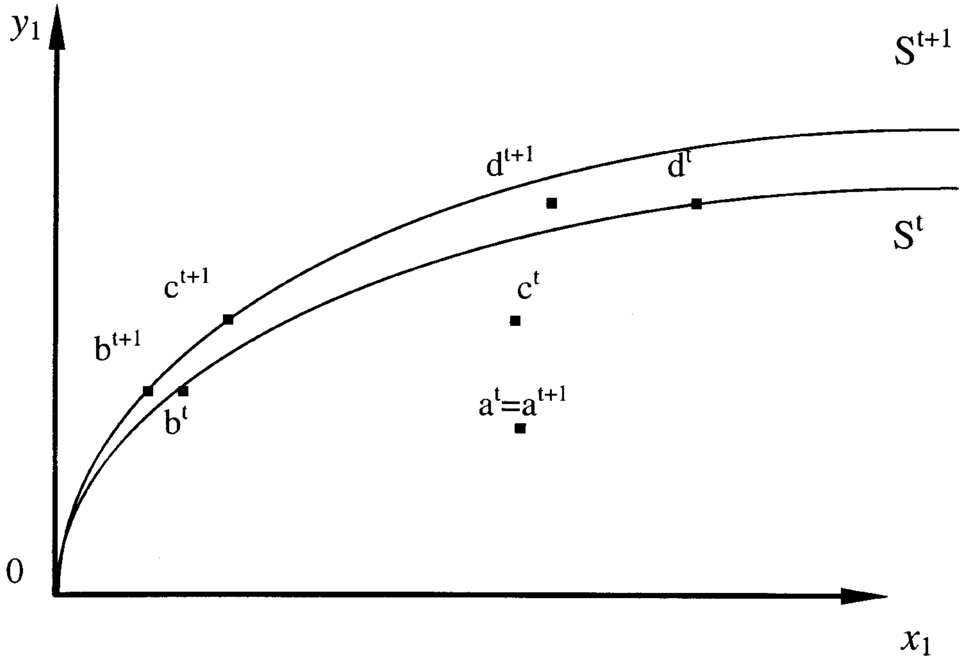

Figure 2. Technical change and technical efficiency.

in periods t (b^t) and $t+1$ (b^{t+1}), implying that it has immediately managed to incorporate technical progress. Firm a on the contrary took no advantage of the new production methods. Being technically inefficient, its performance even deteriorates relative to the production function at $t+1$. Firms c and d illustrate the cases of respectively technically inefficient and efficient firms that have been able to (c) or that have failed (d) to keep up with the new frontier.

For completeness sake, note that the notion of economies of scope, a traditional characteristic of multiple output technologies, can also be evaluated directly using an estimated frontier. Economies of scope or diversification require that the costs of simultaneously producing several outputs are lower than producing each of these outputs separately using a specialized technology.

2.2. *Frontier methodologies and efficiency measurement*

To estimate production or cost frontiers, methods have been developed for analysing time series, cross-section or panel data. Our discussion focuses on cross-section data and on production frontiers. Minor adjustments are needed for cost frontiers and for other types of data. Once frontiers have been estimated, productivity changes can directly be derived from shifts in the frontier over time. Technical inefficiency estimates are readily available as well, as is illustrated below.

Existing approaches to reconstruct production frontiers can be usefully distinguished along the lines below. (A general survey is found in Lovell (1993). Färe *et al.* (1994) overviewed non-parametric methods, while Greene (1997) surveyed parametric frontiers.)

- Parametric versus non-parametric frontier specifications:
  - The parametric approach assumes that the boundary of the production possibility set can be represented by a particular functional form with constant parameters.
  - The non-parametric approach imposes minimal regularity axioms on the production possibility set and directly constructs a piecewise technology on the sample.

- Deterministic versus stochastic frontier specifications:
  - Stochastic methods make explicit assumptions with respect to the stochastic nature of the data by allowing for measurement error.
  - Deterministic methods take all observations as given and implicitly assume that these observations are exactly measured.

Combining these distinctions yields a four-way classification, as illustrated in table 1. Since the literature on stochastic non-parametric frontiers is still burgeoning (recent proposals include resampling (bootstrap), chance constrained programming, etc.) and no consensus has yet emerged, this issue was not pursued here (Grosskopf 1996). The present authors, therefore, focus on a representative selection of methods in the three other cells of table 1.

First, the early literature often used deterministic parametric frontier methods. However, given that they combine the most restrictive assumptions (deterministic and parametric) they are no longer very popular (Lovell 1993). The present authors, therefore, only briefly mention the ‘corrected’ OLS procedure commonly used. This method assumes a particular functional form for the boundary of the production possibility set, i.e. $y = f(x_i;\beta)$, and estimates an average practice frontier using OLS. The estimates resulting from OLS are then transformed into frontier estimates by shifting the intercept of the average function so that one of the transformed residuals is zero, while all others become negative. The results are illustrated in figure 3 for the case of a linear frontier. The average OLS and corrected OLS (COLS) frontiers are respectively given by the lowest and highest curve. Note that relative to the latter, one observation has zero residual (point A), while all other residuals are negative. In the single output case, the radial measure of technical efficiency in the outputs is equal to the ratio of observed to maximal output (TE = y/y^*), where y^* is the fitted frontier output. For example, in figure 3, TE for observation $a = 0b/0d$.

Second, stochastic parametric frontiers allow for composed errors that include both measurement error and technical inefficiency. Observed output can fall short of maximal output by a positive amount u due to technical inefficiency, but in addition

Table 1. Taxonomy of frontier methodologies.

	Measurement error	
Functional form	Deterministic	Stochastic
Parametric	corrected OLS, etc.	frontiers with explicit assumptions (exponential, half-normal, etc.) for the TE distributions
Non-parametric	FDH, DEA-type models, etc.	resampling; chance constrained programming, etc.

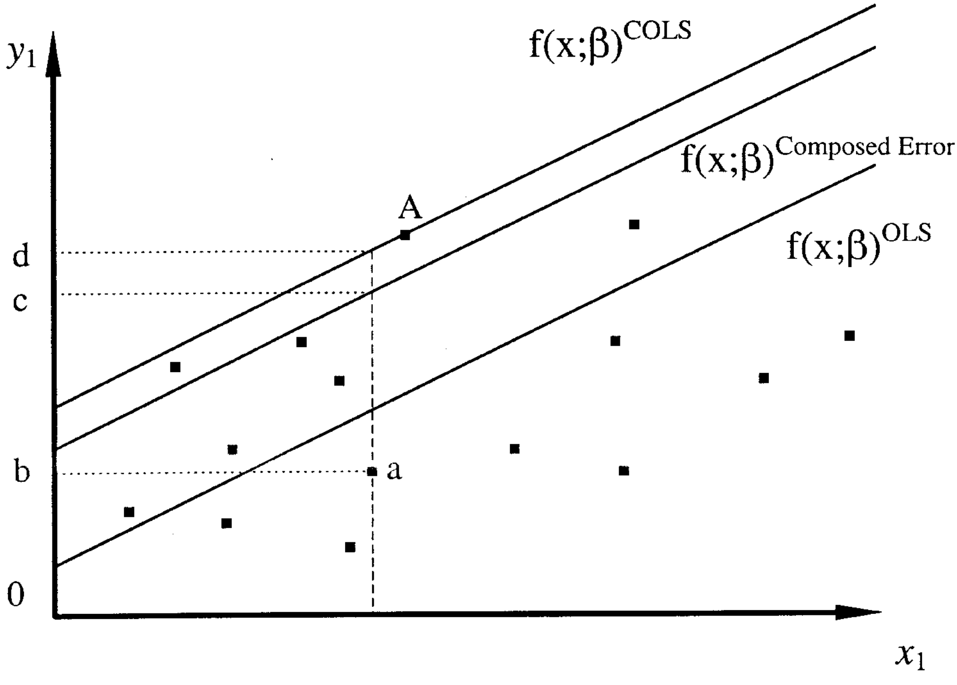

Figure 3. Stochastic and deterministic parametric frontiers.

there is a random error term v with mean zero. In other words, $y \leqslant f(x;\beta) + v$ or $y = f(x;\beta) + v - u$, where $u \geqslant 0$. Both OLS or maximum likelihood (ML) estimators are available by making assumptions on the asymmetric distribution of u. Traditionally the assumption has been that the highest probability is associated with small amounts of inefficiency. Specific distributions for u explored in the literature are, among others, exponential, half-normal, truncated normal and gamma.

In figure 3 the middle curve represents a typical stochastic parametric frontier. It is situated below the corrected OLS frontier, because apart from inefficiency, it also accounts for random measurement errors, so that the degree of technical inefficiency is lower on average. Moreover, technical inefficiency can no longer be directly measured as distance to the frontier, because the latter is partly due to measurement errors. However, given specific assumptions on the distribution of the one-sided technical efficiency component, methods are available to determine technical efficiency for each observation (Lovell 1993). Finally, typically there is no observation being part of the frontier itself.

Third, deterministic non-parametric methods assume no particular functional form for the boundary and ignore measurement error. Instead, the best practice technology is the boundary of a reconstructed production possibility set based upon directly enveloping the observations. These extremal methods use mathematical programming techniques to envelop the data as tightly as possible, in a piecewise linear way, subject to certain maintained production assumptions. These assumptions are generally less restrictive than those used in parametric approaches. The most important technologies of this kind are briefly presented.

A production technology only assuming strong input and output disposability is known as the free disposal hull (FDH). Strong input disposability means that any given level of outputs remains feasible if any of the inputs is increased. Strong output disposability means that with given inputs it is always possible to reduce outputs. Figure 4 shows a section of a FDH production possibility set in a single input and output dimension. The FDH frontier derives its typical staircase form from these maintained assumptions. Observation C, for instance, should produce less output while keeping the same input, simply by wasting part of current production (hence the line from C down south). For this same observation, it is feasible to produce the same level of outputs with more inputs by simply wasting any additional input available (hence the line originating in C and going eastwards). Note that observations c–g are all technically inefficient, since they are situated below the frontier.

An alternative production technology adds convexity to the assumptions maintained by FDH. Convex non-parametric frontiers, known as Data Envelopment Analysis (DEA) models, allow for linear combinations of observed production units. In terms of figure 4, for instance, convexity implies that all linear combinations of observations C and E are feasible. The immediate ramification is that observation D is no longer part of the boundary, since there are linear combinations of C and E that need less inputs to produce the same output level. Figure 4 illustrates two types of DEA-technologies. The first allows for variable returns to scale and is graphically represented by the piecewise linear convex frontier. The second assumes constant returns to scale so that all observed production combinations can be scaled up or down proportionally. The constant returns to scale DEA frontier is simply given by the ray through the origin passing through point E (figure 4).

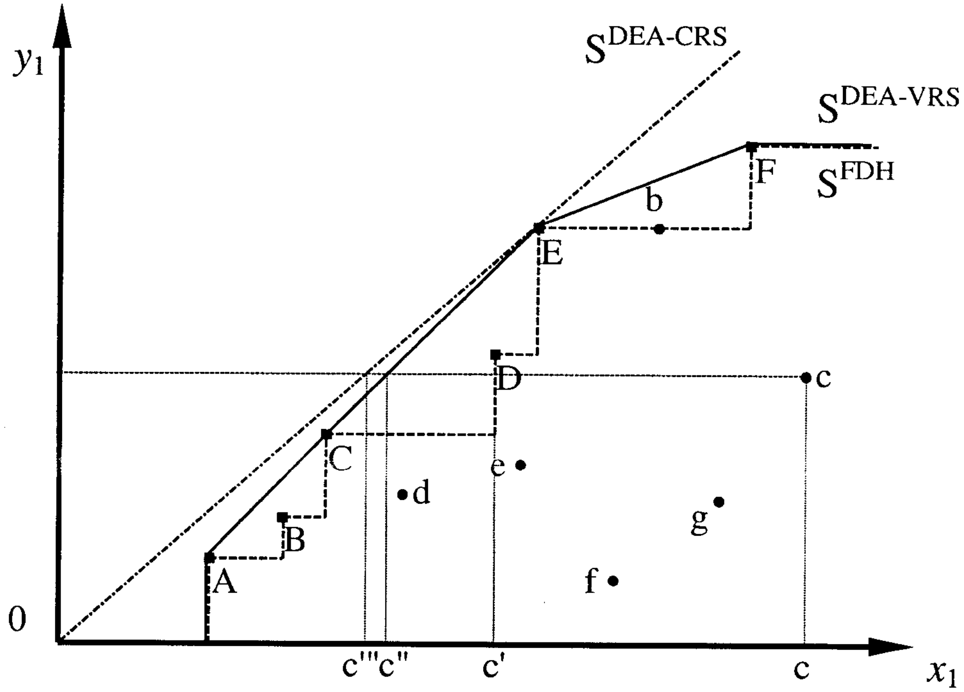

Figure 4. Non-parametric, deterministic frontiers.

The three frontiers discussed are based upon slightly different hypotheses concerning the nature of production possibilities. It is trivial to illustrate that the amount of technical inefficiency depends on the choice of hypotheses. Take observation c as an example, and let us measure efficiency in the inputs. Relative to the FDH frontier this inefficiency is measured by the ratio Oc′/Oc. For the convex, variable returns to scale technology inefficiency is given by the distance Oc″/Oc, while for the constant returns to scale case this ratio equals Oc‴/Oc.

Finally, note that there also exist combinations of methods. For example, a semi-parametric procedure first filters inefficient observations using non-parametric frontiers and then fits a parametric frontier (e.g. Thiry and Tulkens 1992).

3. Urban transit technology specifications

A wide variety of different specifications of the urban transit technology has been used in the frontier literature. An overview of parametric and non-parametric frontier studies is given in tables 2 and 3. The tables contain information about the authors, the type of data used (i.e. cross-section, time-series or panel data), sample size (number of operators and years analysed), transport modes being analysed, and the countries concerned. If nothing to the contrary is mentioned, then the data have a yearly periodicity. Moreover, summary information is provided on the precise parametric or non-parametric methodology being used, and on the definition of inputs (input prices in case of a cost approach), outputs and (mainly for the parametric studies) technology characteristics included in the analysis.

The tables clearly show that frontier studies are a relatively recent business: the majority of studies have been published during the 1990s. They also illustrate that

Table 2. Performance of urban bus companies: parametric frontier methodologies.

Reference	Database	Sample size	Country, city	Function	Inputs, input prices[1]	Outputs[3]	Characteristics
Bhattacharyya *et al.* (1995)	panel data (yearly) 1983–87	32 bus	India	stochastic translog cost frontier	fuel, prod. labour[2], admin. labour, no. of vehicles on road per day	Pk	% fleet utilization, load factor, vehicle utilization, no. of breakdowns, three types of public ownership analysed
Costa and Markellos (1997)	time series	1970–94	London metro, UK	deterministic CD production frontier	vehicles, staff	Vk	
De Jong and Cheung (1999)	unbalanced panel 1994–95	19 urban and regional	The Netherlands	stochastic CD production frontier	staff, total seats, energy costs	Pk	
Delhause *et al.* (1992)	panel data (yearly) 1978–87	13 all, five urban SNCV (eight geographical units)	Belgium	stochastic cost frontier	staff[2], seats, energy	Sk passengers	
Fazioli, *et al.* (1993)	panel data (yearly) 1986–90	40 bus regional	Emilia-Romagna, Italy	deterministic translog cost frontier	labour, capital	Sk	network length Ownership dummy
Filipini *et al.* (1992)	panal data (yearly) 1986–89	62 bus regional	Switzerland	deterministic translog cost frontier	labour, energy, capital	Sk	no. of stops, time trend
Gathon (1989)	cross-section (yearly) 1984	60 all urban	Europe	deterministic translog production frontier	staff, seats	Sk	
Hanusch and Cantner (1991)	panel data (yearly) 1970–83	13 all urban	Germany	stochastic translog cost frontier	labour, capital	Sk	
Kumbhakar and Bhattacharyya (1996)	unbalanced panel (yearly) 1983–87	28/28/24/31	India	stochastic translog cost frontier	fuel, prod. labour, admin. labour, no. of vehicles on road/day	Pk	% fleet utilization, load factor, four types of public ownership analysed

(continued)

Table 2. Continued.

Reference	Database	Sample size	Country, city	Function	Inputs, input prices	Outputs	Characteristics
Lijesen (1998)	unbalanced panel (yearly) 1992 – 94	14/16/9	The Netherlands	deterministic translog cost frontier	labour	Pk, revenues, other activities	% urban public transport
Levaggi (1994)	cross-section (yearly) 1989	55 bus urban	Italy	stochastic translog cost frontier	labour, fuel, other materials	Pk	network length, average speed, no. of vehicles, load coefficient, population density
Loizides and Giahalis (1995)	Time series (yearly) 1971 – 89	one bus regional	Greece	stochastic Cobb Douglas production and cost frontier	staff, capital and other services	Pk	
Matas and Raymond (1998)	panel data (yearly) 1983 – 95	nine bus	Spain	stochastic translog frontier	labour	Vk, passenger trips	network length
Sakano and Obeng (1995)	cross-section (yearly) 1988	84 bus	USA	stochastic translog production frontier	labour, fuel, vehicles	Vk	average bus age, network size
Sokano *et al.* (1997)	unbalanced panel (yearly) 1983 – 92	439 bus in total	USA	stochastic translog production frontier	labour, fuel, vehicles	Vk	population density, network size
Thiry and Tulkens (1992)	time series (monthly) 1977 – 85/79 – 85	three all	Belgium	translog production frontier	labour, energy, seats	Sk	
Viton (1986)	cross-section (yearly) 1979/80	67 bus	USA	stochastic translog production frontier	drivers labour, other labour, fuel vehicles	Vk	average bus age, peak – base ratio

(continued)

Table 2. Continued.

Reference	Database	Sample size	Country, city	Function	Inputs, input prices	Outputs	Characteristics
Viton (1992, 1993)	Cross-section (yearly) 1984–86	289	USA	flexible, quadratic cost frontier	labour	vehicle miles/ mode	average speed/mode, peak and base ratio, modal dummy

[1] Contains inputs in case of a production approach and input prices when a cost function is estimated. [2] *Staff* is number of workers; *labour* is hours of work. [3] Pk, passenger-km; Sk, seat-km; Vk, vehicle-km.

Table 3. Performance of urban bus companies: non parametric frontier methodologies.

Reference	Data type	Sample size	Country	Function	Inputs	Outputs[2]
Button and Costa (1999)	Panel	nine bus 1985 – 93	Italy	DEA	vehicles staff	Vk (Sk)
Chang and Kao (1992)	time series	one bus 1956 – 88	Taiwan	DEA	vehicles staff fuel	Vk Vk revenue bus trips revenue
	panel	five bus 1970 – 88	Taiwan	DEA	vehicles staff fuel	Vk Vk revenue bus trips revenue
Chu *et al.* (1992)	cross-section	86 bus 1986	USA	DEA	vehicle operating expenses maintenance expenses general/administion expenses other expenses	revenue vehicle hours
					revenue vehicle hours population density % of households without car subsidy passenger	unlinked passengers trips
Costa and Markellos (1997)	time series	1970 – 94 London metro	UK	DEA	vehicles staff	Vk
Costa (1998)	time series	one underground Madrid 1981 – 92	Spain	DEA	staff vehicles energy	Vk
					network route length	passengers

(continued)

Table 3. Continued.

Reference	Data type	Sample size	Country	Function	Inputs	Outputs
Cowie and Asenova (1999)	cross-section	141 all urban 1995 – 96	UK	DEA	staff vehicles ● 35 seats vehicles < 35 seats	operating revenue
Gathon (1989)	cross section	60 all urban 1984	Europe	FDH	staff seats	Sk
Kerstens (1996, 1999)	cross-section	114 bus 1990	France	DEA a,b FDH a	vehicles staff energy	VK (Sk) passengers
Levaggi (1994)	cross-section	55 bus 1989	Italy	DEA	load staff costs fuel costs other variable costs kms of route population density vehicles same without load	Pk load
Loizides and Giahalis (1995)	time series (yearly) 1971 – 89	one bus regional	Greece	DEA	staff capital and other services	Pk
Nolan (1996)	panel	25 bus 1989 – 93	USA	DEA	staff fuel vehicles	Vk

(continued)

Table 3. Continued.

Reference	Data type	Sample size	Country	Function	Inputs	Outputs
Nolan *et al*. (2000)	panel	20 bus 1990 – 95	USA	DEA	staff fuel vehicles	vehicle revenue miles vehicle miles
					inout efficiency score (previous model)	non-diesel fuel staff safety incidents route miles
Nollet *et al*. (1988)	time series (monthly)	one all 1977 – 98	Belgium	FDH	labour energy seats	Sk
Obeng (1994)	cross-section	73 bus 1988	USA	DEA	labour fuel vehicles operational subsidies capital subsidies price of capital	vehicle miles
Tone and Sawada (1990)	cross-section	207 bus 1985	Japan	DEA	vehicles staff	Vk
					operating expenses	vehicles staff
					vehicles staff	operating income
					vehicles staff	density of service

(continued)

Table 3. Continued.

Reference	Data type	Sample size	Country	Function	Inputs	Outputs
Tulkens *et al.* (1988)	time series (monthly)	three bus 1979 – 1985	Belgium	FDH	labour energy seats	Sk
Tulkens (1993)	time series (monthly)	one all 1977 – 89	Belgium	FDH	seats labour energy	Sk
Tulkens and Vanden Eeckaut (1995)	time series (monthly)	one all 1977 – 91	Belgium	FDH (benchmark)	seats labour energy	Sk
Tulkens and Wunsch (1994)	time series (monthly)	one all 1977 – 92	Belgium	FDH	seats labour energy	Sk
Viton (1997)	cross section	217 bus motor bus (MB) or demand-responsive (DR) 1990	USA	DEA	labour (four types per MB/DR: transportation, maintenance, administrative, other) fuel (MB/DR) vehicles (MB/DR) tires + materials cost (MB/DR) services cost utilities cost insurance cost average speed (MB/DR) average fleet age directional miles	vehicle miles passenger trips

(continued)

Table 3. Continued.

Reference	Data type	Sample size	Country	Function	Inputs	Outputs
Viton (1998)	panel	183 bus in 1988, 169 bus in 1992, motor-bus (MB) or demand-responsive (DR)	USA	DEA (Malmquist)	labour (four types per MB/DR: transportation, maintenance, administrative, other) fuel vehicles average fleet age directional miles	vehicle miles (MB/DR) passenger trips (MB/DR) vehicle hours (MB/DR)
Wunsch (1994)[1]	cross-section	53 all 1988–93	Europe	DEA FDH	costs	Vk Sk

[1] The published version (Wunsch 1996) does not contain the frontier efficience results. [2] Pk, passenger-km; Sk, seat-km; Vk, vehicle-km.

the majority of frontier analyses deal with European or US data, but that there are many exceptions. For example, Bhattacharyya *et al.* (1995) provided an Indian application, Chang and Kao (1992) used data from Taipe (Taiwan) and Tone and Sawada (1990) studied Japanese data. Unfortunately, very few studies have taken a comparative international perspective (Gathon (1989) and Wunsch (1994) are relevant exceptions.) In terms of methodology, tables 2 and 3 suggest that most parametric studies employ some kind of flexible functional form, often the translog. In the non-parametric cases both DEA and FDH are widely used.

Probably the most striking feature of tables 2 and 3 is the wild variability in the use of inputs and outputs in urban transit technology specifications. Most studies use labour and capital as inputs (Viton 1992 has an application using labour as the only input), but not all of them include energy (e.g. Hanusch and Cantner 1991, Fazioli *et al.* 1993). Some applications include environmental variables to provide more detail on input quality. For instance, Levaggi (1994) included a load factor, population density and network length as inputs, Chu *et al.* (1992) considered revenue, population density and percentage of households without a car, Tone and Sawada (1990) included operating expenses, and Costa (1998) included the network route length of the metro operator. A similar wide variety of indicators is observed at the output side. Parametric studies mainly use supply-oriented indicators such as seat-km or vehicle-km, with the exceptions of Bhattacharyya *et al.* (1995) and Levaggi (1994) who both selected passenger-km as appropriate measure of output. In non-parametric work, there is a broader choice of outputs, although the vehicle-km or seat-km specifications are still the most common. Levaggi (1994) also considered passenger-km, Tone and Sawada (1990) had four applications with outputs of different nature. While most studies include a single output, Chang and Kao (1992), Costa (1998), Tone and Sawada (1990) and Wunsch (1994) included applications with multiple outputs.

This variability in the input and output measures simply suggests that there is no generally accepted set of relevant variables in the bus industry. Traditionally, the input set in the transport sector consists of capital, labour and energy. Of course, differences between operators may exist in terms of quality and composition of inputs, since the latter may be highly heterogeneous. For example, not only should one in principle differentiate between driving and non-driving labour, but in addition the definition of 'effective' labour time may be quite difficult for drivers due to interrupted shifts, etc. Rolling stock, a major part of capital, typically consists of various vintages used at different intensities, implying different patterns of depreciation. Moreover, some variability exists in terms of fuels used. Conceptually, however, if the data incorporate the required information, it is quite feasible to correct for input quality differences.

The definition of outputs is much more problematic. The early literature focused on the distinction between pure supply indicators (vehicle-km or seat-km) and output measures that at least to some extent reflect the demand for transit services (e.g. passenger-km and number of passengers). Arguments for and against either specification are found in the literature.

Many authors have argued that demand-related output specifications are very relevant when evaluating the firm's effectiveness (Chu *et al.* 1992, Costa 1998, Tone and Sawada 1990), but that one should be careful when the focus is on costs and productivity. The main arguments have been nicely summarized by, for example, Berechman (1993). First, inputs do not vary systematically with demand-related output measures so that they do not allow an adequate description of transit technology. Second, supply-related output indicators are to a larger extent under the control of operators than demand-related outputs. In unregulated environments they actually would be direct decision variables for the transit companies. Even if this is not the case due to government regulation of service levels, operators do have some control via the negotiation process with the authorities. Third, independent of the achievement of broader goals defined in terms of passenger transit services actually consumed, supplying bus services in the least costly way may be considered a reasonable requirement for operators. Therefore, the argument continues, when measuring productivity and efficiency the focus should be on pure supply indices. This is especially the case since it is unlikely that all parties involved can agree upon possible broader objectives. For instance, regulators may be interested in the efficient use of their funds, while operators may be inclined to stress effectiveness in terms of the number of passengers, service availability, etc.

A simple but powerful counter-argument, suggesting that demand factors should play a relevant role in output definitions, is that any realistic output measure should take into account the objectives of the firms under consideration. Since passengers or passenger-km at least partially capture the economic motive for providing the services, such demand-oriented output measures must indeed be relevant. After all, if one ignores demand altogether, then the most cost efficient and productive bus operators may be the ones not servicing any passengers. Furthermore, it is now better understood that there is a strong interdependency between the characteristics of demand, the spatial and quality attributes of supply and the appropriate specification of technology for the purpose of performance evaluation. When these issues are accounted for, the early debate between demand or supply oriented output measures loses much of its significance.

To substantiate this claim, first note that there is no overall consensus on the goals of transit firms. Normative models (e.g. Rees 1984) have put forward traditional public enterprise objectives resulting from welfare maximization (efficiency, equity, deficit finance, and macro-economic objectives). In positive models, by contrast, actual objectives result from the interaction between operator preferences, the political and regulatory environment, and pressure groups (Berechman 1993: 95–8). In any case, the proper objective function of transit firms is intimately related to its surrounding social, political and regulatory environment. For instance, when the regulator implicitly stimulates over-hiring labour, then cost minimization at observed input prices is an inappropriate benchmarking model yielding highly misleading results. Second, there is now a general recognition of the heterogeneity of transport output in terms of temporal, spatial and quality characteristics. To illustrate, networks may be dense or sparse; companies may always offer a complete range of services or have distinct services during peak and off-peak periods; their services may differ in quality as reflected in speed, punctuality, frequency, travel links, etc. These characteristics should be an integral part of the technology description.

The above discussion implies some practical problems for both parametric and non-parametric approaches. For parametric specifications, in particular flexible functional forms, the number of parameters to be estimated quickly becomes very large. To circumvent this problem, Spady and Friedlaender (1978) suggested the specification of hedonic output composites correcting the generic output vehicle-km for variations in the above characteristics. These are estimated jointly with the structure of the technology. A completely different solution is to define outputs in a very disaggregated way, namely at the level of individual origin-destination flows per period (Jara Díaz 1982). However, this method raises questions about the relation between technology characteristics (e.g. returns to scale) and the underlying origin-destination flows.

In the case of non-parametric technology specifications, the problem is more severe. Defining hedonic outputs does not seem feasible because, if a large number of dimensions representing the characteristics are added to inputs and outputs, then this almost automatically increases efficiency and leads to a larger number of efficient observations (Kerstens and Vanden Eeckaut 1995). With only few test procedures to guide the selection of additional dimensions, this could ultimately undermine the discriminatory power of the analysis. (This topic undoubtedly deserves more systematic attention. Developments in including non-discretionary environmental (e.g. categorical) variables can provide a solution. Another possibility is to construct parametric hedonic outputs in a first stage, and to measure efficiency based on inputs and hedonic outputs in a second step (Obeng 1995). This semi-parametric approach reverses the combination of non-parametric and parametric methods relative to Thiry and Tulkens (1992).) Another fruitful approach, however, is to ignore the characteristics in the frontier specification itself, but to include them into a second explanatory stage. The assumption underlying this second phase is that the characteristics only affect the distance to the frontier, but do not influence its shape (Lovell 1993).

Despite these difficulties, over the past decade many empirical models have incorporated various output quality characteristics, several of which are demand-related (e.g. Filippini *et al.* 1992, Hensher 1992, Prioni and Hensher 1999). If both demand and supply attributes are appropriately accounted for, the discussion with respect to the choice of demand versus supply related indicators is no longer crucial. Of course, to the extent that service quality indicators map into both supply and demand characteristics it seems desirable to analyse their impact on performance within the framework of a joint demand–supply equation system (Prioni and Henscher 1999).

4. Urban transit efficiency and productivity: results from frontier studies

The main findings of the literature on urban transit performance are outlined in tables 4 and 5 for parametric and non-parametric approaches respectively. The present authors consecutively discuss efficiency and productivity results, returns to scale and scope, the relation between efficiency and effectiveness, the impact of ownership, subsidies and contracts, and the role of environmental variables and network characteristics.

4.1. *Efficiency and productivity*

When interpreting results note that, to some extent, the distribution of efficiency is determined by the methods employed. For example, differences in underlying assumptions imply that deterministic non-parametric and stochastic parametric methods may generate efficiency scores that substantially diverge. Similarly, among non-parametric approaches it is well known that the FDH specification is more conservative than a DEA model, automatically resulting in higher efficiency scores. Finally, it should be borne in mind that frontier methods only yield relative efficiencies. Efficiency scores are relative to the sample considered and are not based

Table 4. Performance of urban bus companies: parametric frontier findings.

Author	Efficiency findings	Returns of scale and scope	Observations
Bhattacharyya *et al*. (1995)	average TE per operator: between 51 and 98% form of public ownership and management structure influence efficiency (to some extent related to no. of breakdowns) performance deteriorates over time (need to replace old fleet) nationalized companies are the least efficient (had problems at origin) units directly run by government transport department are most efficient	decreasing returns to scale	three types of public ownership analysed: government owned corporations; nationalized units and ownership and operation in state depart. average technical regress: 3.7%, due to the rapid deterioration of vehicles (poor state of the roads)
Costa and Markellos (1997)	average TE: 99% small range of values inhibits detecting any temporal pattern		
De Jong and Cheung (1999)	average TE: 80% regional firms more efficient than urban firms	economies of scale	compute subsidy allocation models: subtracting the inefficiency components would lead to drastic changes
Delhausse *et al*. (1992)	average TE per operator: between 92 and 94% important efficiency differences among companies average TE improves over time		reduced substitution possibilities between labour and energy substitution possibilities between the variable factors (work and energy) and no. of vehicles
Fazioli *et al*. (1993)	average cost efficiency per operator between 78 and 100% ownership irrelevant to efficiency (absence of effective competition and strong regulation)	important economies of scale and density (merger policy justified) importance of scale economies decreases with operator size	no evidence of technical change positive influence of network length on costs

(continued)

Table 4. Continued

Author	Efficiency findings	Returns of scale and scope	Observations
Filippini *et al.* (1992)	efficiency positively correlated to degree to which Cantons subsidize deficits and to compensating payments for public services, and negatively correlated with size and alpine region	important economies of scale and density (selective mergers welcomed) importance of scale economies decreases with operator size	no neutral technical change regress efficiency measures on explanatory variables:
Gathon (1989)	TE per operator: between 58 and 100% TE postitively correlated with operational speed	increasing returns to scale	
Hanusch and Cantner (1991)	efficiency levels on average between 80 and 84% depending on the specification		average technical progress of 1% per year
Kumbhakar and Bhattacharyya (1996)			technical regress for 55% of operators (need to replase old fleet) form of public ownership influences productivity: nationalized companies grow best; and units directly run by government transportation departments perform worst over time
Lijesen (1998)	average TE per operator: 74%	almost constant returns to scale no economies of scope	speed and load reduce costs network length and population density increase costs
Levaggi (1994)	average cost inefficiency per operator: between 14 and 40% all firms are inefficient in their use of capital (excess capacity)	short and long run economies of passenger (load) density (low level of capacity utilization) short run scalse economies and network, density economies, but not in the long run	speed and load reduce costs network length and population density increase costs

(continued)

Table 4. Continued

Author	Efficiency findings	Returns of scale and scope	Observations
Loizides and Giahalis (1995)	no TE or AE found		average total factor productivity decline of 2%, but no systematic tendencies
Matas and Raymond (1998)	average efficiency increases with size efficiency decreases with average bus age, the range of night routes, network-km per vehicle efficiency negatively related to share of subsidies and average speed	increasing returns to density almost constant returns to scale, with slight diseconomies for the largest companies	
Sakano and Obeng (1995)	average TE per operator: 94% AE: capital is 68% overallocated relative to labour and 123% relative to fuel negative impact of subsidies on TE, no effect on AE no relation between size and TE, but AE decrease with size	on average increasing returns to scale	efficiency decreases with average fleet age
Sakano *et al*. (1999)	average TE per operator: 83% average AE: 23% excess labour relative to capital, 88% excess fuel relative to capital, and 65% excess fuel relative to labour impact of subsidies on AE is small relative to other internal factors	mild increasing returns to scale	positive impact of network size and population on performance

(continued)

Table 4. Continued

Author	Efficiency findings	Returns of scale and scope	Observations
Thiry and Tulkens (1992)	TE of the three firms are quite different: STIB: minimum of 79%; STIL: minimum of 98%; and STIC: minimum of 90%	increasing returns to scale (only slightly for STIB-Brussels)	progress up to 10% in a single year semi-parametric methods: frontier estimated after elimination of inefficient observations with FDH filtering improves the quality of the estimation (more consistent with economic theory)
Viton (1986)	TE per operator ranges between 30 and 100% no systematic pattern of AE TE and AE are correlated substantial TE prevails inefficiency unrelated to size	decreasing returns to short and long run scale economies (fleet sizes smaller than optimal)	little substitution possibilities large transit monopolies have no cost advantages over an industry with many firms
Viton (1992, 1993)		consolidations of scale depend on output levels provided by the merging parties consolidations of scope depend most heavily on resulting systemwide wage, and also on outputs and modes	operators analysed have 100 or more vehicles anaomalous results in the sign of peaking Viton (1993) provides details on all possible consolidations of seven operators in the San Francisco Bay Area (1988 data)

Table 5. Performance of urban bus companies: non parametric frontier findings.

Author	Database	Inputs	Outputs	Efficiency findings	Observations
Button and Costa (1999)				average TE increases over time, except for smallest operators evolution probably explained by pressure to reduce subsidies	also discuss the Madrid metro results earlier reported in Costa (1998)
Chang and Kao (1992)	time series 1956–88	vehicles, staff, fuel	Vk	public operator increased efficiency after liberalization in 1969	
			Vk revenue	public operator decreased efficiency after liberalization	
			bus trips	public operator did not increase efficiency after liberalization	
	panel data 1970–88	vehicles, staff, fuel	Vk Vk revenue bus trips revenue	lower efficiency scores of public operator	one public operator and four private operators private firms more flexible in adopting different technologies (when considering three outputs)
Chu *et al*. (1992)				average input TE: 85% average input effectiveness: 65%	identification of peer groups performance evaluation must distinguish TE and affectiveness measures
Costa and Markellos (1997)				average TE: 90% average SCE: 87%[1] efficiency improves over time	also use multi-layer perception neural networks model: detect congestion

(continued)

Table 5. Continued.

Author	Database	Inputs	Outputs	Efficiency findings	Observations
Costa (1998)				average input TE: 91%[1] (range: 83 – 100%) average input effectiveness: 89%[1] (range: 78 – 100%) reorganization of public transport affected efficiency and effectiveness, but in different forms	measures of efficiency and effectiveness must be distinguished
Cowie and Asenova (1999)				low average TE: 68% public companies less efficient than private ones TE far more important than SCE most companies operate under IRS (up to about £12 million operating revenue, or about 250 buses)	
Gathon (1989)				average TE: 96% (range: 65 – 100%) only 15 of 60 firms are inefficient	inefficient firms with FDH are classified likewise with COLS
Kerstens (1996, 1999)				average TE varies between 82 and 89% for DEA depending on specification private ownership, risk-sharing and duration of contracts have a positive effect on TE subsidies diminish TE; French transportation tax promotes performance heterogeneity of network justifies some differences in TE (length of line with positive effect and distance between stops with a negative effect) TE is most important, followed by SCE and congestion	importance of outliers (in particular for SCE) explanation of TE scores on limited subsample (33 observations) small operators experience IRS; large operators DRS efficiency and effectiveness are uncorrelated impact of measurement orientation

(continued)

Table 5. Continued.

Author	Database	Inputs	Outputs	Efficiency findings	Observations
Levaggi (1994)				average TE between 41 and 59%, depending on model plausible causes of inefficiency: use of excess capital; large spare capacity; and high share of labour input	slacks are usually related to labour and load high rank correlation between DEA and parametric results
Loizides and Giahalis (1995)				TE between 94 and 100% AE between 94 and 100%	
Nolan (1996)				average TE scores per year: around 1.10 TE more important than SCE peak – base ratio, share of maintenance personnel to total employees, average fleet age have negative effect on TE state subsidies diminish TE, while federal subsidies improve it network heterogeneity justifies some differences in performance (average speed with positive effect)	
Nolan *et al.* (2000)		staff fuel vehicles	Veh. revenue miles veh. miles	TE ranges between 76 and 100%	
		Input effic. score (prev. model)	Non-diesel fuel staff safety incidents route miles	'social efficiency' ranges between 20 and 100% 'social efficiency' largets among large firms no trend in 'social efficiency' after 1991 (= introduction of ISTEA legislation)	ISTEA legislation did not make a change (lack of proper monitoring mechanisms)
Nollet *et al.* (1988)				more efficient in the beginning (1977) and in the end (1985), after a significant decline	minor technical progress

(continued)

Table 5. Continued.

Author	Database	Inputs	Outputs	Efficiency findings	Observations
Obeng (1994)				TE between 54 and 100% operating and capital subsidies enhance TE TE declines with size	conclusions may be affected by a methodological error
Tone and Sawada (1990)	cross-section 1985	vehicles staff	Vk	service efficiency concept public companies use excess inputs	initial data set had 207 bus operators (163 provate and 44 public) conclusion are taken with 22 (16 private and six public) companies operating in North Kyushu
		operating expenses	vehicles staff	cost efficiency concept results reflect differences in labour costs between urban and rural regions	
		vehicles staff	operating income	income efficiency concept urban companies more efficient than rural companies in urban areas public companies are more efficient than private (reverse in rural areas)	
		vehicles staff	density of service	public service efficiency concept rural companies more efficient	public transport in rural (urban) areas should be publicly (privately) managed the more a company is income efficient, the less it serves the area
Tulkens *et al.* (1988)				Liège (STIL) most efficient; Charlerloi (STIC) in between; Verviers (STIV) least efficient STIL and STIC: input and output orientated efficiency measures give similar results; STIV: results diverge	Liège: low value of technical progress Charlerloi: more frequent technical progress Charlerloi: more frequent technical progress Verviers: no technical progress

(continued)

Table 5. Continued.

Author	Database	Inputs	Outputs	Efficiency findings	Observations
Tulkens (1993)				TE ranges between 82 and 100% 73 months are inefficient 61 months are sequentially inefficent	FDH calcualted sequentially to detect technical progress 19 observations exhibit technical progress: occasionally in 1984–85–86; and systematically in 1988–89 maximum technical progress in November 1988 (11.7%)
Tulkens and Vanden Eeckaut (1995)				occasional progress between 1985 and 1988 substantial and consistent progress after 1989	
Tulkens and Wunsch (1994)				TE increases in recent period, especially after a formal contract between operator and regulator has been agreed	results less favourable concerning effectiveness (Pk/Sk)
Viton (1997)				mean input and output TE of 0.96 resp. 1.06 (Russell index) majority of units operates under CRS; less under IRS; few under DRS	conclusions w.r.t congestion are affected by a methodological error
Viton (1998)				mean input and output productivity change of 2.1% resp. 2.6% productivity change mainly due to changes in relative TE (catching up)	uses Russell index to compute Malmquist productivity index
Wunsch (1994)				TE ranges from 43–100% (FDH) and from 26–100% (DEA) UK operators perform very well	identification of role models per mode

[1] Own computations. Pk, passenger-km; Sk, Seat-km; Vk, vehicle-km.

on some absolute (e.g. engineering) standards, making, for instance, comparisons of efficiency levels between studies impossible.

With these caveats in mind, we turn to the main findings. First, in terms of technical efficiency most studies report substantial remaining inefficiency among urban transit operators in the different countries (e.g. Viton 1986, Hanusch and Cantner 1991, Fazzioli *et al.* 1993, Levaggi 1994, Bhattacharyya *et al.* 1995, Kerstens 1996, Lijesen 1998). Of course, though the evidence of substantial technical inefficiencies among urban transit operators in the different countries is undeniable, it is less clear how these performance results compare to other sectors of the economy. Second, comparative work of transit operators in different countries (e.g. Gathon 1989, Wunsch 1994) reveals a huge variability in technical inefficiency, both across and within countries. This observed variation captures differences in the regulatory framework, in managerial quality and in operating environment, among others. For example, operators in the UK appear to do very well compared to other countries, a phenomenon that has been attributed to recent regulatory changes. As observed by Glaister (1997), deregulation in the UK brought about drastic cost reductions for at least two reasons. One was that it introduced productivity enhancing working practices and led to reduced wage rates. The other cost-reducing factor was the requirement that the remaining subsidized (social) bus services should be subjected to competitive tendering, i.e. a bidding process for the monopoly right to supply a predefined service at a particular spatial level during a particular period. This is believed to have lowered subsidies by ~20%.

Third, the available efficiency studies clearly illustrate the relative nature of best-practice comparisons and the importance of underlying assumptions. The Brussels public bus operator, for example, seems to perform reasonably well when studied in isolation using time series techniques (e.g. Tulkens and Wunsch 1994), but in a comparative perspective it turns out to perform far below average.

Frontier methods have also been used to study scale, structural (congestion) and allocative efficiency. From the scarce available literature it appears that scale inefficiencies are no major source of poor performance (Nolan 1996, Cowie and Asenova 1999, Kerstens 1999). Results on congestion are mixed. In Kerstens (1999) structural inefficiencies do not appear to be important for a sample of French public transport operators, but Costa and Markellos (1997) found evidence of congestion for the London Underground. Finally, the few studies (Viton 1986, Loizides and Giahalis 1995, Sakano and Obeng 1995) considering allocative inefficiencies suggest that the nature of these inefficiencies strongly depend on the regulatory environment. On the one hand, the existence of capital subsidies encourages capital-intensive production methods. On the other hand, union influence and managerial preferences may induce excessive labour utilization in producing transport services (e.g. De Borger 1993). Empirically, only Sakano and Obeng (1995) reported important allocative inefficiencies. They suggest that the subsidy effect dominates in their sample and that production is relatively too capital intensive. Some of these findings are less pronounced in their sequel work (Sakano *et al.* 1997), where allocative inefficiency is mainly caused by factors internal to the firm instead of subsidies.

Regarding productivity change, a mixed picture arises. Most of the time, very small or negligible rates of productivity growth are obtained. In a sense, this is expected given the mature nature of bus technology and its operating environment. Technology is well established, since major improvements in fuel efficiencies were achieved in the past, and further labour efficiency improvements are unlikely given that one man-one bus operation are nowadays the rule. Moreover, increasing traffic congestion decreases commercial speeds and lowers performance despite counteracting measures (e.g. exclusive lanes, etc.).

There are, however, both negative and positive exceptions to the general picture of small or zero productivity changes. For example, Bhattacharyya *et al.* (1995) and Kumbhakar and Bhattacharyya (1996) reported very large productivity declines over time in Indian urban transit. On the positive side, Delhausse *et al.* (1992), Button and Costa (1999), Costa and Markellos (1997), Tulkens and Wunsch (1994) and Viton (1998) all reported positive productivity changes. In the case of the Belgian and UK studies these are partially attributed to regulatory changes that increased financial responsibility. According to Viton (1998), productivity growth was apparently largely due to a catching-up effect (i.e. an improvement in technical efficiency over time) and not so much due to technological frontier shifts.

4.2. *Returns to scale and scope*

Although the frontier studies reviewed are not specifically designed to study returns to scale and scope, they often produce interesting results as an automatic by-product. While numerous frontiers point at economies of scale and density (e.g. Filippini *et al.* 1992, Thiry and Tulkens 1992, Fazioli *et al.* 1993, Sakano *et al.* 1997), most recent studies provide evidence in favour of the classical U-shaped cost functions. These imply increasing returns to scale for the smaller operators, then constant and finally decreasing returns to scale for big companies. Examples of non-parametric studies reporting these firm-specific scale economies are Viton (1997) and Kerstens (1999). These findings are largely consistent with earlier non-frontier work surveyed in Berechman (1993). He points out that, in the very short-run, i.e. holding both network structure and fleet size constant, there appear to be large economies of capital stock utilization. In addition, most studies find that bus technology is characterized by economies of traffic density so that more intensive use of a given network reduces the cost per vehicle-km. This appears to be true in the short run because of the aforementioned capital stock utilization economies, but it is also valid in the medium run when fleet size can be adjusted. Finally, the overall picture in terms of scale economies is one of a U-shaped relation between average cost per vehicle-km and output expressed in vehicle-km, with a very broad range of constant returns to scale. Berechman (1993) also argued that small firms typically experience increasing returns to scale; while medium-sized companies face limited increasing or constant scale returns; and that the large systems are subject to decreasing returns to scale. The transition point from increasing to decreasing returns to scale seems to be situated somewhere between 250 and 400 buses (cf. Berechman 1993, Cowie and Asenova 1999).

Viton (1992, 1993) is the only work reporting in detail on economies of scope. In particular, an attempt is made to answer the question whether a consolidation operation could lead to cost savings for the seven companies in the San Francisco Bay Area. It turns out that the answer to some extent depends on the modes being offered by the potentially merging companies and by the number of companies being merged. In general, benefits fall as the number of companies involved increases.

4.3. *Relation between efficiency and effectiveness*

An interesting question is the extent to which efficiency and effectiveness are related. This issue directly bears on the specification of the appropriate objectives for public transit firms. If empirical studies do not sufficiently incorporate output characteristics reflecting these objectives (e.g. demand-related output attributes), then efficiency and effectiveness may be not at all, or even negatively, related. Empirical evidence corroborates this view. Kerstens (1999) noticed that there is almost no correlation between technical efficiency and effectiveness, and finds that conclusions regarding performance are strongly conditional on output specification. Casual evidence reported by others even points at negative relations between efficiency and effectiveness (e.g. Tone and Sawada 1990, Chu *et al.* 1992, Schinnar 1993). (In fact, Schinnar (1993) cites a study of 145 public bus companies reporting such a negative association.)

There is also mixed evidence with respect to the time pattern of efficiency and effectiveness, again suggesting that both concepts may be unrelated unless all relevant characteristics are accounted for. For example, Costa (1998) considered a single operator (Metro de Madrid) over a small time span and finds a simultaneous improvement in efficiency and effectiveness after the introduction of organizational reforms. Equally focusing on a single operator, Tulkens and Wunsch (1994), by contrast, find a temporal pattern of improving efficiency and deteriorating effectiveness.

4.4. *Ownership*

A popular informal argument states that productivity and efficiency are higher in the private than in the public sector. Earlier surveys for the transit sector by Perry *et al.* (1988) and Berechman (1993) do not provide much support for this view. Variations in ownership and management systems are little correlated with performance, although it turns out that public operators generally offer higher service levels in general as well as during peak hours.

Frontier studies provide some more detailed and recent information on this highly controversial issue. Although Fazioli *et al.* (1993) found no relation between technical efficiency and ownership among Italian urban transit firms, most studies do suggest positive associations between efficiency and private ownership. For example, Tone and Sawada (1990) and Chang and Kao (1992) reported a better performance of private operators in Japan and Taiwan, respectively. Cowie and Asenova (1999) found public companies to be less efficient than private ones in the deregulated UK markets, while Kerstens (1996) also finds a positive effect of private ownership on efficiency in France.

However, ownership comes in different kinds and this may well prove important. This is illustrated by Bhattacharyya *et al.* (1995) analysing several types of ownership in India. They conclude that the form of public ownership and management structure affect efficiency levels. Nationalized firms experience the highest degree of inefficiency, but one reason could be that that the nationalization only affected units with problems right from the outset. Furthermore, the autonomous public transport corporations are less efficient than the operations organized directly by the transport department itself. Kumbhakar and Bhattacharyya (1996), in a complementary study, reported that the nationalized firms grow fastest while the units run directly by the government transport department perform worst over time. Hence, static and dynamic efficiency patterns need not coincide.

Unfortunately, the evidence provided by frontier studies in favour of private sector provision should be weighted against the fact that almost none of these studies controls for the degree of competition and the nature of government regulation in the sector. Indeed, it is often argued that for strongly regulated markets (in terms of entry and exit, pricing, etc.) like urban transit, ownership is of little relevance on its own, though the market structure and the nature of competition is. The deregulation of UK and US bus markets led to an extensive discussion regarding the impact of potential competition in urban transit (Mackie *et al.* 1995, De Borger and Kerstens 2000).

4.5. *Subsidies and contract design*

Filippini *et al.* (1992) reported that government subsidies can have positive or negative impacts on performance, depending on the political proximity of the regulator and on whether the regulator can or cannot control company information. More specifically, they argue that local or regional government bodies are better able to monitor the performance of urban transit operators than the central government. Nolan (1996), however, did not find support for this view for the USA: state subsidies diminish efficiency, while federal subsidies improve it. Kerstens (1996) and Matas and Raymond (1998) found a clear negative relation between subsidies and urban transit performance, but do not control for the sources of subsidies. (Obeng (1994) concluded that operating and capital subsidies enhance technical efficiency. However, the latter inference is based on comparing a DEA model with and without subsidies. Since efficiency measurement is sensitive to the number of dimensions, his result may be an artefact of the methodology (Kerstens and Vanden Eeckaut 1995, Obeng 1995).) Sakano and Obeng (1995) also reported a negative impact of subsidies on technical efficiency, but no effect on allocative efficiency. Sakano *et al.* (1997) found that allocative inefficiencies are mainly caused internally, instead of being induced by subsidies Finally, Tulkens *et al.* (1988) related the bad performance of one Belgian operator to excess capacity resulting from seemingly redundant investments in additional buses. Probably this is linked with investment subsidies.

Kerstens (1996) also reported on other contractual arrangements. The negative effect of subsidies is independent of the risk sharing agreed upon in contracts between operators and public authorities. Sharing risks was found to enhance performance, as did the contract duration. Another important determinant of performance was a locally levied, earmarked tax on the wage bill ('versement transport') that turns out to have a positive impact. It is conjectured that tax rates affect monitoring efforts of citizens and, indirectly, regulators.

The findings on subsidies are certainly in line with the extensive literature documenting that subsidies in fact contribute to cost escalation in the sector, and not the reverse (Pucher 1988), though the size of this effect partly depends on the political proximity of the regulator. Subsidies tend to worsen the performance of urban public transport in a variety of ways: higher costs, fewer revenue-passengers, excessive wage growth, and technical inefficiency. Furthermore, specific capital subsidies tend to create excess capacities. Too few studies have so far empirically looked at the impact of contractual arrangements to derive useful conclusions for regulatory policies.

4.6. *Environmental variables, network characteristics and size*

As to the spatial characteristics affecting performance, Filippini *et al.* (1992) reported a negative effect of the number of stops served in the network, the latter being an element of environmental heterogeneity. Similarly, Fazioli *et al.* (1993) and Levaggi (1994) observed that network length affects performance negatively. Sakano *et al.* (1997), however, came to an opposite conclusion. Gathon (1989) detected that technical efficiency is positively related to average operational speed. This result is confirmed in Levaggi (1994), Nolan (1996) and Viton (1992, 1993). Kerstens (1996) noticed that the distance between stops has a negative effect on efficiency levels, while the length of lines is positively related to performance. Matas and Raymond (1998), Nolan (1996) and Sakano and Obeng (1995) reported that efficiency decreases with average fleet age. By contrast, Kerstens (1996) and Viton (1986) found no significant capital-vintage effects. Tone and Sawada (1990) discovered that urban companies perform better than rural operators.

Temporal service characteristics also affect the performance of transit operators. Not surprisingly, Nolan (1996) observed a negative relation between peak-to-base ratios and technical efficiency. Matas and Raymond (1998) detected a similar negative relation between efficiency and the range of night routes, i.e. another determinant of network heterogeneity. A bit surprisingly, Viton (1986) found no effects at all of the peak-to-base ratio on urban transit performance, while Viton (1992, 1993) even reported the anomalous result that higher peak-to-base ratios lower costs.

For a sample of US urban transit operators Obeng (1994) reported efficiency levels declining with size. Analysing Swiss companies Filippini *et al.* (1992) discovered that technical efficiency is negatively related to the size of companies, which is interpreted as evidence of bureaucratic influences. This finding can also qualify their above-mentioned conclusion that the prevalence of economies of scale and density may legitimate a selective merger policy. Viton (1986) is less conclusive since inefficiency is unrelated to the size of operations. Sakano and Obeng (1995) found no relation between size and technical efficiency, but report allocative inefficiencies decreasing with size.

5. Conclusions and policy implications

The survey on urban transit performance suggests that frontier methodologies have substantially increased our knowledge of the determinants of productivity growth and efficiency changes in the sector. In particular, technical inefficiencies are widespread and technical progress is low and not unequivocal. Frontier studies even suggest that technical inefficiency is the dominant source of poor performance, rather than congestion, inefficiencies in scale or allocative inefficiency. Furthermore, the finding of a negative relation between efficiency and effectiveness certainly requires further investigation. One implication for regulatory policies is that the choice between input and output monitoring as well as the precise specification of outputs may demand more reflection.

The frontier evidence clearly shows that the regulatory environment and the characteristics of the network substantially influence efficiency and productivity. With respect to the former, it was found that ownership, the risk sharing properties of contracts between operator and public authority, and the level and nature of subsidies to operators all directly affect public transit performance. (The destructive impact of subsidies may call for making them conditional on performance. De Jong and Cheung (1999) developed a subsidy allocation mechanism net of technical inefficiency.) With respect to the latter, although network characteristics are always found to be highly relevant, their role in explaining efficiency is not entirely obvious. Indeed, while some characteristics influencing efficiency levels are under the control of the companies or the public authorities (e.g. number of stops; network length; length of lines, etc.), others are largely exogenous (e.g. the average operational speed is mainly determined by transport infrastructure and congestion levels). The results do suggest that it may be wise to allow operators some freedom to organize their production to achieve maximum efficiency. Moreover, public authorities can influence the efficiency of transport operations by improvements in the transport network so as to reduce, for instance, congestion levels.

Unfortunately, the overall picture is not entirely positive. Although frontier methodologies have, as highlighted above, enlarged our knowledge on the determinants of efficiency, a large number of basic problems persist. First, many studies suffer from the lack of appropriate data. For this reason, correcting for differences in quality of inputs and outputs remains difficult for these new methods. Second, appropriately accounting for the network structure of transit operators remains a challenge. Again, data on attributes are often unavailable. In addition, as argued above, many relevant characteristics are largely outside the control of operators but imposed by the regulatory environment, or partly determined by demand. This makes it unclear whether such characteristics are part of technology or determinants of performance. Third, in many cases insufficient evidence is available on economically crucial issues. For example, the efficiency effect of improving competitive conditions in the industry has not convincingly been shown. In this respect it is especially unfortunate that few frontier studies have focused on the effects of privatization and regulatory changes (e.g. in countries like the UK). (Interestingly, Nolan *et al.* (2000) showed that regulatory policies should be carefully designed to have efficiency effects. The ambitious Intermodal Surface Transportation Efficiency Act (ISTEA) seems to have very little effect on transit firms' efficiency.) Fourth, it is evident that there is a huge need for comparative international research to provide more details on the relative performance of operators working under different regulatory regimes.

Acknowledgements

Thanks are due to an anonymous referee. The usual disclaimer applies.

References

Berechman, J., 1993, *Public Transit Economics and Deregulation Policy* (Amsterdam: North-Holland).

Bhattacharyya, A., Kumbhakar, S. and Bhattacharyya, A., 1995, Ownership structure and cost efficiency: a study of publicly owned passenger-bus transportation companies in India. *Journal of Productivity Analysis*, **6,** 47–61.

Button, K. and Costa, A., 1999, Economic efficiency gains from urban public transport regulatory reform: two case studies of changes in Europe. *Annals of Regional Science*, **33,** 425–438.

Chang, K.-P. and Kao, P.-H., 1992, The relative efficiency of public versus private municipal bus firms: an application of data envelopment analysis. *Journal of Productivity Analysis*, **3,** 67–84.

Chu, X, Fielding, G. and Lamar, B., 1992, Measuring transit performance using data envelopment analysis. *Transportation Research*, **26A,** 223–230.

Costa, Á., 1998, Public transport efficiency and effectiveness: Metro de Madrid. In K. Button, P. Nijkamp and H. Priemus (eds), *Transport Networks in Europe: Concepts, Analysis and Policies* (Cheltenham: Elgar), pp. 248–264.

Costa, Á. and Markellos, R. N., 1997, Evaluating public transport efficiency with neural network models. *Transportation Research*, **5C,** 301–312.

Cowie, J. and Asenova, D., 1999, Organisation form, scale effects and efficiency in the British bus industry. *Transportation*, **26,** 231–248.

De Borger, B., 1993, The economic environment and public enterprise behaviour: Belgian railroads, 1950–1986. *Economica*, **60,** 443–463.

De Borger, B. and Kerstens, K., 2000, The performance of bus transit operators. In D. Hensher and K. Button (eds), *Handbooks in Transport—Handbook I: Transport Modelling* (New York: Pergamon), 577–595.

De Jong, G. and Cheung, F., 1999, Stochastic frontier models for public transport. In H. Meersman, E. Van De Voorde and W. Winkelmans (eds), *World Transport Research: Selected Proceedings of the 8th World Conference on Transport Research,* vol. 1: *Transport Modes and Systems* (New York: Pergamon), pp. 373–386.

Delhausse, B., Perelman, S. and Thiry, B., 1992, Substituabilité partielle des facteurs et efficacité-coût: l'example des transports urbain et vicinal Belges. *Economie et Prévision*, **32,** 105–115.

Färe, R., Grosskopf, S. and Lovell, C. A. K., 1994, *Production Frontiers* (Cambridge: Cambridge University Press).

Farrell, M., 1957, The measurement of productive efficiency. *Journal of the Royal Statistical Society*, **120,** 253–281.

Fazioli, R., Filippini, M. and Prioni, P., 1993, Cost-structure and efficiency of local public transport: the case of Emilia Romagna bus companies. *International Journal of Transport Economics*, **20,** 305–324.

Filippini, M., Maggi, R. and Prioni, P., 1992, Inefficiency in a regulated industry: the case of Swiss regional bus companies. *Annals of Public and Cooperative Economics*, **63,** 437–455.

Førsund, F., 1992, A comparison of parametric and non-parametric measures: the case of Norwegian ferries. *Journal of Productivity Analysis*, **3,** 25–44.

Gathon, H.-J., 1989, Indicators of partial productivity and technical efficiency in the European urban transit sector. *Annals of Public and Cooperative Economics*, **60,** 43–59.

Glaister, S., 1997, Deregulation and privatisation: British experience. In G. De Rus and C. Nash (eds), *Recent Developments in Transport Economics* (Aldershot: Ashgate), pp. 135–197.

Glaister, S., Starkie, D. and Thompson, D., 1990, The assessment: economic policy for transport. *Oxford Review of Economic Policy*, **6,** 1–21.

Good, D., Nadiri, M., Röller, L.-H. and Sickles, R., 1993, Efficiency and productivity growth comparisions of European and US air carriers: a first look at the data. *Journal of Productivity Analysis*, **4,** 115–125.

Greene, W., 1997, Frontier production functions. In M. H. Pesaran and M. R. Wickens (eds), *Handbook of Applied Econometrics*, vol. II: *Microeconomics* (Oxford: Blackwell), pp. 81–166.

Grosskopf, S., 1996, Statistical inference and non-parametric efficiency: a selective survey. *Journal of Productivity Analysis*, **7,** 161–176.

Hanusch, H. and Cantner, U., 1991, Produktion öffentlicher leistungen: Effizienz und technischer fortschritt. *Jahrbücher für Nationalökonomie und Statistik*, **208,** 369–384.

Jara Díaz, S. R., 1982, The estimation of transport cost functions: a methodological review. *Transport Reviews*, **2,** 257–278.

Kerstens, K., 1996, Technical efficiency measurement and explanation of French urban transit companies. *Transportation Research*, **30A,** 431–452.

Kerstens, K., 1999, Decomposing technical efficiency and effectiveness of French urban transport. *Annales d'Economie et de Statistique*, **54,** 129–155.

Kerstens, K. and Vanden Eeckaut, P., 1995, The economic cost of subsidy-induced technical inefficiency: a methodological postscript. *International Journal of Transport Economics*, **22,** 225–229.

Kumbhakar, S. and Bhattacharyya, A., 1996, Productivity growth in passenger-bus transportation: a heteroskedastic error component model with unbalanced panel data. *Empirical Economics*, **21,** 557–573.

Levaggi, R., 1994, Parametric and non-parametric approach to efficiency: the case of urban transport in Italy. *Studi Economici*, **49,** 67–88.

Lijesen, M., 1998, Analyzing cost structures of public sector activities, with an application to regional public transport. In G. B. K. de Graan and F. G. Volmer (eds), *Performance Budgeting: A Perspective on Modelling and Strategic Planning in the Public Sector in Holland* (Delft: Eburon), pp. 178–187.

Lovell, C. A. K., 1993, Production frontiers and productive efficiency. In H. Fried, C. A. K. Lovell and S. Schmidt (eds), *The Measurement of Productive Efficiency: Techniques and Applications* (Oxford: Oxford University Press), pp. 3–67.

Loizides, I. and Giahalis, B., 1995, The performance of public enterprises: a case of the Greek railway organization. *International Journal of Transport Economics*, **22,** 283–306.

Mackie, P., Preston, J. and Nash, C., 1995, Bus deregulation: ten years on. *Transport Reviews*, **15,** 229–252.

Marchand, M., Pestieau, P. and Tulkens, H., 1984, The performance of public enterprises: Normative, positive and empirical issues. In M. Marchand, P. Pestieau and H. Tulkens (eds), *The Performance of Public Enterprises: Concepts and Measurement* (Amsterdam: Elsevier), pp. 3–42.

Matas, A. and Raymond, J.-L., 1998, Technical characteristics and efficiency of urban bus companies: the case of Spain. *Transportation*, **25,** 243–263.

Nolan, J. F., 1996, Determinants of productive efficiency in urban transit. *Logistics and Transportation Review*, **32,** 319–342.

Nolan, J. F., Ritchie, P. C. and Rowcroft, J. E., 2001, Identifying and measuring public policy goals: ISTEA and the US bus transit industry. *Journal of Economic Behavior and Organization* (forthcoming).

Nollet, C., Thiry, B. and Tulkens, H., 1988, Mesure de l'efficacité productive: application a la société de transports intercommunaux de Bruxelles. In B. Thiry and H. Tulkens (eds), *La Performance Economique des Sociétés Belges de Transport Urbains* (Charlerloi: CIRIEC), pp. 137–170.

Obeng, K., 1994, The economic cost of subsidy-induced technical inefficiency. *International Journal of Transport Economics*, **21,** 3–20.

Obeng, K., 1995, The economic cost of subsidy-induced technical inefficiency: a reply. *International Journal of Transport Economics*, **22,** 231–236.

Oum, T. H., Thretheway, M. and Waters II, W. G., 1992, Concepts, methods and purposes of productivity measurement in transportation. *Transportation Research*, **26A,** 493–505.

Oum, T. H. and Waters, W. G., II, 1996, A survey of recent developments in transportation cost function research. *Logistics and Transportation Review*, **32,** 423–463.

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,** 9–42.

Perry, J., Babitsky, T. and Gregersen, H., 1988, Organizational form and performance in urban mass transit. *Transport Reviews*, **8,** 125–143.

Pestieau, P. and Tulkens, H., 1993, Assessing and explaining the performance of public enterprises. *Finanzarchiv*, **50,** 293–323.

Pucher, J., 1988, Urban public transport subsidies in Western Europe and North America. *Transportation Quarterly*, **42,** 377–402.

Rees, R., 1984, *Public Enterprise Economics*, 2nd edn (London: Weidenfeld & Nicholson).

Sakano, R. and Obeng, K., 1995, Re-examination of inefficiencies in urban transit systems: a stochastic frontier approach. *Logistics and Transportation Review*, **31,** 377–392.

Sakano, R., Obeng, K. and Azam, G., 1997, Subsidies and inefficiency: stochastic frontier approach. *Contemporary Economic Policy*, **15,** 113–127.

Schinnar, A., 1993, Tradeoffs between efficiency and effectiveness in management of public services. In Y. Ijiri (ed.), *Creative and Innovative Approaches to the Science of Management* (Westport: Quorum), pp. 177–190.

Spady, R. and Friedlaender, A., 1978, Hedonic cost functions for the regulated trucking industry. *Bell Journal of Economics*, **9,** 159–179.

Thiry, B. and Tulkens, H., 1992, Allowing for inefficiency in parametric estimation of production functions for urban transit firms. *Journal of Productivity Analysis*, **3,** 45–65.

Tone, K. and Sawada, T., 1990, An efficiency analysis of public vs. private bus transportation enterprises. In H. Bradley (ed.), *Operational Research '90* (New York: Pergamon), pp. 357–365.

Tulkens, H., 1993, On FDH efficiency analysis: Some methodological issues and applications to retail banking, courts, and urban transit. *Journal of Productivity Analysis*, **4,** 183–210.

TULKENS, H., THIRY, B. and PALM, A., 1988, Mesure de l'efficacité productive: methodologies et applications aux sociétés de transports intercommunaux de Liège, Charlerloi et Verviers. In B. Thiry and H. Tulkens (eds), *La Performance Economique des Sociétés Belges de Transport Urbains* (Charlerloi: CIRIEC), pp. 81–136.

TULKENS, H. and VANDEN EECKAUT, P., 1995, Non-frontier measures of efficiency, progress and regress for time series data. *International Journal of Production Economics*, **39,** 83–97.

TULKENS, H. and WUNSCH, P., 1994, Les performance économiques de la STIB au mois le mois. *Revue Suisse d'Économie Politique et de Statistique*, **130,** 627–646.

VITON, P., 1986, The question of efficiency in urban bus transportation. *Journal of Regional Science*, **26,** 499–513.

VITON, P., 1992, Consolidations of scale and scope in urban transit. *Regional Science and Urban Economics*, **22,** 25–49.

VITON, P., 1993, How big should transit be? Evidence from the San Francisco bay area. *Transportation*, **20,** 35–57.

VITON, P., 1997, Technical efficiency in multimode bus transit: a production frontier analysis. *Transportation Research*, **31B,** 23–39.

VITON, P., 1998, Changes in multimode bus transit efficiency, 1988–1992. *Transportation*, **25,** 1–21.

WUNSCH, P., 1994, *Costing Busses: Back to the Basics, Brussels, FUSL* (SMASH Cahier 9405).

WUNSCH, P., 1996, Cost and productivity of major urban transit systems in Europe: an exploratory analysis. *Journal of Transport Economics and Policy*, **30,** 171–186.