3 Transport production and the analysis of industry structure

Sergio R. Jara-Díaz

3.1 Introduction

As transport activities mean displacements of individuals and goods in both time and space, the analysis of transport production involves the assignment of resources to generate trips among many different points in space during many different periods. As a consequence, the microeconomic analysis of transport production is far from a simple extension of the theory of the firm. In this chapter we present the underpinnings of a microeconomic theory of the transport firm, with particular emphasis on the nature of the technical relations between inputs and outputs (production or transformation function) and the use of the cost function as a tool to obtain valuable information for the design of transport policies (for example pricing, regulation)

The chapter is sequentially organised, beginning with the notion of transport production, including the definition of transport output, the role of space, the idea of operating rules, and the concept of scale, all of which are illustrated using simple cyclical systems (section 3.2). Then the cost function and its properties regarding the calculation of marginal costs, economies of scale and economies of scope, are presented and explained within the context of transport systems analysis (section 3.3). A synthesis of the empirical work using transport cost functions is offered, with special emphasis on the adequate treatment of output in its specification, and on the difficulties with the prevailing approach to analyse industry structure, including recently improved procedures to calculate scale economies correctly and a discussion on network density versus economies of scope (section 3.4). The closing section contains a synthesis and directions for research.

3.2 Transport production

In essence, the production of goods and services can be synthetically described using the concepts of inputs, outputs and technology. Inputs have to be acquired by the firm in order to be combined - within the boundaries of process-specific rules - in order to produce outputs. For a given level of outputs, the firm has to choose type and amount of inputs, as well as a subset of combination rules. All feasible input combinations define the technology.

In the case of transport, the firm has to use vehicles, terminals, rights-of-way, energy, labour, and so on, to produce movements - freight or/and passenger - from many origins to many destinations during many different periods. Thus, the output of a transport firm is a vector

$$Y = \left\{ y_{ij}^{kt} \right\} \in R^{KxNxT} \tag{3.1}$$

where each component y_{ij}^{kt} represents the flow of type k moved from origin i to destination j (O-D pair ij), within period t, for example passengers from Paris to Frankfurt during a specific weekend (K, N and T are the the number of flow types, the number of O-D pairs, and the number of time periods considered in Y, respectively).

For a given set of flows in Y, the firm has to make a number of choices: number and capacity of vehicles

(fleet size), design of the rights-of-way (location, flow capacity), design of terminals (location, loading-unloading capacity), route structure (*i.e.* how vehicles would flow on the network), vehicle frequencies, and so on. Some of these decisions involve choosing the characteristics of inputs, and some are related with their use, *i.e.* with the form in which inputs are combined to accommodate the flow vector. We will call these latter types of choices 'operating decisions'.

For a given type of transport firm (for example interurban bus) some of the decisions related with the acquisition of inputs are constrained, because of the existence of common infrastructure (for example the road system) or the rigidity of input markets (for example fleet size). On the other hand, operating decisions are generally made within the boundaries of existing inputs. As a simple example, consider an O-D system with three nodes, a single period and a single flow type.

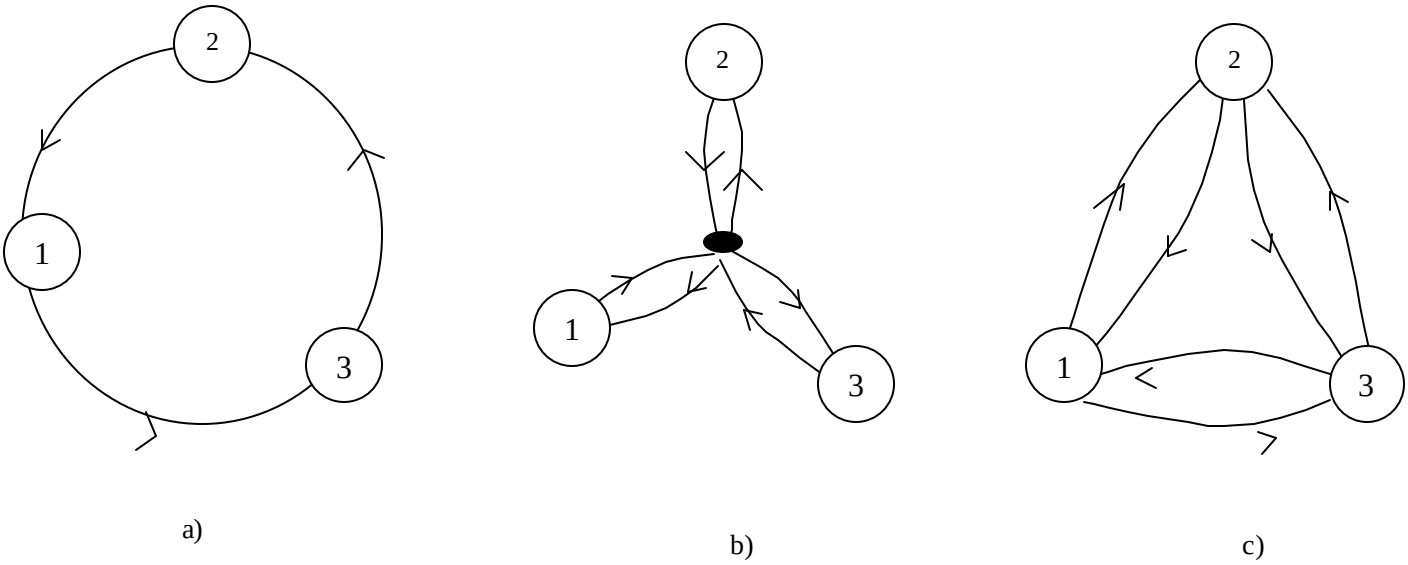


Figure 3.1
Possible route structures for a simple O-D system

For a given set of flows $\{y_{ij}\}$, the appropriate combination of inputs and operating rules would depend on many factors. If, in Figure 3.1, nodes 1, 2 and 3 represent distant cities (airports), then three possible air route structures are a), b) and c). These route structures should be analysed in parallel with aircraft size and frequency in order to make the most convenient choice. If this was either a road or a railway system, the physical structure of the road network would constrain the choice of routes and schedules. Moreover, for a given fleet size (including vehicle capacity), scheduling would be the only decision to make.

The technical relation between inputs and outputs is summarised through the concept of a transformation or production function. Let us give an example using the simplest possible case, *i.e.* a single O-D pair, single product, single period (Gálvez, 1978; Jara-Díaz, 1982b). Let Y be the flow from O to D. If k is load size, B is fleet size, $t(k)$ is travel time as a function of load size and μ is loading-unloading speed, then

$$Y \equiv \frac{Bk}{t(k) + 2\frac{k}{m} + t(0)} \tag{3.2}$$

For a given B and μ, one can find the value of k that maximises Y, k^*. It can be easily proved that k^* would be given by vehicle capacity K, provided the effect of k on travel time is small.

Therefore

$$Y \leq \frac{BK}{t(K) + 2\frac{K}{m} + t(0)} = h(B, K, m) \tag{3.3}$$

where $h(B,K, \mu)$ is the production function which gives the maximum flow for a given value and characteristics of the inputs: B, K and μ. Which combination should be chosen for a *given* value of Y, would depend on the relative prices of vehicles and loading-unloading capacity. In this simple cyclical system, the input choice, their feasible combinations and the operating rule can be clearly distinguished.

Thus, depending on the characteristics of the particular transport system, the transport firm could adjust inputs and operating rules according to the different levels of Y. This concept remains when Y is a vector. The simplest possible version of a multioutput transport firm is one serving a backhaul system with two nodes (*1* and *2*) and two flows (y_{12} and y_{21}) of a single product during a single period (Gálvez, 1978; Jara-Díaz, 1982b). Let us assume for simplicity that the firm operates the same fleet to move both flows. Then vehicle frequency in both directions is the same, and given by the maximum necessary, which in turn depends upon the relative flows; let us assume $y_{12} \geq y_{21}$. Then the technical optimum requires the vehicles in the $1 \rightarrow 2$ direction, to be fully loaded, and frequency will be given by

$$f = \frac{y_{12}}{K} \tag{3.4}$$

and the load size in the opposite direction, k_{21}, will be

$$k_{21} = \frac{y_{21}}{f} = \frac{y_{21}}{y_{12}} K \tag{3.5}$$

The fleet size needed, B, has to be equal to f times cycle time t_c which, under our simplifying assumption, is given by

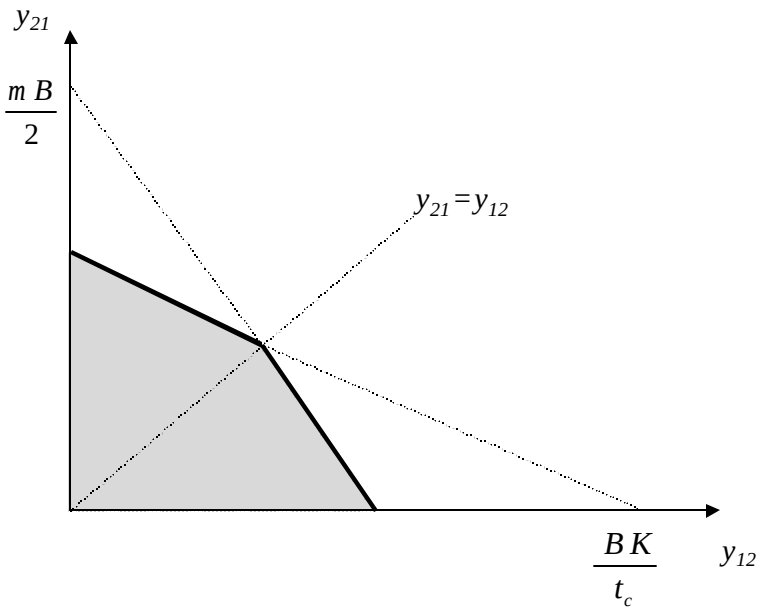
$$t_c = t_{12}(K) + \frac{2K}{m} + \frac{2}{m}\frac{y_{21}}{y_{12}} K + t_{21}(k_{21}) \tag{3.6}$$

Just for the sake of simplicity, let us see the case characterised by vehicle speed v independent of load size and potentially different route distances d_{ij} in each direction. Then using equations (3.4) and

(3.6), the equality $B=f\,t_c$ can be turned into

$$BK = y_{12}\left[\frac{d_{12}}{v} + 2\frac{K}{m} + \frac{d_{21}}{v}\right] + 2\frac{K}{m}y_{21} \tag{3.7}$$

As this is valid for $y_{12} \geq y_{21}$, and there is a symmetric expression for $y_{21} \geq y_{12}$, the general result for the technical relation among flows and inputs is



$$y_{ij} = \frac{mB}{2}\left(\frac{d_{12} + d_{21}}{2Kv} + 1\right)y_{ji}, \forall\, y_{ji} \geq y_{ij} \tag{3.8}$$

Figure 3.2
Production possibility frontier of the backhaul system

It is fairly simple to show that the graphical representation of the backhaul system in the output space looks like Figure 3.2. Equation (3.8) represents the production or transformation function of the system, and the shaded area in the figure represents all the vectors (y_{12}, y_{21}) that can be produced with a given fleet B, and capacities μ and K, but only the boundary represents optimal usage. This boundary

is the production possibility frontier, whose symmetry is derived from the assumption of load independence of speed.

Both the cyclical and backhaul systems are illustrative of the idea of technical feasibility and optimality. One of the most important conceptual points is the distinction between inputs, as fleet or loading-unloading capacities, and operating rules, as frequency, speed or vehicle load. The former is related with things that have to be acquired and the latter are ways to combine the former to produce flows. Roles and relations are clear.

In complex systems, the technical relations can not be obtained in such an explicit form as in equations (3.3) and (3.8), but they can be envisaged as a sort of 'specialised black box' which includes a number of analytical relations dealing with networks, itineraries, routes, frequencies, and so on, trying to aim at the best possible use of resources: fleet, terminals and rights-of-way. This general idea helps understanding the kernel of transport production; changes in the flow vector Y potentially induce changes in input usage as well as in route structures and operating rules in general. It may well be that some of the inputs can not be adjusted, which means that some other inputs will have to be changed in combination with different operating rules. A good example is the restructuring of routes and itineraries for a given fleet of buses facing a change in the passenger volumes in different O-D pairs.

To end this general idea of transport production, let us introduce an important technical concept that can be examined directly from the transformation function: the concept of scale economies. The relevant question is by how much can output be expanded if all inputs are expanded by the same proportion. In the single output case represented by equation (3.3), a local expansion of vehicle capacity (BK through K) and loading-unloading capacity (μ) would allow Y to be increased by the same proportion if speed was unrelated to K; note that in this example the right-of-way input is assumed to be exogenous to the firm. In the two-outputs case represented by Figure 3.2, a similar expansion of inputs moves the production possibility frontier away from the origin, but the 'how much can output be expanded' question becomes ambiguous, as nothing has been said about output combinations. If the concept of scale economies is forced to deal with proportional expansions of output, it is clear that, again, (y_{12}, y_{21}) can be expanded by the same proportion as inputs (same condition as in the earlier case).

In general, if $F(X, Y) \geq 0$ represents the transformation function (*i.e.* all technically feasible combinations of inputs and outputs) where X is the input vector and equality represents technical optimality, the (multioutput) degree of scale economies, S, is defined as the maximum proportional expansion of Y, $\lambda^S Y$, after an expansion of X by λX (Panzar and Willig, 1977). Analytically,

$$F\left(\lambda X, \lambda^S Y\right) = 0 \tag{3.9}$$

which means than in the previous examples S takes the value of one, usually called constant returns to scale. A value of S greater or smaller than one is called increasing or decreasing returns to scale respectively.

3.3 Transport cost functions: the theory

Basic definitions and properties

Technical analysis is not enough to understand the choice of inputs combination by the firm. The

question is which of the combinations in the technical frontier is the most convenient to produce a given output *Y*. The answer is given by one of the most interesting tools in the microeconomics of production: the cost function, which requires input prices to be introduced in the picture. Formally, the cost function *C(w,Y)* gives the minimum expenditure necessary to produce output *Y* at given factor prices *w*. It corresponds to the solution of

$$\min_X \sum_i w_i x_i \quad \text{subject to } F(X,Y) \geq 0 \tag{3.10}$$

The solution for each input x_i is a conditional demand function $x_i^*(w,Y)$, which represents the optimum amount of input. Then the cost function is

$$C(w,Y) \equiv \sum_i w_i x_i^*(w,Y) \tag{3.11}$$

If some inputs x_j are fixed at a level $\overline{x_j}$, then the short run cost function is defined as $C(w_v, \overline{X}, Y)$, where $\overline{X}$ is a vector containing fixed inputs and w_v is a vector-containing variable input prices. The optimisation process represented by equations (3.10) and (3.11) is exactly the same.

Out of the many properties of the cost function, five are particularly relevant for a basic analysis and discussion of production in general and of transport in particular. First, the derivative property or Shephard's lemma, which states that the derivative of the cost function *C(w, Y)* with respect to each factor input price w_i equals the cost minimising amount of $x_i(w, Y)$, that is $x_i^*(w, Y)$. Analytically,

$$\frac{\partial C(w,Y)}{\partial w_i} \equiv x_i^*(w,Y) \tag{3.12}$$

and is very helpful in estimating and interpreting a cost function. Second, the marginal cost specific to product *i*, m_i, is simply

$$m_i = \frac{\partial C(w,Y)}{\partial y_i} \tag{3.13}$$

Next, the (multioutput) degree of scale economies which has been defined on the technology, can be shown (Panzar and Willig, 1977) to be obtainable from the cost functions as

$$S = \frac{C(w,Y)}{\sum_i y_i \frac{\partial C}{\partial y_i}} = \frac{1}{\sum_i \eta_i} \tag{3.14}$$

where is the cost elasticity with respect to output *i*.

Fourthly, the degree of economies of scope relative to a subset *R*, SC_R can be calculated (Baumol et al., 1982) as

$$SC_R = \frac{1}{C(Y)}\left[C(Y_R) + C(Y_{M-R}) - C(Y)\right] \tag{3.15}$$

where Y_R represents vector Y with $y_i = 0, \forall i \notin R \subset M$, with M being the set of all products (we have suppressed w for simplicity). Thus, a positive SC_R - the existence of economies of scope - means that it is cheaper to produce Y with a single firm than to split production into two orthogonal subsets R and M-R.

Finally, a cost function is said to be subadditive for a particular output vector Y when Y can be produced more cheaply by a single firm than by any combination of smaller firms (Baumol *et. al.*, 1982, p. 170). Therefore, a cost function is subadditive if

$$\sum_i C(Y^i) \geq C(Y) \qquad \forall \left\{Y^i\right\} \Big/ \sum_i Y^i = Y \tag{3.16}$$

which is the multioutput notion of natural monopoly. Under this set of definitions and properties, it is very clear that both $S > 1$ and $SC_R > 0$ favour subadditivity, but neither guarantees its presence by itself.

Scale and scope in transport production

With product defined as in equation (3.1) and the notion of scale synthesised in equation (3.9), scale analysis in transport should be conceptually clear. It refers to the behaviour of costs as flows in all markets served by a firm expand proportionally. In order to emphasise space, let us create an example using the O-D structure depicted in Figure 3.1 relative to non-perishable cargo. Imagine that distances i-j are relatively short, that flows y_{ij} and y_{ji} are unbalanced, and that the sum of flows clockwise are approximately equal to the sum of flows counter-clockwise. It may well be that for relatively low volumes, a route structure like a), with complete vehicle cycles involving a homogeneous fleet, is the least cost answer. Imagine output expands proportionally; the firm could accommodate that expansion by increasing frequency (enlarging fleet) and/or using larger vehicles. For further expansions, the hub-and-spoke structure like a) could well become the best answer, making the hub a transfer point and involving vehicles of different sizes. It might be the case that direct services like in c) happen to be the least cost structure for individually large enough flows. If there are scale advantages in loading-unloading activities and in vehicle size, it is very likely that through appropriate scheduling and rerouting, total cost will increase less than proportionally with increases in the flow vector, at least up to a certain scale.

Regarding scope, again Figure 3.1 will prove very helpful. If the six flows are divided into subsets $\{y_{12}, y_{23}, y_{31}\}$ and $\{y_{21}, y_{13}, y_{32}\}$, very possibly the sum of the costs of assigning each subset to a different firm will be greater than that cost of moving all six flows with one firm. The case is not that clear when the partition is $\{y_{12}, y_{21}\}$,$\{y_{13}, y_{31}, y_{23}, y_{32}\}$. In general, the partition of the flow vector could be made in terms of flow type (for example passengers and freight), periods (for example weekends and weekdays) or O-D pairs, as we have done in the example. In this latter case we would talk about economies of spatial scope, when they exist.

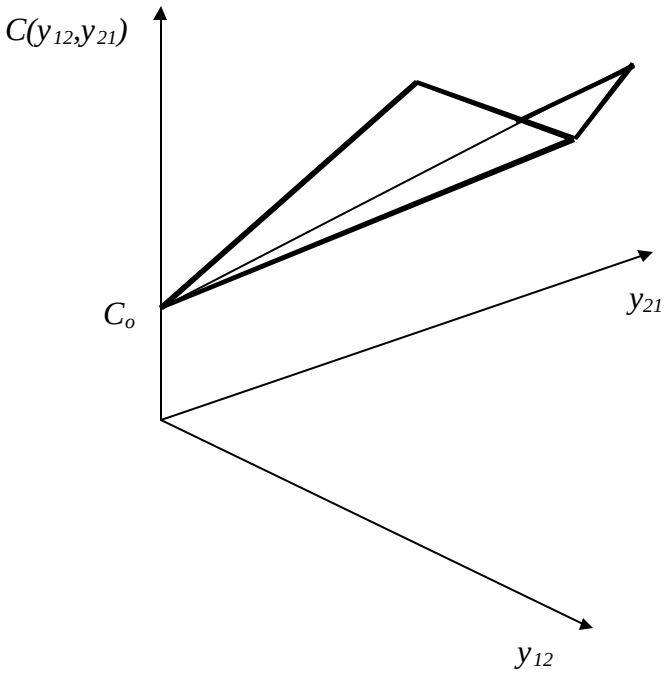


Figure 3.3
Transport cost function of the backhaul system

In order to provide a specific analytical example, let us use the simple backhaul system technically described in Section 3.2 to obtain and analyse the corresponding cost function. The system represented by equation (3.8) can be used to get the number of vehicles as a function of product, $B(y_{12}, y_{21})$. On the other hand, the number of loading-unloading sites, L, is given by $2(y_{12} + y_{21})/m$. Without affecting the conceptual analysis, we can hold d, v, k and μ constant such that inclusive prices for vehicles (P_B) and sites (P_L) can be defined, *i.e.* prices that encompass rent, labour and energy (fuel) necessary to operate one vehicle and one site respectively. Replacing all variables, the multioutput cost function is given by

$$C(y_{12},y_{21}) = C_o + y_{ij}\left[P_B\left(\frac{d_{12}+d_{21}}{vK}+\frac{2}{\mu}\right)+\frac{2P_L}{\mu}\right] + y_{ji}\frac{2}{\mu}(P_B+P_L) \qquad y_{ij} > y_{ji} \qquad (3.17)$$

where C_0 represents costs that are associated with the right-of-way.[1] This is graphically represented

in Figure 3.3.

Although equation (3.17) has been obtained using highly simplifying assumptions, it represents a fairly transparent cost function for the simplest possible multioutput transport system. Its importance becomes apparent when it is used to analyse scale, scope and aggregate output. If the degree of economies of scale is calculated using (3.14), it is quite easy to show that $S=1$ for $C_0 = 0$, which we can name the 'lorry' case, as lorries (or buses) do not pay a fixed cost for road infrastructure. On the other hand, scope analysis can be done for the only partition possible in this case (i.e assigning each of the flows in the backhaul system to different firms). After elementary calculations we get

$$C(y_{12},0) + C(0,y_{21}) - C(y_{12},y_{21}) = C_o + P_B \frac{d_{12} + d_{21}}{vK} y_{ji} \qquad \forall y_{ij} \geq y_{ji} \tag{3.18}$$

which is positive even if C_o is nil. This shows that, under the assumptions made, it is fleet utilisation what causes the existence of economies of scope. Thus, if $C_o=0$, we have constant returns (a case for competition or deregulation) and economies of scope. These latter would cause incentives for merging if two firms are operating, each in one direction. The conclusion is that, as far as costs of production are concerned, competition would be desirable, with each firm operating both markets. It is relevant to mention here that, in their pioneering work on hedonic cost functions, Spady and Friedlaender (1978) verbally explained merging among trucking firms serving different routes when a regulatory regime that had exercised restrictions on routes is dismantled. However, that phenomenon was said to reflect 'economies of density and utilisation', which can not be derived from their cost function specification. With a well-defined output, it is directly explainable as economies of (spatial) scope. This leads to the third interesting aspect that can be explored using this simple example: output aggregation.

Even up to our days, aggregates like passenger- (or ton-) kilometres *(TK)* are used as a basic or synthetic unit to describe transport output both in general and within the context of empirically estimated cost functions. Since the late seventies, its ambiguity began to be addressed, raising the issues of network shape and fleet utilisation, as described in the preceding paragraph. The simplified cost function of the backhaul system can be used to explore the adequacy of the *TK*-index as a representation of transport output.

First we have to recognise that *TK* is indeed a function of the true output as defined in (3.1). In the backhaul system,

$$TK = y_{12} d_{12} + y_{21} d_{21} \tag{3.19}$$

On the other hand, (3.17) can be used to represent the combinations of y_{12} and y_{21} that yield the same expenditure C_i. The resulting iso-cost locus can be shown in the output space as we have done in Figure 3.4, where cost increases with the distance from the origin. The popular 'output' *TK* can be shown in the same space using equation (3.19) as a straight line with a negative slope that depends on the relative value of d_{12} and d_{21}. As evident, all flow combinations within the straight line yield the same value for *TK*. We have represented as TK_o the case of $d_{12}>d_{21}$; as the corresponding line intersects many iso-cost curves, TK_o can not be associated with a single minimum cost figure. If should be noted that this ambiguity remains even if both distances were equal (as represented by TK_1). On the other hand, every pair (y_{12}, y_{21}) corresponds to a single cost value, unambiguously.

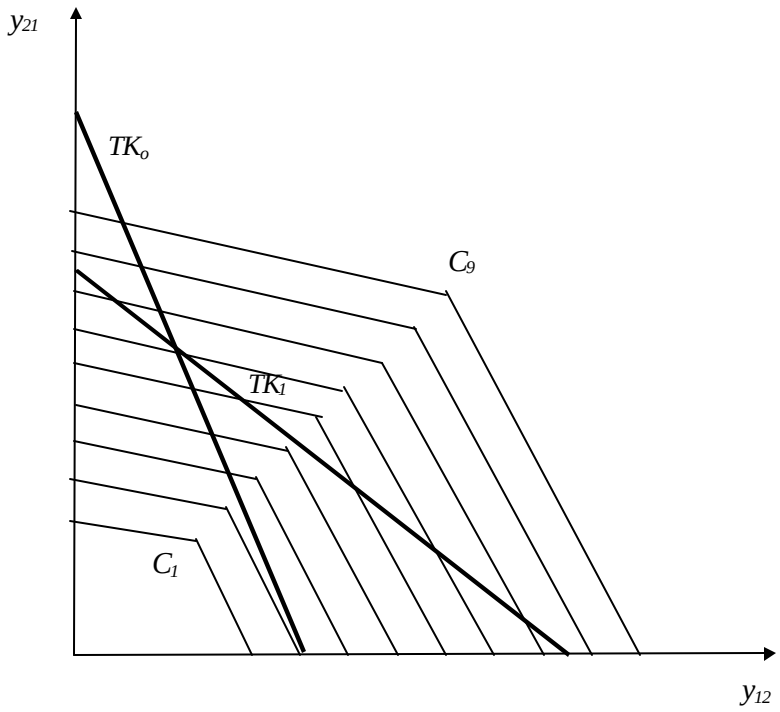


Figure 3.4
Cost ambiguity of aggregate output

The ambiguity of aggregate output is a key aspect in the analysis of industry structure in transport activities by means of a cost function. We have shown that, even in a simple system like the backhaul service developed in this section, an association between expenses C and output TK might yield completely erroneous conclusions. In terms of scale analysis, an expansion of TK by λ corresponds to many possible flow combinations, as shown by equation (3.19).

In terms of scope, the pairs $(0, y_{21})$ and $(y_{12}, 0)$ get reflected as $y_{21}\, d_{21}$ and $y_{12}\, d_{12}$ respectively when converted into TK units. Thus, scope 'turns' into scale, provoking an extremely confusing panorama when trying to obtain conclusions on industry structure.

For synthesis, transport production is a multioutput process where the concepts of scale and scope are very useful for the analysis of industry structure, provided they are properly applied. The degree of economies of scale reflects the behaviour of cost as all flows (e.g. in every O-D pair) expand proportionally. The degree of economies of scope examines the convenience of partitioning transport

services into two mutually exclusive subsets; depending on the type of partition, we will refer to economies of spatial scope, commodity scope, or time scope, whenever the cost of producing the whole set is less than the sum of costs for the partition. Diseconomies of scope reflect the opposite. Within this context, the use of aggregates to describe transport output distorts the analysis of scale and reduces (and sometimes destroys) the possibility of analysing scope. However, transport systems produce passenger and/or commodity trips over many O-D pairs, which makes reduced output description a key issue in empirical studies.

Transport output and the estimation of cost functions

Obtaining an adequate representation of either $C(w,Y)$ or $C(w,\overline{X},Y)$ - the long run and the short run cost functions respectively - is not a simple task in transport activities. As evident, the general idea is to construct a reliable statistical relation between expenses as the dependent variable, and output, input prices and fixed factors as explanatory variables. The statistical data is composed by a series of observations, each one relating production to cost. This series can be feeded by the evolution of a single transport firm in time, by the activity of many firms within a period (cross section), or by observations of many firms during many periods (pool).

The case of a time series is, conceptually speaking, the most transparent one; product (as defined in equation 3.1) is quite precise, as well as factors of production. Let us consider the case of a firm moving a single type of commodity (or passengers) among many points in space during homogeneous periods, and imagine potential observations that include services from two to six O-D pairs as depicted in Figure 3.5. If all observations were associated with an O-D system like (a), output would be a two-dimensional vector. Output would be a six-dimensional vector if *all* observations were related with movements like those represented in (c). How to represent output if observations included all three cases? The answer is straightforward: the output vector should have six components and some of them will be nil for observations including flows like in (a) or (b). Formally,

$$y = \{y_{12}, y_{21}, y_{13}, y_{31}, y_{23}, y_{32}\}$$

$$y^a = \{y^a_{12}, y^a_{21}, 0, 0, 0, 0\}$$

$$y^b = \{y^b_{12}, y^b_{21}, 0, 0, y^b_{23}, y^b_{32}\}$$

$$y^c = \{y^c_{12}, y^c_{21}, y^c_{13}, y^c_{31}, y^c_{23}, y^c_{32}\}$$

where y^n_{ij} corresponds to actual flow in O-D pair ij for observation (period) n. The case is very similar if observations correspond to transport firms operating on the same spatial setting.

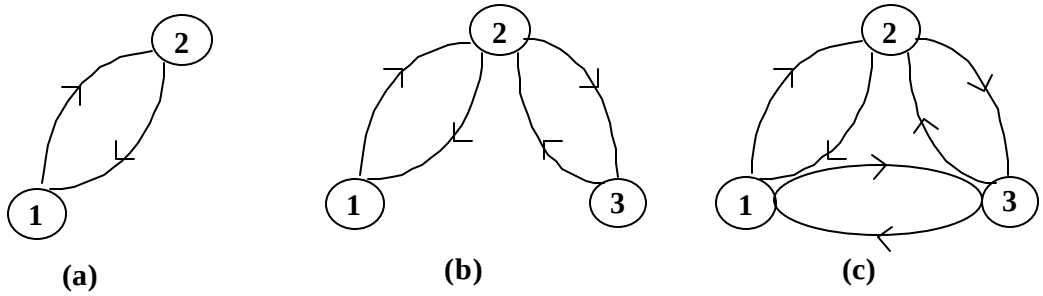
Figure 3.5
Transport output in a three nodes system

On the other hand, observations of firms serving different O-D systems correspond, in fact, to different products. This does make a difference regarding other production processes observed through a cross-section, as the optimal combination of resources to produce a given amount of an output bundle (say shoes, bags and belts) at given input prices, is likely to be equal across firms if all of them have access to the same technology. But the optimal combination of vehicles, terminals and rights-of-way (by means of routes, frequencies and load sizes) will depend upon the characteristics of the underlying physical network and the actual configuration of each O-D system. Nevertheless, it is true that an external observer (transport analyst) should be able to obtain some information regarding cost structure form observations of different transport firms performing similar services on different spatial settings (for example interurban rail, urban transit, international flights, and so on). But this requires a careful analysis in order to make the correct inferences on policy and industry structure.

Thus, transport output description within the context of the estimation of cost functions, implies a challenge at least in two dimensions. First, when output is well defined, the number of components is usually huge and certainly unmanageable in detail for statistical purposes. Second, cross-sectional observations usually involve different products. How to aggregate flow components and how to introduce product equivalency or homogeneity across different systems, are indeed problems to solve; neither, however, changes the strict definitions of scale and scope which are unambiguous with a well defined transport output.

3.4 Transport cost functions: the empirical work

Functional form



The estimation of cost functions for different transport industries has been the preferred tool for the analysis of industry structure, regulation, technical change, productivity, and so on. Within the period 1970-1997, the empirical work on transport cost functions has experienced a series of improvements. Perhaps the most evident is the use of flexible forms for the functional specification of the function, the translog form being the most popular one (see Christensen *et al.*, 1973). In order to understand analytically this form, it is useful to view first another flexible specification called the quadratic.

Conceptually, the quadratic corresponds to a second order Taylor expansion of $C(w,Y)$ around a point (w^0,Y^0), which is usually the mean of input prices $(\overline{w}_i)$ and flows $(\overline{y}_i)$ in the data set. Analytically, the stochastic expression for the quadratic cost function is

$$\begin{aligned} C(w,Y) = {} & A_0 + \sum_{i=1}^{n} A_i (w_i - \overline{w}_i) + \sum_{i=1}^{m} C_j (y_i - \overline{y}_i) \\ & + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} (w_i - \overline{w}_i)(w_j - \overline{w}_j) \\ & + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} B_{ij} (w_i - w_j)(y_j - \overline{y}_j) \\ & + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} (y_i - \overline{y}_i)(y_j - \overline{y}_j) + \varepsilon, \end{aligned} \tag{3.20}$$

where the system considers n inputs and m outputs; ε is the error tem. The translog form is analogous to equation (3.20) with $C(w, Y)$, w_i and y_i in logs. Both forms are flexible in the sense that no a priori functions are postulated either for technology or costs.

Each of these flexible forms has its own advantages. The translog facilitates the analysis of the properties corresponding to the underlying technology, *i.e.* homogeneity, separability, scale economies and non-joint production, by means of relatively simple tests on the adequate set of parameter estimates.[2] Its first order coefficients are the cost elasticities of output calculated at the mean, and their summation yields an estimate of the inverse of S as shown in equation (3.14). Further, this form makes it easy to impose homogeneity of degree one in factor prices.

On the other hand, the plain quadratic form is extremely adequate to directly obtain marginal costs evaluated at the mean of observations, C_i, and the elements of the Hessian C_{ij}, which are essential for analysing sub-additivity. In addition, equation (3.20) is well defined for zero output levels (while the translog is not); this not only represents an advantage for the estimation process, but also allows for the calculation of economies of scope, which involve output vectors with some zero components. Nevertheless, adequate transformations of output (for example Box-Cox) allow for nil values of output using the translog form as well.

One of the shortcomings of flexible forms is the fairly high number of coefficients to be estimated, which requires a substantially larger number of observations for statistical relevance of the estimated function. However, the application of Sheppard's Lemma to equation (3.20) generates as many additional equations as factor prices included, involving part of the coefficients from the original equation. Thus, inputting some usually available information (factor usage, factor expenditure or factor cost share),[3] the number of 'observations' can be multiplied, generating a system of equations, which increases the efficiency in parameter estimation. This problem is particularly relevant in transport analysis, because the usually high dimension of Y is further magnified by squared and interaction terms. This makes output aggregation an extremely important aspect: it is required to make estimation feasible in most real systems, but it should not distort the analysis thus obtaining irrelevant or misleading results.

Output aggregates

Aggregation of output over any dimension (commodity, time or space) involves loosing information associated with the transport processes generated by the system in reference, as illustrated earlier in

the backhaul system. As is evident, spatial aggregation destroys information on the geographical context of the origin-destination system in which a transport system operates. Aggregation of output over time may cause distortions when estimating cost functions if periods of distinctive mean flows are being averaged. Finally, commodity aggregation may affect cost estimation since the (minimum) cost of moving the same aggregate weight or volume will generally depend on the composition of that output.

In summary, the loss of information due to aggregation over any dimension may cause serious problems of coefficient interpretation when estimating a cost function. The bulk of the empirical work, however, has not been developed with full awareness of the problem. Most reported transport cost functions use a basic output aggregate (for example ton-kilometres or total passenger trips) together with other 'output' variables or, as called in the literature, 'output characteristics'. In other words, we do not find efforts to construct appropriate aggregates from disaggregated information on output[4]. The usual procedure is to add other aggregates that should somehow control for the ambiguity of the single output index.

Thus, seasonal and 'traffic condition' dummies are in fact trying to capture the effect of the implicit time aggregation on costs. Similarly, variables like traffic mix or insurance value try to grasp commodity aggregation. The first effort to somehow counterbalance spatial aggregation was the use of mean haul length as part of output description within a 'hedonic' treatment (Spady and Friedlaender, 1978). In the last twenty years, the literature on transport cost functions includes an enormous variety of output descriptions. Unfortunately, this has not led yet to a universally accepted form of output treatment, mainly due to an implicit reluctance to try to understand transport technology, which is a fairly complex construct as suggested at the beginning of the chapter. In order to clarify this, let us use the synthesis presented in Table 3.1, where we have included studies covering more than twenty years of evaluation.[5]

	MODES	**OUTPUT**	**ATTRIBUTE**
Berechman (1983)	bus	REV	
Berechman (1987) Berechman and Giuliano (1984)	bus	VK, PAS	
Ying (1990) Ying *et al.* (1991)	lorries	RTK	ALH,%LTL,AL,AS, I
Caves *et al.* (1984) Gillen *et al.* (1990) Windle (1991)	air	RPK Scheduled services RTK Charter services	ALH,LF,NC
Daughety *et al.* (1985) Friedlaender and Bruce (1985) Kim (1987) (*) Spady and Friedlaender (1978) Wang and Friedlaender (1984)	lorries	TK	ALH,AS,AL,%LTL, I CU(*)
Gagné (1990)	lorries	TK and N	ALH,AS,UT,IN

Caves *et al*. (1980, 1981, 1985)	rail	TK PK	ALH, ATL
Filippini and Maggi (1992) Formby et. al. (1990) Keeler and Formby (1994) Tauchen *et al*. (1983) Koshal and Koshal (1989) Braeutigam *et al*. (1980) Keaton (1990)	air bus lorries railways	SK VK LCK	LF,ALH, TD, NC
Harmatuck (1981, 1985, 1991)	lorries	NTL NLTL	ALH,AS (TL), AS (LTL)

Table 3.1
Output description in transport cost functions

TK:	ton-kilometres	%LTL:	percentage of less-than-truckload services
PK:	passenger-kilometres	AL:	average load
PAS:	passengers-trips	AS:	average shipment size
RTK:	revenue ton-kilometres	IN:	average cargo loss-and-damage insurance per dollar of cost
RPK:	revenue pax-kilometers	LF:	load factor
REV:	revenue per pax-kilometer	CU:	capacity utilisation
VK:	vehicle-kilometres	TD:	traffic density
SK:	seat-kilometres	NC:	network characteristics (for example points served, hub, etc.)
LCK:	loaded car-kilometres		
N:	number of shipments		
NTL:	number of truckload shipments		
NLTL:	number of less-than-truckload shipments		
ALH:	average length of haul (freight)		
ATL:	average trip length (passengers)		

From the Table we can verify that in addition to full aggregation of flows (for example passengers) or distance-weighted flows (for example ton-kilometres), the list of accompanying variables is varied: average load, average trip length, percentage of less-than-truckload services, number of shipments, average shipment size, and so on. It is important to note that these variables are sometimes called outputs, sometimes output characteristics, and sometimes quality dimensions. The most sophisticated variables appeared during the eighties, and they are related with network shape and size. And here we have a new source of confusion: network as infrastructure, (*i.e.* a fixed factor associated with the rights-of-way) and network as route structure, which is an endogenous, operating decision for many modes or transport systems (for example the cyclical system or the hub-and-spoke in Figure 3.1).

Scale and scope from aggregates

Whenever a cost function is specified in terms of one or more output aggregates, the analyst obtains a series of coefficients that can be given a microeconomic interpretation by simple association with properties (3.13) or (3.14). If the function is a translog-around-the-mean, first order coefficients are 'output' elasticities, and the inverse of their sum could be offered as an estimate of the degree of scale economies. This is a procedure that has been frequently applied in the literature with some qualifications. Just as an example, Caves et. al (1980) included passenger-kilometres, ton-kilometres,

average length of haul (freight) and average trip length (passengers) in their translog specification, and then calculated the degree of scale economies in various ways, always using the cost elasticities (obtained directly from the coefficients). Cost elasticities for ton and passenger-kilometres were always used, but the average-distance elasticities were left out in one of the measures of $\tilde{S}$ and included in other. The reason offered was that ton or passenger-kilometres might increase due to either more or longer trips. In fact, a coefficient of 0.5 on the elasticity of the average trip distance variables was suggested as a compromise. The thing is that an increase in the mean distance travelled necessarily requires that flows in the more distant O-D pairs have to increase more than flows in the relatively closer ones, and this violates the condition for scale analysis which relates to *proportional expansions* of output. Failure to look at *S* properly is in fact the main cause of ambiguity in this example. As said, *S* is related with proportional expansions within the vector of flows *Y*, and not directly to changes in ton- or passenger-kilometres. And when all flows increase by a factor of λ, then average distance remains constant, which means that their elasticities should be left out always.

The fact that aggregates make the calculation of *S* obscure was highlighted by Gagné (1990) and by Ying (1992). Both observed that aggregates are usually interrelated, for example ton-kilometres is equal to total flow times average length of haul, a fact that had not been taken into account when making calculations of *S*. Our view is different: we should look at the behaviour of *C(w,Y)* as the basic flow variables increase, but this operates through the aggregates. Let $\tilde{Y}$ be the vector of aggregates with components $\tilde{y}_j$ (for example ton-kilometres, total flow, less-than-truckload movements, etc). The key fact is that most of these $\tilde{y}_j$'s are implicit constructs from the components of *Y*. This is evident in the case of ton-kilometres (equation 3.19) or total flow (for example total passengers in a period) which is simply the summation over all y_i. Thus, if $\tilde{y}_j$ is an implicit function of *Y*, then the estimated $\tilde{C}(w,\tilde{Y})$ is an implicit representation $\hat{C}$ of *C(w, Y)* because (Jara-Díaz and Cortés, 1996)

$$\tilde{C}(w,\tilde{Y}) \equiv \tilde{C}\left[w,\tilde{Y}(Y)\right] \equiv \hat{C}(w,Y). \tag{3.21}$$

Then the correct calculation of an estimate $\hat{S}$ for *S* can be obtained through direct application of equation (3.14) using the y_i's as arguments. It can be easily shown (Jara-Díaz and Cortés, 1996) that

$$\hat{S} = \left[\sum_i \frac{\partial \hat{C}}{\partial y_i}\frac{C}{y_i}\right]^{-1} = \left[\sum_j \alpha_j \breve{\eta}_j\right]^{-1} \tag{3.22}$$

where $\breve{\eta}_j$ is the cost elasticity associated with aggregate *j* in $\tilde{C}$, and

$$\alpha_j = \sum_i \frac{\partial \tilde{y}_j}{\partial y_i}\frac{y_i}{\tilde{y}_j}. \tag{3.23}$$

In summary, the correct estimate is not necessarily equal to the inverse of the sum of the aggregate's elasticities, $\breve{\eta}_j$, unless the α_j's are all equal to one.

The procedure to use $\tilde{C}$ correctly, rests upon the relation between the $\tilde{y}_j$'s and the y_i's. But, according to equation (3.22), this applies to all arguments of $\tilde{C}$ which are functions of *Y*, no matter

how they are called (*i.e.* characteristics, attributes or outputs). Thus, equation (3.23) provides a test for the inclusion of any aggregate elasticity in the calculation of $\hat{S}$. Just as an example, we show here the coefficients α_j which correspond to a ton-kilometres variable (*TK*) and an average length of haul variable (*ALH*). This requires to make it explicit that

$$TK = \sum_i y_i d_i \tag{3.24}$$

and

$$ALH = \left[\sum_i y_i d_i \right] \Big/ \sum_i y_i \tag{3.25}$$

From this, one can easily show that α_{TM} is equal to one and α_{ALH} is nil. Therefore, the elasticity of *TK* should be used always in the calculation of $\hat{S}$, and the elasticity of *ALH* should never be used. These are simple cases to illustrate how to proceed with a $\tilde{C}(w, \tilde{Y})$ function. A fairly complete analysis of nearly all forms of output description and their role in the calculation of $\hat{S}$ is contained in Jara-Díaz and Cortés (1996). It is relevant to note that the α_j's are not necessarily equal to either zero or one. Just to illustrate the point, consider the case of an output index, which is in fact related with transport supply, like vehicle-kilometres. The relation between this index and the flow vector is dependent on the manner in which frequency and average load is adapted following an increase in the flows. It can be shown that a pure frequency adjustment makes α_j =1 and a pure load adjustment (which has a limit) makes α_j = 0; most cases would be in between, making $0 \le \alpha_j \le 1$.

Before moving into scope analysis, it is useful to introduce a concept that has been in the transport economics literature since the late seventies: economies of density. This concept coincides with the notion of scale economies, except for the fact that the physical network is held constant (originally, it was either total track or road length what was held constant). In the literature, the degree of economies of density ED has been always calculated from a $\tilde{C}(w, \tilde{Y}, \overline{X})$ type cost function, including some index or variable representing either the network as a fixed factor (for example track length) or a network 'characteristic' related to operations (for example number of points served). The usual procedure (which is actually nearly a definition) is to calculate ED as the inverse of the sum of all cost elasticities except those related with the network. An estimate of *S*, on the other hand, would include all elasticities. Again, failure to think in terms of *Y* prevents the true analysis from surfacing. The key issue is whether the O-D system varies or not with the variable representing network shape or operations, because if it does, the associated elasticity is not related with flow expansions but with the addition of new flows. And this is not related with scale but with scope analysis.

The best example of what we have exposed in the previous paragraph is the number of points served, PS, a variable that is usually part of the output description in the analysis of the airline industry. Is a variation of PS related with scale? If PS increases by one, the number of O-D pairs can increase up to two times PS, because the new point is a potential new destination for PS origins, and a potential new origin for PS destinations. In other words, a change in PS means a change in the number of O-D pairs which, by definition, is a matter of scope. This is an issue in the transport economics literature, that has been just recently addressed (Jara-Díaz, Cortés and Ponce, 1997).

Finally, transport cost functions using output aggregates do allow for some type of scope analysis whenever the value of the aggregates can be recovered when some of the components of Y go to zero. Trivially, this can be done when, for instance, we have passenger and freight movements distinctly represented; in this case, the presence of economies of scope between both type of services simply requires making zero each aggregate at a time and calculating SC as in (3.15).

3.5 Synthesis

In this chapter we have presented the main concepts of a microeconomic framework for the analysis of transport production and industry structure. The theory of transport production involves two key aspects: transport output, which is a vector of flows with many dimensions, and operating rules, which are the forms of input combinations to produce a flow vector. The main elements here are frequency, load size, route structure, and so on, which are operating decisions. On the other hand fleet size, vehicle capacity, loading-unloading capacity, rights-of-way design, and so on, are decisions related with input acquisition. Both types of decisions are related, but the former is taken within the boundaries of the latter.

Thus, the theory of multioutput production provides the appropriate framework for the study of transport industries, where the analysis of both scale *and* scope economies are necessary for an assessment of the optimal industry structure. This is done through the estimation of transport cost functions. However, we have shown that attempts to simplify matters by using output aggregates introduces a non-negligible degree of ambiguity. Economies of scale are clearly defined on the original output description, as the concept examines the behaviour of costs as all flows in all O-D pairs expand by the same proportion, but this clean interpretation is darkened by aggregation. Since output is usually a vector of huge dimensions, the empirical literature shows a variety of aggregate output indices that, placed in groups of three or four, are used for the estimation of cost functions in an effort to capture the complexity of transport services. We have summarised here a method to calculate correctly the degree of scale economies from such transport cost functions, which is based upon the recognition of aggregates as constructs from the original flow components.

Economies of scope are difficult to analyse from the transport cost functions reported in the literature, unless distinct aggregate output variables are used for different movement types (for example passenger-km and ton-km). The main cause of the problem, as has been explained in Section 3.4, is that making zero some of the flows has an unknown impact on the value of each aggregate. This further complicates when the transport cost function includes 'network' variables (which *are* important in an aggregate analysis), because their variation implies a variation in the number of O-D pairs; thus, scope becomes somehow related with scalar variables. This is indeed a topic for further research.

Acknowledgements. This research was partially funded by Fondecyt, Chile. The collaboration of C. Cortés is appreciated.

References

Baumol, W.J., J.C., Panzar and R.D. Willig (1982), *Contestable markets and the theory of industry structure*, New York: Harcourt Brace Jovanovich.

Berechman, J. (1983), 'Costs, economies of scale and factor demand in bus transport', *Journal of Transport Economics and Policy*, 17, 7-24.

Berechman, J. (1987), 'Cost structure and production technology in transit', *Regional Science and Urban Economics*, 17, 519-534.

Berechman, J. and G. Giuliano (1984), 'Analysis of the cost structure of an urban bus transit property', *Regional Science and Urban Economics*, 17, 519-534.

Braeutigam, R.R., A.F. Daughety and M.A. Turnquist (1980), 'The estimation of a hybrid cost function for a railroad firm', *Review of Economics and Statistics*, 62, 394-403.

Caves, D.W., L.R. Christensen and J.A. Swanson (1980), 'Productivity in U.S. railroads, 1951-1974', *Bell Journal of Economics*, 11, 166-181.

Caves, D.W., L.R. Christensen and J.A. Swanson (1981), 'Productivity growth, scale economies, and capacity utilisation in U.S. railroads, 1955-74', *American Economic Review*, 71, 994-1002.

Caves, D.W., L.R. Christensen, and M.W. Tretheway (1984), 'Economies of density versus economies of scale: why trunk and local service airline costs differ', *Rand Journal of Economics*, 15, 471-489.

Caves, D.W., L.R. Christensen, M.W. Tretheway and R.J. Windle (1985), 'Network effects and the measurement of returns to scale and density for U.S. railroads', in A.F. Daughety, ed., *Analytical Studies in Transport Economics*, Cambridge: Cambridge University Press, 97-120.

Christensen, L.R, D.W. Jorgenson and L.J. Lau (1973), 'Transcendental logarithmic production frontiers', *Review of Economics and Statistics*, 55, 28-45.

Daughety, A.F. ed. (1985), *Analytical Studies in Transport Economics*, Cambridge: Cambridge University Press.

Daughety, A.F., F.D. Nelson and W.R. Vigdor (1985), 'An econometric analysis of the cost and production structure of the trucking industry', in A.F. Daughety, ed., *Analytical Studies in Transport Economics*, Cambridge: Cambridge University Press, 65-95.

Filippini, M. and R. Maggi (1992), 'The cost structure of the Swiss private railways', *International Journal of Transport Economics*, 19, 307-327.

Formby, J.P., P.D. Thistle and J.P. Keeler (1990), 'Costs under regulation and deregulation: the case of US passenger airlines', *Economic Record*, 66, 308-321.

Friedlaender, A.F. and S.S. Bruce (1985), 'Augmentation effects and technical change in the regulated trucking industry, 1974-1979', in A.F. Daughety, ed., *Analytical Studies in Transport Economics*, Cambridge: Cambridge University Press, 29-63.

Gagné, R. (1990), 'On the relevant elasticity estimates for cost structure analyses of the trucking industry', *Review of Economics and Statistics*, 72, 160-164.

Gálvez, T. (1978), *Análisis de operaciones en Sistemas de Transporte* (Operations analysis in transport systems), Publicación ST-INV/04/78, Santiago de Chile: Universidad de Chile, Departamento de Obras Civiles.

Gillen, D.W., T.H. Oum and M.W. Tretheway (1990), 'Airline cost structure and policy implications', *Journal of Transport Economics and Policy*, 24, 9-34.

Harmatuck, D.J. (1981), 'A motor carrier joint cost function', *Journal of Transport Economics and Policy*, 21, 135-153.

Harmatuck, D.J. (1985), 'Short run motor carrier cost function for five large common carriers', *Logistics and Transportation Review*, 21, 217-237.

Harmatuck, D.J., (1991), 'Economies of scale and scope in the motor carrier industry', *Journal of Transport Economics and Policy*, 25, 135-151.

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

Jara-Díaz, S.R. (1982b), 'Transportation product, transportation function and cost functions', *Transportation Science*, 16, 522-539.

Jara-Díaz, S.R., P. Donoso and J. Araneda (1991), 'Best partial flow aggregation in transportation cost functions', *Transportation Research B*, 25, 329-339.

Jara-Díaz, S.R., P. Donoso and J. Araneda (1992), 'Estimation of marginal transport costs using the flow aggregation function approach', *Journal of Transport Economics and Policy*, 26, 35-48.

Jara-Díaz, S. and C. Cortés (1996), 'On the calculation of scale economies from transport cost functions', *Journal of Transport Economics and Policy*, 30, 157-170.

Jara-Díaz, S., C. Cortés and F. Ponce (1997) Número de puntos servidos y economías de diversidad espacial en funciones de costo en transporte aéreo. (Number of points served and economies of spatial scope in air transport cost functions*), Actas del Octavo Congreso Chileno de Ingeniería de Transporte*, Santiago, Chile: Universidad Católica de Chile, 133-145.

Keaton, M.H. (1990), 'Economies of density and service levels on U.S. Railroads: an experimental analysis', *Logistics and Transportation Review*, 26, 211-227.

Keeler, J. and F. Formby (1994), 'Cost economies and consolidation in the U.S. airline industry', *International Journal of Transport Economics*, 21, 21-45.

Kim, M. (1987), 'Multilateral relative efficiency levels in regional Canadian trucking', *Logistics and Transportation Review*, 23, 155-173.

Koshal, R.K. and M. Koshal (1989), 'Economies of scale of state road transport industry in India', *International Journal of Transport Economics*, 16, 165-173.

Oum, T.O and W.G. Waters II (1996), 'A survey of recent developments in transportation cost function research', *Logistics and Transportation Review*, 32, 423-463.

Panzar, J.C and R.D Willig (1977), 'Economies of scale in multioutput production', *Quarterly journal of Economics 91*, August: 481-493.

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

Tauchen, H., F.D. Fravel and G. Gilbert (1983), 'Cost structure in the intercity bus industry', *Journal of Transport Economics and Policy*, 17, 25-47.

Wang Chiang, S.J. and A.F. Friedlaender (1984), 'Output aggregation, network effects, and the measurement of trucking technology', *Review of Economics and Statistics*, 66, 267-276.

Windle, R.J. (1991), 'The world's airlines', *Journal of Transport Economics and Policy*, 25, 31-49.

Ying, J.S. (1990), 'The inefficiency of regulating a competitive industry: productivity gains in trucking following reform', *Review of Economics and Statistics*, 72, 191-201.

Ying, J.S. (1992), 'On calculating cost elasticities', *Logistics and Transportation Review*, 28, 231-235.

Ying, J.S. and T.E. Keeler (1991), ' Pricing in a deregulated environment: the motor carrier experience', *Rand Journal of Economics*, 22, 264-273.

Notes

1. 1.A complete analytical derivation of cost functions for both the simple cyclical system (equation 3.7) and the backhaul system (equation 3.8) can be found in Jara-Díaz (1982b).

2. For a condensed overview of the technical analysis based upon the coefficients obtained from the translog specification of $C(w, Y)$, see Spady and Friedlaender (1978).

3. Sheppard's Lemma states that $\frac{\partial C(w,Y)}{\partial w_i} = X_i$; this can be manipulated to obtain either $w_i \frac{\partial C}{\partial w_i}$ (factor expenditure) or $\frac{w_i}{C}\frac{\partial C}{\partial w_i}$ (factor share). This third form is particularly appropriate when using the translog form.

4. Possible exceptions are two pieces on output aggregation published by the author (Jara-Díaz, Donoso and Araneda, 1991, 1992).

5. Note that this is not intended as a review of techniques and results. The reader might want to look at two fairly complete studies: the period 1970-1980 is reviewed in detail in Jara-Díaz, 1982a; the remainder is analyzed by Oum and Waters (1996).