Beyond the Mohring effect: scale economies induced by transit lines structures design.

Andrés Fielbaum, Sergio Jara-Diaz and Antonio Gschwender
Universidad de Chile and Instituto Sistemas Complejos de Ingeniería (ISCI)

Abstract

In this paper we study how the spatial arrangement of transit lines (lines structure) influences scale economies in public transport. First we show that the degree of scale economies (*DSE*) increases discretely whenever passenger volume induces a change in lines structure. The technical elements behind this are explained by using a new three-dimensional concept called directness, encompassing number of transfers, number of stops and passenger route lengths. This is first exemplified in a simple ad-hoc network, and then applied to examine the structural changes that occur in the design of transit lines in a fairly general representation of a city. We show that directness increases whenever lines structure changes as a response to larger demand volumes - increasing *DSE* at the particular value of flow where this change occurs - because systems with more direct lines for each OD pair diminish in-vehicle times while increasing waiting times mildly, such that users are benefited by lower travel times and operators are benefited by lower idle capacity. After the change, however, *DSE* decreases within the demand range where the new line structure is maintained, just as in the one line model. The possibility of deciding the line structure introduces directness as a new source of economies of scale which are finally exhausted after full directness is achieved.

Keywords: public transport; scale economies; lines structures; directness

1. Introduction

Cost functions and economies of scale are economic concepts that are quite relevant for the normative analysis within production theory, including industry structure and optimal pricing policies. Behind cost functions lies the technical process of conversion of inputs into outputs, such that cost functions can capture scale economies (a technical property indeed). In transport, the main technical elements are frequencies, vehicle sizes and the organization of lines in space. In this paper we aim at understanding the relations behind the third design element and scale economies in public transport.

The provision of public transport services exhibits various technical characteristics that have been shown to affect its degree of scale economies. First of all, the so-called Mohring effect, where an increase in patronage makes optimal frequency larger and waiting times lower. Mohring (1972) found the frequency of the service to be proportional to the square root of the demand when only this effect is modeled in an isolated public transport line. In addition to this waiting time effect, as demand increases the system can also be adapted by incorporating new lines, thus reducing another component of users' cost, namely the walking time. This has been modeled for a bus feeder system by Hurdle (1973), in a rectangular area by Kocur and Hendrickson (1982) for a single period, by Chang and Schonfeld (1991) for multiple periods, and by Small (2004), who analyzed the impact of road pricing on public transport. All of them obtain a cube root formula for both the optimal frequency of each line and for the optimal number of routes.

A third variable that can be adapted according to the demand level is the size of the vehicle, which also increases with patronage. As operators' cost per passenger diminish with vehicle size (due to fixed costs per vehicle), this is also a source of scale economies. However, when vehicles size increases the time spent at each stop also increases because more passengers board to, and alight from, each single vehicle, thereby increasing cycle time - which affects operators' cost as a larger fleet is needed - and users' in-vehicle time. Both effects reduce the degree of scale economies. Including these effects in his model of an isolated public transport line, Jansson (1980) obtained a modified square root formula for optimal frequency. In all models the adjustment of frequency and vehicle size generates scale economies that, nevertheless, diminish as flow increases.

An important element of design that responds in a discrete way to increases in flow is lines structure, i.e. the way in which vehicles serve a number of routes in order to move a given set of flows (product). Such a structure can be optimized together with fleet and vehicle sizes, admitting many possible arrangements in space, with public transport lines organized as, for example, cyclical, hub-and-spoke, feeder-trunk or direct services. As flows grow these arrangements might evolve in a way that should be studied specifically; understanding the evolution of design including lines structure and analyzing its impact on total costs and scale economies is the main objective of the paper. Considering operators' costs only, Basso and Jara-Díaz (2006b) study the difference in the analysis of scale economies when lines structures are fixed or a variable to be optimized. Kraus (2008) formulates the problem including users and operators' costs over a cost minimizing network, which in public transport would imply that users choose system optimal routes rather than individually optimal ones. In this paper we analyze scale economies looking at the evolution of lines structure design as total flow grows considering total costs and recognizing that users choose individually optimal routes. The main conclusions are that changes in lines structure induce scale economies at the particular value of total flow where this change occurs; that the technical elements behind this are the reduction of stops, transfers and route lengths; and that vehicle sizes and frequencies grow as well, as in single line models.

The paper is organized as follows. In the remainder of this Section we summarize the various ways in which scale has been studied in transport. Section 2 contains a discussion of what it means to introduce lines structure in scale analysis, showing that the degree of scale economies (*DSE*) - the ratio between average and marginal cost - increases discretely whenever lines structure changes as a response to a continuous proportional increase in flows. In Section 3 we use a simple network to illustrate this property and to introduce the multi-dimensional concept of directness that helps describing the evolution of lines structure as flows grow. This concept is used to present, in Section 4, a more general case that rests on a parametric description of a city; most importantly, the technical elements that help explaining the change in lines structure as patronage grows are presented in detail. Section 5 concludes, emphasizing the role of directness in scale economies analysis of public transport systems.[1]

Although transport processes usually involve many inputs and outputs, the engineering technology has been usually formulated using aggregates, where product was described using single scalar measures as ton or passenger-miles until mid-eighties, and by means of a vector of a very small dimension thereafter, including flows related variables, service quality variables and network description variables. The compact description of output prompted two definitions in the literature around the analysis of scale economies, both referring to proportional expansions of output: returns to density (called RTD) and returns to scale with variable network size (called RTS). The former considered a proportional expansion of outputs keeping network size fixed, while the latter considered a simultaneous expansion of both flows and the network by the same proportion (Caves *et al*., 1984; Keaton, 1990). However, using aggregate output descriptions blurs the technical relations with inputs and has some unpleasant consequences in the analysis of economies of scale in transport activities.

Behind any compact description of transport output lays the true output of any transport firm: a vector of origin-destination (OD) flows of different things during different periods (Jara-Díaz, 1982a). In very simple transport systems the analytical derivation of the technical relations between inputs and flows - the production function – can be done, such that the corresponding cost functions can be obtained analytically as well.[2] This approach

[1] Scale economies in public transport have also been reported in other dimensions. Tirachini *et al*. (2010a), for example, show that when crowding discomfort is considered diseconomies of scale are found for high levels of patronage, a result that vanishes when more than one line is considered (Tirachini *et al*., 2010b). Tirachini and Hensher (2011) and Jara-Díaz and Tirachini (2013) have studied the impact of the boarding-alighting-paying methods, finding yet another source of economies of scale. Considering different modes also impacts the analysis, as shown by Tirachini and Hensher (2012) or Basso and Jara-Díaz (2010, 2012).

[2] See for example the analysis of the backhaul transport system involving two flows only (Jara-Díaz, 1982b) or the three-nodes system studied by Jara-Díaz and Basso (2003) involving a discrete decision regarding the spatial arrangement of the vehicles (service structures).

proved very useful to show that the use of aggregates introduced ambiguity in the economic analysis in transport because, for example, the same amount of passenger-miles could require very different types and amounts of inputs depending on how this passenger-miles are distributed in space. Most importantly, scale economies should be studied holding the origin-destination system constant, as introducing new OD pairs means introducing new products, which would require the analysis of economies of scope; this means that "economies of scale with variable network size" is actually an ill-defined concept, as shown by Basso and Jara-Díaz (2006a) while "economies of density" is better suited to the definition of economies of scale.[3] A corollary from this story is that more attention has to be paid to the transport production process itself in order to fully understand scale economies. This is the main objective of this paper.

2. The impact of the discrete nature of lines structure choice on *DSE*.

In this section we analyze the general relation between the adjustment of lines structures and scale economies in transit networks. Let us formally define a "lines structure" as a set of spatially organized transit routes that operates on a given network serving all flows. A simple example is shown in Figure 1, where a three nodes network (**a**) - with potentially six OD pairs - can be served in different ways, such as a single line running counterclockwise (**b**), or with two lines each one circulating between two nodes (**c**). How to decide which lines structure is best for a given origin-destination (OD) flow matrix $\boldsymbol{Y}$? In transport production this is part of the search for the optimal input combinations that yield the minimum total cost, so choosing the best structure has to be done together with other design variables like frequencies and vehicle sizes in order to find the smallest value of the resources consumed (VRC) provided those design variables are technically able to produce $\boldsymbol{Y}$; for short, a cost function has to be obtained, which requires finding the optimal input demand functions depending on product and input prices, noting that the VRC includes all resources, i.e. operators' and users'. This requires a certain procedure which we now summarize for a general case.

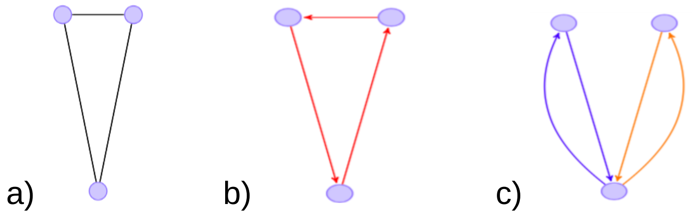


Figure 1. Network (a), and two alternative lines structures.

[3] Sometimes RTD has been defined adding the condition that route structure is unchanged after an increase in flows (Basso and Jara-Díaz, 2006b).

Consider a physical network (e.g. streets of a city) and a given OD matrix of flows $\boldsymbol{Y}$. For this setting, each candidate lines structures E_i is composed of a series of transit lines that, altogether, are capable to serve all trips[4]. In turn, each of the lines l that form a lines structure has to be assigned a frequency f_l and a vehicle size K_l. In order to find the optimal values of these variables for a given lines structure i composed of m_i lines, one has to minimize the total value of the resources consumed VRC_i, that depends on the set of frequencies $(f_1, \dots, f_{m_i}) = \boldsymbol{f}$ and vehicle sizes $(K_1, \dots., K_{m_i}) = \boldsymbol{K}$ of all lines in structure i provided $\boldsymbol{f}$ and $\boldsymbol{K}$ can carry flows $\boldsymbol{Y}$. VRC_i can be expressed as the sum of the resources consumed by operators VRC_{iO} and users VRC_{iU}, i.e.

$$VRC_i(\boldsymbol{f}, \boldsymbol{K}, \boldsymbol{Y}) = VRC_{iO}(\boldsymbol{f}, \boldsymbol{K}) + VRC_{iU}(\boldsymbol{f}, \boldsymbol{K}, \boldsymbol{Y}) \qquad (1)$$

VRC_i is a function of $\boldsymbol{Y}$ directly because users' costs increase with $\boldsymbol{Y}$. The optimal values of $(\boldsymbol{f}, \boldsymbol{K})$ for a given $\boldsymbol{Y}$, denoted $(\boldsymbol{f}^*, \boldsymbol{K}^*)$, are those that minimize VRC_i subject to technical feasibility constraints, as explained in Appendix A. Then solving (1) we get $\boldsymbol{f}^* = \boldsymbol{f}^*(\boldsymbol{Y})$ and $\boldsymbol{K}^* = \boldsymbol{K}^*(\boldsymbol{Y})$. When these optimal values are plugged back into VRC_i one obtains the conditional cost function $C_i(\boldsymbol{Y}) \equiv VRC_i(\boldsymbol{f}^*(\boldsymbol{Y}), \boldsymbol{K}^*(\boldsymbol{Y}), \boldsymbol{Y})$ as defined in Jara-Diaz (2007), i.e. the minimum VRC to serve flow $\boldsymbol{Y}$ for a given lines structure i. Finally, the best lines structure for each $\boldsymbol{Y}$ is given by the $argmin_i C_i(\boldsymbol{Y})$. This way the optimal lines structure for a given $\boldsymbol{Y}$ is found together with the optimal frequencies and vehicle sizes that were found in the previous step. Note that $f_l^* = f_l^*(\boldsymbol{Y})$ and that, for some $\boldsymbol{Y}$ values, some frequencies can become nil and some can become positive.

In equation (1) it is assumed that operators' costs increase with frequencies and bus sizes, while users' costs decrease. For the proposition below, these expressions can be general; we only need them to be differentiable and to include at least one component inversely related with frequency. For the sake of clarity, in what follows we will use waiting cost as representative of this component[5].

The question we want to address is whether adjusting line structures contributes to scale economies in transport networks. First, we have to recall that the degree of scale economies DSE is defined as the ratio between average and marginal costs such that there are scale economies iff $DSE > 1$. The proposition formulated and proved below states that the DSE increases when the lines structures changes (unless common lines

[4] In practice, the total number of possible lines structures is huge and cannot be obtained. Nevertheless, this is not needed for the analysis in this section. In sections 3 and 4, where specific networks are analyzed, we will work with a set of pre-conceived lines structures.

[5] There are more components of VRC inversely related with frequency, as in-vehicle users' cost (because of time at stops and crowding) or bus size-related operators' cost (Jara-Díaz and Gschwender, 2009; Hörcher and Graham, 2018).

exist everywhere[6]). The proof will be based on the discrete change in lines structure when passengers choose their routes minimizing their individual costs (an example using the network in Figure 1 is offered and analyzed in detail in the next Section). If we are using a set of predefined lines structures, this change is obviously discrete (and the first part of the proof below is not necessary); if all the lines are always candidates to appear, the crucial fact is that nobody will wait for a line with a frequency that is extremely low when there are no common lines.

Let us define a vector of OD flows $\boldsymbol{q}$ as a *threshold point* if there exists at least one line l such that optimal frequency $f_l^*(\boldsymbol{q}) = 0$ and $f_l^*(\boldsymbol{q} \cdot (1+\varepsilon)) > 0 \; \forall \; \varepsilon > 0$ with no common lines for some of its passengers (see footnote 6). This means that when $\boldsymbol{Y}$ just exceeds $\boldsymbol{q}$ at least one new line appears because it minimizes VRC and becomes best for some users[7]. Note that a vector $\boldsymbol{Y}$ at which a change in lines structure occurs is, by definition, a *threshold point*, in which new routes emerge for some users (and other inferior ones can disappear).

The main idea behind the proposition that follows is that the emergence of a new line (i.e. a line whose optimal frequency goes from zero to a positive value) triggers an upward jump in the *DSE*.

Proposition: Consider a network served by a public transport system. Then at every threshold point the DSE increases discretely, i.e. $\underset{\varepsilon \to o^+}{lim} DSE(\boldsymbol{q}_{\boldsymbol{\varepsilon}}) > \underset{\varepsilon \to o^-}{lim} DSE(\boldsymbol{q}_{\boldsymbol{\varepsilon}})$, with $\boldsymbol{q}_{\boldsymbol{\varepsilon}} = \boldsymbol{q} \cdot (1+\varepsilon)$.

Proof: the proof has two parts.

1) First, let us show that when some f_l^* becomes strictly positive due to a growth of $\boldsymbol{Y}$ at $\boldsymbol{q}$, it increases in a discontinuous way from zero. Define f_0 such that if f_l<f_0, then the waiting cost for passengers using line l will induce a travel time cost that is larger than the current total user cost of any other route in the system. In that case no passenger chooses l and the optimal frequency is zero; in other words, f_l will never be in the interval $(0, f_0)$, but jumps in a discontinuous way from 0 to some positive value $f_l^* \geq f_0$.

[6] In the literature the case known as "common lines" appears when for some portions of the route, the passenger is indifferent to choose within a certain set of lines because they all make almost the same trip. Using Figure 1 as an example, if both line structures (b and c) coexist, passengers travelling from the bottom node to the upper-right one could use the line of structure (b) or the right line of structure (c). For them, both are common lines. Passengers travelling from the bottom node to the upper-left one could face common lines or not, depending on the relative frequencies and travel times of the line of structure (b) and the left line of structure (c). For a precise definition, see Chriqui and Robillard (1975) or Cominetti and Correa (2001).

[7] The potential adaptation of route structure following a growth in flows is at "the kernel of transport production; changes in the flow vector $\boldsymbol{Y}$ potentially induce changes in input usage as well in route structures and operating rules in general" (Jara-Diaz (2007).

As an example, suppose that the “south” node is the only origin in Figure 1. If we only had the red long line from Figure 1b, and the blue line from Figure 1c was added (starting from a different bus stop, such that users have to decide which line to take in advance), this blue line will be used by those passengers whose destination is the upper-left node only if its frequency is large enough to compensate for waiting plus the longer in-vehicle time if using the red line.

What we have just shown is that when a change in line structure occurs, there is a discrete jump in the value of at least one frequency: the choice of an optimal line structure is essentially discrete. Then the production of flows can be represented by a function that involves the choice of a discrete variable (the lines structure i, indicating which lines are present in the system) and several variables that are continuous for a given discrete variable (such as the frequencies and vehicle capacities of each of the lines in i).

The second part of the proof applies to any kind of cost function that results from the optimal choice of a discrete variable Z (in our case the lines structure i) and several variables $\boldsymbol{X}$ (in our case the frequencies and vehicle capacities of each of the lines in i) that are continuous for a given Z. As scale analysis deals with expansions of output vector $\boldsymbol{Y}$ along a ray $\mu\boldsymbol{Y}$, with $\mu \geq 1$ (Baumol et al, 1982), it is equivalent to a single product like analysis. Then, in the rest of the paper, $\boldsymbol{Y}$ will be treated as a scalar given by the total sum of the flows, i.e. $Y = \sum y_i$; note that this means that $\mu\boldsymbol{Y} = \sum \mu y_i = \mu Y$.

2) Consider $\boldsymbol{X}^*, Z^*$ optimal to produce Y. Consider Y_0 such that any increase leads to a change from $Z^* = Z_1$ to $Z^* = Z_2$ (equivalent to a threshold point above). Then $\lim\limits_{\varepsilon \to o^+} DSE(Y_0 + \varepsilon) > \lim\limits_{\varepsilon \to o^-} DSE(Y_0 + \varepsilon)$.
To prove this, consider the conditional cost functions C_1 and C_2 associated to Z_1 and Z_2 respectively obtained by optimizing $\boldsymbol{X}$ only given Z_i (in our case this means optimizing frequencies and vehicle sizes for a given lines structure). Let us look at the average and marginal costs for C_1 and C_2 at Y_0. As C_1 and C_2 are continuous functions, then $C_1(Y_0)/Y_0 = C_2(Y_0)/Y_0$. Regarding the marginal costs, the derivative of the cost with respect to Y verifies $\frac{\partial C_2}{\partial Y} < \frac{\partial C_1}{\partial Y}$, because C_2 becomes lower than C_1 when Y grows. As average costs are equal and marginal costs are lower for C_2, it is direct to conclude that the ratio between average and marginal cost, i.e DSE, increases.

$$\lim_{\varepsilon \to o^+} DSE(Y_0 + \varepsilon) = \lim_{\varepsilon \to o^+} DSE_2(Y_0 + \varepsilon) = DSE_2(Y_0) > DSE_1(Y_0) = \lim_{\varepsilon \to o^-} DSE_1(Y_0 + \varepsilon) = \lim_{\varepsilon \to o^-} DSE(Y_0 + \varepsilon)$$

Q.E.D.

The Proposition is represented in Figures 2, where average and marginal costs for C_1 and C_2 are shown. At the exact point where the two average costs coincide (i.e. where the optimization process induces a change from Z_1 to Z_2), a black arrow shows that a) the marginal cost is lower for Z_2 and b) the global DSE increases discretely.

As a conclusion, the structural design of public transport systems involves variables as frequency f (and the associated fleet B), density D and vehicle capacity K, that can be treated (or approximated) as continuous[8], and lines structure, which has a discrete nature and introduces technical novelties that are worth studying. In the following Section we will introduce a multidimensional concept that helps analyzing the technical relations between lines structure and scale economies.

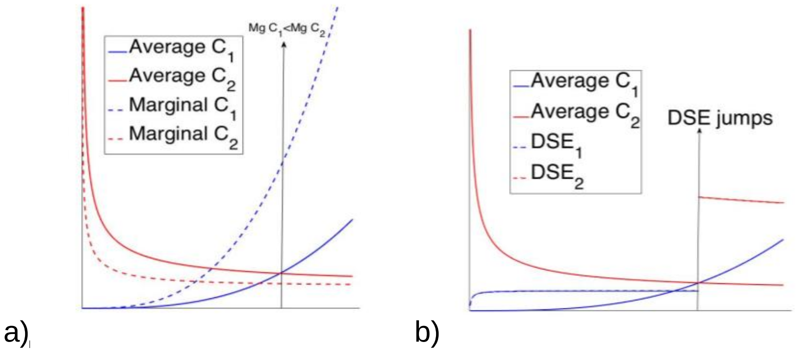


Figure 2. Change in *DSE* due to (discrete) change in lines structure

3. Directness in lines structures: a source of scale economies.

3.1 Introducing directness.

We have proved that changes in line structures always lead to a discrete (local) increase in scale economies. This general result, however, says nothing about what exactly are the transport-related technical elements that help understanding what lies behind this. We do know from the literature that increasing demand induces higher frequencies, larger vehicles and an increase in the density of lines. As a result, waiting, access and egress times diminish (scale economies) while in-vehicle and cycle times increase (scale diseconomies). What is the equivalent technical effect that links overall demand with lines structure and scale economies? And how do scale economies behave once a change in lines structure has occurred?

These are quite complex questions because frequency, vehicle size or lines density can be represented each by a single, well-defined variable, whereas a lines structure, e.g. feeder-trunk or hub-and-spoke, can be conceptually described with some precision by a generic description but cannot be represented by a single variable. Further, changes in line structures are not continuous but discrete, occurring at some specific levels of total

[8] Frequency and capacity are discrete variables, although the former can be fractional when looked at on a per hour basis. The latter, though, is constrained by commercially available vehicles.

patronage. Both elements not only increase the mathematical complexity of the associated optimization problem, but also add new challenges to scale economies analyses.

Generally speaking the literature on lines structures in the last fifteen years shows that, for low levels of overall demand distributed in space, those structures involving transfers tend to be appropriate, e.g., hub-and-spoke or feeder-trunk systems[9]. As patronage increases, lines get organized along the idea of routes that follow more closely the origin-destination pattern avoiding transfers, increasing what can be called "directness", such that each new passenger generates positive externalities on the rest of the passengers because a) transfers diminish, b) distances travelled diminish, and c) number of stops diminish[10]. Element a) has a clear positive impact on users, b) diminishes in-vehicle-time for all, and c) diminishes in-vehicle-time for users and cycle time for operators. So these elements seem to contribute to increase the DSE through the reduction of average users' costs, but all effects should be analyzed. In order to represent directness in a more precise way, we propose the following three (continuous) indices: average transfers required per trip, average stops required per trip (including the extremes) and the average across all passengers of the ratio between their traveled distance and the length of the shortest path that link their origin and destination (relative distance). Note that these flow-related indices can also be defined as averages across OD pairs, such that these new "network indices" can be calculated irrespective of the assignment of flows.

The concept of directness has an extreme case in non-stop services (which have been called exclusive in previous papers), where each OD pair is served by one line only, providing a service similar to a private car but with lower operating costs per passenger and larger waiting times. From this viewpoint, as directness increases the number of passengers with different origins and destinations sharing the same vehicle diminishes. It is worth noting that a connection between patronage and directness has emerged in the transit network design literature. For example, Fielbaum *et al*. (2018) studied the heuristics proposed by Dubois *et al*. (1979) and Ceder and Wilson (1986), that are built around direct services: both create spatial arrangements of lines depending on a parameter σ that controls the maximum admissible deviations from the shortest paths, i.e., represents exactly the trade-off between more directness ($\sigma = 0$) and bus-sharing; σ is inversely related with directness. When searching for the best σ, Fielbaum *et al*. (2018)

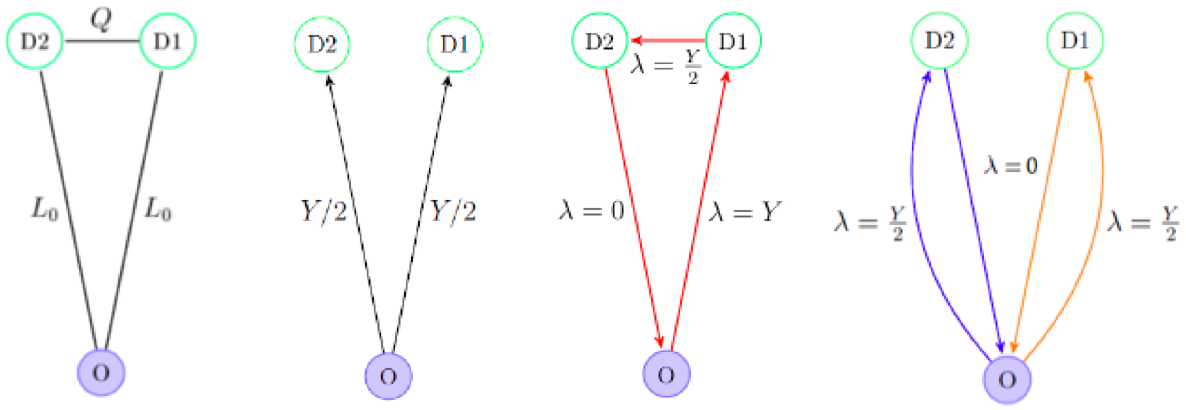
[9] Gschwender *et al*. (2016), for example, study a Y-shaped city. They show that as the patronage increases, the optimal structure changes in one of the following ways (depending on trip distribution): from No transfers to No stops, from Feeder-trunk to No stops, or - the only odd case - from No transfers to Feeder-trunk. Daganzo (2010) studies a grid city served with direct lines within an internal region and with hub and spoke from the external region, optimizing the size of the internal region; he shows that the larger the patronage, the larger the zone served with direct lines (internal region). Badia *et al*. (2014) extend the paper by Daganzo (2010) and this conclusion remains valid; also, the set of lines becomes denser when the number of passengers increases.

[10] This is an extension of the concept of OD-directness originally defined by Laporte *et al*. (2011) on the lines network as the fraction of the OD-pairs that can be joined without transfers.

found systematically that small values for this parameter were optimal when patronage was large, i.e. increasing directness was the optimal response to demand increases.

3.2 An illustrative model

In order to illustrate in a simple way what has been discussed above, let us consider the network we introduced in Section 2 (Figure 1), with network and flow characteristics as represented in Figure 3, where two destinations are located at the same distance L_0 from a single origin, forming an isosceles triangle; the distance between the destinations is Q (Figure 3a). The total number of passengers in the system is Y – half on each OD pair as represented in Figure 3b – and the question is whether it is better to have only one line carrying all the passengers (full bus sharing, Figure 3c), or two lines, one for each destination (full direct, Figure 3d); λ represents the load of the lines on each directed arc. The directness indices are shown in Table 1 (note that in this case the flow indices and the network indices coincide, as there is only one flow assignment option).



(a) (b) (c) (d)

Figure 3. Network (a), transport demand (b) and alternative service structures: one shared line (c) and two direct lines (d).

SERVICE STRUCTURE DIRECTNESS INDICES	Bus-Sharing	Direct
Number of transfers	0	0
Number of stops	2.5	2
Distance traveled/Minimum distance	$1 + \frac{Q}{2L_0}$	1

Table 1. Indices of directness for the alternative service structures.

Let us represent the value of the resources consumed necessary to serve total flow Y by each system as

$$VRC = B(c_0 + c_1 K) + p_w Y t_w + p_v Y t_v \ , \tag{2}$$

where p_v and p_w are the values of in-vehicle and waiting times respectively, and the parameters c_0 and c_1 define the operators' cost per bus.

Model (2) follows Jansson (1980), Jara-Díaz and Gschwender (2009) and Fielbaum et al (2016), among others; it has been shown to be enough to capture the most relevant aspects of public transport costs in a system without transfers. In this approach, fleets, capacities, waiting times and in-vehicle times for each of the two systems can be expressed as functions of the corresponding frequency f - that becomes the (only) design variable to be optimized -, the vehicle speed V, boarding-alighting time t, Y, Q and L_0. A simple analysis (in Appendix B) yields the expressions shown in Table 2. Note that in the two-lines case lines are symmetric and exhibit the same frequency.

	One line (Bus-sharing)	**Two lines (Direct)**
Bus capacity K	Y/f	$Y/2f$
Fleet B	$\frac{f(2L_0+Q)}{V} + 2tY$	$\frac{4fL_0}{V} + 2tY$
Waiting time t_w	$Y/2f$	$Y/2f$
In-vehicle time t_v	$\frac{1}{2}\left(\frac{L_0}{V} + \frac{1}{4f}tY\right) + \frac{1}{2}\left(\frac{L_0+Q}{V} + \frac{3}{4f}tY\right)$	$\frac{L_0}{V} + \frac{tY}{4f}$

Table 2. Elements of the alternative service structures as a function of frequency.

Replacing the respective functions from Table 2 into equation (2), the optimal frequencies are obtained from the first order conditions (as shown in Appendix B). Both optimal frequencies and capacities are shown to increase with Y (as in Jansson, 1984), such that the scale effects (explained in section 1) are preserved. By plugging optimal frequencies back into VRC we obtain the cost function C_i for each system:

$$C_1 = 2\sqrt{\frac{c_0(2L_0+Q)}{V} Y\left(2c_1 tY + \frac{p_w + p_v tY}{2}\right)} + 2c_0 tY + \frac{c_1 Y(2L_0+Q)}{V} + \frac{p_v Y}{2}\frac{(2L_0+Q)}{V} \tag{3}$$

$$C_2 = 2\sqrt{\frac{c_0 4L_0}{V} Y\left(c_1 tY + \frac{p_w + p_v tY/2}{2}\right)} + 2c_0 tY + \frac{2c_1 Y L_0}{V} + p_v Y \frac{L_0}{V} \tag{4}$$

Note that C_1 and C_2 can be written as $C_i(Y) = \sqrt{\alpha_i Y^2 + \beta_i Y} + \varepsilon_i Y,$ with $\alpha_1 > \alpha_2,\ \varepsilon_1 > \varepsilon_2$ and $\beta_1 < \beta_2$. For high values of patronage, α and ε dominate, such that C_2 is smaller, i.e. the two-lines structure (full directness) is better; shorter routes are good for both users (through p_v) and operators (through c_1). On the other hand, when Y is small, β dominates, such that the system with only one line (full bus sharing) is better due to the lower waiting times (through p_w). The average costs resulting from C_1 and C_2 are shown in Figure 4a using the parameters shown in Appendix B. DSE is represented in Figure 4b for each system, with the solid lines representing DSE for the optimal structure.

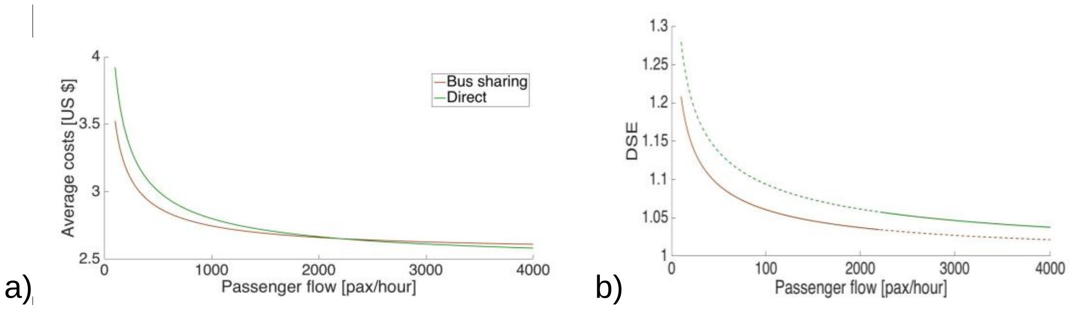


Figure 4. Average costs (a) and *DSE* (b) for Bus-sharing and Direct services.

The general property advanced in Section 2 and Figure 2 emerges very clear: the DSE "jumps" when Y reaches a certain volume that makes the direct lines superior, which is explained because of more direct routes and fewer stops. What about DSE after the lines structure changes? Using the short notation introduced above DSE can be expressed as

$$DSE_i = 1 + \frac{\beta_i}{2\alpha_i Y + \beta_i + 2\varepsilon_i Y\sqrt{\alpha_i + \beta_i / Y}} \qquad (5)$$

This expression shows that economies of scale are always present, but $\lim\limits_{Y\to\infty} DSE = 1$, suggesting that the positive externalities induced by each of the elements that constitutes "directness" in this model get exhausted in spite of the upward jump in DSE induced by the change in lines structure: eventually everybody travels along the shortest possible route and with no intermediate stops[11].

4. Directness and scale economies in a representative urban setting.

Transit systems can be spatially organized in many (and complex) ways. In order to visualize the technical elements that intervene in the relation between lines structure and scale economies, a better representation of the underlying spatial setting is needed such that lines could be structured following many possible arrangements. To do this we will

[11] This refers to scale economies induced by directness. If the number of passengers gets too large, new sources of economies (or diseconomies) of scale might emerge, such as congestion or a change in technology (e.g. metro).

apply the lines structure analysis by Fielbaum *et al*. (2016) over the simplified parametric urban model introduced by Fielbaum et al (2017) shown in Figure 5, where trips go from n peripheries P (that only generate trips) to both the CBD and the n subcenters SC. There are also trips from the subcenters to other subcenters and to the CBD. In other words, peripheries only generate trips and the CBD only attracts trips, representing a simplified morning peak situation. The proportions of total trips Y departing from the peripheries and from the sub-centers, and the proportions going from the peripheries to the CBD, own sub-center and other sub-centers are treated parametrically, such that all types of cities can be represented (monocentric, polycentric and dispersed).

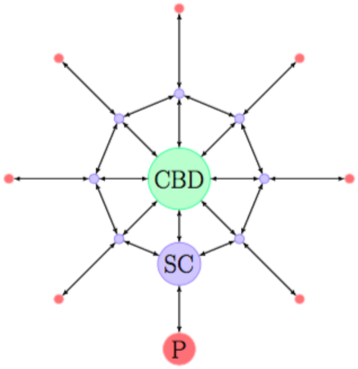
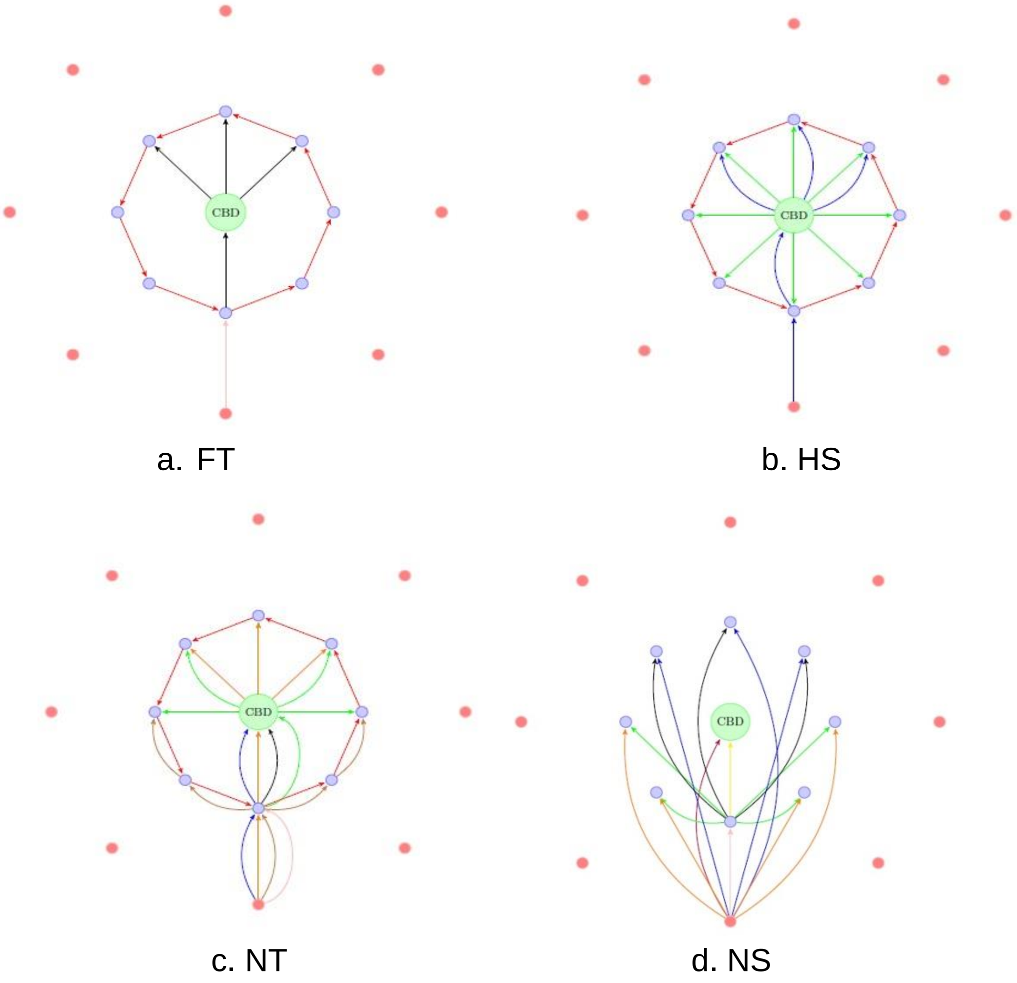


Figure 5. A simplified parametric urban model (Fielbaum *et al*., 2017).

Under this setting, one can search for the best lines structure as total trips Y grow keeping trip distribution constant[12], in order to visualize a relation between lines structure, demand and scale economies. Finding an optimal set of routes is an NP-Hard problem, which is the reason why this evolution can only be analyzed over a reasonable set of strategic line structures[13]. We will consider four traditional generic schemes with different directness indices: Feeder-Trunk (FT), Hub and spoke (HS), No Transfers (NT, or “direct”), No Stops (NS, or “exclusive”). The lines that belong to each structure are represented in Figure 6; as all structures are radially symmetric, only lines emerging from one zone are shown. Each type of line (e.g. radial, circular) is represented by a different color. Lines of the same type that share one link are grouped (such as the three black lines in 6a).

[12] In Fielbaum *et al*. (2017) the distribution of total flow in trips from the peripheries and subcenters to the CBD and (other) subcenters is represented by three parameters. These proportions (i.e. the parameters) are held constant in the analysis of scale, looking only at the effect of Y (ray analysis).

[13] Informally, a problem is NP-Hard when any algorithm that seeks the exact solution would take absurdly long times. Quak (2003), Schöbel and Scholl (2005), and Borndörfer *et al*. (2007) have shown that finding an optimal set of routes is an NP-Hard problem for various specifications.



a. FT

b. HS

c. NT

d. NS

Figure 6. Four strategic line structures

A brief description of each of these four generic line structures establishes a connection with the three indices that describe directness as defined above:

- FT: each periphery is connected with its subcenter, and subcenters are linked by direct trips that follow shortest paths, that could follow the circular line or one of the lines connecting each subcenter with the 3 “opposite” subcenters. The number of transfers is always 1 for trips from a periphery unless the destination is the own subcenter. Trips are the shortest possible. Buses stop at each node.

- HS: all peripheries are connected to the CBD that acts as a hub, and there are two additional circular lines (clockwise and counterclockwise; only one is shown in Figure 6) connecting the ring of subcenters. The number of transfers is 1 for most trips that do not finish at the CBD. Trips may be longer than the shortest path but only for a small fraction of the trips. Buses stop at each node.

- NT: nobody needs to transfer. Trips may be longer than the shortest path but only for a small fraction of the trips. There are specific lines connecting each OD-pair (some of them vanish because their optimal frequency is nil). Buses stop at each node.
- NS: nobody needs to transfer. As buses are OD specific, their routes are as short as possible and each bus travels non-stop from start to end (only 2 "stops" per trip).

The trips paths followed by the passengers are not known a priori because they depend on optimal frequencies (some of which could be zero) that in turn depend on Y. In order to characterize the structures in terms of directness independent of Y, Table 3 shows the network indices of the four structures calculated as averages across OD pairs – instead of passenger trips – in a city with eight zones (n=8, 136 OD pairs)[14]. Directness increases from FT to HS, then to NT and finally to NS.

Structure	**FT**	**HS**	**NT**	**NS**
Number of transfers	0.47	0.35	0	0
Number of stops	3.06	3.06	3.06	2
Distance traveled/Minimum distance	1	1	1	1

Table 3. Network indices describing directness for each lines structure.

In Fielbaum *et al*. (2016), frequencies of lines within each structure are optimized, minimizing a VRC function similar to equation (2) studied in section 3.2, but now including a penalty p_T for each of the T transfers in the system, as shown in equation (6). As in the illustrative model, users are assumed to be homogeneous regarding time valuation, crowding and congestion are not considered, and the number of users is exogenous (i.e min VRC for a given Y, which yields a cost function).

$$VRC = B(c_0 + c_1 K) + p_w Y t_w + p_v Y t_v + p_T T \quad (6)$$

Because of the complexity of the network, users now may have more than one route to choose from. All passenger routes are assumed to have the same fare such that assignment of passenger to routes are commanded only by the operational characteristics of the system; their choices depend on frequencies and frequencies depend on choices (as formulated in Appendix A), which prevents analytical solutions. Therefore, an iterative procedure is needed to find the optimal frequency and vehicle size for each line within a given lines structure, where each iteration rests on finding a relation between (B, K, t_w, t_v, T) and the vector of frequencies.Using the parameters shown in

[14] Note that whenever some lines vanish as a result of the optimization process (zero frequency) the flow directness indices may increase.

Appendix C, the optimal vector of frequencies is obtained for each structure[15]; again, both frequencies and bus sizes increase with patronage. Plugging these back into VRC the cost function C_i for each lines structure is obtained.

Figure 7a shows the results of Fielbaum *et al*. (2016) regarding the average cost of each line structure; as Y increases the optimal structure changes from hub and spoke, to no transfers and finally to no stops, i.e., directness increases (and feeder-trunk is never optimal)[16]. In Figure 7b this evolution is shown by means of the corresponding DSE of the optimal structure for each level of the total flow: scale economies indeed increase after each change (including a change within HS when the circular line emerges), and decrease thereafter. For synthesis, *the possibility of deciding the line structure introduces directness as a new source of economies of scale which are finally exhausted after full directness is achieved.*

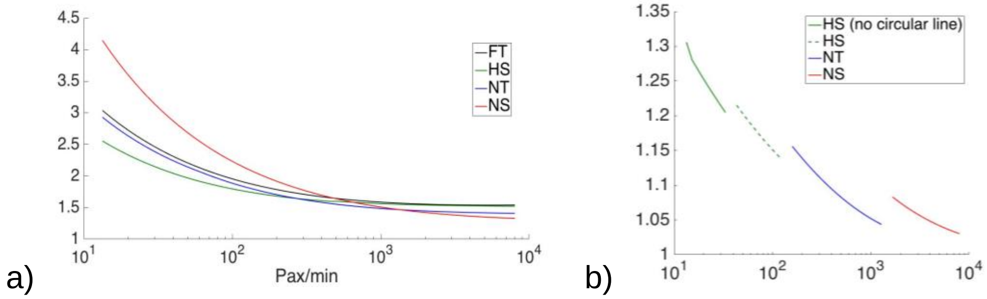


Figure 7. Average costs and overall *DSE* as directness increases.

Let us now address the central question: which design elements lie behind these results regarding scale economies? Having found the superior structures, an analysis of directness can be made taking into account the passengers' trips. Figures 8 show the evolution of each of the three flow indices that define directness as a function of the number of passengers whose growth induces lines structure changes from HS to NT to NS.

As represented in Figure 8a, transfers occur only for low values of Y where the hub and spoke structure dominates; the emergence of a new line (whose frequency jumps discretely from zero) within the HS design that connects directly some OD pairs (a new lines structure rigorously speaking) generates a reduction in the number of transfers and also in the number of stops and distance traveled, which shows up in Figures 8b and 8c. The average stops per trip decreases down to 2 when the no-stops structure dominates

[15] Parameters were chosen from Fielbaum *et al*. (2016), including meaningful values for trip distribution. For example, 80% of the trips depart from the peripheries and half of them go to the CBD.

[16] We use a logarithmic scale to be able to represent both low and high volumes of passengers and the corresponding dominant structures (this scale will also be used for all subsequent figures). Flow is shown in passengers per minute.

for high values of Y (Figure 8b). The ratio between the distance traveled and the minimum distance possibly required (called “detour” in Figure 8c) generally decreases except when changing from hub and spoke to the no-transfers structure, as some passengers experience longer trips because some short lines disappear in favor of longer ones that collect more passengers; note that this is counterbalanced by the reduction in transfers, showing that sometimes there is a trade-off between the different components of directness.

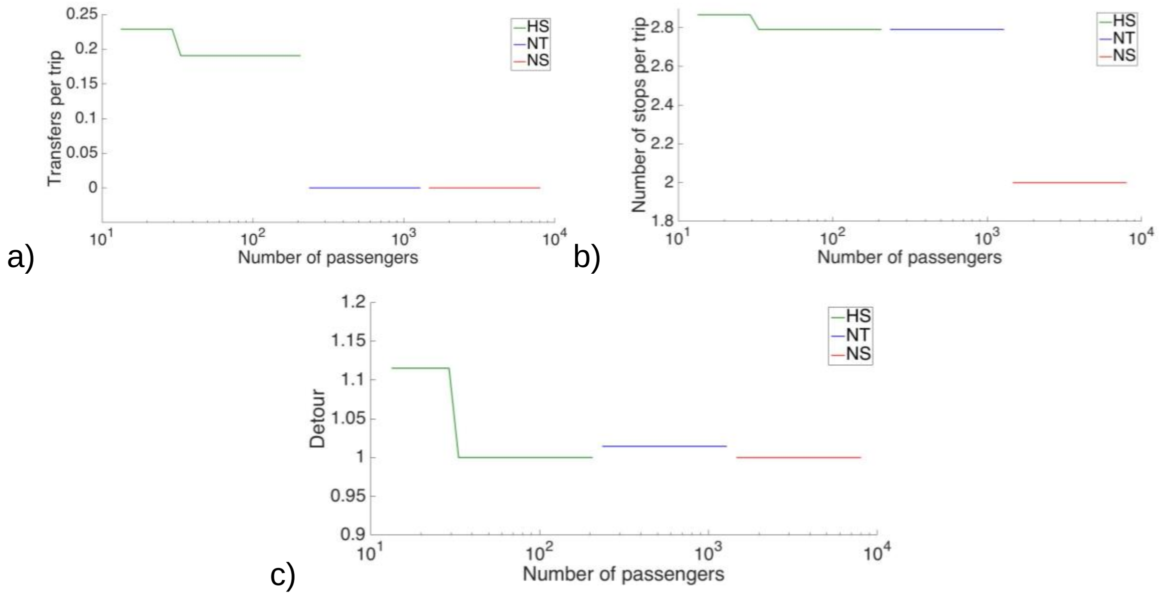


Figure 8. The three flow indices of directness as a function of patronage.

The physical measures of directness translate into users’ time and users’ costs, which are shown in Figures 9. Figure 9a summarizes the “equivalent time” associated to each of the directness indices: length of the routes translates into time-in-motion, the number of stops (together with vehicle load) translates into time at stops, and each transfer is valuated as 24 minutes in motion (as in Fielbaum *et al*., 2016). Their sum is the total equivalent time (TET) presented at the top of Figure 9a, and it synthesizes the total effect of directness on users; the fact that TET diminishes when lines structure changes clearly shows that increasing directness as patronage increases, contributes to scale economies. The slight increase of TET within each structure is caused by the larger time at stops induced by larger vehicles, an effect that is almost irrelevant when compared with the rest including the reduction in the number of stops each time the structure changes. Note that the more than 10 minutes reduction of TET is mostly explained by the reduction in time-in-motion and transfers (some 4 minutes each) against the 2 minutes reduction in time at stops.

Figure 9b shows the average costs per passenger due to in-vehicle time, waiting time and transfers, which are the three components of the users' cost function. Looking at the points where lines structure changes, it becomes evident that increasing directness makes in-vehicle time and transfer cost decrease, but there is a local increase in waiting time because directness diminishes bus-sharing and each passenger now has less lines to choose from. This local increase in waiting times, however, is more than compensated by the frequency growth as patronage increases within each structure (Mohring effect).

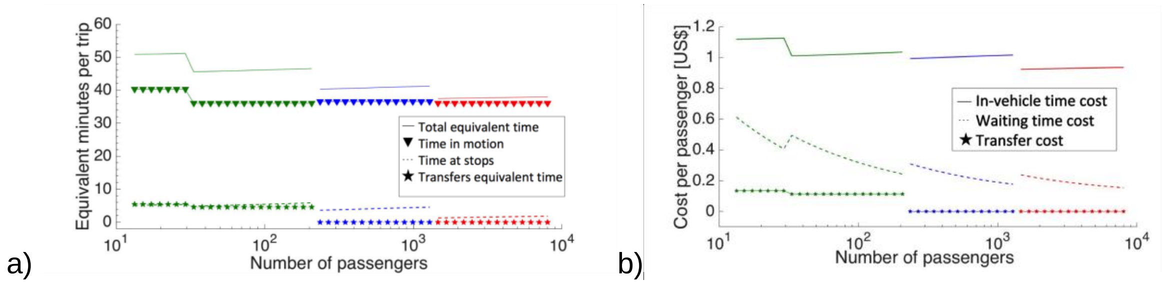


Figure 9. Effects of directness on equivalent users' times and users' costs.

So far, we have interpreted scale economies in terms of users' costs; what about operators' costs? Which are the effects of directness? To tackle these questions, let us recall that total operators' costs are given by $c_0B + c_1\Sigma$ where B is total fleet and $\Sigma = BK$ is total number of seats. Let us analyze both variables.

In Figures 10 we show (a) number of seats per passenger and (b) number of buses per passenger as a function of patronage. Seats per passenger drop significantly when lines structure changes. This effect occurs because bus-sharing diminishes (when directness increase) reducing the idle capacity of buses as we now explain in detail: the size of the buses for a line is given by its most loaded segment, such that idle capacity is present in the rest of the arcs used by the line; only in the NS structure buses are always full. On the other hand, within a given structure increasing Y increases cycle time through boarding-alighting time, which makes Σ/Y an increasing function of Y[17].

Figure 10b reveals that the number of vehicles per passenger decreases nearly in a continuous way, which shows that the effect of the change in lines structure over total fleet as Y grows is less important than the increase in bus size. In other words, when Y increases, optimal frequencies and vehicle capacities increase as well, but frequency grows at a decreasing rate precisely because the capacity grows making fleet per capita decrease.

[17] This (novel) result can be obtained analytically in the one-line case using the expressions for optimal frequency and capacity in Jansson (1984) or Jara-Díaz and Gschwender (2009).

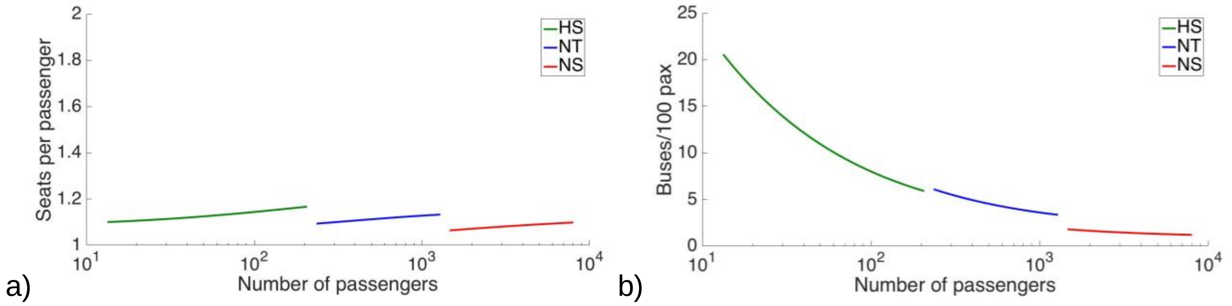


Figure 10. Effect of directness on the components of operators' costs.

In summary, including lines structures as part of the (optimal) design of public transport services in an urban space introduces yet another source of scale economies which has been defined here as directness, a concept that encompasses many elements summarized by three indices that capture transfers, routes length, and stops; as directness increases the total equivalent time for users decreases, approaching the (time related) characteristics of a private car trip. All in all, searching for the optimal lines structure is both affected by scale effects (such as the Mohring effect) and triggers new scale sources.

5. Conclusions.

In this paper we have introduced a basic structural design element - the spatial arrangement of transit lines - in the analysis of scale economies in public transport systems. We have shown that the discrete change from one structure to another as patronage increases is a source of scale economies. This change occurs because at the threshold point average costs are equal but the marginal cost of the new structure is lower (which justifies the change), such that the degree of scale economies increases at that point. The difficulty in the detailed analysis of what lies behind the effect on scale economies emerges due to the lack of a single variable that captures the evolution of lines structures as flows grow, which makes a substantial difference with the analysis of frequency and vehicle capacity in a single line.

In order to understand the engineering aspects behind the relation between lines structure and scale economies, we have proposed a three-dimensional concept called directness encompassing number of transfers, number of stops and passenger route lengths. We have shown in a very simple network as well as in a fairly general representation of a city, that all these indices improve (diminish) when a change in lines structure takes place due to an increase in passenger volume. Grossly speaking, as more passengers use public transport, it is possible to evolve towards systems with more direct lines for each OD pair, diminishing in-vehicle times while keeping reasonable waiting times, such that users are benefited by lower travel times and operators are benefited by lower idle capacity.

The change in lines structure occurs at specific levels of patronage, such that there are segments of demand where the same lines structure remains as the best. Within those segments scale economies analysis replicates the case of the single-line models, i.e. frequencies and bus capacities increase with patronage, such that waiting times for users diminish (Mohring effect), and average cost for operators diminish, which outbalances the diseconomies of scale induced by larger times at bus stops. For synthesis, the degree of scale economies increases locally when lines structure changes and diminishes afterwards until the next change occurs. And this happens until full directness is achieved; from then on frequencies and vehicle sizes increase until scale economies are exhausted.

Next steps in the analysis should take into account that, besides frequency and vehicle size, another relevant source of scale economies has also emerged from simple models: lines density, which has been represented as parallel lines that provide the same service (i.e. same frequency and bus size) affecting access time for a single OD pair. When dealing with lines structures, the introduction of density would require the inclusion of yet another variable in the model, namely the number of actual streets represented by each arc in the city network. As patronage continues increasing it is very likely that the density of lines running between each pair of nodes in the parametric representation of the city should increase as well. Future research should consider the joint evolution of density, frequencies, bus sizes, and lines structures as patronage grows.

Finally, the analysis presented here involves only the variation of total patronage keeping trip distribution constant; in the terminology created by Baumol *et al*. (1982) within a multi-output framework, this is a ray analysis, where flows in every OD pair grow by the same proportion. When a city exhibits an evolution of flows that involves a change in its basic urban structure, e.g. from monocentric to polycentric, the scale effects recognized and analyzed in this paper should be complemented with the study of economies of scope.

Acknowledgements

This research was partially funded by Fondecyt, Chile, Grant 1160410, and CONICYT PIA/BASAL AFB180003.

References

Badia, H., Estrada, M., and Robusté, F. (2014). Competitive transit network design in cities with radial street patterns. *Transportation Research Part B: Methodological*, *59*, 161-181.

Basso, L., and Jara-Díaz, S. (2006a). Are returns to scale with variable network size adequate for transport industry structure analysis? *Transportation Science*, *40*(3), 259-268.

Basso, L., and Jara-Díaz, S. (2006b). Distinguishing multiproduct economies of scale from economies of density on a fixed-size transport network. *Networks and Spatial Economics*, *6*(2), 149-162.

Basso, L., and Jara-Díaz, S. (2010). The case for subsidisation of urban public transport and the Mohring effect. *Journal of Transport Economics and Policy*, 44(3), 365-372.

Basso, L., and Jara-Díaz, S. (2012). Integrating congestion pricing, transit subsidies and mode choice. *Transportation Research Part A: Policy and Practice*, 46(6), 890-900.

Baumol, WJ, Panzar, JC and Willig, RD (1982) *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace Jovanovich, New York, USA.

Borndörfer R, Grötschel M, Pfetsch ME (2007) A column-generation approach to line planning in public transport. *Transportation Science* 41(1):123–132.

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

Ceder A. and Wilson N. (1986) Bus network design. *Transportation Research Part B: Methodological* 20(4):331–344

Chang, S., and Schonfeld, P. (1991). Multiple period optimization of bus transit systems. *Transportation Research Part B: Methodological*, *25*(6), 453-478.

Chriqui, C., and Robillard, P. (1975). Common bus lines. *Transportation Science*, *9*(2), 115-121.

Cominetti, R., and Correa, J. (2001). Common-lines and passenger assignment in congested transit networks. *Transportation Science*, *35*(3), 250-267.

Daganzo, C. (2010). Structure of competitive transit networks. *Transportation Research Part B: Methodological*, *44*(4), 434-446.

Dubois D., Bel G. and Llibre M. (1979) A set of methods in transportation network synthesis and analysis. *J. Oper. Res. Soc.* 30(9):797–808.

Fielbaum, A., Jara-Díaz, S., and Gschwender, A. (2016). Optimal public transport networks for a general urban structure. *Transportation Research Part B: Methodological*, 94, 298-313.

Fielbaum, A., Jara-Díaz, S., and Gschwender, A. (2017). A parametric description of cities for the normative analysis of transport systems. *Networks and Spatial Economics*, 17(2), 343-365.

Fielbaum, A., Jara-Díaz, S., and Gschwender, A. (2018). Transit Line Structures in a General Parametric City: The Role of Heuristics. *Transportation Science*, 52(5), 1092-1105*.*

Gschwender, A., Jara-Díaz, S., and Bravo, C. (2016). Feeder-trunk or direct lines? Economies of density, transfer costs and transit structure in an urban context. *Transportation Research Part A: Policy and Practice*, 88, 209-222.

Hörcher, D. and Graham, D. (2018). Demand imbalances and multi-period public transport supply. *Transportation Research Part B: Methodological*, 108, 106-126.

Hurdle, V. (1973). Minimum cost locations for parallel public transit lines. *Transportation Science,* 7(4), 340-350.

Jansson, J. O. (1980). A simple bus line model for optimisation of service frequency and bus size. *Journal of Transport Economics and Policy,* 14, 53-80.

Jansson, J. O. (1984). *Transport system optimization and pricing*. John Wiley & Sons, Chichester, UK.

Jara-Díaz, S. (1982a). The estimation of transport cost functions: a methodological review. *Transport Reviews*, *2*(3), 257-278.

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

Jara-Díaz, S. (2007). *Transport Economic Theory*, Elsevier, Amsterdam, The Netherlands.

Jara-Dıaz, S., and Basso, L. (2003). Transport cost functions, network expansion and economies of scope. *Transportation Research Part E: Logistics and Transportation Review*, *39*(4), 271-288.

Jara Díaz, S. and Gschwender, A. (2009). The effect of financial constraints on the optimal design of public transport services. *Transportation* 36(1), 65-75.

Jara-Díaz, S., and Tirachini, A. (2013). Urban bus transport: open all doors for boarding. *Journal of Transport Economics and Policy (JTEP)* 47(1), 91-106.

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

Kocur, G., and Hendrickson, C. (1982). Design of local bus service with demand equilibration. *Transportation Science*, *16*(2), 149-170.

Kraus, M. (2008). Economies of scale in networks. *Journal of Urban Economics*, *64*(1), 171-177.

Laporte, G., Mesa, J.A., Ortega, F.A., Perea, F., (2011). Planning rapid transit networks. *Socio-Econ. Plann. Sci.* 45 (3), 95-104.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *The American Economic Review*, *62*(4), 591-604.

Quak C.B. (2003). Bus line planning. A passenger-oriented approach of the construction of a global line network and an efficient timetable. Master's thesis, Delft University, Delft, Netherlands.

Schöbel A. and Scholl S. (2005). Line planning with minimal traveling time. *5th Workshop Algorithmic Methods Models Optim. Railways*, Palma de Mallorca, Spain.

Small, K.A. (2004). Road pricing and public transport, in G. Santos (eds.), Road Pricing: Theory and Evidence, Research in Transportation Economics, Vol. 9, Elsevier Science, 133-158

Tirachini, A., Hensher, D. and Jara-Díaz, S. (2010a). Restating modal investment priority with an improved model for public transport analysis. *Transportation Research Part E: Logistics and Transportation* Review, 46(6), 1148-1168.

Tirachini, A., Hensher, D., and Jara-Díaz, S. (2010b). Comparing operator and users costs of light rail, heavy rail and bus rapid transit over a radial public transport network. *Research in transportation economics*, 29(1), 231-242.

Tirachini, A. and Hensher, D. (2011). Bus congestion, optimal infrastructure investment and the choice of a fare collection system in dedicated bus corridors. *Transportation Research Part B: Methodological*, 45(5), 828-844.

Tirachini, A. and Hensher, D. (2012). Multimodal transport pricing: first best, second best and extensions to non-motorized transport. *Transport Reviews*, 32(2), 181-202.

Appendix A: Procedure to optimize $(\boldsymbol{f}, \boldsymbol{K})$ for a given lines structure and a given $\boldsymbol{Y}$.

The problem is

$$Min_{\boldsymbol{f},\boldsymbol{K}}\, VRC_i(\boldsymbol{f},\boldsymbol{K},\boldsymbol{Y}) = VRC_{iO}(\boldsymbol{f},\boldsymbol{K}) + VRC_{iU}(\boldsymbol{f},\boldsymbol{K},\boldsymbol{Y}) \qquad \text{(A.1)}$$

where the combinations of frequencies and vehicles sizes can generate flows $\boldsymbol{Y}$ and the minimization must respect capacity constraints: if we denote the number of passengers that use each line l at each segment e as λ_{le}, then $K_l \geq \lambda_{le}\ \forall e, l$ (A.5). In order to find these λ_{le}, for each OD-pair w all the possible routes $r \in R_w$ need to be identified[18] with their corresponding users costs c_r (note that VRC_{iU} is then defined by the costs of the selected routes), such that the Y_w passengers use the less costly one[19] (A.3), which will be identified by $x_r = 1$ (A.2 and A.6); note that in (A.3), when $x_{r'} = 1$ the corresponding cost must be the minimum, and when $x_{r'} = 0$ the inequality is trivially fulfilled. The load of each line at each arc is obtained as the sum of the passengers over the set of routes that use line l in arc e defined as S_{le} (or a portion $\theta_{lew}(\boldsymbol{f},\boldsymbol{K})$ of them, if common lines are present) as written in (A.4). Formally, the program being solved for each lines structure is the following:

$$\min_{\boldsymbol{f},\boldsymbol{K}} VRC(\boldsymbol{f},\boldsymbol{K})$$

subject to

$$\textstyle\sum_{r \in R_w} x_r = 1\ \ \forall\, w \qquad \text{(A.2)}$$

$$x_{r'} c_{r'}(f, K, x_{r'}) \leq c_r(f, K, x_r)\ \ \forall r, r' \in R_w \qquad \text{(A.3)}$$

$$\lambda_{le} = \textstyle\sum_w Y_w \sum_{r \in R_w \cap S_{le}} x_r \theta_{lew}(f, K)\ \ \forall e, l \qquad \text{(A.4)}$$

$$K_l \geq \lambda_{le}\quad \forall l, e \qquad \text{(A.5)}$$

$$\boldsymbol{f} \geq 0, \boldsymbol{x} \in \{0,1\} \qquad \text{(A.6)}$$

[18] By definition, when there are common lines in a segment, they are recognized as part of a single route.

[19] Other criteria for passenger assignment to routes can be used without affecting the analysis, provided high cost routes are discarded; one possible such criteria could be some distribution across routes with similar low costs.

Appendix B. Optimal frequencies and capacities in the simple model (Section 3.2)

Let us solve first the one-line system. Each of its components can be expressed as a function of frequency:

- Bus capacity (K): total passengers per unit time Y use f buses per unit time, such that the load of each bus is $K = Y/f$.
- Cycle time (t_c): regarding vehicle in motion, each bus needs to travel across a path whose length is $2L_0 + Q$, taking a time of $(2L_0 + Q)/V$; regarding time at stops, each passenger needs $2t$ to board and alight a bus whose load is Y/f passengers, which makes a total of $2tY/f$. Total cycle time is the sum of these two terms: $\frac{2L_0+Q}{V} + 2t\frac{Y}{f}$.
- Fleet (B): recalling that $f = \frac{B}{t_c}$, it becomes apparent that $B = f\frac{2L_0+Q}{V} + 2tY$.
- Waiting time (t_w): passengers arrive at an homogeneous rate to the bus stop, and buses exhibit a constant headway such that on average each passenger will wait half the headway $(1/2f)$.
- In-vehicle time (t_v): it needs to be calculated as the average between two types of OD-passengers. Passengers that alight from the bus at the first stop travel a distance L_0 such that time in-motion is L_0/V. At the first stop the bus stays $\frac{Y}{2f}t$, and users that alight there spend on average half of that time. Passengers that alight at the second stop travel a distance $L_0 + Q$; they stay in the vehicle $\frac{Y}{2f}t$ at the first stop, and - on average - half that time at the second stop. The average in-vehicle time for passengers is then $\frac{1}{2}\left[\left(\frac{L_0}{V} + \frac{Y}{4f}t\right) + \left(\frac{L_0+Q}{V} + \frac{Y}{2f}t + \frac{Y}{4f}t\right)\right]$.

Replacing these expressions in $VRC = B(c_0 + c_1K) + p_wYt_w + p_vYt_v$ yields

$$VRC = (f\frac{2L_0+Q}{V} + 2tY)(c_0 + c_1\frac{Y}{f}) + p_wY\frac{1}{2f} + p_vY\frac{1}{2}\left[\left(\frac{L_0}{V} + \frac{Y}{4f}t\right) + \left(\frac{L_0+Q}{V} + \frac{Y}{2f}t + \frac{Y}{4f}t\right)\right] \quad \text{(A.7)}$$

Making the derivative with respect to f equal to zero yields:

$$f^* = \sqrt{\frac{YV(\frac{p_w}{2}+2tY[c_1 + \frac{p_v}{4}])}{2c_0(L_0+Q)}},\ \ K^* = \sqrt{\frac{Y2c_0(L_0+Q)}{V(\frac{p_w}{2}+2tY[c_1 + \frac{p_v}{4}])}} \quad \text{(A.8)}$$

Both expressions increase with Y, with f^* tending to a linear function, and K^* tending to some constant when $Y \to \infty$.

The solution for the two-lines system is the following:

- Bus capacity (K): total passengers per unit time per line are now $Y/2$, and use f buses per unit time, such that the load of each bus is $K = Y/2f$.
- Cycle time (t_c): each bus travels across a path whose length is $2L_0$, so time in motion is $2L_0/V$; regarding time at stops, each passenger needs $2t$ to board and alight a bus

whose load is $Y/2f$ passengers, which makes a total of tY/f. Cycle time is the sum of these two terms: $t_c = \frac{2L_0}{V} + t\frac{Y}{f}$.

- Fleet (B): there are two identical lines so $B = 2ft_c = \frac{4fL_0}{V} + 2tY$.
- Average waiting time (t_w): as in the one line system $t_w = 1/2f$.
- Average in-vehicle time (t_v): Passengers spend in motion $\frac{L_0}{V}$. At stops, each bus spends $\frac{1}{2f}tY$, such that passengers spend on average half that time, which yields $t_v = \frac{L_0}{V} + \frac{tY}{4f}$.

Replacing these expressions in $VRC = B(c_0 + c_1K) + p_wYt_w + p_vYt_v$ yields

$$VRC = \left(\frac{4fL_0}{V} + 2tY\right)\left(c_0 + c_1\frac{Y}{2f}\right) + \frac{p_wY}{2f} + p_vY\left(\frac{L_0}{V} + \frac{tY}{4f}\right) \quad \text{(A.9)}$$

Making the derivative with respect to f equal to zero yields:

$$f^* = \sqrt{\frac{YV\left(\frac{p_w}{2} + tY\left[c_1 + \frac{p_v}{4}\right]\right)}{4L_0c_0}}, \;\; K^* = \sqrt{\frac{YL_0c_0}{V\left(\frac{p_w}{2} + tY\left[c_1 + \frac{p_v}{4}\right]\right)}} \quad \text{(A.10)}$$

Again, both expressions increase with Y, with f^* tending to a linear function, and K^* tending to some constant when $Y \rightarrow \infty$.

Appendix C. Definitions and values of the parameters for simulations.

Symbol	Meaning	Value
α	Fraction of trips starting at the peripheries that go to the CBD.	0.5
β	Fraction of trips starting at the peripheries that go to the own subcenter.	0.25
a	Fraction of trips that start at the peripheries.	0.8
$\tilde{\alpha}$	Fraction of trips starting at the sub-centers that go to the CBD.	0.67
c_0	Unitary cost per bus per period of time.	0.17 [US$/min]
c_1	Unitary cost per seat per period of time.	0.0034 [US$/min]
g	Distance periphery-subcenter/distance subcenter-CBD.	0.33
n	Number of zones in the city.	8
p_T	Users' cost of a transfer.	0.59 [US$]
p_v	Value of in-vehicle time.	1.48 [US$/h]
p_w	Value of waiting time.	2.96 [US$/h]
T_0	Vehicle in-motion time between a subcenter and the CBD.	30 [min]
L_0	Distance from origin to each destination in triangle city	30 [km]
Q	Distance between destinations in triangle city	2 [km]
V	Commercial speed of the buses	13 [km/h]