REQUIEM FOR RETURNS TO SCALE WITH VARIABLE NETWORK SIZE IN TRANSPORT INDUSTRY STRUCTURE ANALYSIS

Leonardo J. Basso and Sergio R. Jara-Díaz

Departamento de Ingeniería Civil – Universidad de Chile
Casilla 228-3, Santiago, Chile. lbasso@cec.uchile.cl – juradiaz@cec.uchile.cl


Abstract

It is customary to analyze transport industry structure using two indices: economies of density and economies of scale with variable network size (RTS). The first is calculated as the inverse of the sum of the product elasticities of cost, while the second adds the network elasticity. Here we analyze in detail what lies behind the concept of RTS. By taking into account the spatial aspects underlying aggregate output descriptions, we show this concept is ambiguous in terms of what exactly is being analyzed when it is used. We conclude this index does not permit an adequate analysis of industry structure. (JEL D40, L11, L91).
In the last 25 years, transport industries have been subject to numerous detailed econometric analyses designed to study deregulation, productivity and other important economic aspects. Undoubtedly, the preferred microeconomic tool has been the estimation of transport cost functions. The econometric work has made enormous advances like the utilization of flexible forms (translog and quadratic) and the efficiency gains from systems of equations generated through Shepard’s Lemma, both of which are now standard practice. On the other hand, many articles on multioutput theory (synthesized by William J. Baumol et al., 1982) promoted the discussion regarding transport product and its description, an issue that was originally stressed by Richard Spady and Ann F. Friedlaender (1978) who showed how sensitive the industry structure analysis was to product specification within transport cost functions. Presently, most of the econometric work on these functions (generically named multiproduct) include a variety of product descriptions, attributes, and network indices. Although there has been an increasing agreement on what theoretically transport output is, no set of output descriptions for econometric purposes has been really universally accepted.

Starting with the empirical work by Douglas W. Caves et al. (1984), it became customary to analyze transport industry structure using two indices: economies of density and economies of scale with variable network size. The first is calculated as the inverse of the sum of the product elasticities of cost, while the second includes, in addition, the elasticity of cost with respect to the network variable. The difference rests on whether the network is allowed to vary along with product growth. These two concepts have attained a status of almost textbook definitions and have been (and still are) extensively used in empirical work. However, in our view, their application to transport industry structure analysis has not been as clear as desired. The fact that usual (aggregate) transport output descriptions include components that are interrelated has induced a number of questions regarding what is left constant and what is not when calculating density and scale. The key aspect in the debate has been the identification and use of transport output description, as reflected by the discussion in the articles by Robert Gagné (1990), John S. Ying (1992), Kefeng Xu et al. (1994), Sergio R. Jara-Díaz and Cristian E. Cortés (1996), Tae H. Oum and Yimin Zhang (1997) and Jara-Díaz et al (2001). We believe that at least some agreement has been reached regarding the association between economies of density as used in
transport analysis and *economies of scale* as usually measured in the context of manufacturing industries (see Jara-Díaz and Cortes, 1996, and Oum and Zhang, 1997).

However, debate regarding the other index, namely *economies of scale with variable network size*, persists. On one hand, authors like Oum and Zhang (1997) propose improvements to this index by taking into account the relations between the so called attributes and the variables representing the network. On the other hand, Jara-Díaz et al. (2001) showed that the elasticity of cost with respect to the network variable, which has been used to make the distinction between *density* and *scale*, is in fact related with a specific type of economies of scope, spatial scope. The fact is that, to date, it is still unclear what ‘*economies of scale with variable network size*’ mean in terms of what exactly is being analyzed when it is used. In this article we answer this question by taking into account transport output, production and costs, concluding that this index does not permit an adequate analysis of transport industry structure.

The outline of the paper is as follows. In the next section, we offer an interpretative discussion of the current state of the debate, which will provide a useful background for readers less familiar with the transport cost function literature, while allowing us to stress issues that will play a central role in our analyses and to introduce both terminology and notation. Section 3 contains the analytical formulation of what *RTS* represents within the context of cost functions as presented in the literature, which feeds the closer examination detailed in sections 4 and 5. Section 6 contains our conclusions and discussion on further research needs.

I. THE CURRENT STATE OF THE DEBATE

A. PRODUCT, SCALE AND SCOPE IN TRANSPORT

Cost functions can be used in multioutput industry structure analysis to detect the presence of economies of scale and scope. Economies of scale exist if a proportional expansion of the products contained in a vector \( \mathbf{Y} \) causes a less than proportional increase in cost. Economies of scope exist if it is cheaper to produce \( \mathbf{Y} \) with one firm than to split production into two orthogonal subsets. In other words, scale analysis deals with the (proportional) growth of all products while scope analysis is related with the addition of new products to the line (John C. Panzar and Richard D. Willig, 1977, 1981).
Analytically, the degree of economies of scale, $S$, can be calculated as the inverse of the sum of cost-product elasticities 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)

\[
SC^R = \frac{C(w, Y^R) + C(w, Y^{M-R}) - C(w, Y)}{C(w, Y)}
\]

where $C(w, Y)$ is the cost function, $w$ is the vector of input prices and $Y^R$ 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 case of transport, a firm produces movements of different things between many origins and destinations (OD pairs), during different periods. Strictly, then, the product of a transport firm is a vector $Y = \{y_{ijt}\}$, where $y_{ijt}$ represents flow of type $k$ (goods or persons), between origin $i$ and destination $j$, during period $t$ (Jara-Díaz, 1982a,b; Clifford Winston, 1985; Braeutigam, 1999). For simplicity, we will keep the OD sub-indices only. Thus, the multiproduct cost function $C(w, Y)$ represents the minimum expenditure necessary to produce the OD flow vector $Y$ at input prices $w$. Scale economies exist if $C$ increases less than $\lambda$ percent when all flows increase by $\lambda$ percent, while there are economies of scope if it is cheaper to produce all OD flows $Y$ with a single firm than to specialize production spatially.

In order to illustrate these concepts, let us consider a firm that moves freight on six OD pairs (see figure 1). If there are economies of scale at $Y$, it is not convenient from a cost viewpoint to have various firms, each one serving a fraction of $Y$, competing on the six OD pairs. If there are economies of scope for the partition $\{y_{12}, y_{21}, 0, 0, 0, 0\}$, $\{0, 0, y_{23}, y_{32}, y_{13}, y_{31}\}$, this means that is more cost convenient to have a single firm producing all six flows, than to have two firms each one serving one of the subsets (see Jara-Díaz, 2000). Thus, it may well happen that a firm...
exhibits $S=1$ and $SC>0$ simultaneously (as illustrated in Jara-Díaz, 1982b, and Jara-Díaz and Leonardo J. Basso, 2003), in which case it would be cost convenient to have many firms competing, each one serving all flows.

![Figure 1. OD structure.](image)

It is important to note that the OD structure represented in figure 1 does not represent a physical network nor a route structure but the product vector. The physical network—a necessary input for production—is the set of connected links on which vehicles can circulate to reach every node, while a route structure is the set of links chosen by the firm to actually reach the nodes using the vehicle fleet. In other words, the physical network is (together with the OD structure) exogenous information for the cost minimizing firm, while the route structure is a fundamental endogenous decision (as is the number and size of the vehicles). In fact, the optimal route structure—that which generates the minimum production cost—will vary depending on $\mathbf{Y}$ (OD structure) and on the characteristics of the physical network. These concepts and how economies of spatial scope relate to them are studied in Jara-Díaz and Basso (2003).

As evident, the detailed multioutput nature of transport product cannot be captured completely in applied work due to the huge amount of OD pairs usually served, which makes aggregation necessary for the econometric estimation of cost functions. Unfortunately, there is no universally accepted description of product and many aggregates have been used in the literature, each one representing a partial aspect of the disaggregated product ($\mathbf{Y}$). Some of these aggregates are named ‘products’, as passenger-kilometers, seat-kilometers, total tons, number of shipments or vehicle-kilometers, and others are named attributes or characteristics, as average length of haul or load factor. The latter are elements that try to capture somehow the heterogeneity of transport product, diminishing the inherent ambiguity of single aggregates as those that reflect volume-
distance. Since the mid eighties, some elements describing the network have been added to this
type of description. Route miles, \( RM \), and number of points served, \( PS \), have been the most
popular (for a synthesis of the main aggregates used in the literature, see Jara-Díaz, 2000). Both,
the strict definition of transport output and the need of aggregation for econometric purposes,
have been frequently recognized in the literature\(^2\). From now on, following Jara-Díaz (1982a) and
Winston (1985), we will call vector \( Y \) the *true transport product*, as a way to distinguish it from
aggregates.

Then, on one hand, the concepts of scale and scope are crystal clear when the true vector \( Y \) is
considered. On the other hand, the variety of variables representing products, attributes or
network, used in different forms when cost functions are estimated, makes it difficult to obtain
clear inferences regarding optimal industry structure in the transport sector. This poses two types
of challenges. The first can be called a pre-estimation problem, namely the search for the best
aggregates (e.g. Spady, 1985, Andreas Antoniou, 1991, Jara-Díaz *et al.*, 1991). The second is a
post-estimation problem, namely how to use estimated cost functions in order to make
meaningful and reliable interpretations from the obtained parameters (e.g. Jara-Díaz y Cortés,
1996). In this article we face the second challenge.

**B. **RETURNS TO SCALE AND RETURNS TO DENSITY FROM AGGREGATED COST FUNCTIONS

Consider an estimated function \( \tilde{C}(\tilde{Y}; N) \) representing long run costs, where \( \tilde{Y} \) is the vector of
aggregated product descriptions (including the so-called attributes) and \( N \) is the variable
representing the network. Factor prices will be suppressed for simplicity. Since the early eighties,
most of the studies that used cost specifications as the one above have relied upon two indices for
the transport industry structure analysis: returns to density (\( RTD \)) and returns to scale (\( RTS \))
(Caves et al., 1981). They differ because the network is regarded as constant for \( RTD \) and
variable for \( RTS \). 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”. Analytically
(our notation)
\( RTD = \frac{1}{\sum_{h \in H} \eta_h} \)

\( RTS = \frac{1}{\sum_{h \in H} \eta_h + \eta_N} \)

where \( \eta_h \) is the elasticity of \( \tilde{C}(\tilde{Y}; N) \) with respect to aggregate product \( h \) and \( \eta_N \) is the elasticity with respect to \( N \), fulfilling \( RTS < RTD \) since the network elasticity is positive\(^3\). The conclusions on industry structure that follow from the estimated values of \( RTS \) and \( RTD \) appear in many case studies and reviews (e.g. Gillen et al., 1990; Pels and Rietveld, 2000). Increasing returns to scale (\( RTS > 1 \)) 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 (\( RTD > 1 \)) would indicate that traffic should be increased keeping network constant. Note that \( RTS < RTD \) poses the first glimpse of a problem: a firm that, from a cost viewpoint, has both optimal density and optimal network size cannot be described. Moreover, if network size was optimal (\( RTS=1 \)) the firm must exhibit increasing returns to density. This is unpleasant indeed, to say the least.

As explicitly stated in equations (2) and (3), the sum of the product elasticities is made on a subset \( H \) of the aggregates, which we have not defined so far. This, in fact, is an issue not resolved in the technical literature. Many articles do not include the so-called attributes in \( H \). For instance, Caves et al. (1984) and Gillen et al. (1990) do not include the elasticities of average length of haul nor load factor in their air transport studies. In other articles, however, the authors argue that the addition of certain elasticities will depend on how the product is expanded. This is the case of Caves et al. (1985), who consider the average length of haul elasticity, and Caves and Christensen (1988) who include the load factor elasticity both for urban buses and air transport analysis. In other cases, the problem is not even mentioned as no attribute was considered (e.g. Massimo Filippini and Rico Maggi, 1992; Anna Matas and Josep Lluis Raymond, 1998). The definition of the elements in \( H \) does not follow a general rule.
The issue of what elasticities should be considered in equations (2) and (3), and how they should be treated, acquired a relevant status in the nineties, as this affected the calculated values and the corresponding conclusions or recommendations on industry structure. The clearest examples of this discussion are the articles by Gagné (1990), Ying (1992) and Xu et al. (1994), who investigated the impact of the interrelation between aggregates (‘products’ and ‘attributes’). In essence, their idea was to choose one aggregate as the generic output, making its elasticity dependant on the other aggregates as well.

Jara-Díaz and Cortés (1996) proposed a new look at the problem, based upon the true definition of output, as it permits an unambiguous analysis of scale and scope. The idea was that such analysis can be preserved even if aggregates are used. The key point is that most elements in vector \( \tilde{Y} \) are, in fact, implicit constructions using the elements of the true vector \( Y \), i.e. \( \tilde{Y} \) is actually \( \tilde{Y}(Y) \). For instance, total flow (total tons in a period) is simply the sum of the \( y_{ij} \) over all OD pairs, ton-kilometers are a sum of distance-weighted-OD-flows and average length of haul is the ratio between the latter and the former. Therefore, \( \tilde{C} \) is an implicit function of \( Y \), which can be used to calculate the elasticities of cost with respect to the true product using the chain rule. Following this procedure, the expression for \( S \) happens to be (see Jara-Díaz and Cortés, 1996, for the full derivation)

\[
S = \frac{1}{\sum_h \alpha_h \cdot \tilde{\eta}_h}
\]

where \( \tilde{\eta}_h \) is the elasticity of cost with respect to the \( h^{th} \) argument in \( \tilde{C} \), noted \( \tilde{y}_h \), and

\[
\alpha_h = \sum_i \varepsilon_{hi} = \sum_i \frac{y_i}{\tilde{y}_h} \cdot \frac{\partial \tilde{y}_h}{\partial y_i}
\]

where \( y_i \) is flow on OD pair \( i \). In short, \( \varepsilon_{hi} \) is the elasticity of aggregate \( h \) with respect to \( y_i \) and \( \alpha_h \) corresponds to the degree of homogeneity of aggregate \( h \) with respect to \( Y \).
Result (4) shows that a correct estimation of the degree of scale economies will not be equal to the inverse of the sum of the aggregate elasticities unless the $\alpha_h$ are all equal to one, which does not happen for many aggregates. To have an idea of how this works analytically, consider ton-kilometers ($TK$) and average length of haul ($ALH$), which can be written as

\begin{align}
TK &= \sum_i y_i \cdot d_i \\
ALH &= \frac{\sum_i y_i \cdot d_i}{\sum_i y_i}
\end{align}

where $d_i$ is distance traveled to move flow $i$. Using equation (5), and assuming for simplicity that $d_i$ is invariant to a proportional expansion of $Y$ (which might not be the case, as shown by Jara-Díaz and Basso, 2003), one obtains that $\alpha_{TK} = 1$ and $\alpha_{ALH} = 0$. Note that the intuition behind these results follows the basic definition of scale, i.e. if all components of output $Y$ are expanded proportionally (to $\lambda Y$), then $TK$ expands by the same proportion while $ALH$ remains constant. In some cases, $\alpha_j$ can be between zero and one, depending on how the firm operates. This is the case of vehicle-kilometers, for which $\alpha_h$ will be one if the firm accommodates flow increases only through frequency adjustments; if the firm acts over frequency and load size, then $0 < \alpha_h < 1$.

A fairly complete analysis of the most frequently used aggregates can be found in Jara-Díaz and Cortés (1996).

With the method synthesized above, the decision about which aggregates should be considered in $H$ (be it products or attributes) is no longer arbitrary, although the way in which the elasticities should be considered in the calculation is not a matter of zero or one. On the other hand, the $Y(\tilde{Y})$ approach makes it evident that the problem of relations between aggregates (as depicted in the early nineties) is just the consequence of implicit relations between aggregates and $Y$. Oum and Waters (1996) described the method as “a more rigorous reconsideration” of the problem, pointing out, however, that it corresponds to the calculation of $RTD$ rather than $RTS$ because the network is not allowed to vary. Looking at $\tilde{C}(\tilde{Y}; N)$, this is in fact true, but it is also true that the method was rigorously developed to calculate $S$, the degree of economies of scale at the disaggregated $Y$ level, which considers proportional expansions of the true products (flows) and,
therefore, keeps the network served constant as no new products (new OD pairs) should be considered. Therefore, $RTD$ (as defined in the literature) and $S$ (as defined rigorously) become equivalent. As foreseen by Filippini and Maggi (1992, p. 308), “economies of scale, as they are usually measured in the context of manufacturing industries, simply represent economies of density in a network industry”; this is now apparent.

The question that remains then is what $RTS$ represents. Jara-Díaz and Cortés (1996) explicitly point out that their methodology is not intended to give an interpretation of a scale measure within the context of a varying network, because $N$ does not vary with equi-proportional changes of the true output, the flow vector $Y$. Moreover, they indicate that increasing the number of points served, $PS$, a frequently used network index, would imply an increase in the number of OD pairs; “this cannot be studied with scale analysis, because new products are being added to the line of production, which requires scope analysis by definition” (p. 170). The fact that $N$ is related more with scope than with scale had been suggested by some authors. For example, Gloria J. Hurdle et al. (1989, p.121) pointed out that the definition of density is closer to what is generally considered as scale, while that of scale resembles scope; Andrew F. Daughety (1985, p.473) noted that “size economies [$RTS$] depend on product-specific economies, scope economies and subadditivity”; and Severin Borenstein (1992, p. 59-60) directly called $RTS$ ‘economies of scope’. Recently, Jara-Díaz et al. (2001), have shown that the $PS$ elasticity of cost, that makes the difference between $RTS$ and $RTD$, can be used to obtain an approximated estimation of the degree of economies of (spatial) scope. But, if spatial scope ($SC$) is hiding behind $RTS$ and density is in fact scale, then a first problem arise: $S=1$ and $SC>0$ cannot be represented by $RTD=1$ and $RTS>1$ because the former has to be always larger than the latter by virtue of equations (2) and (3).

The key issue is, then, that in spite of the enormous advancements in the econometric tools that support the estimation of transport cost functions in the last twenty five years, and in spite of the unambiguous meaning of scale and scope at a disaggregated output level, the true meaning of $RTS$ and what its calculation implies remain quite obscure. In the following sections we analyze this index rigorously by using the disaggregated view of product, which proved to be a fruitful one when inspecting $RTD$ (Jara-Díaz and Cortés, 1996). We will show that the conditions behind
the definition of RTS as an analytical construct are such that this index is not adequate to analyze transport industry structure.

II. THE DIRECT INTERPRETATION OF RTS ON THE AGGREGATES

As seen, when transport product is described in detail (vector $Y$), $S$ and $SC$ play unambiguously complementary roles: $S$ takes care of cost behavior as output expands proportionally keeping OD pairs constant and $SC$ deals with costs when OD pairs are added. With the aggregate indices, something similar is intended: $RTD$ is aimed at analyzing what happens with costs as (aggregate) product increases keeping the network fixed ($density$ varies), while $RTS$ analyzes cost behavior as both product and the network increase, keeping $density$ constant (see Caves et al., 1984, or Braeutigam, 1999). Grossly speaking, $density$ is understood as the ratio between (aggregated) product and the network index$^4$.

To begin making a rigorous interpretation of these concepts it is convenient to recall that the degree of scale economies, whatever the intended meaning, corresponds to the degree of homogeneity of the cost function in those arguments that are involved. Applying this property to $RTS$ as defined in equation (3), the following identity holds (see appendix 1):

$$\tilde{C}(\gamma^{RTS} \hat{Y}_H, \hat{Y}_K, \gamma^{RTS} N) \equiv \gamma \tilde{C}(\hat{Y}_H, \hat{Y}_K, N)$$

This means that, behind $RTS$ as defined in the literature, we have the variation in costs provoked by a proportional variation of all aggregates in $H$, $\hat{Y}_H$, and by the variation in the network index $N$ by the same proportion, keeping the aggregates not in $H$, $\hat{Y}_K$, invariant. Note that identity (8) implies that the network grows keeping $density$ constant. For the purpose of this paper, we want to emphasize the following interpretation of this identity: in the calculation of $RTS$ it is analytically imposed that, after a network expansion, aggregates in $H$ will vary in that same proportion while those not in $H$ (subset $K$) will not vary.

In order to explore the implications of the definition of $RTS$ under this perspective, let us consider a firm $A$ whose aggregated product vector and network size are described by $\hat{Y}^A$ and $N^A$
respectively. A cost function $\bar{C}(\bar{Y}; N)$ that reproduces industry costs well is available. The cost of firm A’s production is $\bar{C}^A = \bar{C}(\bar{Y}^A, N^A)$. In order to analyze scale with variable network size as defined in (8), let us consider an expansion of both the network and the aggregates in $H$ by a certain proportion $\lambda$; aggregates not in $H$ have to keep the original level. To simplify notation, let us call $L$ the expanded firm $A$, such that $N^L = \lambda N^A$, $\bar{Y}^L = (\bar{Y}^A_H, \bar{Y}^A_K)$ and $\bar{C}^L = \bar{C}(\lambda \bar{Y}^A_H, \bar{Y}^A_K, \lambda N^A)$. With these elements, let us try to form a first idea of what $RTS$ is. To begin with, the elasticity of cost with respect to $N$, $\eta_N$, can be approximated as

$$\eta_N = \frac{\partial \bar{C}}{\partial N} \approx \frac{\bar{C}^L - \bar{C}^A}{N^L - N^A},$$

$$= \frac{\bar{C}^L - \bar{C}^A}{\lambda N^A - N^A},$$

(9)

Note that the same approximation is valid for the elasticities of cost with respect to aggregates in $H$, as these aggregates vary by the same proportion $\lambda$. Replacing the elasticities in equation (3) for $RTS$, and calling $H^0$ the cardinality of subset $H$, we obtain

$$RTS = \frac{1}{\sum_{h \in H} \eta_H + \eta_N} \approx \frac{\lambda - 1}{(H^0 + 1)} \frac{\bar{C}^A}{\bar{C}^L - \bar{C}^A},$$

(10)

Equation (10) represents an explicit relation between $RTS$ and the costs of the expanded firm, $\bar{C}^L = \bar{C}(\lambda \bar{Y}^A_H, \bar{Y}^A_K, \lambda N^A)$. Two things are evident from this approximation. First, $\frac{\partial RTS}{\partial \bar{C}^L} < 0$, which says that the smaller the cost increase after a network (and product) expansion, the larger the value of $RTS$, an intuitively appealing result that is in line with the interpretation that it was meant to have. Second, the value of $RTS$ depends on the number of aggregates whose elasticities are taken into account, an unpleasant feature indeed; its impact, however, is uncertain because the cost of the expanded firm depends on the aggregates considered as well. A deeper examination of $RTS$ is required.
In order to have a complete view of what can be actually analyzed through RTS, we identify three aspects that have to be carefully studied. First, the role of the specific variables used to describe the network. Second, the properties that have to be fulfilled by $Y_L$, the true (disaggregate) product of firm $L$. Finally, what are the conclusions on industry structure that can be obtained from the analysis of RTS. These three issues are examined in the following section.

III. A DISAGGREGATE EXAMINATION OF RTS

A. THE NETWORK VARIABLE

As mentioned in section two, there are two frequently used network variables in the literature: the number of points served, $PS$, used mainly in air transport, and route miles, $RM$, mostly used in land transport. The definition of $PS$ is simply the number of nodes in the system (number of airports connected in the case of air transport). The definition of $RM$, however, is less transparent. Sometimes $RM$ is understood as the total length of the physical network available, and sometimes it is taken as the total distance used within the physical network: the operated network. As recently explained from a technological perspective by Jara-Díaz and Basso (2003), the first definition corresponds to an input in the cost minimizing process of a transport firm, while the second is a result of such process, as the used route length depends on the route structure chosen, a key endogenous decision for most transportation firms.

Thus, the information behind each network index is different. On one hand, $PS$ measures the number of nodes covered by the firm, information that is related with the number of OD pairs served, which means that $PS$ is ultimately related with the number of elements in the fully described product vector (OD flows). Therefore, necessarily, when $PS$ grows the number of products grows. On the other hand, route miles have two interpretations. Under the “length of the physical network” definition, the relation between $RM$ and production is unclear, as it could happen that the network gets larger but the number of nodes connected stays the same (OD pairs). This would imply that the only change possible is that of the route structure, adjusting (reducing) costs on the new network while generating the same product vector (flows vector $Y$). Thus, in this case, a network expansion will not be associated with an increase in the number of products as with $PS$. Moreover, as stated by Mark H. Keaton (1990) in his railroad study, if $RM$ does not change, density might change in two ways: increasing product (car kilometers in his study) in the
existing OD pairs, or increasing the number of OD pairs for the given network. Then, it might well happen that the same process is described as scale (RTS) if PS is used, or as density (RTD) if RM is used as network variable. As evident, this is a source of ambiguity regarding the use of N within the context of RTS, as economies of scale with variable network size do not have a unique meaning in terms of production; it depends on what network descriptor is used.

B. The true flow vector

In what follows, we will keep PS as the variable that describes the network, as it is not only unambiguous but also has a direct interpretation in terms of the true flow vector. To fully understand what RTS permits to analyze, we will try to unveil the characteristics of firm L, the expanded firm, from the point of view of the actual product, the YL vector. Note first that behind the aggregate description of firm A’s production, YA, there is a true vector YA, from which the aggregates are implicitly obtained as YA(YA), e.g. equations (6) and (7). Similarly, behind the aggregate description of firm L’s product, there is a true flow vector YL whose properties we want to examine. The problem can be formulated as follows: assume YA is known and that PS and part of YA(YA) expands by λ, forming YL = [λYH A(YA), YK A(YA)]. Then the question is: What are the characteristics or properties of YL, the true flow vector underlying YL?

Schematically

\[
(Y^A; PS^A) \Rightarrow \left[ Y_H^A (Y^A), Y_K^A (Y^A); PS^A \right] \xrightarrow{\text{PS expansion (RTS)}} (\lambda Y_H^A, \lambda Y_K^A; PS^A) \equiv (Y_L^A; PS_L^A) \xleftrightarrow{\text{RTS}} (Y_L^A; PS_L^A)
\]

Recall that increasing PS, even by one node, implies new flows in the system (2PS potential new flows in that case). Then, the expanded firm L will serve two types of flows: the original ones and new ones whose origin or destination is one of the new nodes. How do they change after the network expansion within the RTS context? Nothing can be inferred from the equations describing RTS (3 or 8) and little has been said in the literature regarding this issue. In our opinion, Braeutigam (1999) gives the more explicit interpretation of what appears to be the meaning of the relation between flows and RTS in many articles (see also Pels and Ritveld,
2000). Breutigam uses points served and ton-kilometers as depicted in figure 2 to explain: “The size of the network has been increased by adding another node (5) and a link to serve that node. There are now five nodes ($PS=5$), so the number of nodes has been increased by 25 percent, but the volume of traffic over the existing links has not been increased (the density of traffic is unchanged). In other words, the size of the network is increased, but the density of traffic movements is unchanged. This is the type of output expansions usually envisioned in studies of economies of scale.” (Braeutigam, 1999, p.75)

In order to make a rigorous analysis of this interpretation, we need to formulate the problem in terms of the true product, which, as explicitly stated by the author himself (see footnote 2), is the underlying vector of OD flows. Then we have to study the cost minimizing decisions of the firm (route structure), which in turn requires an explicit physical network. An ample view of the actual physical network behind figure 2 would be one in which all nodes can be potentially connected directly (as in the airline case). If this was the case, serving new flows might translate into a more convenient new route structure. It could even happen that one of the links become no longer necessary (Jara-Díaz and Basso, 2003). But let us assume that the actual physical network includes only the links explicitly represented in figure 2, i.e. those between node 1 and each of the others, and let us call generically $d_{ij}$ the distance between nodes $i$ and $j$. On the original network, we have twelve potential OD flows $y_{ij}$. Clearly, the total ton-kilometers on link $1j$, $TK_{1j}$, is given by the total flow that travels through that link times $d_{1j}$. On the original network, then

![Figure 2. A 25% network expansion, constant density (Braeutigam, 1999)](image-url)
Note that in equation (11) we are not imposing nor assuming anything in particular for the OD flows. After the network is expanded, eight new OD pairs could be potentially served. Only two of these new flows, namely 1–5 and 5–1, will use the new link; the other six will necessarily use some of the other original four links, adding $y_{15} + y_{51}$ to the parenthesis of equation (11). As $TK_{1j}$ remains constant for all $j$ after node 5 is added, either the original OD flows diminish or they are kept constant and the six potential new OD flows are nil. Besides this evident ambiguity, unpleasant by itself, neither alternative seems particularly useful as a basis for a rigorous cost and industry structure analysis as explained in section 2. First, diminishing the original OD flows (diminishing true product) in order to construct an index for the analysis of the advantages of increasing production seems simply paradoxical. Second, analyzing the effects of a network growth imposing that most of the flows in the new OD pairs are zero, is unjustifiable. Thus, the hidden conditions on the OD flows behind the constancy of ton-kilometers on each link seem unreasonable within this interpretation of $RTS$.

Another explicit interpretation of the relation between flows and $RTS$ is given by Oum and Zhang (1997) within the context of their discussion regarding the inclusion of the average length of haul elasticity into the calculation of $RTS$. They explicitly state that “if an increase in output is accompanied by a change in network size, that is, if the number of origin-destination pairs also increases with traffic flows in each route, then theoretical consistency would require investigation on the correlation between length of haul and network size” (p. 310, emphasis added). This implies growth of traffic flows on all links, which evidently contradicts the graphical example and the interpretation provided by Braeutigam (1999). As a conclusion, the literature does not provide a definitive answer regarding the form of variation of the flows after the network expansion.

So, besides the definition of a network variable, we have a second problem with $RTS$, namely the form in which the true output changes is unknown: some articles provide mutually conflicting alternatives and most do not even touch on the problem. In our opinion, the adequate approach to
The dilemma is to use the distinction between the flows on the original OD pairs and those on the added ones. Recalling the discussion of Braeutigam’s example, the first relevant question is related with the magnitude of the original OD flows after the network expansion: Do they change? We believe that the most reasonable assumption is one that is compatible with a meaningful cost analysis of industry structure when one is dealing with new products in addition to old ones, namely that the original OD flows (i.e. those served before the network expansion) keep their level, which seems to have been the idea behind figure 2, on the “wrong” output. What about the added flows?

C. THE CONSTANT DENSITY ASSUMPTION

After the discussion on the original flows produced, we need to analyze the conditions that the new products (flows) have to fulfill within the context of RTS. Let us take a look at relatively simple cases involving two of the most popular aggregates in the literature: ton- kilometers (TK), usually an output, and average length of haul (ALH), usually an attribute. Network will be described by the number of points served, PS. Let us work with the simple network represented in figure 3, and consider a network expansion from two nodes (1,2) to three nodes (1,2,3).

![Figure 3. Physical network](image)

The true output of firms $A$ and $L$ will be

$$Y^A = \{y_{12}, y_{21}, 0, 0, 0\} \quad \text{and} \quad Y^L = \{y_{12}, y_{21}, y_{23}, y_{32}, y_{13}, y_{31}\}$$

where $y_{ij}$ represents tons per unit time moved from origin $i$ to destination $j$. So, $PS^A=2$ expands to $PS^L=3$ such that the network variable increases by $\lambda=3/2$ (note that OD flows on the originally served OD pairs are kept constant).

Using equations (6) and (7), the values of the aggregates are
\( TK^A = y_{12} \cdot d_1 + y_{21} \cdot d_1 \)  
(12) 
\( TK^A = y_{12} \cdot d_1 + y_{21} \cdot d_1 + y_{32} \cdot d_2 + y_{13} \cdot (d_1 + d_2) + y_{31} \cdot (d_1 + d_2) \)  
(13) 
\( ALH^A = \frac{y_{12} \cdot d_1 + y_{21} \cdot d_1}{y_{12} + y_{21}} = d_1 \)  
(14) 
\( ALH^A = \frac{y_{12} \cdot d_1 + y_{21} \cdot d_1 + y_{32} \cdot d_2 + y_{13} \cdot (d_1 + d_2) + y_{31} \cdot (d_1 + d_2)}{y_{12} + y_{21} + y_{32} + y_{13} + y_{31}} \)  
(15) 

Now we are fully prepared to analyze the conditions imposed by RTS on the new flows for various cases, each one defined by the elasticities considered in the calculation of RTS (which we have called set \( H \)).

**Case 1:** \( \bar{C} = \bar{C}(TK, PS) \) and \( RTS = \left[ \eta_{TK} + \eta_{PS} \right]^{-1} \)

According to the definition of RTS in equation (8), \( TK \) should expand by the same proportion as \( PS \), this is \( TK^A = (3/2)TK^A \). From equations (12) and (13) after some manipulation we get

\[
\frac{d_1}{d_2} = \frac{y_{23} + y_{32} + y_{13} + y_{31}}{0.5 \cdot (y_{12} + y_{21}) - (y_{13} + y_{31})}
\]
(16)

As distances \( d_1 \) and \( d_2 \) are known, what equation (16) does is to impose restrictions on the new flows. For example, the denominator on the right hand side has to be positive, which implies \( 0.5 \cdot (y_{12} + y_{21}) > (y_{13} + y_{31}) \). If distances were equal, it should hold that

\[
\frac{1}{2} (y_{12} + y_{21}) = y_{23} + y_{32} + 2 \cdot (y_{13} + y_{31})
\]

i.e. the four new flows should add up to less than half the original ones. Evidently, both conditions are quite restrictive on the flow magnitudes.

**Case 2:** \( \bar{C} = \bar{C}(TK, ALH, PS) \) and \( RTS = \left[ \eta_{TK} + \eta_{PS} \right]^{-1} \)
Now $ALH$ is a variable in the cost function, but is not considered in the calculation of $RTS$. This means that $TK$ should increase by $3/2$ while $ALH$ remains constant. Therefore, two analytical conditions have to hold simultaneously: $TK^L=(3/2)TK^A$ and $ALH^L = ALH^A$. Imposing this on equations (12) to (15) we get

\[
\frac{1}{2}(y_{12} + y_{21}) = y_{23} + y_{32} + y_{13} + y_{31}
\]

which imposes the stringent condition that the four new flows served after the expansion have to add up to half the total flow previously served. But the following case is most surprising.

**Case 3:**

\[
\bar{C} = \bar{C}(TK, ALH, PS) \text{ and } RTS = \left[\eta_{TK} + \eta_{ALH} + \eta_{PS}\right]^{-1}
\]

In this case, both $ALH$ and $TK$ belong to $H$ and are considered in the $RTS$ calculation. Thus, $RTS$ imposes that both should increase by $3/2$ simultaneously, i.e. $TK^L=(3/2)TK^A$ and $ALH^L = (3/2)ALH^A$. Using equations (12) to (15) we get

\[
y_{23} + y_{32} + y_{13} + y_{31} = 0
\]

which directly translates into

\[
y_{23} = 0 ; \ y_{32} = 0 ; \ y_{13} = 0 ; \ y_{31} = 0
\]

In this case, then, internal consistency in the calculation of $RTS$ after a one node network expansion imposes that nothing should enter nor depart from the new node. If the process had been looked at exclusively in terms of both $ALH$ and $TK$, this shocking condition would not have been made apparent.

As shown by equations (17) and (18), cases 2 and 3 above impose conditions on the new flows that do not depend on the link distances. This suggests that the conditions found might not be dependent on the route structure (an endogenous decision) neither on the physical network.
(exogenous information). To explore whether they can be generalized beyond the specific network depicted in figure 3, note that in general

\[(19) \quad T = \sum_i \sum_j y_{ij}\]

\[(20) \quad ALH = \frac{TK}{T}\]

where \(T\) is the number of total tons moved. Then **case 2** can be reformulated as

\[ALH^L = ALH^A \quad \Rightarrow \quad \frac{TK^L}{T^L} = \frac{TK^A}{T^A} \quad \Rightarrow \quad \frac{\lambda TK^A}{T^L} = \frac{TK^A}{T^A} \quad \Rightarrow \quad T^L = \lambda T^A .\]

If \(y'\) and \(y\) are the average of the new flows and original flows respectively, then it can be shown that their ratio has to fulfill (see appendix 2)

\[(21) \quad \frac{y'}{y} = \frac{1}{2} \left( \frac{PS - 1}{PS} \right)\]

which depends only on the (initial) number of points served, \(PS\). Equation (21) has been constructed assuming constant density, understood as the ratio between output \((TK)\) and network \((PS)\). From this assumption we have shown that the new flows, in average, have to vary between one fourth \((PS=2)\) and one half \((PS\) very large) the average value of the original flows, a strange condition indeed.

If we apply the same analysis to **case 3** the implications are even more dramatic. The conditions in this case are \(ALH^L = \lambda ALH^A\) and \(TK^L = \lambda TK^A\), but

\[ALH^L = \lambda ALH^A \quad \Rightarrow \quad \frac{TK^L}{T^L} = \frac{\lambda TK^A}{T^L}\]
which, combined with the condition on $TK$ implies $T^L = T^A$. This directly implies that the new flows $y^*_{ij}$ to be served after the network expansion have to fulfill

$$\sum y^*_{ij} = 0 \quad \Rightarrow \quad y^*_{ij} = 0 \,.$$ 

In other words, all new flows have to be zero, irrespective of the physical network, the original number of nodes or the route structure. Furthermore, this result could hold for other network variables as well, provided the expansion of $N$ brings new OD pairs into the picture.

These simple examples and their generalizations show that looking at the properties of RTS in terms of conditions on the aggregates, as done in most of the literature, in fact hides stringent restrictions that have to be fulfilled by the underlying true product, the vector of OD flows. These restrictions seem particularly objectionable when analyzing the new flows after a network expansion. But this is not all. As seen here, the conditions on the components of $Y$ do depend on the aggregates considered (both those that are in set $H$ and those that are not), which makes evident yet another important limitation of RTS, namely that it is not uniquely defined in terms of what it analyzes but is heavily dependent on the specification of variables within the cost function. Note that this is not an econometric problem related with either misspecification of the cost function or with biased parameter estimates because of inclusion of too many or omitted variables; it is an analytically induced ambiguity. And the larger the number of aggregates used to describe transport product in the cost function, the more acute the problems just described, because the number of simultaneous constraints increase, irrespective of their consideration or not in the calculation of RTS. All this evidently erodes, if not completely destroys, the usefulness of RTS as an instrument to analyze industry structure.

Two final observations are important before finishing this closer examination of the implications behind the constant density assumption. First, the presence of this type of problems does not seem to depend on the inclusion or absence of attributes elasticities in that calculation, something that has drawn extensive attention in the literature. Second, although we have focused on the number of points served as the network variable because of its unambiguity, these problems subsist when other network descriptors are used. For example, if route miles were used, the
proportion of network growth would be given by $\lambda=(d_1+d_2)/d_1$ and some different restrictions would appear (unless $d_2$ equals one half of $d_1$). Note that for different values of $d_2$, different conditions on the flows are obtained, something that does not appear to be justifiable. Even if other network descriptions are used, the basic important problem of $RTS < RTD$ subsists in any case.

IV. AN ATTEMPT TO SOLVE THE PROBLEM

The problem described in the previous section occurs, in our opinion, because of the failure to think in terms of $Y$ when dealing with aggregates within transport cost functions in the empirical work. In fact, understanding the ambiguity that arise when aggregates are used to study scale and density, has required the explicit recognition of their definitions in terms of $Y$. Dealing with this requires a careful look at the variation of both the aggregates and the actual product as the network increases, a la Jara-Díaz and Cortés (1996). Note, however, that the problem seems to admit a different formulation, namely that forcing aggregates in $H$ to vary exactly as the network does while those not in $H$ do not vary, is the primal cause of the difficulties. In other words, the constant density assumption of $RTS$ could be blamed for the failure. Some authors have taken this view of the problem, with an approach that seems very reasonable: why not examining empirically how the aggregates vary with the network?

The work by Oum and Zhang (1997) moves exactly in the direction of understanding how the aggregates respond to a network expansion. They argue that the average length of haul might change after a change in network size and, therefore, “theoretical consistency would require investigation of the correlation between length of haul and network size”, a relation that had been previously suggested by Caves et al. (1985). Formally (our notation), Oum and Zhang explicitly define $ALH=f(N)$, which means that a variation of $N$ by $\lambda$ would cause an effect on $ALH$ such that

$$N^A \rightarrow \lambda N^A$$

$$ALH^A \rightarrow f(\lambda N^A)$$

Then they argue that $RTS$ should be calculated as
\[ \text{(22)} \]

\[ \text{RTS} = \left[ \eta_{\text{TK}} + \eta_{\text{ALH}} \cdot \eta_{N}^{\text{ALH}} + \eta_{N} \right]^{-1} \]

where \( \eta_{N}^{\text{ALH}} \) is the elasticity of \( \text{ALH} \) with respect to the network index. \( \text{RTS} \) should fulfill

\[ \text{(23)} \]

\[ \tilde{C}(\gamma^{\text{RTS} \text{TK}}, \text{ALH}(\gamma^{\text{RTS} N}), \gamma^{\text{RTS} N}) = \gamma \tilde{C}(\text{TK}, \text{ALH}(N), N) \]

This proposition seems to contribute to the solution of the problems presented in the preceding section, as \( \eta_{N}^{\text{ALH}} \), which multiplies the cost elasticity, permits a controlled representation of the true effect of the network expansion (as the authors use a log-linear form for \( f \), the elasticity is directly estimated). Nevertheless, even if the proposed \( f \) function captured exactly the variation of \( \text{ALH} \) after variations in \( N \), we would still have to face the problem of the other aggregates that have to vary in the same proportion as the network (the constant density assumption). Expanding Oum and Zhang’s procedure to those aggregates seems like an obvious way to continue. Let us examine this.

If we do not want the unpleasant conditions that showed up in the previous section, we could obtain econometrically estimated representations of the aggregates as a function of \( N \), such that the total derivative can be taken as the network expands. Thus, if the network grows by \( \lambda \), aggregates like \( \text{TK} \) and \( \text{ALH} \) would vary according to the estimated relations, i.e.

\[ N \rightarrow \lambda N \]
\[ \text{TK} \rightarrow g(\lambda N) \]
\[ \text{ALH} \rightarrow f(\lambda N) \]

Then, in order to obtain an improved version of \( \text{RTS} \), here defined as \( \overline{\text{RTS}} \), we would have to calculate

\[ \text{(24)} \]

\[ \overline{\text{RTS}} = \left[ \eta_{\text{TK}} \cdot \eta_{N}^{\text{TK}} + \eta_{\text{ALH}} \cdot \eta_{N}^{\text{ALH}} + \eta_{N} \right]^{-1} \]
which is equivalent to postulate that costs increase by, say, $\gamma$ when $N$ expands by $\gamma^{\text{RTS}}$, i.e.

\[
\hat{c}(g(\gamma^{\text{RTS}} N), f(\gamma^{\text{RTS}} N), \gamma^{\text{RTS}} N) = \gamma \hat{c}(TK(N), ALH(N), N)
\]

We can safely assume that functions $g$ and $f$ can capture correctly the actual variation of $TK$ and $ALH$ when the network expands. In this case the problems arising with the unpleasant underlying variation of the true flow vector do vanish, because the values of the aggregates have to follow what is actually happening with the components of $Y$ as $N$ varies across firms or in time. However, the extended interpretation of Oum and Zhang’s (1997) proposition that makes $\text{RTS}$ a good candidate to solve the problem of the weird behavior of $Y$ within the context of $\text{RTS}$, leaves a fundamental question unattended: What are the conclusions regarding industry structure that can be obtained with this method?

Let us formulate the new question more precisely. The original, widely used version of $\text{RTS}$ does not permit reliable conclusions regarding industry structure due to unwanted conditions imposed on the actual product. But, what can we say about $\text{RTS}$, a well-behaved index from the point of view of $Y$? Note the following: with $\text{RTS}$ we would be examining the behavior of cost as network expands, and this expansion induces, by means of the addition of new OD pairs, variations in the level of aggregates (products and attributes). With this in mind, recall that policy conclusions can be related with the estimated value for $\text{RTS}$ as follows:

- $\text{RTS} > 1$ implies that cost increases less than proportionally with network size; increasing network size is convenient.
- $\text{RTS} = 1$ implies that cost increase by the same proportion as the network; increasing network size is neutral regarding cost.
- $\text{RTS} < 1$ implies that cost increases more than proportionally with network size; a network expansion is not convenient.
Are these policy implications reasonable? Note that we are comparing network size with total cost. In many cases, it will happen that costs will increase more than the number of points served (\( RTS < 1 \), do not grow) but the expansion of the network might be cost convenient because the incremental cost (i.e. the cost of adding the vector of new flows to the line of production) is smaller than the cost of serving those new flows with a different firm. For example, when the number of points served increases from 2 to 3, the network expands by 3/2, but four new flows enter the picture, tripling the number of products previously served. Costs will very likely increase by more than 1.5 but the network expansion might well be convenient\(^6\). For synthesis, the problem of the unpleasant properties of \( Y \) behind \( RTS \) can be solved, but the “improved” version fails to properly account for its ultimate goal, namely to provide useful information for adequate policies or conclusions regarding industry structure.

V. CONCLUSIONS

In this article we have examined critically the analysis of the transport industry that has been made from econometrically estimated cost functions using aggregate output and the concept of returns to scale with variable network size. First we have seen that when the transport industry structure is analyzed in terms of the true output, \( Y \), using the degree of economies of scale, \( S \), and the degree of economies of scope, \( SC \), the conclusions are unambiguous as opposed to what happens when product is described in a synthetic manner. We have explained that even if aggregates descriptions are mandatory in the applied work (as is always the case), recognition of these aggregates as implicit functions of \( Y \) permits the correct calculation of \( S \) using the method by Jara-Díaz and Cortés (1996), which yields an improved version of returns to density as defined in the transport literature. We have seen that the concept of returns to scale (\( RTS \)) as used in the literature is related with changes in the network, which implies the addition of flows in new OD pairs to the vector of products of the firm, a process that is related with economies of scope when looked from the perspective of \( Y \). But \( RTS \) is calculated by adding the elasticity of cost with respect to the network variable. This causes the main question, which we have tried to face in this article: what exactly can be analyzed with an analytical construction such as \( RTS \). In order to answer this, a detailed analysis of this index has been done, keeping permanently an interpretation of variables, functions and expansions in terms of the detailed product, the vector of OD flows. The main conclusion has been that \( RTS, economies of scale with variable network \)
size, as defined and used in the literature, is not adequate to analyze the transport industries structure within the context of varying networks, because of at least three problems:

(a) The variable used to describe the network, whose elasticity is used to distinguish density from scale, has a non-negligible impact on what can be studied. First, the same process of production growth can be seen as density related if route miles are used, or as scale if the number of points served is used instead. Second, whether that variable is endogenous or exogenous to the firm is something that usually remains without discussion.

(b) It is not possible to make inferences regarding the variation of the true flow vector served by a firm, after a network expansion. We can only make interpretations. The few hints that can be rescued in the literature contradict each other and some are not consistent with the true (disaggregated) definition of product.

(c) Imposing a constant density, implicitly defined in the RTS equation, translates into very strong conditions on the flows in the new OD pairs, those that are added after the network expansion. As shown here, in some cases these conditions might imply that all new flows are zero. These restrictions are even more complicated the more aggregates and/or attributes are considered, because the number of simultaneous conditions that have to be fulfilled increases. This also implies that what is analyzed with RTS depends analytically, not only econometrically, on the specification of variables within the cost function.

Even if the three problems are apparently solved by, for instance, using PS as network variable, assuming that the original flows stay at the original levels, and extending Oum and Zhang’s method, the resulting scale index is essentially commanded by a synthetic relation between cost and network size, and its interpretation becomes completely inadequate when making inferences on the optimal transport industry structure.

We believe that what has prevented the detection of these problems has been the failure to think in terms of the true product Y. This definition has not been questioned in the literature. In fact, within the last two decades many studies have emphasized that the main reason to use aggregates have been both lack of data and the huge dimension of the resulting product vector. However, not being able to estimate cost functions empirically with product precisely described, does not imply
that we should forget or abandon such definition when using those function to make economic inferences. We think that this is the main problem: aggregates are seen and treated as well-defined entities with names such as “generic product”, “attributes” and so on. They are in fact synthetic representations of what a transport firm produces, but they are given properties that do not necessarily hold. Aggregates are necessary, and we are not challenging their use for econometric purposes. What we are defending is a rigorous view of transport production, which means, among other things, that the interpretation of cost functions for the purpose of industry structure analysis should be done in terms of the true product. This is something we have shown to be both feasible and rewarding as it permitted to show that scale analysis can be done from the aggregates (Jara-Díaz and Cortés, 1996) while RTS can not be rescued in a similar way.

The obvious remaining question is how to analyze transport industry structure within the context of varying networks, which was the intention behind RTS. We think the answer is direct: economies of spatial scope, which is a concept that can be used very easily to identify the optimal spatial structure of the industry when product is correctly understood. However, the real challenge is how to do this from cost functions that include aggregates to describe production, as they are the only feasible outcome form empirical studies. This is indeed a key aspect in a research agenda, where encouraging results –in terms of methodology, applicability and explanatory power– are being obtained (Jara-Díaz et al., 2001; Basso and Jara-Díaz, 2003).

RTS has been widely used in case studies and estimates have been used for policy purposes broadly in the last twenty years. The fact that RTS did not do a good job in forecasting or explaining some observed industry behavior, as network growth, has now an explicit explanation. Certainly, once the new methods to replace RTS are discussed and established, policy conclusions using RTS will have to be re-examined.
APPENDIX A

Proposition: equation (3) for RTS can be obtained from

\[ \tilde{C}(\gamma^{RTS} \tilde{Y}_h, \tilde{Y}_k, \gamma^{RTS} N) = \gamma \tilde{C}(\tilde{Y}_j, \tilde{Y}_k, N) \]  

Proof: deriving both sides of (A.1) with respect to \( g \)

\[ R^{RTS-1} \left[ \sum_{h \in H} \frac{\partial \tilde{C}(\gamma^{RTS} \tilde{Y}_h, \tilde{Y}_k, \gamma^{RTS} N)}{\partial \gamma^{RTS} \tilde{Y}_h} \tilde{y}_h + \frac{\partial \tilde{C}(\gamma^{RTS} \tilde{Y}_h, \tilde{Y}_k, \gamma^{RTS} N)}{\partial \gamma^{RTS} N} N \right] = \tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N) \]

Doing the same with respect to \( \tilde{y}_h \) with \( h \in H \), multiplying times \( \tilde{y}_h \) and adding over \( h \)

\[ \gamma^{RTS-1} \sum_{h \in H} \frac{\partial \tilde{C}(\gamma^{RTS} \tilde{Y}_h, \tilde{Y}_k, \gamma^{RTS} N)}{\partial \gamma^{RTS} \tilde{Y}_h} \tilde{y}_h = \sum_{h \in H} \frac{\partial \tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N)}{\partial \tilde{y}_h} \tilde{y}_h \]

Taking the derivative of both sides of (A.1) with respect to \( N \) and multiplying times \( N \)

\[ \gamma^{RTS-1} \frac{\partial \tilde{C}(\gamma^{RTS} \tilde{Y}_h, \tilde{Y}_k, \gamma^{RTS} N)}{\partial \gamma^{RTS} N} N = \frac{\partial \tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N)}{\partial N} N \]

Adding (A.3) and (A.4) we obtain

\[ \gamma^{RTS-1} \left[ \sum_{h \in H} \frac{\partial \tilde{C}(\gamma^{RTS} \tilde{Y}_h, \tilde{Y}_k, \gamma^{RTS} N)}{\partial \gamma^{RTS} \tilde{Y}_h} \tilde{y}_h + \frac{\partial \tilde{C}(\gamma^{RTS} \tilde{Y}_h, \tilde{Y}_k, \gamma^{RTS} N)}{\partial \gamma^{RTS} N} N \right] = \sum_{h \in H} \frac{\partial \tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N)}{\partial \tilde{y}_h} \tilde{y}_h + \frac{\partial \tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N)}{\partial N} N \]

from which, after dividing (A.2) into (A.5), we can solve for RTS as

\[ R^{Ts} = \frac{\tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N)}{\sum_{h \in H} \frac{\partial \tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N)}{\partial \tilde{y}_h} \tilde{y}_h + \frac{\partial \tilde{C}(\tilde{Y}_h, \tilde{Y}_k, N)}{\partial N} N} = \frac{1}{\sum_{h \in H} \tilde{\eta}_h + \eta N} \]

Q.E.D.

APPENDIX B

Let \( PS_A = PS \). Then \( PS_L = PS + 1 \) and \( \lambda = (PS+1) / PS \). Then \( T^A \) and \( T^L \) can be written as
\[ T^A = \sum_{i=1}^{PS} \sum_{j=1}^{PS} y_{ij} \quad T^L = \sum_{i=1}^{PS+PS+1} \sum_{j=1}^{PS} y_{ij} \quad \text{with} \quad j \neq i \]

As \( T^L = \lambda T^A \) and the original flows do not vary, then

\[
\sum_{i=1}^{PS+PS+1} \sum_{j=1}^{PS} y_{ij} = \sum_{i=1}^{PS} \sum_{j=1}^{PS} y_{ij} + \sum_{i=1}^{PS} y_{i(PS+1)} + \sum_{i=1}^{PS} y_{i(PS+1)} = \left( \frac{PS+1}{PS} \right) \cdot \sum_{i=1}^{PS} \sum_{j=1}^{PS} y_{ij} \quad \text{with} \quad j \neq i \]

(B.1)

\[
\Rightarrow \sum_{i=1}^{PS+1} y_{i(PS+1)} + \sum_{i=1}^{PS} y_{i(PS+1)} = \left( \frac{PS+1}{PS} - 1 \right) \sum_{i=1}^{PS} \sum_{j=1}^{PS} y_{ij} \]

The 2-PS new flows are added on the left hand side, while the PS (PS-1) original flows are added on the right hand side. If \( y \) is the average of the original flows and \( y' \) is the average of the new ones, then equation (a) can be re-written as

(B.2)

\[
2 \cdot PS \cdot y' = \left( \frac{PS+1}{PS} - 1 \right) \cdot PS \cdot (PS - 1) \cdot y
\]

from which the ratio \( y' / y \) is obtained as

(B.3)

\[
y' = \frac{1}{2} \cdot \frac{PS - 1}{PS}
\]

REFERENCES


Endnotes

1 Summaries of published estimates can be found in Tae H. Oum and Willam G. Waters (1996), Ronald R. Braeutigam (1999) and Eric Pels and Piet Rietveld (2000).

2 For example, David W. Gillen et al. (1990, p.13) state that “ideally one would treat every type of airline service in every city pair market as a separate output. The impossibility of estimating a cost function with thousands of outputs requires some aggregation of the data. After aggregation, output attribute variables are introduced to control for some of the aggregation bias”. Braeutigam (1999, p.68) states that “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”. Other examples are Winston (1985, p.64-65), Andrew Daughety et al. (1985, p.71) and Kenneth A. Small (1992, p.50).

3 In every empirical study we are aware of, the estimated network elasticity has been positive. Since in many cases flexible functional forms are used, it might happen that in some points away from the deviation point, the network elasticity turns out to be negative. In any case, a nil or negative network elasticity of cost would be something somewhat unexpected. Authors often say the network elasticity has the expected sign when they obtain a positive estimate.

4 For instance, James P. Keeler and John Formby (1994) used the ratio between revenue passenger enplanements and PS as an explicit measure of traffic density, which they introduced directly in the cost function.
This way of calculating $RTS$ have been justified by some authors who explain that, since the network is allowed to change, it is possible that the length of haul also changes. See for example Caves et al. (1985).

By constructing a transport cost function analytically from the technology, Jara-Díaz and Basso (2003) are actually able to develop this case: the network expands from 2 to 3 points served, minimum cost increases more than four times ($RTS < 1$), but the network expansion is cost convenient because of better fleet utilization.