UCI-ITS-WP-82-1

Cost Functions and Economies of Scale in Bus Transit: A Critique

UCI-ITS-WP-82-1

Joseph Berechman [1]
Genevieve Giuliano [2]

[1] Visiting Associate Professor, School of Social Sciences and
Institute of Transportation Studies
University of California, Irvine. On leave from Tel-Aviv University.

[2] Institute of Transportation Studies
University of California, Irvine

January 1982

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

Abstract

The issue of returns to scale in bus transit continues to be a subject of debate among transportation analysts. From a public policy perspective, returns to scale are relevant to many policy areas such as transit service pricing, cost allocation, subsidization, and optimal firm size. This paper argues that conclusions regarding economies of scale are in large part the result of confusion regarding the concept of scale economies, variable definition, assumptions about the shape of the cost function, and certain characteristics of the data base. It is suggested that generalized cost functions with very few a priori economic restrictions can better represent the cost structure of the industry and therefore are more appropriate for measurement of scale economies.

1. INTRODUCTION

The issue of economies of scale in urban bus transport continues to attract the attention of transportation analysts. Over the past decade, a consensus has developed that bus transit is characterized by constant returns to scale for most bus firms and by decreasing returns for the largest operations (Oram, 1979; McGillivray, Kemp, and Beesley, 1980). However, recent econometric studies indicate that these conclusions may not hold. In fact, the results of these studies indicate economies of scale over a wide range (Viton, 1981; Williams and Dalal, 1981; Berechman, 1982).

The absence or existence of scale economies has important implications for policy issues such as optimal pricing, subsidization, and organization of transit services. If, for example, an expansion of transit services is contemplated, it might be asked whether existing bus companies should increase their level of operation, or whether the additional service should be supplied by separate operators. The existence of scale economies would suggest the former, while the absence of scale economies might suggest the latter, recognizing of course that many factors other than economies of scale would be taken into account in the decision-making process. Thus, conflicting conclusions regarding the economic characteristics of the bus transit industry merit investigation.

A review of the empirical studies which have tested for economies of scale indicates that differences in results stem from both analytical and empirical issues. That is, while the degree of scale economies (i.e., the numerical value of the cost elasticity parameters) is an empirical

question, its measurement is an analytical question which affects empirical results. In particular, confusion regarding the concept of scale economies, the theoretical assumptions made regarding the shape of the cost function, the definition of the output variable, as well as certain characteristics of the data base can affect results. The purpose of this paper is to critically examine these factors and show how they affect empirical conclusions about economies of scale in bus transit.

The paper begins with a discussion of the relevance of scale economies to public transit policy formulation and the particular importance of assuming constant returns to scale. Once some of the policy issues have been discussed, a formal definition of economies of scale is presented. This provides the basis for the discussion of analytical and empirical problems in the four sections which follow. The paper concludes with a summary statement and suggestions for a more appropriate approach to measuring economies of scale.

2. IMPORTANCE OF SCALE ECONOMIES TO POLICY ISSUES

There are many transportation policy issues for which the existence or absence of economies of scale is important. While ample empirical evidence of constant returns may be cited, it can also be argued that researchers are strongly motivated to accept the constant returns hypothesis because it simplifies the analysis of both theoretical and empirical problems. Four such problem areas will be briefly discussed here.

Cost Allocation

Bus firms must estimate the cost of expanding or curtailing services. In order to do so, relevant costs must be identified and measured. In particular, fixed and variable costs must be differentiated in order to correctly assess the cost of the new service. It is a much more complicated problem to estimate costs of incremental service changes under increasing returns to scale, since expenditures on additional factors of production (e.g., labor and rolling stock) are not only a function of their prices but also of the level of output. Most cost allocation studies have therefore adopted the assumption of constant return to scale in developing cost allocation methodologies (Cherwony and Mundle, 1979; McGillivary et al., 1980). If scale economies in bus transport do exist, cost allocation formulaes based on constant returns are at best only approximation and might lead to incorrect incremental cost calculations.

Pricing

An issue which is related to the cost allocation problem is that of pricing bus services. Given the monopolistic status of transit properties, the general theoretical approach is to consider the transit firm as having to set fares and service levels so as to maximize a net revenue function subject to a budget constraint (Nelson, 1972).[1] This

---

[1]Notationally, the decision model is setting fare, P, and service level, S, to maximize: $f[D(P,S),P,S,C]$, with respect to D, demand, or S, subject to the budget constraint: $PD(P,S) - CS = 0$, where C is the constant per unit service cost.

model assumes that demand for services is known, that the transit firm can set the price or the service level or both, that transit firms actually consider budget constraints, and that the cost per unit of service is indeed constant. Pricing strategies have been proposed for cases where the first three assumption may not hold, but if increasing returns to scale in fact prevail in bus operation, no simple normative pricing rule can be derived (Manski, 1979).

A similar problem arises if peak/off-peak pricing is considered. Differential pricing of peak and off-peak services is frequently proposed as a strategy to cover the firm's total costs. When constant returns to scale is assumed, a theoretically satisfactory solution is to set prices where $D = MC$ for each demand period (Mohring, 1976). However, when increasing returns to scale are assumed no simple solution exists, and in fact only heuristic solutions are offered in the literature (Coase, 1946; Mohring, 1970).

Productivity Measurement

Because of the financial difficulties within the U.S. transit industry, productivity is receiving increased attention among analysts (Fielding, Glauthier, and Lave, 1978; Meyer and Gomez-Ibanez, 1975; Tomazinis, 1975). In principle, productivity can be analyzed in a straightforward manner only if unit costs are independent of the scale of operation. Thus, the assumption of constant returns to scale becomes necessary to compare derived productivity rates across transit firms. That is, under conditions of increasing returns, productivity indices, including the performance indicators currently in wide use, will be

relevant only for a given level of output, because productivity itself is a function of the level of output (Caves, Christiansen and Swanson, 1980).

Optimal Output of the Transit Firm and Subsidy Policy

Increasing returns to scale has frequently been proposed as a reason for subsidizing transit firms (Mohring, 1972). Under these conditions, optimal policy requires marginal cost pricing and a subsidy per output unit equal to the difference between average cost and marginal cost at the equilibrium level of output.[2] This pricing and subsidy solution requires that the demand, average cost and marginal cost functions be known by the controlling agency, a situation which is unlikely to prevail. Therefore, unless constant returns are assumed, implementation of such pricing policies is impossible.

The above discussion gives some indication of the importance of constant returns to scale for transit industry policy issues. Without assuming constant returns, existing cost allocation formulae are likely to be invalid, economically sound pricing policies cannot be established in a straightforward manner, and productivity cannot be measured. Thus a clear incentive exists to assume constant returns to scale. The important question is, of course, whether this assumption is valid. While the ultimate answer is an empirical one, it is nevertheless important to examine first the analytical basis of scale economies measurement.

---

[2]Some authors argue that the presence of scale economies does not necessarily lead to the subsidy solution described here. Alternative approaches to the subsidization of natural monopolies have been suggested in the literature (see, for example, Cornell 1980).

3. FORMAL DEFINITION OF SCALE ECONOMIES

Neoclassical economics defines scale economies (or increasing returns to scale) as the case where the total cost elasticity (C.E.) with respect to a change in level of output is less than unity. Formally, letting C be total cost, $\underline{P}$ be a vector of input prices and Q be output,

$$C = f(\underline{P},Q) \quad . \tag{1}$$

where f is a continuous, nondecreasing and concave function of $\underline{P}$. The cost elasticity (C.E.) is:[3]

$$C.E. = \frac{dC}{dQ} = \frac{MC}{AC} \tag{2}$$

where MC and AC are marginal and average cost, respectively. The degree of scale economies can then be measured by using unity minus the

---

[3]To see this, we totally differentiate both sides of (1) to obtain,

$$CdC = Qdf(\underline{P},Q)$$

Dividing both sides by dQ, $(dQ > 0)$, the elasticity of C with respect to Q, dC/dQ, is

$$\frac{dC}{dQ} = \frac{Q}{C}\frac{df(\underline{P},Q)}{dQ} \tag{3}$$

but $Q/C = 1/AC$ and $\frac{df(\underline{P},Q)}{dQ} = MC$; thus $C.E. = \frac{MC}{AC}$.

cost elasticity (Caves et al, 1980). This term (2) is what Griliches (1972) defines as the percent variable, and for C.E. < 1, scale economies exist.

The explicit functional form of $f(\underline{P},Q)$ determines how C.E. changes at various levels of output. In the case of a linear total cost function, MC is constant and AC decreases as output increase (except in the unlikely event that there are no fixed initial costs). Thus C.E. also varies with output. If total cost is a non-linear function of Q, both MC and AC vary with output, and so also does C.E.. For cost functions which are linear in the log of C and Q (i.e., $C = kQ^{\alpha}\underline{P}^{\beta}$), C.E. is constant for all levels of output. Thus, for all functional forms except the last, observed economies of scale depend upon the output level at which they are measured.

Another important point is that if the cost function is linear in output, as many cost studies assume, then as Q increases, $AC \rightarrow MC$ and $dC/dQ \rightarrow 1$[4]. Notice also that in estimating a linear cost model, if the intercept is negative, C.E. > 1, implying diseconomies of scale for every level of output.

The above definition of scale economies is theoretically consistent only if the following conditions are met: no changes in factor proportions, either over time for one firm or among firms for cross-sections analysis; no changes in the product mix or set of service produced,

---

[4]To see this let $C(\underline{P},Q) = G + \underline{P}Q$ where G represents fixed cost. In this case $\frac{\partial C}{\partial Q} = MC = \underline{P}$, and $AC = \frac{G}{Q} + \underline{P}$. As Q increases the term G/Q declines, and, in the limit, $AC = \underline{P} = MC$.

and no changes in the underlying technology. Since economies of scale is basically a long-run concept, these conditions are restrictive indeed. It has been argued that the only increase in capacity permissible under the formal definition is one where an identical producing unit is added, and under these circumstances there is no reason to expect economies of scale (Gold, 1981). On the other hand, empiricists argue that in the real world these conditions are never met, that most changes in scale are associated with other technical adjustments, and consequently that the formal definition is effectively irrelevant. The result of this controversy has been a great deal of confusion over what scale economies means or should mean and what is actually being measured in empirical studies. In response to the confusion, efforts have recently been made to redefine the concept and clarify some of the problems with existing empirical research (Harris, 1977; Gold, 1981). Economic analyses of the bus transit industry also reflect these "state of the art" problems, but the nature of the industry itself presents additional difficulties in defining and measuring economies of scale, as will be discussed in the following sections.

4. THE MEASUREMENT OF OUTPUT AND ECONOMIES OF SCALE IN BUS TRANSIT

The basic concept of scale economies refers to the response of unit cost when the scale of output is increased; thus an appropriate measure of output is necessary for the proper measurement of scale economies. Conceptually, the output of a transit firm is the aggregate of services provided. These services may be described as a set of routes with

varying service characteristics such as service of frequency, travel speed, hours of operation, etc. That is, units of output (however measured) are not homogenous in the aggregate. This presents a number of problems for an empirical cost study. First, the bus property as a whole is the unit of analysis. Thus the overall level of output for the firm must somehow be determined, and therefore some means of aggregating nonhomogenous outputs must be utilized. Secondly, costs of different types of services cannot be clearly differentiated. That is, total cost is a nonlinear and nonseparable function of the level of service on all routes. Moreover, although it is analytically possible to estimate a cost function with multiple outputs, it is likely to be empirically impossible because of data limitations. A unique measure of output which will both approximate service characteristics and allow aggregation for the entire system is therefore required for an empirical estimation of the firm's cost function.

The literature provides two broad categories of such output measures. The first, which might be labeled technical measures, includes measures such as bus-miles and bus-hours. The second category is comprised of demand related measures such as passenger-trips or passenger-miles. Both categories of measures are but crude approximations to the actual services supplied by the bus firm, and thus neither are completely satisfactory.

Data on technical measures such as bus-miles are easily obtained as they are routinely collected by all bus properties. Major cost items like labor and fuel are highly correlated with such measures, and thus

can be expected to provide good statistical results in cost function estimations. Moreover, when conducting a cross-section analysis, such measures are more comparable than demand-related variables and provide good common denominators of output levels of properties which otherwise might be very different from each other. It must be noted, however, that the greater degree of comparability is obtained because such measures do not reflect local service differences. For example, when measured in bus-miles, a highly dense network such as New York City and a relatively sparse network such as Los Angeles, might be indistinguishable, even though service characteristics and presumably production conditions and costs are very different. A second problem with technical measures is that they do not reflect the economic motive for providing the service--the carrying of passengers. Technical measures by themselves do not provide any information on the utilization of bus services. Thus care must be taken not to consider such measures as surrogates for the economic and welfare contribution of the services provided.

Demand related measures have contrasting advantages and disadvantages. Passenger-trips or passenger-miles are of course directly related to actual market transactions and consequently are easily amenable to economic interpretations. These measures not only reflect local service differences, but also reflect differences in the demand environment as well. Input items such as labor and fuel costs may not vary systematically with demand related output measures. Consequently, the use of these measures may rule out the use of many cross-section samples, as the estimation of cost functions using observations on bus properties

with quite different demand environments and consequently production conditions is rather questionable. Furthermore, it has also been argued that since passengers contribute time when making trips, their time must be explicitly included as a factor of production when estimating cost functions based on demand related ouput measures (Mohring 1972). Finally, demand related measures are not collected with the same degree of accuracy as technical measures. Passenger counts are generated either by spot survey techniques, or by applying a passenger mix formula to farebox revenue. Passenger-miles are generally unavailable, and are not computed in a consistent manner across firms.

It is worthwhile at this point to discuss the relationship between the output measure used and economies of scale. The bus firm might be conceptualized as providing service on a given "bundle" of routes. If measured on the basis of bus-miles or bus-hours, economies of scale would depend upon the management and coordination requirements associated with the size of the route bundle, as well as the extent to which routes overlap (the density of the route structure) so as to provide opportunities to utilize input factors more productively, such as by decreasing deadhead time. It is reasonable to expect that "route bundle" economies exist over some range, but that they are likely to be exhausted fairly quickly if, following Gold (1981), changes in firm size (i.e., the scale of operations) are essentially the addition or subtraction of identical ouput units produced under identical production conditions. Indeed, most studies based on technical output measures have reported

constant returns to scale (Nelson, 1972; Veatch, 1973; Wabe and Coles, 1975).[5]

If measured on the basis of passenger-trips, however, one would expect increasing returns if the number of trip possibilities increases more than proportionately with service increases. Under these conditions, ridership should increase more than proportionately as well. In fact, a recent study using a demand related variable reported increasing returns to scale (Berechman, 1982). Moreover, when user time costs are taken into account, Mohring (1972) has shown that economies of scale exist because waiting time (and thus the full trip cost) declines more than proportionately with service frequency increases. Boyd, Asher and Wetzler (1978), following a similar approach, have also reported service-related scale economies in which increases in patronge result in greater frequency, less waiting time costs and lower supplier costs per passenger. It can thus be seen that part of the controversy over the existence of economies of scale may be traced to the use of different output units.

5. ECONOMIES OF SCALE AND ECONOMIES OF DENSITY

A somewhat different approach to the concept of economies of scale is taken by Harris (1977). In his analysis of the railroad industry, Harris makes the distinction between economies of scale and economies of density. Economies of scale measure the relationship between unit cost

---

[5]The exceptions are Viton, 1981; and Williams and Dalal, 1981. Their cost models, however, were markedly different than those of all other studies. See below.

and changes in capacity (the shape of the long-run cost curve relative to size of the firm), while economies of density measure the relationship between unit costs and changes in the intensity of utilization of capacity. In the railroad industry, capacity is measured in terms of routes or route-miles. The more intensively the route system is utilized (i.e., as traffic density inreases), the lower unit costs become. Using the ratio of revenue-ton-miles to miles of route as his measure of density, Harris found significant economies of density in the railroad industry.

Viton (1981) applied the concept of economies of density to bus transit. In this case, economies of density is defined for the short run, and it measures the effect of increased production on unit cost with one fixed factor (rolling stock). Using vehicle-miles as his measure of output, Viton found statistically significant short-run economies of density.

It would appear that the distinction between economies of scale and economies of density is appropriate to the analysis of long-run cost function as well. Analogous to railroad service, the "scale" of the transit firm could be measured in route-miles, and density could be measured in terms of route utilization. In general, it might be assumed that the spatial structure of the route system, as well as the level of service provided, is determined by the level and distribution of demand. Thus different levels of density, or intensity of route system utilization would be associated with different levels of demand. Under these conditions, more intensive use of resources would take place, leading to greater productivity and hence lower unit costs per unit output.

6. THE STRUCTURE OF THE ESTIMATED COST MODEL

The structure of cost models and estimation procedures used in econometric studies of economies of scale merit examination, because the implications of model specification have not always been taken into account. Theoretically, the cost function contains all the relevant economic information about the underlying technology of production (Varian, 1978). Implied in the specification of the cost model is the structure of the production function and its empirical characteristics. For example, if $C(\underline{P},Q)$, where $\underline{P}$ is a vector of input prices and Q is output, is written as $Q^{\alpha}C(\underline{P})$, then for $\alpha = 1$, the production technology $V(Q)$ must exhibit constant returns to scale.[6]

From economic theory, a cost function, $C(\underline{P},Q)$, must meet the following conditions:

a) Continuous function of $\underline{P}$ for $\underline{P}$ greater than zero.

b) Nondecreasing in $\underline{P}$. If $P_2 \geq P_1$, $C(P_2,Q) \geq C(P_1,Q)$.

c) Homogenous of degree one in $\underline{P}$. For $t > 0$, $C(t\underline{P},Q) = tC(\underline{P},Q)$.

d) Concave in $\underline{P}$. For $0 \leq t \leq 1$, $C\ tP_1 + (1 - t)P_2,Q \geq tC(P_1,Q) + (1 - t)C(P_2,Q)$.

If the cost models used in estimating scale economies do not meet these conditions, their estimated parameters, including estimates of scale economies, cannot be regarded as correct. For example, for some functional forms (e.g., Cobb-Douglas, or logarithmic) a necessary and

---

[6] If a cost function $C(\underline{P},Q)$ can be written as, $\phi(Q)C(\underline{P})$, it is said to be a homothetic function. Initial specification of a homothetic cost model is therefore a necessary (but not sufficient) condition in deriving constant returns to scale.

sufficient condition for the homogeneity in prices requirement, condition (c) above, is that the sum of the coefficients of the factor prices equals 1.0. A number of studies which estimated a long-run cost function report results which do not meet this requirement (Fisher and Viton, 1974, p. 74; Pozdena, 1975, p. 44; Nelson, 1972).

More generally, the estimation of economies of scale is but one result of many necesssary to fully characterize the cost structure of the industry. Also of interest are the direct estimation of marginal cost; demand for factors of production, including own and cross-price elasticities of factor demand; elasticity of substitution between factors of production; homotheticity of the cost function, and separability of factor prices. A cost model general enough to place very few a priori restrictions on the economic characteristics of the underlying production process is required in order to obtain these results.

Most of the cost models reported in the literature include many a priori restrictions which, in turn, may affect their economies of scale estimates. To illustrate, most of the cost studies report linear models of the type,

$$C = \alpha_0 + \sum_i \alpha_i P_i + \beta Q \qquad (3)$$

where α_i is the coefficient of the price of the ith factor of production, P_i and Q is output (see, for example, Koshal, 1970 and 1972; Lee and Steedman, 1970; Wabe and Coles, 1975). In some cases (e.g., Nelson 1972; Veatch 1973) the function is linear in the logarithms,

$$\log C = \log \alpha_0 + \sum_i \alpha_i \log P_i + \beta \log Q \quad . \qquad (4)$$

As mentioned earlier, in the case of linear models like equation (3), marginal cost, $\frac{\partial C}{\partial Q}$, is constant and equal to β, and the economies of scale parameter (C.E., Eq. 2) will have a different value for each level of output. For log-linear models like equation (4), marginal cost is, $\beta \frac{C}{Q}$, and the scale parameter C.E. = β; thus scale economies are independent of the level of output.

The specification of models like equation (3) or (4) also implicitly assumes an underlying production technology with zero or unit factor elasticities of substitution.[7] It might be quite undesirable to a priori fix the production technology to have these properties. Furthermore, under these conditions the factor demand price elasticity parameters, ε_{ij} $(i,j = 1,\ldots,n)$, are equal to zero for linear models, and equal to α_i for $i = j$ and zero otherwise, for log-linear models.[8] These substitution

---

[7] The elasticity of substitution, σ, between, say, labor and capital, is defined as $\sigma = \dfrac{d\left(\frac{K}{L}\right) / \left(\frac{K}{L}\right)}{d\left(\frac{P_L}{P_K}\right) \Big/ \left(\frac{P_L}{P_K}\right)}$ where K,L are capital and labor respectively, P_L, P_K, are their unit prices. Uzawa (1962) has shown that given the cost function C, $\sigma = \dfrac{C\left(\frac{\partial^2 C}{\partial P_L \partial P_k}\right)}{\left(\partial C/\partial P_L\right)\left(\partial C/\partial P_K\right)}$.

[8] Let X_i, P_j, be quantity of factor i and price of factor j, respectively. Then $\varepsilon_{ij} = \partial \text{ Ln } X_i / \partial \text{ Ln } P_j$ $(i,j = 1,\ldots,n)$, with quantities and prices of all other factors constant. Allen (1957) had shown $\varepsilon_{ij} = \sigma_{ij} \cdot s_j$, where s_j is the share of factor j in total cost.

and elasticity properties imply a very restricted factor demand function which is unlikely to accurately characterize the bus transit operation. As noted by Gold (1981), changes in various unit cost categories cannot be evaluated independently of other unit costs because of the interconnected effects generated by increases in output, yet these types of models rule out such effects.

In recent years important developments have occurred in the economic theory of production, in particular, the duality theory of production. These developments have led to the utilization of generalized cost functions which place very few a priori restrictions on the underlying production conditions (Fuss et al, 1978). Such cost models have been applied to the bus transit industry only recently in three different studies, namely Viton (1981), Williams and Dalal (1981), and Berechman (1982).[9] All three studies have utilized versions of the generalized translog cost function, the general structure of which is given by equation (5).

$$\text{Ln } C(\underline{P},\underline{Q}) = A + \sum_{i}^{m} \alpha_i \text{ Ln } Q_i + \sum_{i}^{n} \beta_i \text{ ln } P_i + \frac{1}{2} \sum_{i}^{m} \sum_{j}^{m} \delta_{ij} \text{ Ln } Q_i \text{ Ln } Q_j \qquad (5)$$

$$+ \frac{1}{2} \sum_{i}^{n} \sum_{j}^{n} \gamma_{ij} \text{ Ln } P_i \text{ Ln } P_j + \sum_{i}^{m} \sum_{j}^{n} \rho_{ij} \text{ Ln } Q_i \text{ Ln } P_j$$

---

[9] Interestingly these new approaches to production analysis were extensively applied to other modes of transportation like rail and freight. See, for example, Caves et al., 1980, and Friedlander and Spady, 1981.

with the symmetry conditions: $\delta_{ij} = \delta_{ji}$; $\gamma_{ij} = \gamma_{ji}$. Sufficient conditions for C to be homogenous of degree one in $\underline{P}$ are $\sum_{i}^{n} \beta_i = 1.0$;

$\sum_{j}^{n} \gamma_{ij} = 0$, $(j = 1,\ldots,n)$; $\sum_{i}^{n} \rho_{ij} = 0$, $(i = 1,\ldots,n)$.

While it is beyond the purpose of this paper to discuss the theoretical underpinnings of this type of cost function, a few general comments should be made. This model is more flexible than traditional cost models because it does not require the assumptions of homotheticity, separability of factor prices, constant factor elasticity of substitution, and zero or constant price elasticity of factor demand. In general, the translog cost function can be viewed either as a second order approximation to an arbitrary cost function at some point of approxima- tion or as an exact cost function. In either case, it allows the testing of hypotheses regarding the major characteristics of the underlying production structure without imposing any restrictive conditions on the structure, unlike the models previously discussed.

With regard to economies of scale there are several observations to be made. First, if more than one output measure is used, it is necessary to define economies of scale with respect to the particular output variable. Following the discussion in Section 5, it is possible, at least theoretically, to obtain increasing economies of scale for demand related output variables and constant returns for technical output variables. Secondly, given the type of output variable, scale economies are a function of all factor prices as well as the level of output. To illustrate, assume one output ($Q_i = Q$ for all i) and compute $\partial \text{ LnC}/\partial \text{ LnQ}$ for equation (5).

$$\frac{\partial \text{ Ln } C}{\partial \text{ Ln } Q} = \alpha + \delta \text{ Ln } Q + \sum_{j}^{n} \rho_j \text{ Ln } P_j \ . \qquad (6)$$

Noting that ∂ Lnc/∂ LnQ is the cost elasticity of output (the scale economies parameter), equation (6) shows that under this cost function, unlike the traditional cost models, scale economies is a direct function of output and all input factor prices, as well as an indirect function of the demand for factors as reflected by the parameters ρ_j. It is therefore quite possible that findings of constant economies of scale reported in the studies reviewed above (and accepted by many bus transit policy studies) are questionable because they were based on inappropriate cost models. In contrast, all three studies which did use the generalized translog cost function model indeed discovered economies of scale. Even though their data bases were quite different and they used different types of variables (in particular, different output variables), they nevertheless have reported sizeable scale economies over a considerable range of the sample observations.

7. THE DATA BASE

Thus far, the discussion of cost models used in the estimation of scale economies has focused on analytical and conceptual issues. One of the most critical elements in deriving sound and meaningful empirical results, however, is the quality of the data base, and several problems associated with the nature and quality of the data bases of the studies reviewed above were found.

Almost all of the studies discussed above utilized cross-sectional data. Cross sectional analysis implicitly assumes that transit firms are comparable, and that observations are therefore homogeneous. However, there is a great deal of variation among transit firms; they not only operate in different markets facing quite different demand environments, but they may also utilize different technologies to produce transit services. For example, the peak/base ratio is generally high in major metropolitan areas, while extra peak service is almost negligible in small and semi-rural urban areas. Since the costs of producing peak and base period services have been shown to differ significantly (Oram, 1979; Chomitz and Lave, 1981), this ratio may have an important effect on input factor demand and elasticity of factor substitution. If such factors as the peak/base ratio account for cost differences between firms, then they must be explicitly entered into the cost model if specification errors are to be avoided (Griliches, 1972).

While it is difficult to determine the extent of heterogeneity in the cross sections used in various studies, there are some indications that differences between firms are substantial. For example, most cross sections included a very wide range of firm sizes. Viton (1981) used a sample of 54 bus systems ranging in size from 88.5 million vehicle miles (VM) per year to 168 thousand VM per year. Lee and Steedman (1972) used data on British municipal bus firms which ranged in size from 1649 vehicles to 28 vehicles. Wabe and Coles (1975), also using British data, included in their sample operators ranging in size from 2000 vehicle-hours (VH) per day to 25 VH per day. Size of the firm is linked

with the type of environment in which the firm operates (Giuliano, 1981). Large firms generally operate in large metropolitan areas where the level of congestion is high, trips are shorter, more peak service is provided, etc. Transit firms may vary in a number of other ways that might affect the results of an economic analysis. The effect of using a cross sectional sample, then, is to include transit firms which may not be comparable, especially when important differences between firms are not controlled. Under these conditions, the use of cross sectional samples may violate the requirement that output units must be homogeneous when analyzing production properties.

The estimation of an economies of scale factor is also problematic with cross-section data. Economies of scale are measured at a point. In nonlinear models, both average and marginal cost change with the level of output. Recalling that C.E. = MC/AC, the question is, at what point should scale economies be measured? One conventional approach (e.g., Koshal, 1972b) is to use the sample mean as the output level for computing AC, although the average is strongly affected by extreme values. An alternative approach is to compute the economies of scale parameter for each transit firm in the sample. However, interpretation of the results is somewhat obscure. First, the cost coefficients are derived from a data base which, as explained above, may contain very heterogeneous production units. Secondly, using output levels specific to different firms produces a measure of <u>local</u> scale economies which cannot be generalized in a straightforward manner for the entire long run cost function.

The results obtained by Williams and Dalal (1981) and Viton (1981) illustrate the influence that the data base can have on the measurement of scale economies. Utilizing the same cost model (translog) and the same output variable (vehicle-miles per annum), but very different data, Williams and Dalal found an inverted U-shaped average cost curve which they interpreted as economies of scale for larger firms, while Viton found a U shaped average cost curve, from which he concluded that very large firms realize diseconomies of scale. The Williams and Dalal sample was composed of small firms, and the Viton sample was composed of small to very large firms. What seem to be the contraditory results of the two studies may be explained at least in part by these differences in the data base.

Finally, it is sometimes argued that the problem of heteroscedasticity may be encountered with a cross section of different size firms. While heteroscedasticity does not affect the consistency of the estimated coefficients, it does reduce their efficiency (Johnston, 1972). For this reason, some studies deflate the observations by a size factor so that the variance associated with the error term will be reduced.[10] For example, Wabe and Cole (1972) deflated by the number of buses, and Lee and Steedman (1972) deflated by annual vehicle miles. Two problems arise with regard to this approach. First, it is not clear that the error variance is in fact a function of size, especially when vehicle miles are

---

[10] If heteroscedasticity exists, the larger the observation the larger is its associated error term. That is, $e_i = z_i S_i$, where e_i is a random error, proportional to size S_i, and $z_i \sim N(0,\sigma^2)$. Thus, $E(e_i^2) = \sigma^2 S_i$, and the division by S_i will reduce the associated variance.

used as the proxy for size. Secondly, as pointed out by Griliches (1972), if indeed the random error term is proportional to size, then, in the simple two variable case, the correct weighted regression procedure is equivalent to minimizing $\sum_i \frac{C_i}{S} - \frac{a}{S} - b\frac{X_i}{S}^2$, where C_i equals cost of ith firm, S equals the size deflator, X_i is the independent variable and a is the intercept. In contrast, estimates using deflated data are derived by minimizing $\sum_i \frac{C_i}{S} - a - b\frac{X_i}{S}$. Unless $a = 0$, the estimated coefficients in each case will be quite different, and will thus lead to different estimates of economies of scale.

8. CONCLUSIONS

It can be concluded from this paper that the issue of economies of scale in bus transit has yet to be resolved. While the assumption of constant returns to scale simplifies the analysis of policy issues, such an assumption is not warranted. Existing studies of economies of scale in the bus transit industry present conflicting results. These differences in results stem from the definition and selection of the output measures, the use of inappropriate econometric cost models, heterogenous data samples, as well as from a variety of statistical problems, such as the standardization of cost function variables by measures of size.

The principal conclusions of this paper are as follows. First of all, the concept of economies of scale in bus transit must be carefully defined. It may be useful to distinguish between economies of density and economies of scale. The nature of transit service is such that

economies of density might reasonably be expected for a wide range of output levels, while economies of scale depend upon the type of service increment associated with different scales of operation.

Secondly, cost models based on recent developments in production theory, such as the generalized translog model, are less restrictive than other previously used cost models and thus more appropriate. These models provide greater flexibility with regard to factor substitution and price elasticity, and they do not require initial restrictive assumptions on the underlying production conditions. Moreover, because more than one output measure can be used with such models, transit output differences can be explored.

Third, proper examination of economies of scale requires more completely specified models. For example, factor price differences which affect unit costs, such as spread time penalties associated with peak services, might be explicitly entered into the cost function. Although Viton (1981) found variables representing the peak/base ratio and number of route-miles to be insignificant, other studies have found such variables to be significant (Miller, 1970; Foster, 1973). More research is necessary on the question of transit service differences and their relationship to unit costs.

In addition to a more explicit approach to service differences, the problem of heterogeneity in the data sample might further be reduced by utilizing time-series data (e.g., Berechman 1982). While other problems may be encountered with a time-series approach, such an analysis might serve to illuminate the extent to which environmental differences have affected cross section results.

Finally, economies of scale should be examined within the context of a long-run cost model if policy implications are to be derived. Thus the long-run objectives of the firm and its regulatory environment must be taken into account. In summary, then, an analysis of the cost structure of bus transit requires an approach which is both more sensitive to the actual characteristics of the industry and embedded within a sound analytical framework.

REFERENCES

Allen, R.G.D. Mathematical economics. London: MacMillan, 1957.

Berechman, Joseph. Analysis of costs, economies of scale and factor substitution in bus transport. Journal of Transport Economics and Policy, 1982.

Boyd, J. H., Asher, N. J., and Wetzler, E. S. Nontechnological innovation in urban transit, a comparison of some alternatives. Journal of Urban Economics, 5, 1978, 1-20.

Caves, Douglas W., Christensen, Laurits R., and Swanson, Joseph A. Productivity in U.S. railroads, 1951-1974. Bell Journal of Economics. 11(1), 1980, 166-181.

Cherwony, Walter, and Mundle, Subhash. Transit cost allocation model development. ASCE Transportation Engineering Journal, January 1980, 106(TE1), 31-42.

Chomitz, Kenneth M., and Lave, Charles A. Part-time labor, work rules, and transit costs. Final report prepared under contract #UMTA-CA-11-0018. Irvine, Calif.: University of California, Institute of Transportation Studies and School of Social Sciences, January 1981. (NTIS #PB 81-180556).

Coase, R.H. The marginal cost controversy. Economica, August 1946, 169-181.

Cornell, N.W. Rate-of-return regulation: protecting whom from what? Regulation, November/December 1980, 36-41.

Fielding, Gordon J., Glauthier, Roy E., and Lave, Charles A. Performance indicators for transit management. Transportation 7(4), December 1978, 365-379.

Foster, J. R. The determinants of costs and the cost function of urban bus transit 1960-1970 (Doctoral Dissertation, Syracuse University, 1973). Dissertation Abstracts International, 1974, 35, 1288-A.

Fisher, Peter, and Viton, Philip. The full cost of urban transport. Part I: Economic efficiency in bus operations: Preliminary intermodal comparisons and policy implications. (Monograph #19). Berkeley, Calif.: University of California, Institute of Urban and Regional Development, 174. (NTIS #PB 248-145).

Friedlander, A. F., and Spady, R. H. Freight transport regulation. The MIT Press, 1981.

Fuss, M. McFadden, D., and Mundlak, T. A survey of functional forms in the econometric analysis of production. In Fuss, M., and McFadden, D. eds. Production economics. Amsterdam: North-Holland Publishing Co., 1978.

Giuliano, Genevieve. The effect of environmental factors on the efficiency of public transit service. Transit Planning and Management, Transportation Research Board, 1981, 11-16.

Gold, Bela. Changing perspectives on size, scale, and returns: An interpretive survey. Journal of Economic Literature, March, 1981, 19, 5-33.

Griliches, Zvi. Cost allocation in railriad regulation. Bell Journal of Economics, 1972, 3(1), 26-41.

Harris, Robert G. Economies of traffic density in the rail freight industry. Bell Journal of Economics, 1977, 8(2), 557-564.

Johnston, J. Econometric methods. (2nd ed.) New York, N.Y.: McGraw-Hill, 1972.

Koshal, Rajindar K. Economies of scale in bus transport: Some Indian experience. Journal of Transport Economics and Policy, January 1970, 4(1), 29-36.

Koshal, Rajindar K. Economies of scale. II. Bus transport: Some United States experience. Journal of Transport Economics and Policy, May 1972, 6(2), 151-153.

Lee, N. and Steedman, I. Economies of scale in bus transport. I. Some British municipal results. Journal of Transport Economics and Policy, January 1970, 4(1), 15-28.

McGillivray, Robert, Kemp, Michael, and Beesley, Michael. Urban bus transit costing. (Working Paper #1200-72-1). Washington, D.C.: Urban Institute, September 1980.

Manski, Charles F. The zero elasticity rule for pricing a government service: A summary of findings. Bell Journal of Economics, 1979, 10(1), 211-223.

Meyer, John R., and Gomez-Ibanez, Jose A. Measurement and analysis of productivity in transportation industries. (Discussion Paper D75-6). Cambridge, Mass.: Harvard University, Dept. of City and Regional Planning, November 1975.

Miller, David R. Differences among cities, differences among firms, and costs of urban bus transport. Journal of Industrial Economics, November 1970, 19(1), 22-32.

Mohring, Herbert. The peak-load problem with increasing returns and pricing constraints. American Economic Review, September 1970, 60, 693-705.

Mohring, Herbert. Optimization and scale economies in urban bus transportation. American Economic Review, September 1972, 62, 591-604.

Mohring, Herbert. Transportation economics. Cambridge, Mass.: Ballinger, 1976, Chapter 6.

Nelson, Gary R. An econometric model of urban bus transit operations. (Paper P-863). Arlington, Va.: Institute for Defense Analysis, 1972.

Oram, R. L. Peak-period supplements: The contemporary economics of urban bus transport in the U.K. and U.S.A., Progress in Planning, 1979, vol. 2, part 2, 81-154.

Pozdena, Randall J. A methodology for selecting urban transportation projects. (Monograph No. 22). Berkeley, Calif.: University of California, Institute of Urban and Regional Development, 1975.

Tomazinis, Anthony R. Productivity, efficiency, and quality in urban transportation systems. Lexington, Mass.: D.C. Heath/Lexington, 1975.

Uzawa, H. Production functions with constant elasticities of substitution. Review of Economic Studies, 1962 291-299.

Varian, H.R. Microeconomic analysis. New York, N.Y.: W.W.Norton, 1978.

Veatch, James F. Cost and demand for urban bus transit. Unpublished doctoral dissertation, University of Illinois, 1973. (Available from Xerox University Microfilms as #74-5723).

Viton, Philip A. A translog cost function for urban bus transit. Journal of Industrial Economics, March 1981, 29(3), 287-304.

Wabe, J.S., and Coles, O.B. The short and long-run cost of bus transport in urban areas. Journal of Transport Economics and Policy, May 1975, 9(2), 127-140.

Williams, M.L., and Dalal, A. Estimation of the elasticities of factor substitution in urban bus transportation: A cost function approach. Journal of Regional Science, 1981, 21(2), 263-275.