

1 Why Cities Exist

1.1 Introduction

In most countries, the population is highly concentrated in a spatial sense. For example, cities occupy only about 2 percent of the land area of the United States, with the rest vacant or inhabited at very low population densities. Even in countries that lack America's wide-open spaces, spatial concentration of the population can be substantial, with much of the land vacant. This chapter identifies some forces that lead to the spatial concentration of population. Thus, it identifies forces that help to explain the existence of cities.

Depending on their orientation, different social scientists would point to different explanations for the existence of cities. A military historian, for example, might say that, unless populations are concentrated in cities (perhaps contained within high walls), defense against attack would be difficult. A sociologist might point out that people like to interact socially, and that they must be spatially concentrated in cities in order to do so. In contrast, economic explanations for the existence of cities focus on jobs and the location of employment. Economists argue that certain economic forces cause employment to be concentrated in space. Concentrations of jobs lead to concentrations of residences as people locate near their worksites. The result is a city.

The two main forces identified by economists that lead to spatial concentration of jobs are scale economies and agglomeration economies. With scale economies, also known as "economies of scale" or "increasing returns to scale," business enterprises become more efficient at large scales of operation, producing more output per unit of input than at smaller scales. Scale economies thus favor the formation of large enterprises. Since scale economies apply to a single business

establishment (say, a factory), they favor the creation of large factories, and thus they favor spatial concentrations of employment.¹

Whereas scale economies operate within a firm, without regard to the external environment, agglomeration economies are external to a firm. Agglomeration economies capture the benefits enjoyed by a firm when it locates amid other business enterprises. These benefits include potential savings in input costs, which may be lower when many firms are present, as well as productivity gains. A productivity effect arises because inputs (particularly labor) may be more productive when a firm locates amid other business enterprises rather than in an isolated spot. The mechanisms underlying these effects will be explained later in the chapter.

Transportation costs also influence where a firm locates, and they can lead to, or reinforce, spatial concentration of jobs. This chapter explains several different ways in which transportation costs affect the formation of cities. The last section explores a special kind of agglomeration force: the kind that causes the clustering of retail establishments and the creation of shopping malls.

1.2 Scale Economies

The role of scale economies in the formation of cities can be illustrated with a simple example. Consider an island economy that produces only one good: woven baskets. The baskets are exported, and sold to buyers outside the island. The inputs to the basket-weaving process are labor and reeds. With reeds growing everywhere on the island, the basket-weaving factories and their workers can locate anywhere without losing access to the raw-material input.

The basket-weaving production process exhibits scale economies. Output per worker is higher when a basket-weaving factory has many workers than when it has only a few. The reason is the common one underlying scale economies: division of labor. When a factory has many workers, each can efficiently focus on a single task in the production process, rather than carrying out all the steps himself. One worker can gather the reeds, another can prepare the reeds for weaving, yet another can do the actual weaving, and still another can prepare the finished baskets for shipment.

1. When used with “scale” or “agglomeration,” the word “economies” means “savings” or “benefits.”

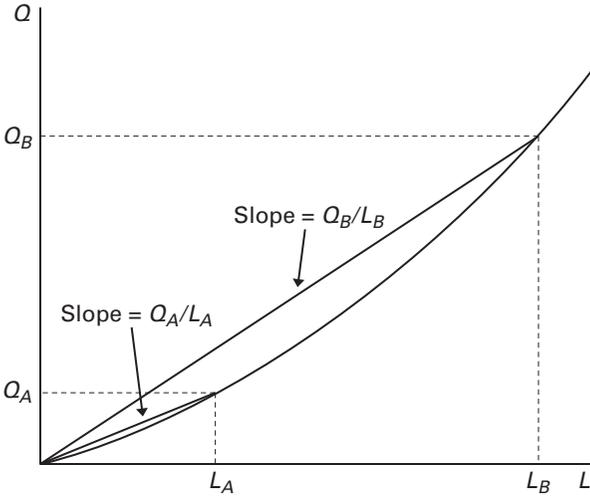


Figure 1.1
Scale economies.

The production function for the basket-weaving process is plotted in figure 1.1. The factory’s output, Q , is represented on the vertical axis, and the number of workers, L , on the horizontal axis. Since the reed input is available as needed, it does not have to be shown separately in the diagram. Because the curve shows basket output increasing at an increasing rate as L rises, scale economies are present in basket weaving. This fact can be verified by considering output per worker, which is measured by the slope of a line connecting the origin to a point on the production function. For example, the lower line segment in the figure (call it A) has a slope of Q_A/L_A , equal to the rise (Q_A) along the line divided by the run (L_A). The higher line segment (call it B), which corresponds to a factory with more workers (L_B as opposed to L_A), has slope Q_B/L_B , equal to output per worker in the larger factory. Since line B is steeper than line A, output per worker is higher in the larger factory, a consequence of a finer division of labor.

Using this information, consider the organization of basket weaving on the island economy. If 100 workers are available, the question is how these workers should be grouped into factories. Consider two possibilities: the formation of one large 100-worker factory and the formation of 100 single-worker factories (involving “backyard” production of baskets). The natural decision criterion is the island’s total output of baskets, with the preferred arrangement yielding the highest output.

Table 1.1

Basket output of the island economy.

Production arrangement	Number of factories (a)	Workers per factory (b)	Output per worker (c)	Output per factory ($b \times c$)	Total output ($a \times b \times c$)
Backyard factories	100	1	α	1α	$100 \times 1\alpha$ $= 100\alpha$
One large factory	1	100	β	100β	$1 \times 100\beta$ $= 100\beta$

The answer should be clear: given the greater efficiency of workers in large factories, the 100-worker factory will produce more output than the collection of 100 backyard factories.

But it is useful to verify this conclusion in a more systematic fashion. Let α be output per worker in a factory of size 1, equal to the slope of a line like A in figure 1.1 with $L_A = 1$. Let β be output per worker in a factory of size 100, equal to the slope of a line like B with $L_B = 100$. From the figure, it is clear that $\beta > \alpha$. Now consider table 1.1.

The island economy's total output of baskets equals (number of factories) \times (workers per factory) \times (output per worker). Table 1.1 shows that this output expression is largest with one large factory. It equals 100β , which is larger than the total output of 100α from the collection of backyard factories, given $\beta > \alpha$.

Since the island economy gains by having one large basket factory, the economy would presumably be pushed toward this arrangement, either by market forces or by central planning (if it is a command economy). But once one large factory has been formed, the basket workers will live near it, which will lead to the formation of a city.

This story is highly stylized, but it captures the essential link between scale economies and city formation, which will also be present in more complicated and realistic settings. But something is missing from the story. It can explain the formation of "company towns," but it cannot explain how truly large urban agglomerations arise.

To see this point, consider a more realistic example in which the production process is automobile assembly. This process clearly exhibits scale economies, since assembly plants tend to be large, typically employing 2,000 workers or more. Thus, an assembly plant will lead to a spatial concentration of employment, and these auto workers (and their families) will in turn attract other establishments designed to serve their personal needs—grocery stores, gas stations, doctor's offices, and

so on. The result will be a “company town” with the auto plant at its center. But how large will this town be? In the absence of any other large employer, its population may be limited in size, say to 25,000. The upshot is that, while scale economies by themselves can generate a city, it will not be as large as, say, Chicago or Houston. In order to generate such a metropolis, many firms must locate together in close proximity. For this outcome to occur, *agglomeration economies* must be present.

1.3 Agglomeration Economies

Agglomeration economies can be either pecuniary or technological. Pecuniary agglomeration economies lead to a reduction in the cost of a firm’s inputs without affecting the productivity of the inputs. Technological agglomeration economies raise the productivity of the inputs without lowering their cost. Simply stated, pecuniary economies make some inputs cheaper in large cities than in small ones, while technological economies make inputs more productive in large cities than in small ones.

1.3.1 Pecuniary agglomeration economies

The labor market offers examples of pecuniary agglomeration economies. Consider a big city, with a large concentration of jobs and thus a large labor market, where many workers offer their labor services to employers. Suppose a firm is trying to hire a specialized type of worker, with skills that are rare among the working population. With its large labor market, a few such workers might reside in a big city, and one presumably could be hired with a modest advertising effort and modest interviewing costs. However, the labor market of a small city probably would contain no workers of the desired type. This absence would force the employer to conduct a more costly search in other cities, and to bring job candidates from afar for interviewing. The firm might also have to pay relocation costs for a hired worker, adding to its already high hiring costs. Thus, by locating in a big city, a firm may lower its cost of hiring specialized labor. The existing employment concentration in the big city thus attracts even more jobs as firms locate there to reduce their hiring costs. The big city then becomes even bigger as a result of this agglomeration effect.

Locating in a big city could also reduce the cost of inputs supplied by other firms (as opposed to labor inputs supplied by individual workers). For example, the large market for commercial security

services in a big city would support many suppliers of security guards. These suppliers would compete among one another, driving down prices and thus lowering the cost of security services used in protecting office buildings or factories. Big-city competition could also reduce the prices of other business services—legal and advertising services, commercial cleaning and groundskeeping, and so on. The same effect could also arise in the context of locally supplied physical inputs, such as ball bearings for a production process or food inputs for the company cafeteria, where competition among big-city suppliers would reduce input costs. As in the case of hiring costs, the job concentration in a big city is self-reinforcing: once jobs become concentrated, even more firms will want to locate in the big city to take advantage of the lower input costs it offers.

In some cases, a small-town location might mean that a particular business service is entirely unavailable locally, just like the specialized worker discussed above. For example, a firm might require specialized legal services (help with an antitrust issue, for example), and there might be no local law firm with such expertise. The firm would then have to pay high-priced lawyers to travel to its headquarters from their big-city offices, or else would have to develop the required expertise “in house” at high cost. In this case, higher input costs arise from the sheer local unavailability of the service in the small city rather than from the low degree of competition among local suppliers.

Firms also purchase transportation services, which are used in shipping output to the market and in shipping inputs to the production site. A firm can reduce its output shipping costs by locating near its market, and it can reduce its input shipping costs by locating near its suppliers. A big city, with its many households, is a likely market for output, and it may also host many of a firm’s input suppliers. In this situation, the firm can minimize its shipping costs for both output and inputs by also locating in the big city. Note that the resulting pecuniary agglomeration benefit is slightly different from those discussed above. Instead of facing a lower unit price of transportation services in the big city (a lower cost per ton-mile), the firm benefits from being able to purchase *less* of these services because of its proximity to the market and to suppliers.²

2. Paying high-priced big-city lawyers to travel to a firm’s small-city location fits into this scenario. The firm is transporting its legal input, an outcome that could be avoided by locating in the big city.

Another observation is that this scenario “stacks the deck” in favor of the transportation-cost argument for a big-city location. Suppose instead that input suppliers are in a different location, while the market is still in the big city. Then, locating there will save on output shipping costs, but the inputs will have to come from afar. In this case, the location with the lowest total transport cost may not be in the big city, so that transport-related agglomeration economies may not be operative. This case is discussed in detail later in the chapter.

1.3.2 Technological agglomeration economies

Technological agglomeration economies arise when a firm’s inputs are more productive if it locates in a big city, amid a large concentration of employment, than if it locates in a small city. To understand how such an effect can arise, suppose that a high-technology firm spends substantial sums on research and development. This spending leads to new technologies and products, which the firm can then patent, allowing it to earn revenue from licensing its discoveries. Suppose for simplicity that the firm’s output is measured by the number of patents it generates per year, and that the only input is labor, measured by the number of engineers the firm employs.

The production function is plotted in figure 1.2, with the output (patents) again on the vertical axis and the input (engineers) on the

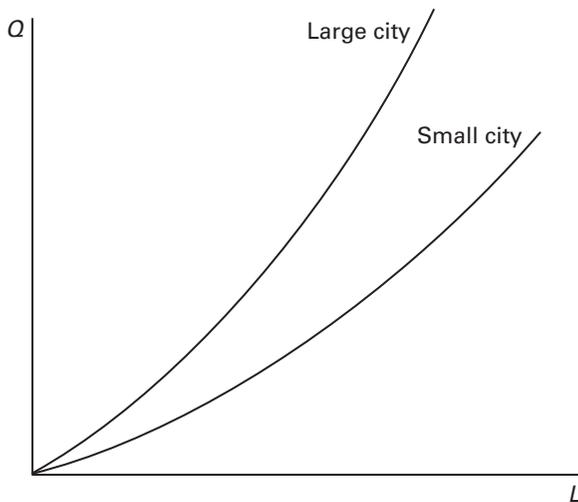


Figure 1.2
Technological agglomeration economies.

horizontal axis. The figure shows that research and development exhibits scale economies, although this feature isn't crucial to the story. The figure also shows two different production functions, which apply in different situations. The lower curve is relevant when the firm locates in a small city where few other high-tech firms are present. The upper curve is relevant when the firm locates in a big city amid a large number of other high-tech firms. The greater height of the upper curve indicates that the firm produces more patents, for any given number of engineers, when it locates in a big city. Thus, engineers are more productive, generating more patentable ideas, in the big city.

This beneficial effect could be a result of "knowledge spillovers" across high-tech firms, which are a type of externality. While engineers within a given firm collaborate intensively in producing patentable ideas, spillovers arise when contact between engineers in *different firms* also stimulates this productive process. For example, engineers from different high-tech firms might socialize together, sharing a pitcher of beer at a Friday "happy hour." Although the engineers wouldn't want to divulge their company's particular secrets, the happy-hour discussion might cover more general ideas, and it might get the engineers thinking in new directions. At work the following week, this stimulation may start a process that eventually leads to patents that wouldn't have been generated otherwise. Thus, the engineers end up being more productive because the large high-tech employment concentration allows them to interact with peers doing similar work in other companies.

Although a big city is likely to have a concentration of high-tech employment, allowing knowledge spillovers to occur, some big cities may not have many high-tech firms. Thus, the big city/small city distinction may be less relevant for knowledge spillovers than it was for pecuniary agglomeration economies. Instead, what may matter is the extent of the city's employment concentration in the industry in which such spillovers occur. If a small or medium-size city happens to have a big employment concentration in the relevant industry, it will offer strong technological agglomeration economies for industry firms despite its limited size.

Might some kinds of knowledge spillovers occur across different industries, so that a city where many different industries are represented is also capable of generating technological agglomeration economies? For example, might knowledge spillovers arise between manufacturers of medical equipment and producers of computer soft-

ware? This type of linkage seems possible, and to the extent that it exists, the overall employment level in a city (rather than employment in a firm's own industry) might be the source of technological agglomeration economies. City size may then capture the extent of such economies, as in the case of pecuniary agglomeration economies.

As will be discussed further below, empirical research on agglomeration effects sometimes finds evidence of such a link between and worker productivity and total employment (and thus city size). "Urbanization economies" is a name sometimes given to this effect. But the empirical evidence for a link between productivity and own-industry employment in a city is much stronger (this effect is referred to as "localization economies"). Thus, technological agglomeration economies appear to operate more strongly within industries than across industries.

In addition to knowledge spillovers, several other channels for such an agglomeration effect can be envisioned. When a city's employment in a particular industry is large, the existence of a large labor pool makes replacement of workers easy. As a result, unproductive workers can be fired with little disruption to the firm, since they can be immediately replaced. Recognizing this possibility, employees will work hard, achieving higher productivity than in an environment in which shirking on the job is harder to punish with dismissal.

The large labor pool also gives employers a broad range of choice in hiring decisions, which may make it harder for any individual to secure a first job in the industry. Workers may then have an incentive to improve their credentials via additional education and training, and these efforts would lead to higher productivity. Thus, the existence of a large labor pool may raise productivity for workers trying to get a job as well as for those worried about losing one.

A third channel could arise through the phenomenon of "keeping up with the Joneses." In a city with high employment in a particular industry, workers may be likely to socialize with employees working for other firms in their industry, as was noted above. In addition to making comparisons within their own firm, workers may then judge their achievements against those of friends in other firms. This comparison may spur harder work as employees try to "look good" in the eyes of a broader social set. Thus, in addition to being driven by knowledge spillovers, the higher worker productivity associated with technological agglomeration economies could arise from these three other sources.

A vast empirical literature tests for the existence of agglomeration economies. For a survey, see Rosenthal and Strange 2004. Early work investigated the connection between worker productivity and both own-industry and total employment in a city (see Henderson 1986, 2003). As was noted earlier, such studies often find evidence of an own-industry effect for knowledge-intensive industries, in which spillovers are likely to occur, without finding much evidence of a total-employment effect. Another empirical approach relates worker productivity to employment density in a city (Ciccone and Hall 1996). Yet another approach links the birth of new firms to own-industry employment, on the belief that business startups are more likely to occur in areas offering agglomeration economies (Rosenthal and Strange 2003). A similar approach relates employment growth to own-industry employment and other agglomeration factors (Glaeser et al. 1992). Other research focuses more explicitly on knowledge spillovers by considering patent activity. A patent application must cite earlier related patents, and some work shows that cited patents tend to be from the same city as the patent application, indicating local knowledge spillovers (Jaffe, Trajtenberg, and Henderson 1993). Other research relates the level of patenting activity to a city's employment density (Carlino, Chatterjee, and Hunt 2007). This literature offers an overwhelming body of evidence showing the existence of agglomeration economies, mostly of the technological type.

1.4 Transport Costs and Firm Location

As was seen above, transportation-cost savings can be viewed as a type of pecuniary agglomeration effect, which may draw firms to a large city when both its market and suppliers are located there. However, when the market and suppliers are far apart, the firm's location decision is less clear. To analyze this case, consider a situation in which a firm sells its output to a single market, presumably located in a city, and acquires its input from a single location distant from the market. Viewing the input as a raw material, let the input source be referred to as the "mine." Moreover, suppose that the mine and the market are connected by a road that can be used for shipments (see figure 1.3), and that the firm's factory can be located anywhere along this road, including at the market or at the mine. The mine and the market are D miles apart.



Figure 1.3
Mine versus market.

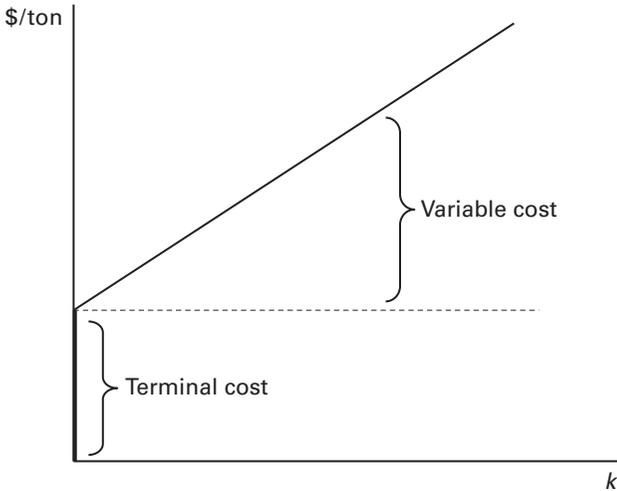


Figure 1.4
Transport costs.

Shipping costs exhibit “economies of distance” in the sense that the cost per mile of shipping a ton of material declines with the distance shipped. Figure 1.4 illustrates such a relationship, with the vertical axis representing cost per ton and the horizontal axis representing shipping distance (denoted by k). The figure shows shipping cost as having two components. The first is “terminal cost,” which must be incurred regardless of shipping distance. This is the cost of loading the shipment onto a truck or a train, and it is represented by the vertical intercept of the line. The second component is variable cost, which equals the product of the fixed incremental cost per mile (equal to the slope of the line) and the distance shipped. Variable cost is thus the height of the line above its intercept. The presence of economies of distance can be seen by drawing a line between the origin and a point on the curve. The slope of this line is equal to cost per mile (analogous to output per worker in figure 1.1). As shipping distance increases, this line becomes flatter, indicating a decline in cost per mile. The reason

for this decline is that the fixed terminal cost is spread over more miles as distance increases.³

For the current analysis, it is convenient to use a somewhat less realistic transport-cost curve that has zero terminal costs but still exhibits its economies of distance. Figure 1.5 shows two such curves, with economies of distance following from their concavity (which reflects declining variable cost). The lower curve represents the cost per ton of shipping the output different distances. To understand the upper curve, suppose that the production process entails refining raw materials, with production of the refined output requiring that a portion of the input be removed and discarded. Then, to produce a ton of output, the refining factory requires more than a ton of input. The upper curve represents the cost of shipping this amount of input various distances. In other words, the curve represents the cost of shipping enough input to produce a ton of output.

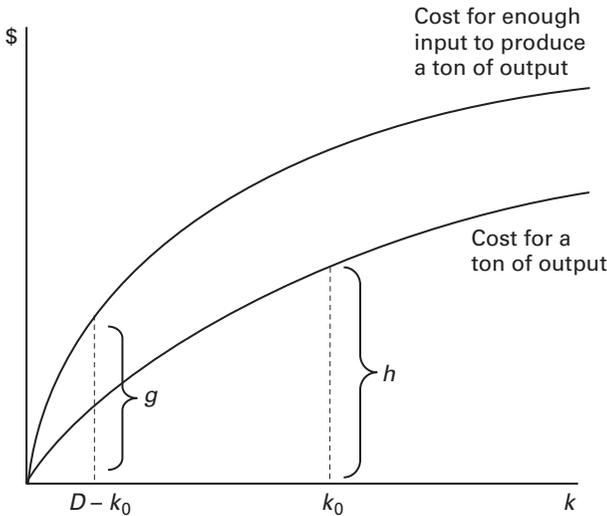


Figure 1.5
Input and output shipping.

3. A diagram like figure 1.4 can be used to illustrate transport mode choice between truck and train. Relative to trains, trucks have low terminal cost (they can be driven directly to a shipment pick-up point), but they have high variable cost, using more fuel and labor per ton shipped than trains. Therefore, the truck line in a diagram like figure 1.4 would start below the train line and eventually rise above it. As a result, a shipper choosing the cheapest mode would select truck for a short-distance shipment and train for a long-distance shipment.

The fact that the unrefined input and the refined output are qualitatively similar (both being, in effect, dirt) is convenient. This similarity means that the cost of shipping a ton of input or output a given distance will be the same, which ensures that the input shipping cost curve in figure 1.5 (which pertains to more than a ton) will be higher.

Things are more complicated if inputs and outputs are qualitatively different. In the baking of bread, for example, the flour input is compact but the output of finished loaves is very bulky, so that a ton of output is more costly to ship than a ton of input. Figure 1.5 would have to be modified to fit this case.

The information in figure 1.5 can be used to compute the best location for the factory. If the factory has a contract to deliver a fixed amount of output to the market, its goal is to minimize the total shipping cost per ton of delivered output. This total includes both the cost of shipping the output and the cost of shipping the input, with the latter cost pertaining to enough input to make a ton of output.

Suppose the factory is located k_0 miles from the market. Then the output must be shipped k_0 miles and the input must be shipped $D - k_0$ miles (recall that D is the distance between the mine and the market). The output shipping cost (per ton) is represented in figure 1.5 by h , and the input shipping cost (per ton of output produced) by g . The total shipping cost per ton of output at location k_0 is then $h + g$. To find the best location, this calculation procedure must be repeated for all k_0 values between 0 and D , with the location yielding the lowest total cost chosen. The required steps are cumbersome, however, and the answer is, so far, unclear.

The answer can be seen immediately, however, if figure 1.5 is redrawn as figure 1.6. This new figure has two origins, one at 0 and the other at distance D , and the input shipping curve is drawn backward, starting at the D origin. Now, total shipping cost (per ton of output) at location k_0 is equal to the height of output curve at that point plus the height of the input curve at the same point, which equals the cost of shipping the input $D - k_0$ miles. Therefore, shipping cost per ton of output is equal to the vertical sum of the two curves at any point, and the resulting curve is the upper hump-shaped one in figure 1.6. By inspection, it is easy to see the cheapest location. It is an endpoint location, with the factory located at the mine. All other locations, including the market and points between the mine and market, result in higher total shipping costs.

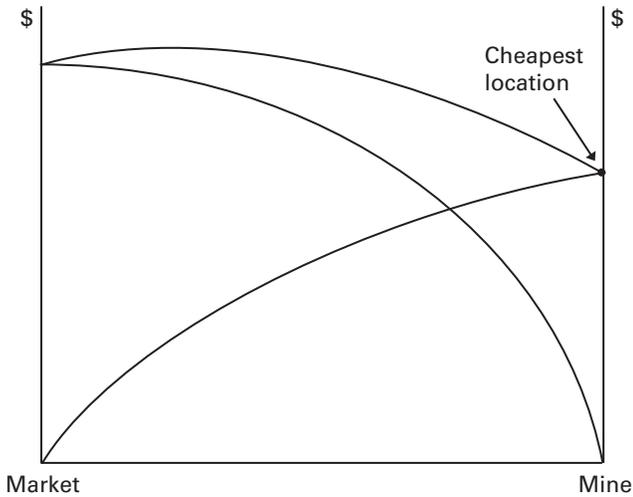


Figure 1.6
Transport-cost-minimizing location.

The mine is the best location because the production process is a “weight-losing” process. In other words, more than a ton of input is transformed into a ton of output. In this case, it doesn’t make sense to ship the input at all, since some of the material will be discarded in the refining process. Only the output should be shipped.

This appealingly simple rationale partially obscures the logic of the solution. The logic has two components. First, because of economies of distance, it isn’t economical to ship both the input and the output. In this case, two intermediate-distance shipments would occur, failing to exploit the lower cost per mile of longer shipments. Thus, either the output should be shipped the entire mine–market distance of D miles or the input should be shipped the entire distance. The best choice is the one that is cheaper, and, given the weight-losing nature of the production process, shipping the output all the way costs less. This argument shows that economies of distance, omitted from the partial explanation above, are crucial to the solution.⁴

4. If transport costs instead exhibited *diseconomies* of distance (with the curves convex), then intermediate-distance shipments would be desirable, and the best location would be an intermediate point between the mine and market. This result, which can be seen by redrawing figure 1.6 for the diseconomies case, shows that weight loss in production isn’t sufficient for a mine location to be best.

Suppose the production process is, instead, weight gaining. An example would be Coca-Cola bottling, in which syrup (evidently produced under secret conditions near the company's headquarters in Atlanta) is combined with water to produce the finished product. In this weight-gaining case, the heights of the input and output curves in figure 1.5 are reversed, as are the heights in figure 1.6. The hump-shaped curve in figure 1.6 then reaches its low point at the market rather than at the mine, making the market the best location for the factory (in this case, the bottling plant). The outcome is natural since it would make little sense to ship finished Coca-Cola long distances when its main component (water) is available everywhere.⁵

In reality, the bottling of soft drinks is indeed a "market-oriented" production process, with bottling plants located in many cities across the United States. The theory says that weight-gaining production processes should all share this feature, while weight-losing processes should be oriented to the mine (or the input). This simple model, however, omits many elements of reality, including the existence of multiple markets and multiple input sites for a given firm, as well as "bulk" differences between inputs and outputs, which (as explained above) complicate the picture.⁶ Nevertheless, the model shows that transport costs will affect where firms locate and thus will influence the formation of cities.

In the simple case analyzed, this influence can be delineated. The market (presumably a big city) will attract weight-gaining production processes as firms seek to minimize transport costs. Therefore, the city's existing concentration of jobs and people will attract additional jobs in weight-gaining industries via this transport-related agglomeration force. Weight-losing industries, however, will shun the market and will instead locate at the source of the raw material. Therefore, the existence of a big-city market may spawn separate employment concentrations at faraway natural-resource sites where factories built to serve the

5. The best factory location may sometimes lie at a "transshipment" point between the mine and the market, where the shipment must be unloaded and reloaded for some reason. Exercise 1.1 considers such a case, assuming that an unbridged river cuts the road between the mine and market, necessitating unloading, transfer to a barge, and reloading of the shipment.

6. With multiple markets and mines, it can be shown that the optimal location is usually some intermediate point, so that a market or mine location is not best.

market find it best to locate. Concentrated employment in one spot may thus cause additional remote employment concentrations to arise as a result of transport-related forces.

1.5 The Interaction of Scale Economies and Transportation Costs in the Formation of Cities

Although the desire to minimize transport costs can pull a factory toward a particular location, these costs can also play a role in the overall organization of production. In particular, transport costs can help determine whether production is centralized in one large factory or divided among a number of smaller establishments. The analysis in section 1.2 showed that a single large factory was best, but transport costs played no role in that analysis.

To illustrate the interaction between scale economies and transport costs, and to show how this interaction can affect the formation of cities, consider a simple model adapted from Krugman 1991. Suppose that the economy has five regions and a total population of N , with $N/5$ people living in each region. The regions are represented as squares in figure 1.7. Each person consumes one unit of a manufactured good, which is produced with scale economies. In this setting, scale economies are best seen via the cost function $C(Q)$, which gives the total cost of producing Q units of the manufactured good. Scale economies mean that the cost per unit of output declines as Q increases, with the factory becoming more efficient as the level of output expands. Thus, $C(Q)/Q$ falls as Q rises.

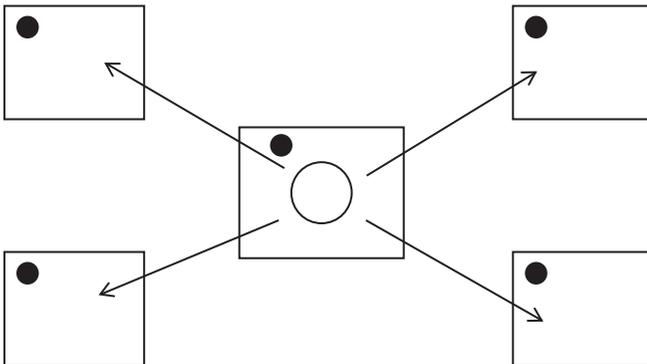


Figure 1.7
Scale economies versus transport costs.

The economy must produce N units of manufactured good to serve the population. Two different arrangements of production are possible: centralized and dispersed. Under dispersed production, a small factory, producing $N/5$ units of output, would be located in each of the five regions. These factories are represented by the small circles in figure 1.7. The cost per unit of output in each factory is $C(N/5)/(N/5) \equiv \lambda$. Under centralized production, one large factory, represented by the large circle, would be located in the central region, producing N units of output at a cost per unit equal to $C(N)/N \equiv \theta$. With scale economies, the cost per unit of output is lower in a large factory than in a small one, with $\theta < \lambda$. The total cost of production for the N units produced in the economy is θN with centralized production and λN with dispersed production, and, since $\theta < \lambda$, total cost is lower in the centralized case.

Although this conclusion led to the superiority of one large factory in the basket-weaving economy (where the output was exported), another consideration comes into play in the present setting. Since the manufactured good is consumed within the economy, adoption of centralized production means that output must be shipped from the large factory to the regions where no manufacturing plant is present. These shipments are represented in figure 1.7 by arrows. The need for shipping is avoided, however, in the dispersed case, since each region has its own factory. Let T denote the cost of shipping a unit of output from the large factory to any of the remote regions. Since four regions lack a factory, a total of $4N/5$ units of the large factory's output (which equals N units) must be shipped, at a cost of $4TN/5$. Note that $N/5$ units of output are consumed within the large factory's region and thus need not be shipped.

The overall cost of the centralized and dispersed arrangements includes both production cost and transportation cost. With transportation cost equal to zero in the dispersed case, the overall cost is just λN . In the centralized case, the overall cost is $\theta N + 4TN/5$. Therefore, the centralized (dispersed) case has a lower overall cost when λN is greater (less) than $\theta N + 4TN/5$. Rearranging, it follows that centralized (dispersed) production is preferred when $\lambda - \theta$ is greater (less) than $4T/5$. The expression $\lambda - \theta$ is the difference between cost per unit of output at the low output level of $Q = N/5$ and the higher output of $Q = N$, a positive difference given the presence of scale economies. If scale economies are strong, so that the small factory is much less efficient than the large factory, then $\lambda - \theta$ will be large, and the first inequality is likely

to hold, for a fixed T . Conversely, for a fixed value of $\lambda - \theta$, the first inequality will be likely to hold when T is small. Therefore, centralized production will be favored when scale economies are strong and transports costs are low. This conclusion makes sense since strong scale economies lead to a substantial production-cost advantage for centralized production, while low transport costs mean that this advantage isn't offset by the cost of shipping the output.⁷

Although the conclusion of the analysis of basket weaving above is reaffirmed in this case, dispersed production will be preferred in an economy in which transport costs are high relative to the strength of scale economies. In this case, the second of the above inequalities will hold. This situation might describe an undeveloped economy with poor transport linkages, which would be unable to exploit potential scale economies. When economic development improves the transportation system, lowering T , production would shift to the efficient centralized arrangement.

Centralized production would presumably lead to concentrated employment for manufacturing workers, who have not been explicitly included in the analysis so far. This concentration would lead to formation of a large city in the central zone. In the dispersed case, manufacturing workers would be scattered across the regions, with no notable job concentration arising. Adding manufacturing workers would require some minor modifications to the model, but the conclusion that scale economies can interact with transportation costs in the location of production (and the hence in the formation of cities) would remain unchanged.⁸

1.6 Retail Agglomeration and the Economics of Shopping Centers

Another type of agglomeration phenomenon is retail agglomeration: the spatial concentration of retail outlets. Cities have long had shopping districts in which stores are concentrated. While such districts were the result of uncoordinated store-location decisions, they have been increasingly supplanted by shopping malls, whose owners "orchestrate" the process of retail agglomeration.

7. Exercise 1.2 provides a numerical example based on this model.

8. A large theoretical literature has modeled the trade-off between scale economies and transportation cost in a more elaborate and sophisticated fashion than the simple approach from above. For a survey and a synthesis of these models, see Fujita and Thisse 2002.

Two forces contribute to retail agglomeration. The first is a desire by shoppers to limit the costs of shopping trips, which include both time costs and out-of-pocket costs (such as automobile expenses, including the cost of gasoline). When a consumer has to visit multiple stores to make a variety of purchases, the cost of the shopping trip is reduced when the stores are in close proximity. Therefore, a multiple-stop shopper would prefer to carry out his or her trip at a shopping district or a mall rather than visiting a sequence of isolated stores. As a result, stores are likely to attract more customer traffic when they are spatially concentrated than when they are dispersed, and this gain can stimulate retail agglomeration.

The second force leading to concentration of stores is the benefit to consumers of comparison shopping, which can arise even when only one purchase is being made. The ability to compare similar products will generate a better purchase decision, raising the benefits from shopping. Comparison shopping, easily done in a shopping district or a mall, is costly when the stores are spatially separated. This added cost may in fact be prohibitive relative to the benefit, making comparison shopping economical only on a visit to a shopping district or a mall. Spatially concentrated stores can thus offer higher shopping benefits than isolated stores, leading to more customer traffic. This gain for stores will again stimulate retail agglomeration.

Price competition between stores selling similar products is also likely to be more intense when the stores are spatially concentrated. This competition, which leads to lower prices, is beneficial for consumers but puts downward pressure on stores' profits. The resulting loss tends to reduce the attractiveness of spatial concentration from the viewpoint of stores, offsetting some of the gains described above. The fact that owners of stores seem to prefer locations in malls and shopping districts, however, suggests that the beneficial forces dominate the loss attributable to greater price competition.

The benefits of agglomeration can be viewed as arising from inter-store externalities. For example, shoppers visiting a shoe store in a mall may also visit a clothing store, and vice versa, so that each of the store types gains from the presence of the other type. Such externalities may be weaker, however, between other types of stores. For example, visitors to clothing or shoe stores may have little reason to visit a toy store or a specialty tobacco and pipe store, and vice versa. Inter-store externalities are illustrated in figure 1.8, where the widths of the arrows represent the strength of the beneficial effects between stores.

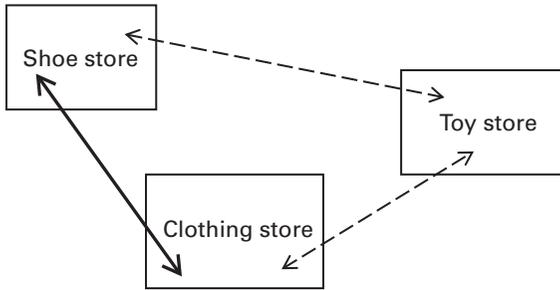


Figure 1.8
Inter-store externalities.

When retail agglomeration is orchestrated by the owner of a shopping mall, the strength and the direction of such externalities are taken into account. Since the mall's owner can charge higher rent to a store when it earns more revenue, and since revenue depends on inter-store externalities, the owner of the mall will want to choose the mix of stores and their sizes taking these externalities into account. The owner will allocate the mall's fixed square footage to stores in a way that "optimizes" the externality flows, in the proper sense. The mall's owner can reap more rental income, and thus earn more profit, by doing so.⁹

1.7 Summary

This chapter has discussed economic reasons for the existence of cities. Scale economies, which favor the creation of large business establishments, are capable of generating a moderate-size company town oriented around a single large factory. But agglomeration economies, which cause separate firms to locate near one another, are required for job concentrations substantial enough to generate a big city. Technological agglomeration economies, which raise worker productivity, can arise from knowledge spillovers among similar firms locating in close proximity. Pecuniary agglomeration economies, in contrast, reduce the cost of inputs without affecting their productivity. One pecuniary effect is the saving on transportation costs when a firm locates in a city that contains both its market and its input suppliers. But when the suppliers are remote, transport-cost considerations may pull the firm

9. Exercise 1.3 provides a stylized numerical example of a shopping-mall owner's optimization problem. For a formal analysis of problems of this kind, see Brueckner 1993.

toward a remote location, generating an employment concentration far from the market. Transport costs may also overturn the employment-concentrating effect of scale economies. When output must be shipped to consumers in dispersed locations, it may be better to forsake the gains from scale by putting small factories in these locations. Finally, retail agglomeration is generated by inter-store externalities, which generate gains for individual stores when they locate in close spatial proximity.