PERGAMON

Transportation Research Part E 39 (2003) 271–288

TRANSPORTATION RESEARCH PART E

www.elsevier.com/locate/tre

Transport cost functions, network expansion and economies of scope

Sergio R. Jara-Díaz *, Leonardo J. Basso

Department of Civil Engineering, Universidad de Chile, Casilla 228-3, Santiago, Chile

Abstract

This paper shows technically that economies of transport network expansion should be viewed through the concept of economies of scope rather than through the concept of economies of scale. The basic technological dimensions that are specific to transport production are identified. The framework is used to derive explicit cost functions for two and three nodes systems in order to show how the potential technical advantages of network expansions (new nodes served) are transferred into transport cost functions. These advantages are shown to translate into economies of scope that can exist even under constant returns to scale.

Keywords: Transport cost functions; Network effects; Industry structure

1. Introduction

The product of a transport firm is a vector of flows of persons and goods, moved during a number of periods and among many origins and destinations (OD pairs) in space (Jara-Díaz, 1982a). This detailed vectorial description of product has not been used in the estimation of transport cost functions although "treating the movement of each commodity from each origin to each destination as a separate product would be desirable. There would be so many outputs, however, that estimating a cost function would be impossible" (Braeutigam, 1999, p. 68). This has made the use of aggregate descriptions necessary, e.g. ton-kilometers, number of shipments or seat-kilometers, each one synthesizing one or more dimensions of the actual product. Other

* Corresponding author. Tel.: +56-2-678-4380; fax: +56-2-689-4206.
E-mail addresses: jaradiaz@cec.uchile.cl (S.R. Jara-Díaz), lbasso@cec.uchile.cl (L.J. Basso).

1366-5545/03/$ - see front matter
doi:10.1016/S1366-5545(03)00002-4

attributes have been considered as well in the estimation of transport cost functions, as is the case of average distance or load factor, which somehow try to capture output heterogeneity.

Econometric cost functions estimated using aggregated output have been used mainly to analyze industry structure. As stated by Oum and Waters (1996), the first purpose of the empirical measurement of transportation cost functions is the study of the "cost structure of the industry, notably whether or not there are economies of scale. These are important for assessing the feasibility of competition between firms of different size, and the long run equilibrium industrial organization of an industry" (p. 425). Thus, estimated cost functions are aimed at understanding for example, how many transport firms will serve a certain region. Will one firm cover the whole network? Would it be less costly to have a few, each one serving part of the network?

When analyzing industry structure, the degree of economies of scale is probably the most important concept that emerges as the center of this type of analysis. In transport, the use of aggregates—that normally obviates the spatial dimension of product—has provoked the need to make a distinction between economies of *scale* and economies of *density*, which have been associated to a varying or constant network size respectively. The usual approach is based upon an estimated cost function $\widetilde{C}(\widetilde{Y};N)$ where $\widetilde{Y}$ is a vector of aggregated product descriptions (including the so-called attributes) and N is a variable representing the network (factor prices are suppressed for simplicity). Returns to density (RTD) and returns to scale (RTS) are defined as

$$\mathrm{RTD} = \frac{1}{\sum_j \tilde{\eta}_j} \tag{1}$$

$$\mathrm{RTS} = \frac{1}{\sum_j \tilde{\eta}_j + \eta_N} \tag{2}$$

where $\tilde{\eta}_j$ is the elasticity of $\widetilde{C}(\widetilde{Y};N)$ with respect to aggregate product j and η_N is the elasticity with respect to N, fulfilling $\mathrm{RTS} < \mathrm{RTD}$ since the network elasticity is positive (as obtained in all empirical studies).

According to Oum and Waters (1996, p. 429), "RTD is referred to the impact on average cost of expanding all traffic, holding network size constant, whereas RTS refers to the impact on average cost of equi-proportionate increases in traffic and network size". From this, the prevailing interpretation on industry structure in the literature is simple. Increasing returns to scale ($\mathrm{RTS} > 1$) suggest that both product and network size should be increased because serving larger networks would diminish average cost. Constant returns to scale together with increasing returns to density ($\mathrm{RTD} > 1$) would indicate that traffic should be increased keeping network constant. This apparently straightforward analysis has an evident limitation, though: as $\mathrm{RTS} < \mathrm{RTD}$ a firm that has both optimal density and optimal network size cannot be described. Moreover, if network size was optimal ($\mathrm{RTS} = 1$) the firm *must* exhibit increasing returns to density. This has generated ad hoc explanations for the observed trend to expand networks in some industries [1] and

[1] For example in the case of air transportation, Brueckner and Spiller (1994) state that "the growth of networks can be understood as an attempt to exploit economies of traffic density, under which the marginal cost of carrying an extra passenger on a non-stop route falls as traffic on the route rises". Note that under this interpretation an optimal network size could never be reached.

a growing controversy regarding what should be measured and what is measured from a cost function in order to develop a meaningful analysis of industry structure (see for example Gagné, 1990; Ying, 1992; Xu et al., 1994; Jara-Díaz and Cortés, 1996; Oum and Zhang, 1997; Jara-Díaz et al., 2001). In our opinion, more effort has been devoted to the development of sophisticated econometric tools than to the understanding of transport technology, which is, after all, what lies behind the properties of a cost function. From this viewpoint, the spatial dimension of product has been particularly neglected in the literature, in spite of the discussions and observations regarding this aspect by authors like Spady (1985), Daughety (1985) or Antoniou (1991).

The objective of this paper is to identify unambiguously the role of the spatial dimension of product in the analysis of the transport industry structure, notably within the context of increasing networks. This is done by means of the analysis of the links between disaggregated product, transport technology and the cost function within the context of varying network size in simple systems whose technology can be reproduced in some detail. The technical description of transport processes in simple systems has been shown to be very useful to clarify some points regarding industry structure, scale and scope. For example, Jara-Díaz (1982b) described the technology and derived the corresponding cost function for a cyclical back-haul system (i.e. two OD pairs) and was able to identify precisely the role of scale and scope within the context of a (fixed) transport network, showing the ambiguity of the most popular aggregate description of product (ton-kilometers). Most important, the technical analysis of transport production for such a simple network was enough to provide an explanation for the incentives to merge between firms serving different portions of a network even under the presence of constant returns to scale, namely the existence of economies of spatial scope. This basic system, however, is not enough to illustrate the effects of a network expansion.

In this paper we identify and define the elements that are essential to characterize the spatial aspects of transport production (decisions of the firm). To do this, we use the two nodes system as a basis to analyze the economics of a network expansion through the examination of both technology and cost function of a three nodes system (six OD pairs). This seemingly simple departure introduces a new dimension to the technical decisions of the firm, namely the choice of a route structure to serve a given set of OD flows. With these tools, the analysis of proportional flow expansions (scale) and the addition of new flows (scope) through network expansions can be performed without ambiguity both technically (efficient frontiers) and economically (cost functions). On this basis, the role of the spatial dimension of product and the so-called network effects flow unambiguously within the context of the study of the transport industry structure. Note that this technical analysis is not done with the intention to recommend the best way to operate such systems, but to understand the relations between production in a transport network and the issues in the analysis of industry structure as viewed with the aim of transport cost functions. The main goal is to obtain elements to estimate, interpret and use empirically estimated transport cost functions, which obviously represent fairly complex networks.

In the next section the decisions of a transport firm are identified with emphasis on the spatially related elements. Then the analytics of production in a two nodes system is summarized in section three in order to set the basis for the expansion to a three nodes system (Section 4), where the potential of some aggregates actually used in the literature on cost functions is highlighted. The analytically obtained cost functions are used in Section 5 to show how the spatial decisions (i.e. how to use the network) translate into the calculation of economies of scale and scope, which are

the main determinants of industry structure. Finally, the challenge on what to do with estimated cost functions for a proper study of industrial organization is synthesized.

2. Decisions of a cost minimising transport firm

The microeconomic description of production processes rests upon the concepts of inputs, outputs and technological feasibility. In the case of transport, a firm produces flows of different things between different origins and destinations along different periods. Thus, transport product is a vector (Jara-Díaz, 1982a) $Y = \{y_{ij}^{kt}\}$ where y_{ij}^{kt} is the amount of flow type k (e.g. persons or goods), between origin i and destination j, during period t. This way, movements of passengers between Santiago and Temuco during the last week of the year and the movement of fruit between Santiago and New York during a week in July, are different products (even if we said passengers instead of fruit). In this paper we will concentrate on the spatial dimension of product, namely its OD structure, as this is the main distinctive feature of transport processes.

For a certain level of production Y, a transport firm has to make decisions regarding quantity and characteristics of inputs (e.g. number of vehicles, number of loading–unloading sites and their respective capacities) and operating rules (e.g. speeds, frequencies, load sizes). Because transport production takes place on a network, a transport firm has to decide, as well, *a service structure*—i.e. the generic way in which vehicles will visit the nodes to produce the flows—and a *link sequence*. These two endogenous decisions define a *route structure*, which has to be chosen using exogenous spatial information, namely the *OD structure* of demand (defined by the vector Y), the *location of the nodes* and the *physical network*. Note that the need to make a decision on a route structure is, finally, a consequence of the spatial dimension of product. Note also that this type of decision is essentially a discrete one. Let us illustrate these new important elements through an example with three nodes and six OD pairs as shown in Fig. 1a, with a physical network as the one represented in Fig. 1b.

Three possible service structures are shown in Fig. 2 (Jara-Díaz, 2000). Structure (a) corresponds to a general cyclical system (Gálvez, 1978), [2]structure (b) corresponds to three simple cyclical systems (direct service) and structure (c), where a distribution node is created, is known as *hub-and-spoke* and is very common in air transport (note that *hub* H can or cannot coincide with an origin or destination node). Regarding vehicle assignment to fleets, which is part of the service structure, there is no choice but one fleet (one frequency) in case (a), three fleets in case (b) and one, two (with three alternatives) or three fleets in case (c). If a cyclical system counter-clockwise like the one in Fig. 2a was chosen, a possible route structure could be the one shown in Fig. 3.

As stated earlier, the decisions of a transport firm are three: quantity and characteristics of the inputs, operating rules and route structure. Given the discrete nature of this latter decision, the underlying cost minimizing process can be seen as a sequence with two stages. First, *for a given route structure* the firm optimizes inputs and operating rules. After establishing the production possibility frontier (technical optimality) input prices are considered and expenses are minimized. A *conditional cost function*, that gives the minimum cost necessary to produce a given output Y for

[2] Obviously, vehicles could circulate clockwise as well.

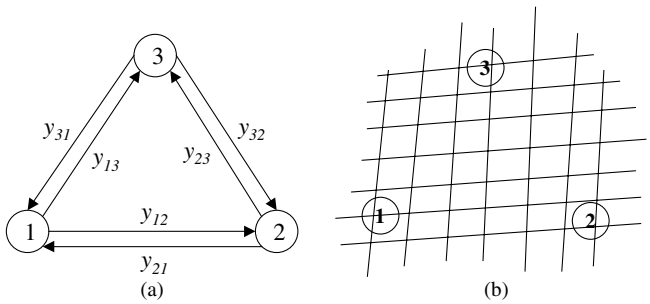

Fig. 1. OD structure and physical network.

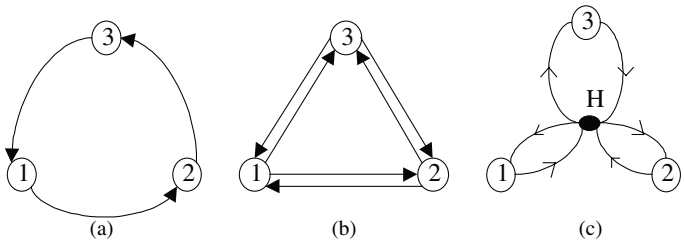

Fig. 2. Service structures.

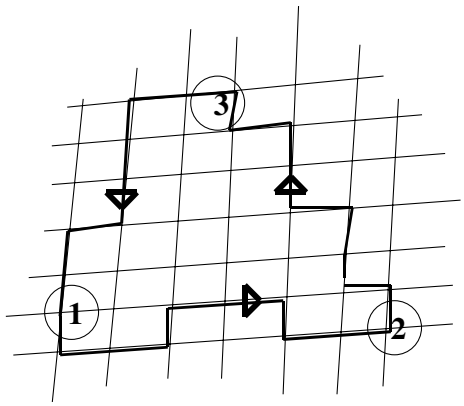

Fig. 3. Route structure.

given output prices and a known route structure, is obtained. In the second stage these conditional cost functions (corresponding to alternative route structures) are compared and the global cost function can be obtained by choosing the cost minimizing route structure.

In the next sections we apply these concepts and the approach to simple systems with two and three nodes, following the two stages optimization process. We will show that varying the exogenous information, i.e. demand or network topology, both the optimal route structure and the

associated global cost function may vary as well. The process from technology to costs will prove useful for the specification of aggregate product as well.

3. The two nodes system—synthesis [3]

The simplest possible version of a multioutput transport firm is one serving a back-haul 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). 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 $y_{12} \geqslant y_{21}$. Then the technical optimum requires the vehicles in the $1 \rightarrow 2$ direction to be fully loaded, which means that the load size in this direction, k_{12} equals the capacity K of the vehicles. Frequency will be given by

$$f = \frac{y_{12}}{k_{12}} = \frac{y_{12}}{K} \tag{3}$$

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

$$k_{21} = \frac{y_{21}}{f} = \frac{y_{21}}{y_{12}} K \tag{4}$$

If loading and unloading are sequentially done at a rate μ, if vehicle speed is v and d_{12} and d_{21} are the distances that have to be traveled between the nodes in each direction, then cycle time t_c of a vehicle is given by

$$t_c = \frac{d_{12}}{v} + \frac{2K}{\mu} + \frac{2k_{21}}{\mu} + \frac{d_{21}}{v} \tag{5}$$

The required fleet size B is obtained as f times t_c. From Eqs. (3)–(5)

$$BK = y_{12}\left[\frac{d_{12}}{v} + \frac{2K}{\mu} + \frac{d_{21}}{v}\right] + \frac{2K}{\mu} y_{21}, \quad \text{with } y_{12} > y_{21} \tag{6}$$

As this is valid only when $y_{12} > y_{21}$ and there is a symmetrical expression for the other case, a general version of Eq. (6) can be written. After some manipulations we get

$$y_{ji} = \frac{\mu B}{2} - \left[\left(\frac{d_{12} + d_{21}}{v}\right)\frac{\mu}{2K} + 1\right] y_{ij} \quad \text{with } y_{ij} > y_{ji}, \;\; i, j = 1, 2, \;\; i \neq j. \tag{7}$$

Eq. (7) represents all vectors (y_{12}, y_{21}) that can be produced efficiently with B vehicles of capacity K, circulating at a speed v and using loading/unloading sites of capacity μ (Jara-Díaz, 2000). [4] Under this frontier we find all combinations (y_{12}, y_{21}) that are technically feasible.

To obtain total expenses in the production of a given vector Y, input prices have to be considered. Let g be vehicle fuel consumption per kilometer, ε and θ the number of men required to

[3] The reader might want to see Jara-Díaz (1982b) for a fully explained first version of this approach.

[4] Note that possible road congestion can be introduced in this technical analysis by making speed dependant on frequency. The uncongested case has been preferred for simplicity only.

operate a vehicle and a loading/unloading site respectively, w is the wage rate, P_g is fuel price, P_K and P_μ are price per hour of a vehicle of capacity K and a loading/unloading site of capacity μ respectively (consider these either as rental prices or depreciation). Let g be independent of load size and speed, let ε and θ be independent of K and μ respectively, and consider the case of $y_{12} > y_{21}$. With these assumptions and variable definitions, vehicle expenses per hour including rent and operation (labor and fuel) are given by

$$P_{\text{veh}} = P_K + w \cdot \varepsilon + P_g \cdot g \cdot \frac{d_{12} + d_{21}}{B \cdot K} y_{12} \tag{8}$$

Regarding loading/unloading sites, the expense per hour P_S and the number of sites needed NS are given, respectively, by

$$P_S = P_\mu + w\theta \tag{9}$$

$$\text{NS} = \frac{2(y_{12} + y_{21})}{\mu} \tag{10}$$

Next, total expenses G can be calculated as the sum of the payments for the right-of-way (C_0), for the vehicles (B times P_{veh}), and for the loading/unloading sites (NS times P_S). Thus, from Eqs. (6), (8)–(10) we get

$$G_{K,\mu,v}(y_{12}, y_{21}) = C_0 + [P_K + w\varepsilon] \cdot B(K, \mu, v, Y) + \frac{P_g \cdot g \cdot (d_{12} + d_{21})}{K} y_{12} + \frac{2 \cdot P_S \cdot (y_{12} + y_{21})}{\mu} \tag{11}$$

where B is given by Eq. (6). To complete the first stage of the cost minimizing process, G should be minimized with respect to K, μ and v. To simplify matters, we will assume that these variables are fixed. This means that vehicle and sites are available in a given size only, and speed is exogenously determined by technical or legal facts. Thus, the optimal fleet size B^* for a given product Y is directly given by Eq. (6), which replaced in Eq. (11) yields the cost function

$$C(y_{12}, y_{21}) = C_0 + y_{12}(d_{12} + d_{21}) \cdot \Lambda + (y_{12} + y_{21}) \cdot \Omega \tag{12}$$

with

$$\Lambda = \left[\frac{P_K + w\varepsilon}{vK} + \frac{P_g g}{K} \right] \quad \text{and} \quad \Omega = \left[\frac{2}{\mu}(P_K + w\varepsilon) + \frac{2P_S}{\mu} \right]$$

with a symmetric expression for $y_{21} > y_{12}$.

It is interesting to note that Eq. (12) includes a flow-distance term and a pure flow term. Following Jara-Díaz (1982b), the latter captures those expenses that occur while product is not in motion, i.e. those due to terminal operations (as evident through Ω), while the flow-distance term captures route expenses (evidently reflected by Λ). This allows us to extract two lessons regarding output description in applied work. First, if only distance related measures are used in the specification of a cost function (for example ton-kilometers and average length of haul), then the real costs of transport production may not be adequately captured. We will highlight in the following sections the importance of this. Second, note that the flow-distance term is, in fact, the *capacity* of the transport system, as only the largest flow appears—i.e. the one that equals frequency times vehicle capacity—and is multiplied by the total distance traveled by each vehicle.

This means that if product is passenger flows, for example, that term will be the total seat-kilometers "produced" by the firm. This provides a justification for the use of such aggregates in the literature on transport cost functions (see, for instance, Keeler and Formby, 1994): once the firm has optimized operations such that the flows are served at a minimum cost, route expenses are directly related with total transport capacity.

As explained, the two nodes system synthesized here is insufficient to show the two stages optimization process completely. The second stage, namely the comparison between cost functions that are conditional on route structure, is not applicable to this single route structure case. However, this case will prove very useful for the explanation of the three nodes system, where a choice on route structure indeed exists, and for the comparison of costs after a network expansion.

4. The three nodes system

4.1. Conditional cost functions

Let us consider a six flows OD structure in a system with three nodes (see Fig. 1) connected by three links of length d_{ij}. Note that for this simple physical network, the choice of a service structure is conveniently coincidental with the choice of a route structure because there is no decision on link sequence. We will keep the simplifying assumptions of the previous section, namely the sequential loading/unloading procedure and known values of K, μ and v. As explained earlier, our goal in the first stage is to find cost functions that are conditional on the route structure. Let us begin with a general cyclical counter-clockwise structure (Fig. 2a), which implies the use of one fleet only. In this case, vehicle load size on each segment of the network k_{12}, k_{23} and k_{31} are defined by

$$k_{12} = \frac{y_{12} + y_{13} + y_{32}}{f}, \quad k_{23} = \frac{y_{23} + y_{21} + y_{13}}{f}, \quad k_{31} = \frac{y_{31} + y_{32} + y_{21}}{f} \tag{13}$$

Assume arbitrarily that link 1–2 carries the largest load. This can be shown to be equivalent to $y_{12} + y_{13} > y_{21} + y_{31}$ and $y_{12} + y_{32} > y_{21} + y_{23}$. Efficiency implies fully loaded vehicles on that segment, such that $k_{12} = K$. Thus, frequency will be

$$f = \frac{y_{12} + y_{13} + y_{32}}{K} \tag{14}$$

which trivially determines load size on the other two segments by replacing (14) into (13). Then, cycle time is given by

$$t_c = \frac{d_{12} + d_{23} + d_{31}}{v} + \frac{2K}{\mu} + \frac{2K}{\mu} \cdot \frac{(y_{21} + y_{23} + y_{31})}{(y_{12} + y_{13} + y_{32})} \tag{15}$$

The production possibility frontier of this route structure is obtained recalling that fleet size B is given by cycle time (Eq. (15)) times frequency (Eq. (14)), from which

$$BK = (y_{12} + y_{13} + y_{32}) \cdot \left[\frac{d_{12} + d_{23} + d_{31}}{v} + \frac{2K}{\mu} \right] + \frac{2K}{\mu} (y_{21} + y_{23} + y_{31}) \tag{16}$$

with

$$y_{12} + y_{13} > y_{21} + y_{31} \quad \text{and} \quad y_{12} + y_{32} > y_{21} + y_{23}$$

Given K, μ and v, Eq. (16) is directly $B^*(Y)$. Just as we did with the two nodes system, the conditional cost function is obtained calculating expenses per vehicle-hour times B^* and adding loading/unloading sites expenses plus the right-of-way cost. Doing this and after some re-ordering, we get

$$C_{CG}(Y) = C_0 + (y_{12} + y_{13} + y_{32}) \cdot (d_{12} + d_{23} + d_{31}) \cdot \Lambda + (y_{12} + y_{13} + y_{32} + y_{21} + y_{31} + y_{23}) \cdot \Omega \tag{17}$$

with

$$y_{12} + y_{13} > y_{21} + y_{31} \quad \text{and} \quad y_{12} + y_{32} > y_{21} + y_{23}$$

that is the *cost function conditional on a cyclical route structure*, with Λ and Ω defined as in Eq. (12).

The similarity between this conditional cost function and the one obtained for the two nodes system (Eq. (12)) is evident. First, we obtain again the distance-flow term with the same meaning, namely the capacity of the system, as the flows involved are those that define frequency. Second, the pure flow term is generated by loading/unloading activities. Finally, note that the obtained function reduces to that of the two nodes system if the four new flows are set to zero and $d_{23} + d_{31}$ is defined as d_{21}.

Let us move to a second possible route structure, the one known as *hub-and-spoke*, very common in air transport. A *hub* is a node that collects and distributes all flows, and usually coincides with one that is origin and destination. Let us assume arbitrarily that the *hub* is in node 2, and that only one fleet (i.e. one frequency) operates. Obviously, other *hub-and-spoke* structures could be considered as well, like a two fleet operation, one in 1–2 and the other in 2–3. We have chosen to develop the one fleet structure because the others can be constructed adequately using the two nodes system, as shown below. With our assumption (see Fig. 4) a vehicle loads flows y_{12} and y_{13} in node 1, unloads y_{12} in 2 and loads y_{23}, then unloads y_{13} and y_{23} in 3, loading y_{32} and y_{31}, goes back to 2 to unload y_{32} and load y_{21} in order to go back to 1 to unload and begin the cycle again.

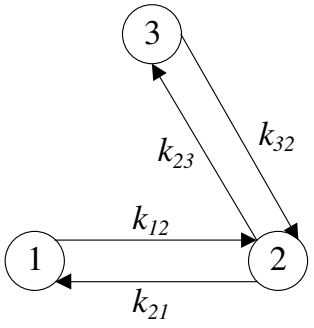

Fig. 4. The *hub-and-spoke* route structure.

In this case, the four load sizes are given by

$$k_{12} = \frac{y_{12} + y_{13}}{f}, \quad k_{32} = \frac{y_{32} + y_{31}}{f}, \quad k_{23} = \frac{y_{23} + y_{13}}{f}, \quad k_{21} = \frac{y_{31} + y_{21}}{f} \tag{18}$$

Again we will assume that total flow in link 1–2 is the largest, which makes k_{12} equal to K and the frequency of the *hub-and-spoke* system happens to be

$$f = \frac{y_{12} + y_{13}}{K} \tag{19}$$

Replacing (19) in (18) the other three load sizes are obtained. Cycle time and fleet capacity are obtained as usual, which yields

$$t_c = \frac{d_{12} + d_{23} + d_{32} + d_{21}}{v} + \frac{2K}{\mu} + \frac{2K}{\mu} \cdot \frac{(y_{21} + y_{23} + y_{31} + y_{32})}{(y_{12} + y_{13})} \tag{20}$$

$$BK = (y_{12} + y_{13}) \cdot \left[\frac{d_{12} + d_{23} + d_{32} + d_{21}}{v} + \frac{2K}{\mu} \right] + \frac{2K}{\mu} (y_{21} + y_{23} + y_{31} + y_{32}) \tag{21}$$

For given values of K, μ and v, Eq. (21) represents $B^*(Y)$ for the *hub-and-spoke* structure. Following the same procedure as in the two previous cases, we obtain the *cost function conditional in the hub-and-spoke route structure*, i.e.

$$C_{HS}(Y) = C_0 + (y_{12} + y_{13}) \cdot (d_{12} + d_{23} + d_{32} + d_{21}) \cdot \Lambda + (y_{12} + y_{13} + y_{32} + y_{21} + y_{31} + y_{23}) \cdot \Omega \tag{22}$$

whose terms have the same interpretation as the ones in Eqs. (12) and (17). Once again, a distance-flow term representing the capacity of the system appears, as well as a pure flow term generated by loading/unloading activities.

So far, we have obtained two conditional cost functions for the three nodes system. By simple analogy, the conditional functions for other three route structures can be obtained as well: the clockwise cyclical system and the *hub-and-spoke* systems with the *hub* in nodes 1 or 3. Additionally, adequately using the cost function for the two nodes system, we can derive conditional cost functions for other cases: the direct service with three fleets, each one serving a pair of nodes cyclically (1–2, 2–3 and 1–3), and the *hub-and-spoke* with two fleets, each one connecting a pair of nodes (with the *hub* at any of the three nodes). Note that in this latter case some flows will have to be loaded and unloaded twice (origin, destination and *hub*), which increases expenses unlike the other cases, but cycle times will be shorter. This is not intended to be an exhaustive list, as other alternative route structures are possible. However, the ones developed and explained here are enough to illustrate that choosing a route structure is a key endogenous element, and to show that the minimum cost is associated with such choice, which is what we do next.

4.2. Global cost function and the role of flows and network

The second stage of the sequential optimization process is the search for the optimal route structure, i.e. the one that minimizes cost in the production of Y and defines the global cost function. This second stage requires the comparison of the conditional cost functions obtained in

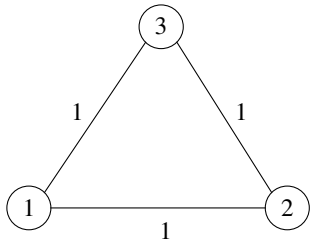

Fig. 5. Physical network.

the first stage. Let us illustrate this process in the three nodes system considering $y_{ij} = y$ (equal flows) and the network shown in Fig. 5.

Let us define the following notation for the alternative route structures.

CG-123: cyclic counter-clockwise.
CG-132: cyclic clockwise.
3F: direct service; three fleets.
HS-i: *hub-and-spoke* with the *hub* in node *i*; one fleet.
2F-i: *hub-and-spoke* with the *hub* in node *i*; two fleets.

Using the conditional cost functions explicitly derived earlier and the definitions of the systems, the conditional cost functions for each one can be constructed and evaluated for the equilateral triangular network in Fig. 5 and equal flows on each OD pair. Omitting C_0 for simplicity the results are

$$\begin{aligned} &C_{\text{CG-123}} = C_{\text{CG-132}} = 9 \cdot y \cdot \Lambda + 6 \cdot y \cdot \Omega \\ &C_{\text{HS-1}} = C_{\text{HS-2}} = C_{\text{HS-3}} = 8 \cdot y \cdot \Lambda + 6 \cdot y \cdot \Omega \\ &C_{\text{3F}} = 6 \cdot y \cdot \Lambda + 6 \cdot y \cdot \Omega \\ &C_{\text{2F-1}} = C_{\text{2F-2}} = C_{\text{2F-3}} = 8 \cdot y \cdot \Lambda + 8 \cdot y \cdot \Omega \end{aligned} \tag{23}$$

Thus, it is optimum to serve the flows directly with three fleets, each one connecting a pair of nodes. It is worth noting that the *hub-and-spoke* structures with two fleets have larger costs than the ones with one fleet *because* of terminal operations, and that the cyclic systems have the largest en-route costs. As we explained in Section 2, the choice of a route structure is dependant on exogenous information, namely the OD structure of demand (the Y vector) and the network topology. Therefore, now we want to examine the effect of a variation of this exogenous information on the choice of an optimal route structure.

First, let us see the effect of changes in the OD structure of demand keeping the same physical network (*flow effect*). Consider the OD structure represented in Fig. 6.

With these new values for the components of Y, we calculated the (conditional) cost for each of the nine proposed route structures. Note that now flows are asymmetric, which induces the need to identify the flows that generate the largest load size, as these are the ones that multiply the constant Λ. The resulting costs are

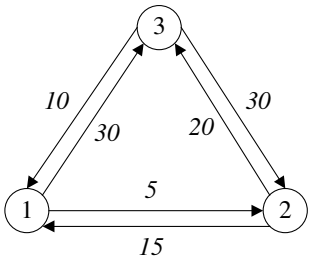

Fig. 6. New OD structure of demand.

$$
\begin{aligned}
&C_{\text{CG-123}} = 195 \cdot \Lambda + 110 \cdot \Omega \qquad C_{\text{HS-1}} = 200 \cdot \Lambda + 110 \cdot \Omega \\
&C_{\text{2F-1}} = 170 \cdot \Lambda + 160 \cdot \Omega \qquad C_{\text{CG-132}} = 165 \cdot \Lambda + 110 \cdot \Omega \\
&C_{\text{HS-2}} = 200 \cdot \Lambda + 110 \cdot \Omega \qquad C_{\text{2F-2}} = 170 \cdot \Lambda + 150 \cdot \Omega \\
&C_{\text{3F}} = 150 \cdot \Lambda + 110 \cdot \Omega \qquad C_{\text{HS-3}} = 140 \cdot \Lambda + 110 \cdot \Omega \\
&C_{\text{2F-3}} = 140 \cdot \Lambda + 130 \cdot \Omega
\end{aligned} \tag{24}
$$

The new optimal route structure is the *hub-and-spoke* with one fleet and the *hub* in node 3, $C_{\text{HS-3}}$. This shows that a different OD structure can generate a new optimal route structure on the same physical network (spatial distribution of nodes and link lengths).

Let us change the network as shown in Fig. 7, keeping the OD structure of Fig. 6, in order to examine the *network effect*.

The new conditional costs functions happen to be

$$
\begin{aligned}
&C_{\text{CG-123}} = 390 \cdot \Lambda + 110 \cdot \Omega \qquad C_{\text{HS-1}} = 300 \cdot \Lambda + 110 \cdot \Omega \\
&C_{\text{2F-1}} = 240 \cdot \Lambda + 160 \cdot \Omega \qquad C_{\text{CG-132}} = 330 \cdot \Lambda + 110 \cdot \Omega \\
&C_{\text{HS-2}} = 500 \cdot \Lambda + 110 \cdot \Omega \qquad C_{\text{2F-2}} = 440 \cdot \Lambda + 150 \cdot \Omega \\
&C_{\text{3F}} = 300 \cdot \Lambda + 110 \cdot \Omega \qquad C_{\text{HS-3}} = 280 \cdot \Lambda + 110 \cdot \Omega \\
&C_{\text{2F-3}} = 280 \cdot \Lambda + 130 \cdot \Omega
\end{aligned} \tag{25}
$$

Unlike the previous cases, now the optimal route structure depends on the values of Λ and Ω. The HS-3 system (which was the best with the previous network) is still superior to both cyclic systems, to 3F, to the other *hub-and-spoke* systems and to the two fleet systems but 2F-1. The *hub-and-spoke* system with two fleets and the *hub* in node 1 could be superior to HS-3 depending on the relative values of Λ and Ω, which are constants defined mostly by prices. Thus, the decision on the optimal route structure will depend, in this case, on how expensive are the loading/unloading

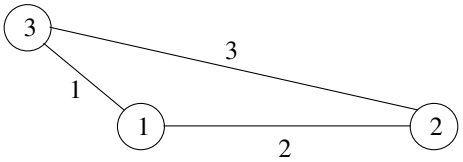

Fig. 7. New network: spatial distribution of nodes and link lengths.

activities relative to activities en-route, which is very reasonable. Note that a proportional growth of distances (links) would increase the difference between expenses en-route of the two structures, keeping the difference between terminal expenses constant, which would increase the attractiveness of route structure 2F-1. This illustrates the network effect on the optimal route structure.

Before closing this section, we find important to stress two points. First, when dealing with real data it is vital to be very precise about what is exogenous and what is endogenous in order to avoid misinterpretations. Our impression is that in the literature, the concept of "network" has been used in fairly different manners, conveying both endogenous decisions and exogenous information. Just as an example, the "network" descriptor *number of route miles* has been defined either as the total miles of the physical network (exogenous information) or the number of miles actually used (endogenously decided). Second, it is worth noting that it have been argued that in the absence of economies of density, a monopoly airline would provide non-stop connections, i.e. direct service, while in the presence of such economies, the airline would save money by operating a hub network (see for instance Brueckner and Spiller, 1991). Our discussion regarding the importance of the exogenous information in the cost minimizing behavior of firms, shows that the selection of a route structure goes beyond the existence of decreasing or constant marginal costs at the arc level. The importance of collecting adequately the costs of loading and unloading activities—which are not related to distances—in the empirical work, is now apparent.

5. Scale, scope and industry structure

Within a multioutput framework, increasing output has two meanings: increasing volume or increasing the number of products. Scale economies are related with the behavior of cost as products expand proportionally (radial analysis). On the other hand, economies of scope are related with the impact on cost of adding new outputs to the line of production. Analytically, the degree of economies of scale, S, can be calculated as the inverse of the sum of cost-product elasticities (Panzar and Willig, 1975), such that a value of S greater, equal or less than 1 shows increasing, constant or decreasing returns to scale respectively, indicating the relative convenience of proportional expansions or reductions of output. The degree of economies of scope relative to a subset R of products, SC_R is calculated as (Panzar and Willig, 1981)

$$\mathrm{SC}_R = \frac{C(Y^R) + C(Y^{M-R}) + C(Y)}{C(Y)} \tag{26}$$

where Y^R is vector Y with $y_i = 0, \forall i \notin R \subset M$, with M being the whole product set. Then, a positive value for SC_R indicates that it is cheaper to produce Y with one firm than to split production into two orthogonal subsets R and $M - R$. It can be easily shown that SC_R lies in the interval $(-1; 1)$.

In the transport case, scale economies are related with the convenience or inconvenience of expanding proportionally the flows in all OD pairs, while economies of scope are related with the potential advantages or disadvantages of serving all OD pairs with one firm (for a full explanation of the concepts of spatial scale and scope, see Jara-Díaz, 2000). Examining the different conditional cost functions derived in the previous sections, we can observe that the choice of a route structure will be invariant to proportional expansions of flows, i.e. the cost minimizing route

structure will be the same. [5] It is quite simple to show that for all the conditional cost functions (including that of the two nodes system) the degree of economies of scale S is given by the ratio between $C(Y)$ and $C(Y) - C_0$. Then $C_0 = 0$ implies constant returns and $C_0 > 0$ generates increasing returns.

Regarding scope analysis, the two nodes system admits only one orthogonal partition. In this case SC examines whether it is convenient to serve the two flows with one firm or to serve y_{12} with one firm and y_{21} with other. It can be easily shown that SC is positive even if $C_0 = 0$ (Jara-Díaz, 1982b, 2000), such that a one firm operation is convenient basically due to the avoidance of idle capacity. Note that for nil right-of-way costs, constant returns to scale coexist with economies of scope, which means that, from a cost viewpoint, it would be convenient to have many firms competing on this system, each one serving both OD pairs.

The analysis of scope in the three nodes system admits various possible orthogonal partitions of the six flows product vector (Fig. 1a). However, the main motivation for the analysis of this system was to study the case in which one node is added or subtracted from a firm service (network expansion or reduction). Thus, let us consider the orthogonal partition $Y^R = \{y_{12}, y_{21}, 0, 0, 0, 0\}$, $Y^{M-R} = \{0, 0, y_{13}, y_{31}, y_{23}, y_{32}\}$, which permits the comparison between one firm serving all six flows against two firms, one serving the two flows between nodes 1 and 2 and the other serving the rest. Note that $C(Y) - C(Y^R)$ is precisely the cost of adding the new node 3 to the network {1,2}, not necessarily equal to $C(Y^{M-R})$ unless SC_R was nil. As seen, the value of the degree of economies of scope depends on the exogenous information (flows and network) because the global cost function does. We will analyze the case depicted in Fig. 6 with link lengths equal to one, with $C_0 = 0$ in order to get rid of the fixed cost effect that influences (increases) both scale and scope. In this case, the global cost function for the firm serving all flows corresponds to that of the HS-3 structure, which yields $C(Y) = 140 \cdot \Lambda + 110 \cdot \Omega$. On the other hand, the global cost function for the two flows system is given by Eq. (10), which yields $C(Y^R) = 30 \cdot \Lambda + 20 \cdot \Omega$ for the values of flows and distances in this case. For Y^{M-R} the three nodes system analysis applies, with two flows set to zero. It can be shown that the optimal route structure could be either 2F-3 or HS-3, both with a cost $C(Y^{M-R})$ given by $120 \cdot \Lambda + 90 \cdot \Omega$. Now we can calculate SC_R from Eq. (26), which yields

$$SC_R = \frac{(30\Lambda + 20\Omega) + (120\Lambda + 90\Omega) - (140\Lambda + 110\Omega)}{C(Y)} = \frac{10 \cdot \Lambda}{C(Y)} > 0. \tag{27}$$

This means that, even if $C_0 = 0$, the six flows will be better served with one firm. Note that in this case savings occur due to expenses en-route, i.e. a single firm permits a better use of the fleet capacity than two firms (vehicle load is larger in average) by means of adjustments in the route structure. Note also that this might not happen for other values of Y or for another physical network. The same applies to loading/unloading activities; in this particular case they are neither a source of economies nor diseconomies of scope because in the three optimal route structures every

[5] We must note, however, that more detailed technical considerations might have an impact on this result: the fact that speed or fuel consumption could be dependent on load size, or that bigger vehicles and loading/unloading sites could have scale advantages, or that congestion could be present, might change the optimal route structure after a proportional growth of all OD flows.

unit is loaded and unloaded once. Under other circumstances, a hub-and-spoke structure with two fleets might be optimal in producing Y, with loading and unloading activities being a source of diseconomies of spatial scope.

The results obtained with this example are enough to show two relevant aspects. First, route structure is an important endogenous cost minimizing variable, which translates into a potential source of economies of scope. As stated by Antoniou (1991), "Networking and the route structure (...) is a well thought-out strategy aimed at taking full advantage of economies of scope and economies of vehicle size" (p. 171). Second, we confirm that incentives to merge might appear because of economies of spatial scope even under constant returns to scale ($C_0 = 0$). This shows that the use of the concepts of economies of scale and economies of (spatial) scope as defined in this article may help solving the problem presented by the couple RTD and RTS, namely that a firm with no incentives to increase its density over a fixed network but with incentives to increase its network may be described. Thus, economies of network expansion should not be looked at through the concept of scale but through the concept of (spatial) scope.

Finally, it is important to emphasize that it was necessary to study a particular case in order to analyze spatial scope correctly, because the characteristics of the physical network and the level of production described as a vector Y, play a key role. Therefore, if these elements are not adequately captured in the empirical work, the analysis of industry structure could be done incorrectly or could be simply unfeasible.

6. Final comments and conclusions

The research described in this article is intended to increase our understanding of the basic technological dimensions that are specific to transport production, in order to improve the analysis of industry structure based upon the estimation of transport cost functions. We feel that the econometrics behind the empirical approach have received much more attention than the production process itself; progress in this latter area is urgently needed. In our opinion, it is precisely the lack of a rigorous view of the key technical aspects, specifically the spatial dimension of product, what is causing difficulties when it comes to make interpretations and inferences on industry structure from estimated cost functions.

This paper shows technically that economies of transport network expansion should be viewed through the concept of economies of scope rather than through the concept of economies of scale. The generalization of the technical approach by Jara-Díaz (1982b, 2000) to a three nodes system presented here, has been instrumental to emphasize the spatial nature of product and to illustrate, in a fairly precise manner, the process that leads a cost minimizing transport firm to the optimization of the route structure as a key endogenous decision. We have been particularly keen on the distinction between what is truly exogenous to the firm—the OD structure of demand, network topology and links description—from what is an endogenous decision as the route structure (service structure and link sequence), showing how the optimal operation of a firm changes as the exogenous variables change. This is a particularly relevant aspect, as all these network related concepts have been treated in a somewhat confusing manner in the empirical work. For example, several variables have been used to describe the network (number of points served, number of route miles, even average length of haul), or the same network descriptor (number of route miles)

has been defined in different ways, either as the total miles of the physical network or the number of miles actually used.

The cost functions obtained for the two nodes system and for various cases in the three nodes system have similar structures. They are separable into two terms: a pure flow term related with expenses on terminal operations, and a flow-distance term directly related with en-route expenses. This latter was shown to represent the optimal transport supply (capacity) offered by the firm (e.g. seat-kilometers), a very reasonable result as en-route expenses depend on the fleet size, frequency and distance traveled, variables that are optimized in order to minimize cost in the production of a given vector Y. Thus, transport supply indices representing product proxies, sometimes used in the empirical work, have received here theoretical support provided the firm is operating in a technically optimal way. The pure flow term was shown to be relevant in the selection of a route structure because some structures imply additional loading and unloading. Thus, given the key role of the route structure in the generation of scope economies, pure flow terms are particularly important for industry structure analysis using aggregated output. When only distance related variables are used to describe output (as is the case in many applied studies) loading/unloading costs may not be adequately captured.

On the other hand, the analysis of scale, scope and industry structure involving relatively simple transport networks technically described in detail, has generated important conclusions. For synthesis, we have been able to show the process through which cost advantages arising from a network expansion (new nodes served) are realized. In essence, fleet capacity can be better used by means of variations in the route structure when new flows are incorporated to production after this type of network expansion. From a microeconomic viewpoint, this translates into economies of scope, which have been shown to exist even under constant returns to scale (in this case, the absence of fixed costs). This provides a new conceptual framework that might help understanding the observed behavior in some transport industries, which is a problem that, so far, has been analyzed as a scale property, an inadequate approach in our view.

Thus, what we can call *increasing returns to* (*spatial*) *scope* are influenced not only by the OD structure of demand—which is already a difficult aspect to cope with in the empirical work—but also by the exogenous spatial information on the network (topology). Therefore, in order to analyze correctly the potential advantages or disadvantages of network expansions through *empirically* estimated cost functions, these elements have to be adequately captured. We believe that this is mainly a problem of an adequate interpretation of "aggregate" cost functions $\widetilde{C}(\widetilde{Y};N)$. So far, this has been treated through ambiguous concepts like economies of scale with variable network size under constant density, a concept that should be carefully examined in order to improve the transport industry structure analysis. The challenge that lies ahead is how to calculate economies of spatial scope from this type of cost function, which is exactly the work in which we are presently engaged.

The two and three nodes networks technically analyzed here have been sufficient to prove the main point, which is that network expansion should be studied through scope rather than scale economies. As stated earlier, this approach was not intended to find optimal service structures for transport firms, but to understand key issues relating cost functions, transport networks and industrial organization. Nevertheless, it is worth asking whether further insight could be gained by expanding this type of detailed analysis to larger networks when feasible. We believe that it might help finding aggregate specifications for transport product that are adequate for the empirical

estimation of cost functions. It could help as well to examine how optimal frequencies and route structures affect aggregates that are used in the literature, as average distance or average load. It should be stressed, though, that finding analytical cost functions for actual transport firms serving many OD pairs over complex networks is simply unfeasible, which is precisely the reason why it is done econometrically. The point is, however, that statistical feasibility should be no excuse to forget neither the true product nor the type of decisions made by a transport firm on a network when using this empirical cost functions for industrial organization analysis.

Acknowledgements

This research has been partially funded by Fondecyt, Chile, Grant 1010687 and the Millennium Nucleus "Complex Engineering Systems".

References

Antoniou, A., 1991. Economies of scale in the airline industry: the evidence revisited. The Logistics and Transportation Review 27 (2), 159–184.

Braeutigam, R.R., 1999. Learning about transport costs. In: Gomez-Ibañez, J., Tye, W.B., Winston, C. (Eds.), Essays in Transportation Economics and Policy. Brookings Institution Press, Washington DC, pp. 57–97.

Brueckner, J.K., Spiller, P.T., 1991. Competition and mergers in airline networks. International Journal of Industrial organization 9, 323–342.

Brueckner, J.K., Spiller, P.T., 1994. Economies of traffic density in the deregulated airline industry. The Journal of Law and Economics 37 (2), 379–413.

Daughety, A.F., 1985. Transportation research on pricing and regulation: overview and suggestions for future research. Transportation Research 19A (5/6), 471–487.

Gagné, R., 1990. On the relevant elasticity estimates for cost structure analysis of the trucking industry. The Review of Economics and Statistics 72, 160–164.

Gálvez, T., 1978. Análisis de Operaciones en Sistemas de Transporte (Transport Systems Operations Analysis), Working Paper ST/INV/04/78, Departamento de Ingeniería Civil, Universidad de Chile, Santiago de Chile.

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., 2000. Transport production and the analysis of industry structure. In: Polak, J., Heertje, A. (Eds.), Analytical Transport Economics, An International Perspective. Elgar, Cheltenham, UK, pp. 27–50.

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

Jara-Díaz, S.R., Cortés, C., Ponce, F., 2001. Number of points served and economies of spatial scope in transport cost functions. Journal of Transport Economics and Policy 35 (2), 327–341.

Keeler, J.P., Formby, J.P., 1994. Cost economies and consolidation in the US airline industry. International Journal of Transport Economics 21 (1), 21–45.

Oum, T.H., Waters II, W.G., 1996. A survey of recent developments in transportation cost function research. Logistics and Transportation Review 32 (4), 423–463.

Oum, T.H., Zhang, Y., 1997. A note on scale economies in transport. Journal of Transport Economics and Policy 31, 309–315.

Panzar, J. Willig, R., 1975. Economies of scale and economies of scope in multioutput production, Economic Discussion paper no. 33. Bell Laboratories.

Panzar, J., Willig, R., 1981. Economies of scope. American Economic Review 71, 268–272.
Spady, R.H., 1985. Using indexed quadratic cost functions to model network technologies. In: Daughety, A. (Ed.), Analytical Studies in Transport Economics. Cambridge University Press, UK, pp. 121–135.
Xu, K., Windle, C., Grimm, C., Corsi, T., 1994. Re-evaluating returns to scale in transport. Journal of Transport Economics and Policy 28, 275–286.
Ying, J.S., 1992. On calculating cost elasticities. The Logistics and Transportation Review 28 (3), 231–235.