Paper 11. Urban economics Dr B Fingleton

These notes are the material providing the basis for lectures delivered in the Land Economy Tripos at Cambridge University. The material is divided into three sections.

1. The Overview
2. The market structure and other assumptions
3. Introducing Externalities

1. The Overview

What are the economics behind the tendency for people and firms to agglomerate together in space? Clearly, there are advantages in agglomeration, in the formation of cities. This is self evident, the share of the world's population living in cities has been rising year-on-year and is, according to a UN report published in 1994, expected to exceed 50% by 2005. People would not be prepared to pay the higher rents and costs of urban living if they were not more than compensated. There is clear empirical evidence (see the evidence cited in Quigley, 1998) that large cities are more productive than smaller cities. This is not simply because they have a greater level of inputs and thus will naturally produce more. The evidence is that they produce more than you would expect from the increase in size. Studies have shown that doubling city size more than doubles output, in other words output per unit of input increases in the range of 3 to 27%, depending on the historical, geographical and industry context. The increase in productivity with city size is accompanied by an increase in wages. Larger cities have higher productivity and larger wages, and therefore attract workers. Also, there are benefits on the consumption side. Larger cities are associated with a greater diversity of goods and services that can be consumed, this again adds to their attraction to workers and is part of the explanation of the rise in urban population.

We can go back to Adam Smith(1776) for an explanation of cities based on economic reasoning. He highlighted the demand side, the existence of a concentration of people, and the supply side, the presence of a cohort of workers whose productivity was enhanced by the division of labour made possible in a large diverse labour pool. High productivity as a result of the division of labour created wealth which attracted people, which boosted both the demand side, since there were more people around earning money to consume the manufactures, and the supply side, since an even larger labour pool enhanced the scope for the division of labour, greater productivity, attracting more migrant labour, and so on. It is easy to imagine how success bred success and why cities, once formed, were fairly unstoppable economic machines. We very quickly enter into a self-reinforcing system of relationships which means that cities emerge, grow and remain a visible element of the economic landscape for centuries. This kind of

self-perpetuating process lies at the center of modern approaches to the economics of urban development and growth.

With rising productivity we get rising wage rates. There is clear evidence that wage rates rise as economic activity becomes denser. This is what we find in cities, the larger the city the more dense is economic activity, and the higher the wage rate. In city centers many firms are clustered together in a restricted space and wage rates are also extremely high, reflecting the enhanced productivity of workers.

There is clear evidence of a wages/density link when we examine the data at the level of local authority districts of Great Britain. There are 408 such authorities, covering the entire country. The following figure shows how the (natural log of) wage rates relates to (the natural log of) employment density, which is defined as the number of employees per unit area. The data are for the year 2000.

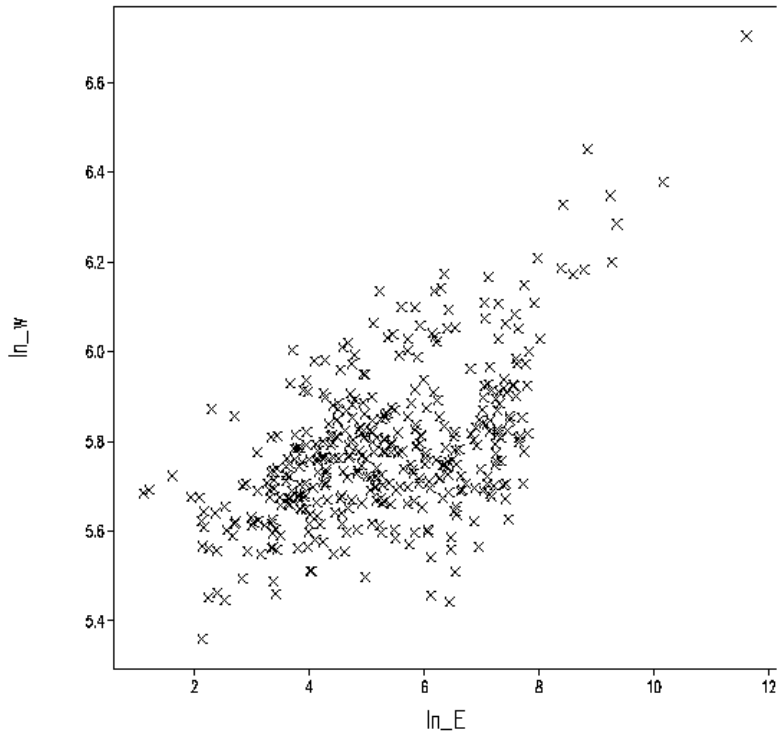
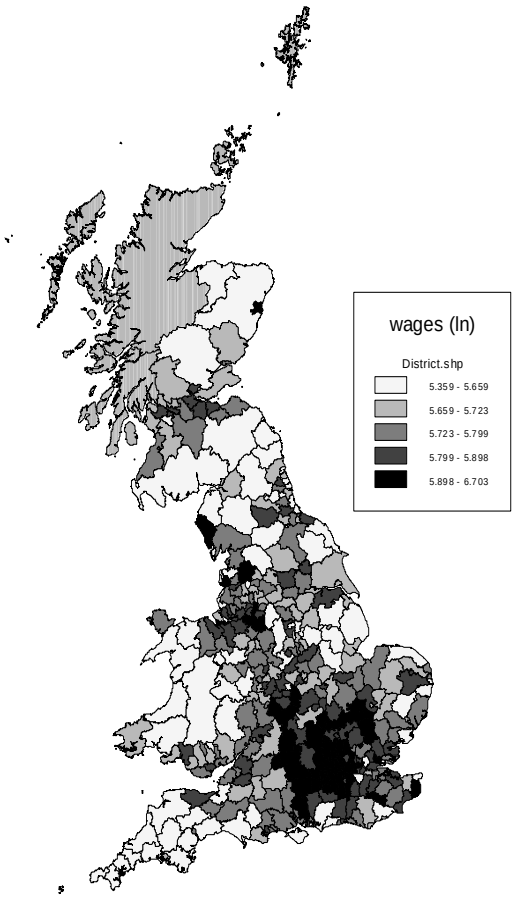


Figure 1 The relationship between the natural log of the wage level and employment density in 408 unitary authority and local authority districts of Great Britain

Where do you have to go to earn big money?



The following are the top ten.

Local authority area	ln (wages)	ln (employee density)
City_of_London	6.703	11.612
Tower_Hamlets	6.451	8.848
Westminster_City_of	6.379	10.168
Islington	6.348	9.251
Hackney	6.328	8.425
Camden	6.284	9.361
Slough	6.208	7.981
Kensington_and_Chelsea	6.200	9.264
Lambeth	6.186	8.389
Hammersmith_and_Fulham	6.183	8.790

The lowest wage rates are associated with lower employment densities. The bottom 10 of the 408 are as follows

Local authority area	ln (wages)	ln (employee density)
Chester-le-Street	5.498	4.978
Torridge	5.494	2.844
Conwy	5.488	3.382
Richmondshire	5.462	2.404
Caradon	5.460	3.421
Weymouth_and_Portland	5.457	6.122
Berwick-upon-Tweed	5.452	2.248
West_Devon	5.445	2.534
Havering	5.442	6.456
Alnwick	5.359	2.140

Of course wages don't depend only on the density of economic activity in an area, otherwise the points on the graph would be on a line, so that we could read off the wage rate given the density. The reason why the relationship is an imprecise one is because of all the other factors that also affect wage rates, for instance educational attainment and skill differences between area, and the access of an area to in-commuters who might be productive. However, the relationship between wage rates and the density of economic activity is an impressive one.

Why then are larger (denser) cities more productive and have higher wage rates as a result? Evidence began to filter through in the period up to the 1990s which was largely based on case studies, but which pointed the way to a general theory which is the main topic of this part of the course. The evidence was that the main causes were .

1) scale economies
2) positive externalities

If we look at each of these in turn in a bit more detail we shall get a good appreciation of the background to contemporary theory

1) scale economies

This is the historical economic rationale for the existence of cities. Without economies of scale, production would be dispersed to save on transport costs. Up to some point when congestion externalities kick in, there are economies of scale for many types of production and also for many public facilities like sports facilities, libraries and museums. With these, the cost per individual declines as the size of the city rises.

2) Positive externalities

There are many factors which are not manifest in the market and are unpriced but which affect urban productivity. On the side of positive externalities we emphasise the role played by knowledge. Knowledge is often created in an urban environment, but its benefits are often not captured completely by the innovator and others free-ride on someone else's effort without paying for it. The potential for knowledge spillovers

are greater in large diverse urban economies. On the negative side, as cities get larger, external effects due to congestion (using the term in a general sense and not simply traffic congestion) also increase. The activity of one firm can negatively impact the activity of another and we see a fall of in productivity. We consider this in detail later.

These factors emphasise the role of urban size and diversity as a reason why large cities are more productive. In recent decades, formal models have emerged, and are still being developed, which attempt to capture these types of effects. The new models are only just beginning to look at externalities – the difficulty being that since they are mainly the result of non-market interactions, they are difficult to embody in a formal model. So I intend to consider first the formal model and then later consider externalities.

We can consider agglomeration economies on both the production side and the consumption side (Rivera-Batiz, 1988). On the production side, the theory emphasizes the role played by economies of scale at the level of firm, industry and market area. Some other theory on the consumption side emphasizes the utility gains that consumers can obtain by concentrating in space. These include economies of scale in the provision of local public goods and amenities and the access to a greater variety of goods and services in larger market areas. Contemporary theory integrates both.

The main thrust on both the production and consumption side is the role played by the service sector in generating agglomeration economies. We can identify two broad sectors within the urban economy, the non-traded producer services sector and traded goods and services. By non-traded producer services I mean local services that provide the inputs to industry and other services. These services are not traded directly (imported or exported) in national or international markets. The supply the needs of the city's industrial base and traded service sector, for instance repair and maintenance services in areas such as water and heating supplies, office equipment, industrial machinery servicing, communications, engineering, legal support, banking and insurance services and so on. In contrast traded goods and services directly supply final demand. They compete in national and international markets. I sometimes use the term industry to represented this sector of the economy operating in competitive markets, and resort to services as a shorthand for the non-traded producers service sector.

My emphasis in these lectures is on the production side. The assumption is that the aggregate production function for industry includes labour, space and a set of specialized producer service input. It is assumed that industrial production entails no internal increasing returns. On the other hand, there are increasing returns in the producer services that translate as an externality affecting the efficiency of industrial production. So there are external effects even here, although of a special kind. The way this occurs is that the agglomeration of producers increases the extent of the market for producer services. Producer services proliferate, become more specialized and this increases the productivity of final goods that uses them. (We can think of this as a localization economy, external to the firms in an industry, but internal to an industry. In contrast urbanization economies are external to each industry in a city, but internal to the city). The important role of producer (and consumer) services is captured by this quote from Mills (given in Rivera-Batiz, 1988),

'large urban areas provide specialized cultural, legal, medical, financial and other services that are not available in small urban areas'.

While that is obvious, it is only fairly recently that theory has developed an adequate framework which explains in formal and precise terms how services are organized so that they enhance industrial productivity. This boils down to a consideration of the market structure.

The market structure for the service sector is assumed to be monopolistic competition. The key aspect of this is the number of service firms, and hence the variety of services available. This number is an endogenous variable, it is an outcome of the logic of the model. In the past, variety has been simply an ad hoc assumption, and therefore not explained. Why should monopolistic competition be the appropriate assumption for the service sector. In fact markets for services are generally highly competitive, and face relatively minor entry and exit barriers, both features of monopolistic competition theory. Also, producers (and consumers) have highly specialized demands so that each service firm becomes differentiated, supplying a specific product. The wide variety of services that are needed to keep the modern industrial complex going, creates a demand for an almost infinite number of different specialisms. The wider the differentiation or variety, the greater the efficiency gains for final goods. To take an example, production will be more efficient if it employs a range of specialized software writers each skilled in his or her own language (C++, Cobol, Fortran, Open GL etc) than if it only employs one programmer who has to learn each different language as is required. It is reasonable to assume that output in a city will be related to the quantities of labour, space and producer services, but regardless of the quantity of producer services, what is also relevant is the number of different producer services. A rise in the number of services available increases industrial output, even if industry keeps the total quantity of services demanded the same. One hundred units of a single service is not of the same value as one unit from each of 100 services[1].

These models have the feature that the size of the city and its labour force will determine the number of specialized producer inputs. On the whole a larger city will have a greater variety of producer inputs. Since the greater variety adds to output, in this type of model larger cities are more productive.

Of course the advantages of size do not go on for ever. The land market and commuting costs means that at some point the increased cost of large cities (higher rents as a result of the competition for space and longer commuting journeys) will offset the production and consumption advantages of diversity. Other costs like noise and pollution will also be higher. Even when we take these into account, the optimal city size will be larger than if we did no allow for the effects of diversity in production and consumption. Urban output will be larger and productivity will be greater. The utility of residents will be greater. Larger cities contribute more than proportionally to national output.

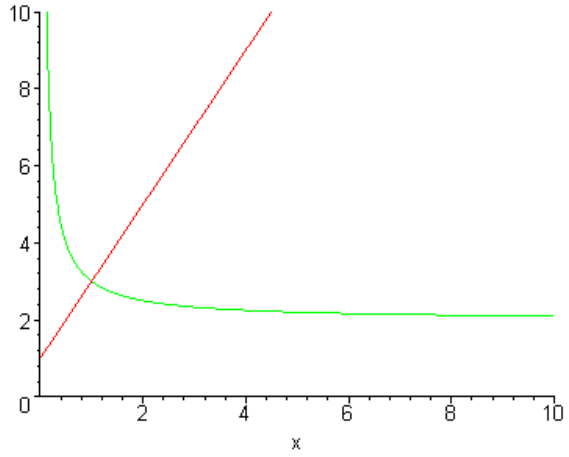
[1] Similar statements can be made about consumer utility benefiting from urban diversity on the consumption side

Now I wish to highlight various concepts we will use to understand the economics behind this phenomenon.

- Increasing returns

Increasing returns is short for increasing returns to scale, meaning lower average costs per unit of output for the firm as the level of output increases. It is manifested as a downward sloping average cost curve. What we see here are internal increasing returns to scale. As output (x) increases, while total costs rise average costs fall.

Figure 2



Scitovsky(1954) distinguished between internal and external economies of scale. Internal economies of scale means that the decrease in average costs is due to an increase in the production level of the firm itself. Since this implies some advantage accruing from size, the implied market structure is imperfect rather than perfect competition. With external scale economies average costs are a function of the level of output of the industry as a whole. External economies can also apply to a sector or the whole economy rather than the firm (also, we have thus far considered static external economies, but dynamic ones may also exist, meaning that average costs per unit of output are a function of the cumulative output of the industry).

It has long been intuitively obvious that increasing returns go hand in hand with city formation. This obvious relationship in fact explains why urban economic was until recently somewhat in the doldrums, Mainstream economists found it difficult to develop formal theory which embodied increasing returns. Since the late 1980s, things have changed and urban economics has moved, along with economic geography, more to the center stage of economics. Elegant theory has been developed which incorporates increasing returns.

The opposite of increasing returns is non-increasing returns. There is no advantage from producing on a larger scale, firms produce at a small scale with no loss of efficiency. Producers then will tend to operate near to where consumers are located, since there are no advantages to be gained from operate on a large scale from a more

remote location and offsetting the transport costs by the efficiency gains. We will see everything that is required produced within a stone's throw of the house. There will be no transport of goods and services in, its all there on your doorstep since production can be anywhere at any scale without loss of efficiency. If population is evenly distributed, we would have a world without cities. This has been called the Robinson Crusoe economy.

- Monopolistic Competition

Monopolistic competition is the basic market structure assumption for the non-traded services sector. It embodies internal increasing returns which is the basis for increasing returns to the scale of city for competitive manufacturing/final goods and services producers. Monopolistic competition is a form of imperfect competition and when we look at what it entails we see that it is ideally suited as a description of the structure of the service sector.

Of the two basic forms of imperfect competition, oligopoly and monopolistic competition, monopolistic competition clearly fits the bill as a model for the service sector. Under oligopoly there are only a few producers, each recognizing that its own price and market share depends on its own output and on the actions of competitors, so there is an element of strategic interaction between firms. Under monopolistic competition, there are very many sellers producing products each producing a different variety, these varieties are to some degree, depending on the precise assumptions of the model, substitutes. Since there are many firms, there is no role for double guessing what rival are up to, in other words no strategic interaction between competing firms. With a large number of small firms, any action taken by an individual firm is of negligible consequence for other firms in the sector. Like perfect competition there is free entry and exit. However, a monopolistically competitive firm has a downward sloping not horizontal demand curve. It cannot sell as much as it wants at the going price.

As an example of monopolistic competition we might consider law firms providing specialist services to industry, since each exhibits product differentiation by specializes in a particular industrial sector. Likewise travel agents providing travel services to businesses will offer differentiated areas of expertise, and so on across a whole range of services. We characterize them as typically comprising numerous small firms providing differentiated services.

While it makes good sense to use monopolistic competition as a model for the structure of the service sector, there is also a sense it which it has become an orthodoxy and a way for a theory to look respectable as a part of mainstream economic theory. The most influential piece of work is the paper by Dixit and Stiglitz published in 1977 in the American Economic Review entitled 'monopolistic competition, and the optimum product diversity'. This revolutionized model-building in several fields of economics, trade theory, industrial organization, growth theory, geographical economics, and urban economics. It provided an elegant and simple way to model production at the firm level benefiting from internal economies of scale operating in a monopolistically competitive market.

- How monopolistic competition in the service sector leads to aggregate increasing returns for final goods.

The most interesting and typical account of the new urban economics based on increasing returns modeled via monopolistic competition is described by Rivera-Batiz (1988) and Abdel-Rahman and Fujita(1990), I rely on the simpler version given by the latter, hereafter AR-F. They consider a city with two sectors, a non-traded specialized producer services sector, referred to in short as the service sector, which supplies services to industry/manufacturing. Services are under monopolistic competition, industry has constant returns to scale so there is no inherent advantage from the technology available to manufacturers for producing on a large scale. Outputs are directly proportional to inputs with no efficiency gain. Paradoxically, despite manufacturers operating under constant returns, there are increasing returns in the labour force for aggregate final goods production of the city.
These increasing returns, that lead to industrial agglomeration, derive from the greater availability in larger cities of more specialized local producer services, such as software and information technology services, legal services, financial and advertising services, communications services, consultancy and so on.
These finely differentiated services are an input to the production function of a firm engaged in final goods production, which is

Equation 1

$$Q = M^{\beta} I^{1-\beta}$$

This is a standard Cobb-Douglas production function with constant returns to scale, since $0 < \beta < 1$. Output (Q) equals final goods labour, or the number of workers (M) raised to the power β, multiplied by the level of composite services (I) raised to the power $(1-\beta)$. It is called composite services because it is an amalgam of all the different varieties of specialized producer services.

The parameter β controls how important workers are versus composite services in determining the level of output Q. With small β, workers are relatively unimportant compared with services in what determines final goods output Q. With large β, then workers are more important and the variation in the level of composite services has less impact on variation in the level of final goods output Q.
Denote $\alpha = 1 - \beta$, when $\alpha + \beta = 1$ we have constant returns, so that doubling inputs doubles output. But when $\alpha + \beta > 1$ then we have increasing returns, doubling inputs more than doubles output. We see this from the following graph.

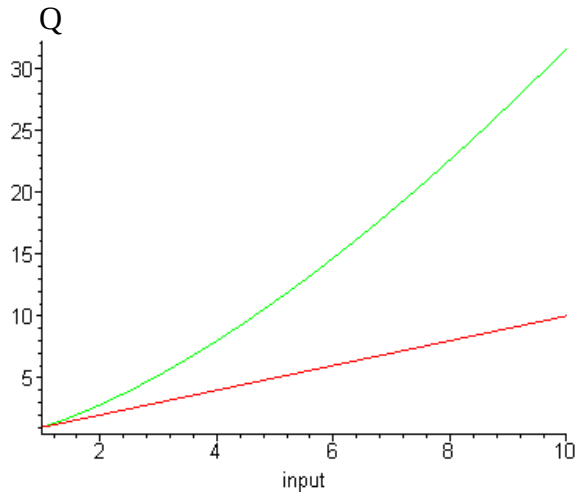


Figure 3 Constant and increasing returns

Figure 3 illustrates that as the value of each input go from 2 to 8 output also goes from 2 to 8 when $\alpha + \beta = 1$. However when $\alpha + \beta > 1 = 1.5$, the fourfold increase from 2 to 8 causes Q to increase from about 2.83 to 22.63, a factor of 8. In fact for final goods we assume that production is competitive, there are no increasing returns due to $\alpha + \beta > 1$, doubling inputs simply doubles outputs. Nevertheless, there are increasing returns to city size. How does this come about?

It is the presence of the level of composite services *I* which produces the result that there are aggregate increasing returns for the aggregate final goods production function for the city.

The real trick in understanding this is to look at what determines the level of composite services. It is not simply a summation of the individual output from each service firm in the city, it is more, it depends on the number of separate varieties that exist. There is assumed to be a 'love of variety' so that a given quantity produced by a large number of differentiated firms rather than a small number boosts the composite level of services available in a city. Taken to the extreme, if all service firms were identical (ie perfectly substitutable), then this is of less value for final goods output that if they are all different. The standard assumption adopted in the literature such as AR-F, is that *I* is a function of the number of different services, denoted by *x*, as follows

Equation 2

$$I = \left[\int_{t=1}^{t=x} i(t)^{1/\mu} dt \right]^{\mu}$$

$$I = [\sum_{t=1}^{x} i(t)^{1/\mu}]^{\mu}$$

in other words we obtain the level of composite services as the sum across *x* varieties. If we have a large enough number of specialized producer services, then the continuous integral is quite closely approximated by the discrete sum. The quantity *i(t)* is the level of output of a typical producer service firm. It is raised to the power

$1/\mu$, and the result for each of the x firms is then added together. The overall sum is then itself raised to the power μ. The reasons for this have yet to be explained, but none the less it is clear that this is not a Cobb-Douglas production function as has been assumed for Q. In fact it is a constant elasticity of substitution (CES) production function and it has the particular property that the elasticity of substitution between the separate varieties, e_s, is constant! This features very prominently in what follows, since it the way in which we model imperfect competition and increasing returns.

Assume for the moment that each firm t produces the same amount of output $i(t)$. In other words their costs are identical and the amount purchased is equal across all service firms. It is then the case that summing over x firms is the same thing as x multiplied by the constant $i(t)$. The rather complicated CES production function then becomes very much simpler, as below

Equation 3

$$I = [\sum_{t=1}^{x} i(t)^{1/\mu}]^{\mu} = [xi(t)^{1/\mu}]^{\mu} = x^{\mu} i(t)$$

Numerically, we see that if $\mu = 2$, $i(t) = 9$, $x = 3$
Then the first version (eqn. 2) gives I = [9**0.5 + 9**0.5 + 9**0.5]**2 = 81
The second version (eqn. 3) gives I = (3 * 9**0.5)**2 = (3**2) * 9 = 81

So the level of producers services available for industry in the city depends on three things, the typical level of output of the service firm $i(t)$, the actual number of service firms in the city x, and a parameter μ. Notice what happens when $\mu = 1$, then the level of services I is simply each firms output i(t) multiplied by the number of firms, which is what you might think it should be.
However, look what happens if $\mu > 1$, it is now the case that I is greater than i(t) times x. There is an extra ingredient boosting the level of services. In fact as μ becomes increasingly large, it follows from the definition of the elasticity of substitution, $e_s = \mu/(\mu-1)$, that the elasticity of substitution falls. A low elasticity of substitution means that the individual varieties of services are not very close substitutes for each other. In other words, with a low elasticity of substitution, doubling the number of firms x results in a more than two-fold increase in I. Variety matters. This is the source of increasing returns for final goods production.

In contrast, if we return to the situation where the elasticity of substitution is very high, this is commensurate with μ close to 1, and the level of composite services tending to the total level of output for services as given by the number of firms times their typical output i(t). One way to think of this is that with μ=1 variety does not matter as a determinant of the level of services and 100 units of one variety gives the same input as one unit of 100 varieties. In this case products are perfect substitutes so that one unit less of one variety can be exactly compensated by one unit more of another variety. Brakman et al (2001 p. 68) provide an explanation of this also.

- Aggregate outcomes

From the foregoing argument it is possible to obtain the aggregate production function for final goods . This is given here without derivation, but we give the mathematics leading to this expression later. It is a nonlinear function involving the size of the city (*N*) and the level of output (*Q*). This works as follows.

$$N \to x \to I \to Q$$

First, if the city increases in size (*N*) it has more workers both in final goods production and in the specialized producer services. We have seen from the Cobb-Douglas production function that increasing final goods workers (*M*) directly increases final goods output (*Q*). Increasing the number of service workers also increases final goods output, but in an indirect way. The increase in the city size increases the number of service varieties (x) available, not the size of the individual firms. There is a direct linear relationship between city size and the number of service firms x. We see from equation (3) that increasing x increases *I* which then plugs into the Cobb-Douglas production function and increases *Q*. Moreover, increasing the number of varieties x has a more than proportionate increase in the level of composite services (*I*). This is due to the way we have defined the level of output for composite services as a CES production function.

Figure 4 shows how the model works to produce a nonlinear relation between x and *I*. As the number of service firms (x) increases, so we get a nonlinear increase in *I*. The precise curvature of the relationship depends on the assumed value of μ , in other words on the elasticity of substitution. As we increase μ, in other words as the elasticity of substitution falls and firms become more monopolistic, the curvature increases. So, we expect μ will control the extent of increasing returns of final goods output to city size. In the figure the upper line is for μ =2, the lower one is for μ=1.1.

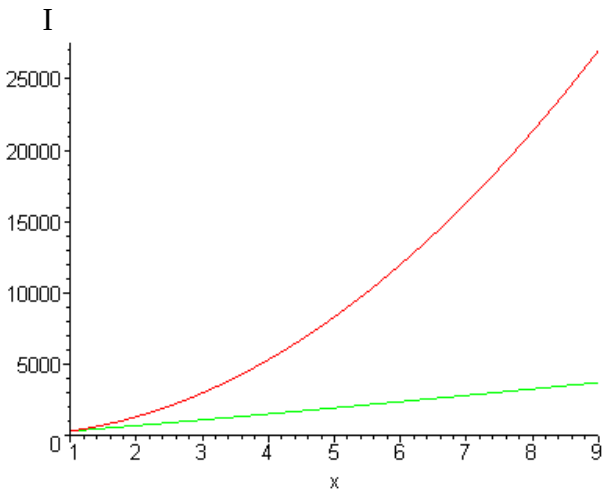


Figure 4 The relation between x and *I* for different μ (2, 1.1)

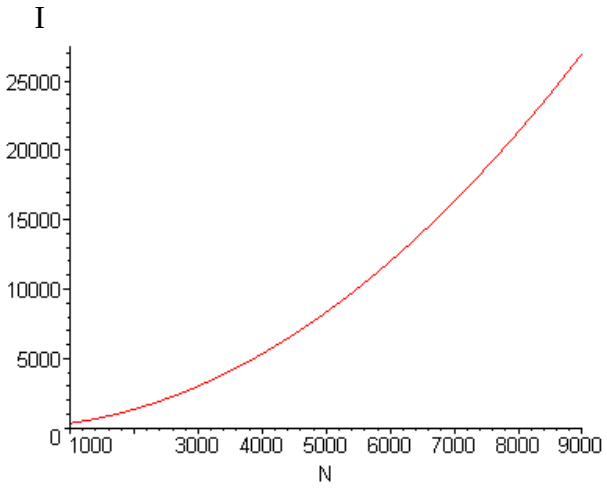


Figure 5 The relationship between *N* and *I* for μ =2

Figure 5 shows that, since the relation between x and *N* is linear, so that x is simply some multiple of *N*, and the relation between x and *I* is nonlinear, the result is that *N* increases, we get a nonlinear increase in *I*.

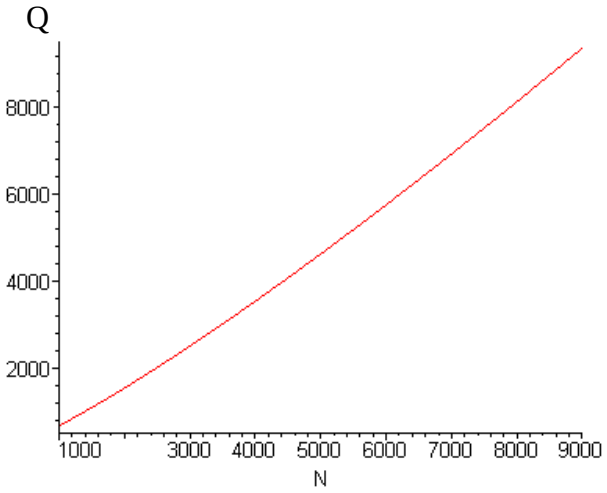
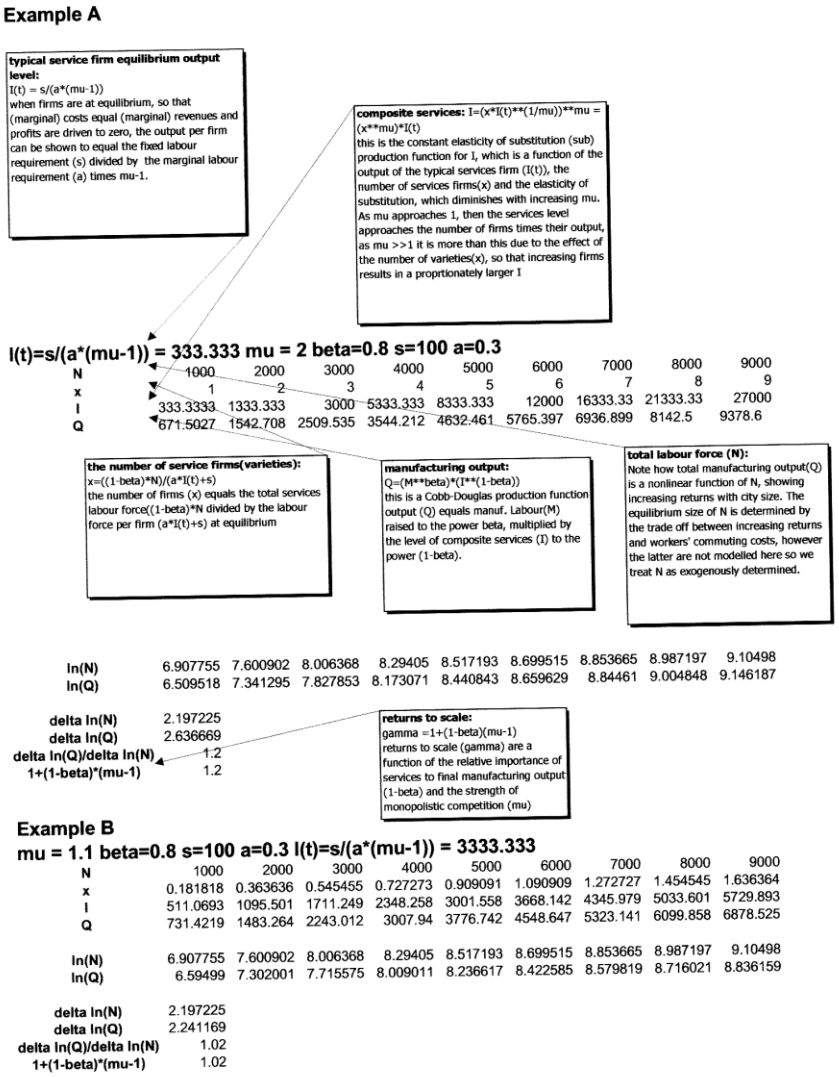


Figure 6 The relationship between *N* and *Q* for μ =2

Figure 6 shows that the increase in *Q* per unit increase in city size (*N*) is (slightly) nonlinear. The reason it is nonlinear is because *Q* is a function of *I*, and we know that the relation between *I* and *N* is nonlinear. The reason why it is only slightly linear is because *Q* depends only partly on *I*.



Example A gives the precise relationships which lead from city size to level of output which correspond exactly to the numerical values used to construct Figures 4, 5 and 6. The key features here are the rows of the spreadsheet labeled *N*, x, *I* and *Q*. These provide the values plotted on the graphs. The way in which we get the values in one row from the values in another row depends on the relationship between the variables which is determined by theory. First, let us look at the way x depends on the city size *Q*. In fact we have to accept for the moment an assumption, that there is an equilibrium fixed size of service firm, so that as the city gets bigger, the size of firm remains constant. If the city gets bigger, then the number of service workers increases, and if each firm employs a constant number, the number of firms not the size of each firm stays the same. This is in line with what we might expect in a large city, very many small firms supplying very specialized services to industry. The equation determining x is given in the box entitled 'the number of service firms'. The numerator is the number of service workers, equal to a proportion β of the total number of

workers *N*. We divide this by the number of workers employed in each firm. This is the equilibrium level of output of each firm i(t) times the marginal labour requirement, in other words the labour needed per unit of output. Even before the firm starts producing however, there is a need for some labour input. This is the fixed labour requirement denoted by s. Together these give the total labour used by each firm, the denominator in the equation in the box . Dividing total service workers by workers per firm gives the number of firms x. As we can see, as the city size rise from 1000 to 9000 (obviously a very small city, but adequate to demonstrate the arithmetic!), the number of firms rises from 1 to 9, a direct linear relationship.

When we know x, we can then calculate *I* by plugging into the CES production function for composite services. We also need to know i(t), the level of output of the typical firm. For the moment you will have to simply accept that this constant has a particular value in equilibrium equal to a function of the fixed and marginal labour requirements and the parameter μ that determines the elasticity of substitution between the different varieties of services. The precise equation is given in the box entitled 'typical service firm equilibrium output level'. Assumed values of the various constants determining i(t) and the resulting value are given in bold. These values are precisely the ones that I have used to calculate the row *I* of example A.

Finally, how do we get the *Q* row. We know that final goods output has a Cobb-Douglas production function, with arguments *M*, the amount of final goods labour in the city, and *I* the level of composite services. We know the value of *I* now, and *M* is simply that share of *N* that are not service workers. The other quantity is β, which determines the worker shares and the relative importance of *I* and *M* in the Cobb-Douglas production function. This we are assuming is equal to 0.8. So, once we have, for the different *I* values, calculated the *Q* row, we can see that the *N* to *Q* relationship is a nonlinear one. As city size *N* increases, output level also increases nonlinearly. The result is Figure 6.

Now we do not actually know the precise values of these constants, but we can play around with them to show the impact of, for instance, increasing the relative importance of composite services *I* compared with workers *M* in determining the level of final goods output *Q*. We would do this by making β smaller. This will in fact increase the curvature of the relationship between *N* and *Q* because *Q* is now weighted more by *I* which is nonlinear in *N*.

If we keep everything the same as in Example A but change β to 0.1. the result is the following relation between *N* and *Q*,

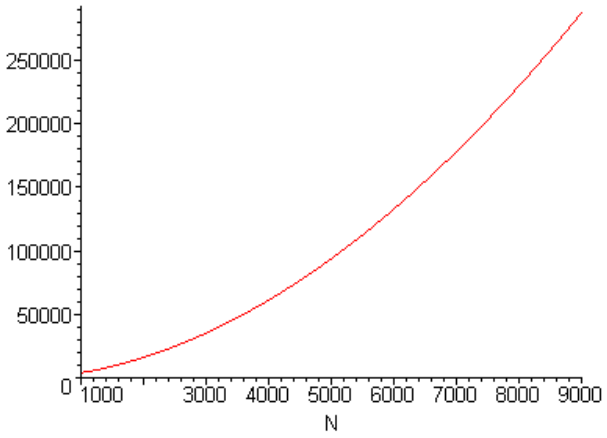


Figure 7 The relationship between *N* and *Q* for μ =2, but with β =0.1 rather than 0.8

Figure 7 is much more curved than Figure 6, showing the heightened effect of *I* in Q's production function.

Another parameter we can change is μ which controls the elasticity of substitution of the different services. The bigger is μ, the lower is the elasticity of substitution and the different service varieties, it controls the relationship between the number of service firms in the city x and the overall level of composite services *I*. If $\mu \approx 1$ then the level is roughly number of firms times output per firm. If it is more, variety counts more and the level of *I* is more than this. In Figure 4 we saw how changing μ changes the curvature of the relationship between x and *I*. This has knock on effects for the relationship between city size *N* and *Q*. This also becomes more curved, other things being equal.

The effect is similar to the effect of diminishing β so I do not need to draw the graph. If instead of increasing μ we move it close to 1, then the opposite takes place. The relationship between *N* and *Q* becomes more like a straight line. When it is exactly linear this means that there are no increasing returns with city size. Example B corresponds closely to this scenario. In this I have reduced μ to 1.1, and kept everything else the same, so β is as in example A and so are the marginal labour requirement per service firm, equal to a, and the fixed labour requirement s. However changing just the constant μ from 2 to 1.1 has the effect of also changing the equilibrium size of each service firm, since this depends on μ (as shown in the box entitled 'typical service firm equilibrium output level'). Again, I haven't explained why this is the case, you will just have to accept it for the moment.

This relationship between *Q* and *N* can be expressed as a precise equation. In order to obtain this equation we need to do a bit of algebra, which is set out in Equation 4. The main things to remember here are that we start out with two production functions. One is the Cobb-Douglas production function for competitive industry, the other is the constant elasticity of substitution production function (CES) for producer services. The latter determines the level of composite services in the city, which feeds into the Cobb-Douglas. We have already seen how x is determined, that is the number of service firms in the city equal to the number of service workers divided by the service

workers per firm, which is constant because firm output i(t) is constant, and marginal and fixed labour requirements per firm are constant. Likewise *M* is a constant, equal to the share β of total employment *N*. Obtaining the mathematical relation between *Q* and *N* is then simply a matter of substitution and gathering together all the constants as a single parameter ϕ.

Equation 4

$$Q = M^{\beta} I^{1-\beta}$$
$$I = x^{\mu} i(t)$$
$$Q = M^{\beta} (x^{\mu} i(t))^{1-\beta}$$
$$Q = M^{\beta} x^{\mu-\mu\beta} i(t)^{1-\beta}$$
$$x = \frac{(1-\beta)N}{ai(t)+s}$$
$$M = \beta N$$
$$Q = N^{\beta+\mu-\mu\beta} \beta^{\beta} (ai(t)+s)^{\mu(\beta-1)} i(t)^{1-\beta} (1-\beta)^{-\mu(\beta-1)}$$
$$Q = N^{\beta+\mu-\mu\beta} \phi$$
$$Q = \phi N^{1+(1-\beta)(\mu-1)}$$

In order to understand equations(4), note that the first 4 lines involve a rearrangement and substitution so that *Q* is a function of *M*, x, i(t) and constants μ and β. We are interested in obtaining *Q* as a function of *N* and some constants. The fifth line is the equation linking the number of firms x to *N*, and the sixth line links the number of final goods workers *M* to total workforce *N*. If we replace x and *M* in line 4 by these functions of *N*, we get line 7. In line 7, everything to the right of *N* is a constant, which we collectively call ϕ, hence line 8. Line 9 simply re-expresses line 8 in a more convenient way. In equations (5), the exponent γ is used to represent the more complicated expression. Note that equation 1.9 on page 26 of Huriot and Thisse(2000) is identical to this, although they have used a different parameterisation.

Equation 5

$$Q = \phi N^{\gamma}$$
$$\gamma = 1 + (1-\beta)(\mu-1)$$

We now have the expression linking *Q* to *N* which is the equation of a curve. It corresponds to the curve given as Figure 6 or 7, depending on our choice for the values of the various constants that define the values of ϕ and γ. How do we know what values of ϕ and γ to use in any real situation, so that we can get a realistic plot of the relation between city size *N* and level of output *Q*?
If we know β and μ we can obtain γ. However, it is often the case that we don't know β and μ, but we do know the sizes of a collection of cities and the amount of final goods output in each of these cities. We can look at how *Q* relates to *N* to estimate if we do indeed have increasing returns to scale from the data, or is it just a theory with no real world validity?

Given real data, in other words the *N* values and the *Q* values, it is a simple matter to obtain and estimate of γ. First, it is better to express the relationship between *N* and *Q* as a straight line, then look at the slope of the line to check if we have increasing returns. If we take natural logs then the relation between ln(*Q*) and ln(*N*) is a straight line relationship. Then it is easy to obtain γ and check its value to see if there are increasing returns.

Hence we take natural logs of equation (5) which gives

Equation 6

$$\ln(Q) = \ln(\phi) + \gamma \ln(N)$$

This is the equation of a straight line (y = mx + c). We now simply need to obtain the slope γ. It is the change in ln(*Q*) per unit change in ln(*N*). We obtain the change in ln(*Q*), Δ ln(*Q*) corresponding to the change in ln(*N*) as we move from a city with *N*=1000 to one with *N*=9000, which is Δ ln(*N*). Dividing Δ ln(*Q*) by Δ ln(*N*) gives the change in ln(*Q*) per unit change in ln(*N*), the slope of the line which is equal to γ. The various steps are set out in equations (7).

Equation 7

$$\ln(Q_1) = \ln(\phi) + \gamma \ln(N_1)$$
$$\ln(Q_2) = \ln(\phi) + \gamma \ln(N_2)$$
$$\ln(Q_1) - \ln(Q_2) = \gamma(\ln(N_1) - \ln(N_2))$$
$$\Delta \ln(Q) = \gamma \Delta \ln(N)$$
$$\frac{\Delta \ln(Q)}{\Delta \ln(N)} = \gamma$$

As it happens, while this works when we only have data for *N* and *Q*, in this instance with examples A and B we do not need to do this in order to find γ, since γ is already known once we know β and μ (see equation (5)). Plugging in β and μ to obtain γ gives exactly the same result as using the steps set out in equations (7).

Thus, to illustrate the empirical method of equations (7) using example A, we first set out the relationship between ln(N) and ln(*Q*), which are the two rows with values obtained by taking the natural logs of the *N* row of values, and the natural log of the *Q* row of values. Below the ln(*N*) and ln(*Q*) rows, we have values for Δ ln(*N*) and Δ ln(*Q*) . To obtain these I have used the first and last values in the rows above, but I could have used any pair of values. Below we get the result of dividing Δ ln(*Q*) by Δ ln(*N*) , we see that the result of this calculation is γ = 1.2. I have labelled this as returns to scale, since that is what is being estimated. It is well known that the slope of a log-log relationship is the constant elasticity, so it is the % change in Q for 1% change in *N*. Hence, *Q* increases by 1.2% for 1% increase in *N*. There are increasing returns to city size.

In this particular example there is no need to go through the steps of equation (7) to obtain the elasticity, we are in the fortunate position of already knowing the precise values of the parameters on which it depends. Since β =0.8 and μ = 2 in example A,

we can simply plug these values in to the equation for γ, we get exactly the same result.
Example B does the same thing as example A, but with a different value for $\mu = 1.1$. If we were to look at the graph plotting the *Q* against the *N* values from example B, we would find that it would be much more linear than even Figure 6, almost straight in fact, reflecting very marginal increasing returns. In fact, given this data, we could be fooled into thinking that there we no increasing returns at all, after all $\gamma = 1.02$ which is very close to 1. A 1% increase in city size produces a 1.02% increase in final goods output.

If we go one step further so $\gamma = 1$ exactly, there are no increasing returns. How might this arise? We see from equation (5), the equation for γ, that when $\beta = 1$, $\gamma = 1$ and there are no increasing returns. This is obvious when we appreciate that increasing returns derive from the service sector. Setting $\beta = 1$ means that composite services have zero weight in the final goods production function. Therefore even though there may be a large variety of specialized services in the city, they don't count when it comes to final goods output.
Another way we go get constant returns is as a result of setting $\mu = 1$, since this also has the effect that $\gamma = 1$. As μ approaches 1, monopoly power to service firms diminishes and the elasticity of substitution tends to infinity. Since the services are in this case completely interchangeable, there is no variety to speak of and therefore no increasing returns coming through as a result of the effect of greater variety in larger cities.

The above equation is the main result of the new urban economic theory. The question arises, is the theory true in the real world. There are indeed some missing factors, really externalities, which should also come into the equation. Things may not be so clear cut because even if we were able to measure *N* accurately, *Q* is bound to be measured with error. We need a method of estimation that allows for the measurement error. This might be crucial if the data indicate that γ is close to 1 but we don't know whether or not this means that the true γ, if we were able to eliminate measurement error, is exactly 1.

In order to do this, it is possible to use regression analysis to regress $\ln(Q)$ on $\ln(N)$ to obtain an estimate of the coefficient γ and test the null hypothesis that $\gamma = 1$.
If we do not reject the null, then there is no evidence for increasing returns. If we reject the null in favour of the alternative hypothesis that $\gamma > 1$, then we have empirical evidence for aggregate increasing returns. In terms of the straight line graph, we are therefore testing whether the slope of the line is significantly greater than 1, acknowledging that the slope we see is contaminated by measurement error other missing effects. The empirical test would come from having data on a set of cities of different sizes (*N*) and different levels of final goods output (*Q*) and plotting the results on a graph, fitting a best fit line to the logarithmic values, and estimating via regression analysis the slope of the line. However, we cannot really expect such a simple theoretical model to fit real data; there are too many other factors, such as externalities, which we have not yet considered.

2. Market structure and other assumptions

There are some assumption we are making in our analysis thus far which have not been fully examined and justified. These are to do with the micro-economics of the service firm. Remember, we assume that each is the same size in equilibrium, each employs the same amount of labour. While this may seem unrealistic, it does at least provide a simplified picture of the very complex reality of the urban economy. We are prepared to sacrifice realism for theoretical elegance at this point in time, although in due course we attempt to make our model more realistic. Given this idealised picture of the urban economy regarding the characteristics of the service firms and their collective market structure, namely monopolistic competition, the outcome is increasing returns to the size of city. We saw this outcome in the form of a nonlinear relationship between the level of final goods output (Q) and city size (N). Bigger cities are more productive.

It is worth noting that (more or less) the same relationship is also an outcome of a different tradition in urban and regional economics, that which has been strongly influenced by people such as Keynes and Kaldor (see Fingleton, 2001). One of the big differences between the two approaches however is that the AR-F model and new economic geographers have insisted on a level of theoretical formalism that goes down to the micro-economic level. In other words the new theory has insisted on a general equilibrium approach and it is this aspect which has pitched it into the mainstream of contemporary economics.

What do we mean by micro-level assumptions and general equilibrium approach? Let us first deal with the micro-level assumptions, which I have glossed over so far. For the producer service sector, the assumption is that they are under monopolistic competition and are non-traded. They supply final goods and services producers and not final consumers. There are no exports and imports of producer services so that the city's supply and demand are equal. In fact explicit assumptions are made regarding the behaviour of the individual service firm, and these assumptions lead to a logical decision on the part of each firm to produce an equilibrium quantity $i(t)$. This then determines the number of varieties (x) and leads ultimately to the aggregate increasing returns for final goods output (Q) with city size (N) that we discussed earlier.

In fact when we talk about monopolistic competition as the market structure for services, we are in fact referring to a specific model developed by Dixit-Stiglitz which has been used throughout different branches of economics as a convenient and useful representation. Under the model, each firm produces a different variety. The reason is simple, there are fixed costs incurred in establishing any one variety in the market place. Many modern producer services can be said to have a significant fixed cost (s) which has to be paid before any service is delivered. In general this can be seen as the cost of 'figuring it out' so that the service is viable and has a market, for example the cost of writing new software or setting up the company with a new logo and developing an advertising pitch and researching the market place. In contrast with things running smoothly as a result of this initial investment, the marginal cost (a) per unit of output may be quite small, often simply the cost of copying the product onto CD for example, and this type of cost may be much less than the initial programme development cost and the research and development that went into the initial set-up. This means that there will be benefits from being larger rather than

smaller, since then average costs will fall. Given a high initial fixed cost, this will deter varieties being shared by companies. The most efficient way to produce a variety is as a single company which once it grows can reap internal increasing returns. There is no point in splitting the production of a variety between different firms since each will have to incur what we will assume will be the same fixed cost and the level of output of each will be smaller, so that the scale economies will be less. Hence logic dictates that it is better to produce each variety in a single firm so as to maximise returns to scale.

So we have each firm producing its own variety of service, and there are fixed and variable costs which mean that it makes sense to be larger since there are internal returns to scale. We can see this if we make the assumption that, since broadly speaking, services are relatively labour intensive, each producer service firm uses only labour as an input. This means that the firm's labour requirements are

Equation 8

$$L = s + ai(t)$$

the labour requirement L needed to produce the output of the firm $i(t)$ is equal to fixed labour requirement(s) plus marginal labour requirement(a) times firm's output $i(t)$.

This is the firm's production function, since we can rearrange it to have i(t) as the output dependent on labour input ($i(t) = (L - s)/a$) . Equation 8 is the function in Figure 1, which is a linear function with constant s and slope a. Figure 1 also gives the downward sloping average labour requirement. The average labour requirement is $L/i(t)$ which falls as $i(t)$ increases, so we obtain the downward sloping curve which we define as internal increasing returns.

Up to now we have firms with internal increasing returns, but why shouldn't they keep on increasing in size, since bigger means better, why is there an equilibrium size to which they converge? The reason is that increasing output increases costs as well as revenues. Revenues increase because sales are higher, but costs increase because more workers are employed. So, there is an equilibrium service firm size at which profits (π), equal to revenues minus costs, are at a maximum.

Equation 9

$$\pi = p_t i(t) - w(ai(t) + s)$$

So we see that price p times quantity sold i(t) equals revenue. Wages (w) times labour equals costs.

In fact, we obtain the typical firm size in a roundabout way, since it is based on the price charged for the service. There is a price at which profits are maximised, and this price in turn determines firm size. How is the price set by the firm? . It is in fact useful to have quite a specific demand function showing how the quantity demanded i(t) changes with price, this is

Equation 10

$$i(t) = kp_t^{-(\mu/(\mu-1))}$$

The demand for a variety i(t) depends simply on the price of that variety and on two constants, k and the constant μ. How can we make the assumption that the demand

function is of this form? Why is this useful? What are the constants that are present in the demand function? What are the implications of assuming that this is the demand function for the firm's services?

First, let us look at the constant k. What we are in effect assuming here is that there is no strategic interaction involving the service firms, in other words the price set by one firm, since it is one of a large number of firms, has no effect on the pricing strategy of its competitors. The quantity demanded depends simply on its own price. Firms are said to be myopic when it comes to strategic interaction, and keep their output the same regardless of the of the price charged by their competitors. There are more complicated expressions given by Rivera-Batiz(1988) and Abdel-Rahman and Fujita(1990) in which k is represented by a function of the level of composite services. If we have strategic interaction, so that other service firms change their output in response to changes in firm t's price, then I will change as p_t changes.

The second constant in the demand function is μ. In fact the term $\mu/(\mu-1)$ is equal to the price elasticity of demand (ped). Using this price elasticity is one of the main advantages arising from using the Dixit-Stiglitz model of monopolistic competition. It simplifies the theoretical analysis. Every firm has the same price elasticity of demand. The reason why $\mu/(\mu-1)$ is the price elasticity is simply because it measures the proportional change in quantity demanded divided by the proportional change in price. More precisely it is the derivative of quantity with respect to price divided by the ratio of quantity to price, hence

$$i(t) = kp^{-\mu/(\mu-1)}$$

$$\partial i(t)/\partial p = -\frac{kp^{-\mu/(\mu-1)}\mu}{(\mu-1)p} = -\frac{i(t)\mu}{(\mu-1)p}$$

$$ped = -\frac{\partial i(t)/\partial p}{i(t)/p} = \frac{i(t)\mu}{(\mu-1)p}\frac{p}{i(t)} = \frac{\mu}{\mu-1}$$

We have not yet addressed the question of why the demand function should have the particular form that it does. In fact it is possible to show that the demand function is the solution to the problem of choosing i(t) so that the final goods profit level $\prod$ is maximized (see Abdel-Rahman and Fujita, 1990, Rivera-Batiz 1988 for an explanation). Let us say that the city's traded goods or manufacturing sector behaves as a single firm. We can make this assumption because competitive firms can be any size, there is no equilibrium size.

Equation 11

$$\Pi = PQ - [Mw + \sum_{t=1}^{x} p_t i(t)]$$

Hence final goods revenue equals price on the world market P times amount of manufactures produced Q. Costs equal final goods wages w times number of workers M plus the prices of each service input p(t) times the amount of each service used i(t), added up across all x services. Solving for i(t) to give us the quantity that maximizes

the final goods firm's profit Π gives us i(t) as a function of price. The solution to this maximization problem gives us the demand function for services.

Let us now look again at the equation for the profit level π of the service firm (equation 9), which included the typical level of service firm output i(t). We have now justified a demand function in which i(t) is a function of the price set. This means we can replace i(t) in the profit equation and write the profits of the typical producer firm in terms of prices also, hence

Equation 12

$$\pi = p_t k p_t^{-(\mu/(\mu-1))} - w(a k p_t^{-(\mu/(\mu-1))} + s)$$

What is the price to be set by the firm to assure that profit is at a maximum? The technique here is to look at how the profit rises as the price rises, and then falls as the price rises even more. The maximum profit is where price rise turns into price fall, at which point the slope of the line will exactly equal zero. We can see this graphically in Figure 8, the maximum profit seems to be at a price of 1.5.

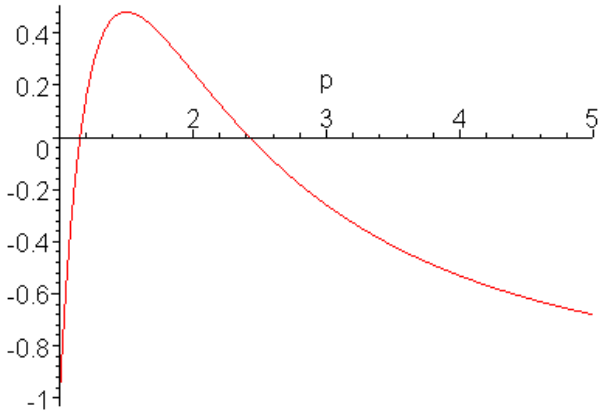


Figure 8 Profits versus price

This is a graph of profit (vertical axis) versus price for some arbitrary values w=1,a=1,s=1, const=10, μ=1.5. Note that profits are at a peak at p(t) = 1.5 = waμ

To find this quantity mathematically, we differentiate π with respect to *p*. This gives us the slope or derivative $\delta\pi/\delta p$ of the curve at any price. If we then work out the price at the peak where he slope is zero, we have the profit maximising price. In fact it is possible to show that when $\delta\pi/\delta p = 0$, the price is exactly equal to

Equation 13

$$p_t = wa\mu$$

So we see that price (p) = wage rate (w) times marginal labour requirement(a) times μ. So under monopolistic competition we see that prices are not equal to marginal cost (wa) but is higher. Since all these three terms are constants, the price charged by all firms is the same, so they all have identical demand and output. This is known as mark-up pricing, since the marginal cost of producing an extra unit of output is wa, and the price the firm charges is a mark-up on this marginal cost by the factor μ. This means that if μ is large, then the price elasticity of demand μ/(μ-1) is small and the mark-up is large. Firms have a lot of monopoly power. If on the other hand μ is small, then the price elasticity of demand μ/(μ-1) is large and then the mark-up is lower.

One of the big assumptions we have been making is that there is an equilibrium firm size. We now show that in the longer run profits are driven to zero because of the ease of entry into the sector. This zero profit situation is an equilibrium and defines the equilibrium firm size. From the graph, we see that firms can obtain profits be charging a price equal to 1.5, which is a mark-up on marginal cost. However, this is only a **temporary situation**. Given these excess profits under monopolistic competition, it becomes very attractive for other service firms to enter the market. We assume that there is free entry and exit from the market, there are no entry barriers imposed for instance by the existing firms, and nothing to stop firms leaving if profits turn into losses. Each new firm that enters provides a new variety of service, so that manufacturers allocate their spending over a wider and wider spectrum of services. Each variety is a substitute for each other, so the entry of new varieties implies that existing firms have their profits eaten into by the advent of the new varieties. This process of entry of new varieties continues until profits fall to zero. In fact the zero profit position is an equilibrium. If profits are negative then firms exit and profits rise to zero.
The zero profit equilibrium has an associated equilibrium level of output for each firm. In order to see this remember profits π are zero at equilibrium. We start with our equation of revenues minus costs written in terms of i(t) and equated to zero. We then rearrange to find an expression for i(t) at equilibrium, hence,

Equation 14

$$\pi = p_t i(t) - w(ai(t) + s) = 0$$
$$p_t i(t) = w(ai(t) + s)$$
$$wa\mu i(t) = wai(t) + ws$$
$$wa\mu i(t) - wai(t) = ws$$
$$(\mu - 1)wai(t) = ws$$
$$i(t) = \frac{ws}{(\mu - 1)wa} = \frac{s}{a(\mu - 1)}$$

Here we see that the equilibrium producer service output is positively related to the fixed labour requirement s. As fixed costs rise, so does the equilibrium output of the service firm. The reason is that higher fixed costs (s) introduce more scope for reaping scale economies, producing on a larger scale will reduce the impact of the fixed cost. On the other hand, high variable labour requirements (a) reduce the output level. This is because increasing a increases the marginal cost of producing at any level of output. Finally, increasing μ reduces the output level. As μ increases, we have more product differentiation, stronger monopoly power and a lower price elasticity of demand, so the market is more characterised by a large number of small producers rather than vice versa.
This equilibrium level of output per firm is precisely the amount given in examples A and B without proof. This is a somewhat strange result, since the equilibrium level of output per firm is fixed no matter what happens. Since s, a and μ are exogenous, meaning they are assumed to be fixed quantities rather than being determined inside the model, i(t) is also a fixed quantity. This means that the equilibrium level of output is unaffected if the size of the service sector expands and is unaffected if it contracts. This means that the service sector as a whole only expands and contracts by the creation of more or fewer varieties. Once we know the equilibrium level of output, then the labour force per firm is determined by the equation

Equation 15

$$L = s + ai(t)$$

The labour requirement equal to fixed labour requirement(s) plus marginal labour requirement(a) times firm's output (i(t))

Since we know the service workers per firm (L), and the total number of service workers in the city's economy, and each firm at equilibrium is the same size, then we can work out the number of firms (x) in the city, in other words the number of varieties of service. That is

Equation 16

$$x = \frac{(1-\beta)N}{ai(t)+s}$$

the number of firms (x) equals the total services labour force($(1-\beta)N$ divided by the labour force per firm (L = ai(t)+s) at equilibrium.

Now we know the number of firms (x) we can use the CES production function to obtain the equilibrium level of composite services, therefore

Equation 17

$$I = [xi(t)^{1/\mu}]^{\mu} = x^{\mu} i(t)$$

this is the CES (constant elasticity of substitution) (sub) production function for *I*, which is a function of the output of the typical services firm (i(t)), the number of services firms(x) and the elasticity of substitution, which diminishes with increasing μ.
And once we have obtained *I*, we can obtain the level of final goods output in the city using the C-D production function, hence

Equation 18

$$Q = M^{\beta} I^{1-\beta}$$

this is a Cobb-Douglas production function. Output (*Q*) equals workers (*M*) raised to the power β, multiplied by the level of composite services (*I*) to the power (1- β).

Interpreting μ

We have seen (equation 13) how μ is the amount of the mark-up on marginal cost pricing that occurs when we have monopolistic competition. We have also mentioned that μ determines the price elasticity of demand, hence

Equation 19

$$ped = \frac{\mu}{\mu - 1}$$

Earlier we also noted that μ is related to the elasticity of substitution of different services, so that

$$e_s = \frac{\mu}{\mu - 1}$$

The higher is the value of μ, the less substitutable are the different services and the less responsive is demand to a change in price (see the Appendix for the mathematical details giving this result).

We now show that μ is a measure of the amount of scale economies associated with equilibrium. Consider our definition of internal economies of scale, that is average costs divided by marginal costs. If marginal costs are lower than average costs, an increase in production will lower the cost per unit.

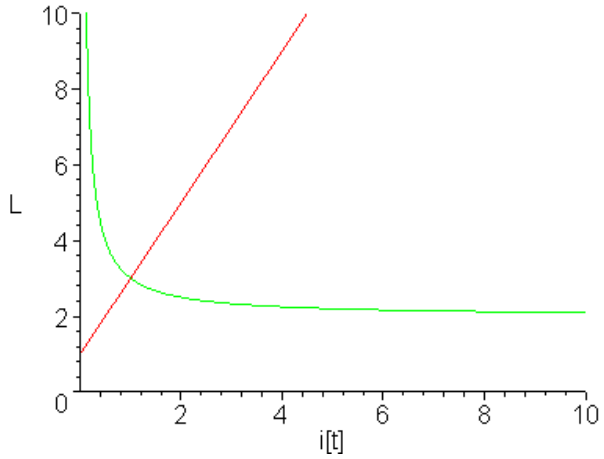


Figure 9

I want to measure my costs in units of labour rather than wages. The actual cost should be number of workers times wages per worker. Of course, labour times wages equals labour if wages per worker per unit of time = 1, so we can choose time units so that wages per worker per unit of time equals one. In the above graph, s = 1 meaning that there is a fixed cost equal to 1 when output is zero. The total cost is given by the upward sloping line L = s + ai(t) so that a = 2 is the marginal cost. There are increasing returns to scale since the curved average cost function L/i(t) is above the marginal cost =2. For example, at i(t) = 1, L= 1 + 2i(t) so that the average cost is 3. However as the level of output i(t) increases, the average costs fall towards the marginal cost of 2. In fact, as the level of output of the firm gets very large, then the internal returns to scale tend to 1, they effectively disappear. This is shown below, as i(t) becomes very large, then we can effectively ignore s and average cost tends to a. Also, as i(t) becomes very large, we can again ignore s and then average cost divided by marginal cost approximates to ai(t)/ai(t) = 1.

Equation 20

$$L = s + ai(t)$$

$$a.c. = L/i(t) = \frac{s + ai(t)}{i(t)}$$

$$a.c. \to a \qquad as \qquad i(t) \to \infty$$

$$m.c. = a$$

$$r.t.s = \frac{a.c.}{m.c.} = \frac{s + ai(t)}{i(t)} \frac{1}{a} = \frac{s + ai(t)}{ai(t)}$$

$$r.t.s \to 1 \qquad as \qquad i(t) \to \infty$$

Likewise we see that if there are no fixed costs incurred, there are no returns to scale. In the following graph a=2, s=0.

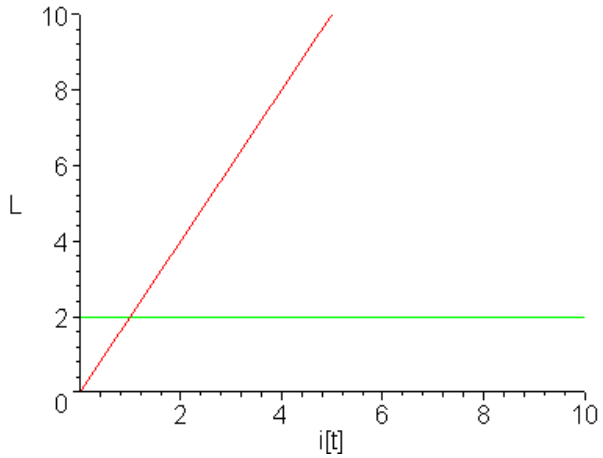


Figure 10

The upward sloping line is the total cost line which goes through the origin, there are no fixed costs. The horizontal line is the a.c. (average cost) line, since there are no fixed costs to start with, we do not see average costs fall as output expands, but they remain constant equal to a. Since a is m.c. = 2, then a.c./m.c. = 1 and there are no returns to scale.

Equation 21

$$L = s + ai(t) = ai(t)$$

$$a.c. = L/i(t) = a = 2$$

$$m.c. = a$$

$$r.t.s = \frac{a.c.}{m.c.} = 1$$

However, in our model neither of these scenarios causing internal returns to scale to fall to 1 will happen. In the first case we are assuming monopolistic competition with fixed costs, s cannot be assumed to be zero. Second, we have already shown that there is a fixed equilibrium size for firms, they do not grow ad infinitum, meaning that i(t) is a finite quantity. In fact we can see that there is a fixed value for returns to scale associated with the equilibrium.

Now let us work out the value of returns to scale in our model, when we have equilibrium. Equilibrium gives the level of output i(t) of each service provider. We therefore substitute for i(t) at its equilibrium value. Hence,

Equation 22

$$i(t) = \frac{s}{a(\mu - 1)}$$

$$L = s + ai(t) = s + \frac{s}{\mu - 1}$$

$$a.c. = \frac{L}{i(t)} = \frac{s + \frac{s}{\mu - 1}}{\frac{s}{a(\mu - 1)}} = a\mu$$

$$m.c. = a$$

$$r.t.s. = \frac{a.c}{m.c} = \frac{a\mu}{a} = \mu$$

The bigger is μ, the greater the internal returns to scale at equilibrium.

So we now see that a high value of μ means that we have

1. low price elasticity of demand,
2. low elasticity of substitution
3. high internal returns to scale
4. large increasing returns to city size since, returning to our aggregate analysis,

$$Q = \phi N^{\gamma}$$

$$\gamma = 1 + (1 - \beta)(\mu - 1)$$

With a low value of μ the reverse is true. As μ approaches 1, the price elasticity and elasticity of substitution become very high, while scale economies are low. We see the way the elasticity Of substitution and price elasticity of demand become very large as μ approaches 1 in the following graph.

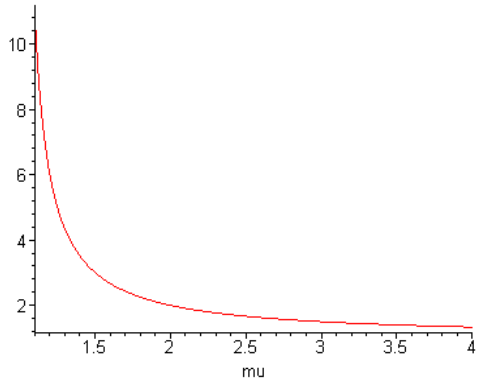


Figure 11

As μ approaches 1, the varieties become ever more perfect substitutes one for the other, so that ultimately with μ = 1 they are indistinguishable. Notice that in this situation, internal returns to scale are at their minimum, average and marginal costs approach equality. In the following equation we see as shown previously that at equilibrium a.c goes to aμ, but as μ approaches 1 that means a, so dividing a.c by m.c gives returns to scale going to 1.

Equation 23

$$\mu \to 1$$
$$a.c. = \frac{L}{i(t)} = a\mu \to a$$
$$m.c. = a$$
$$r.t.s. = \frac{a.c}{m.c} \to 1$$

Of course, this means also that there are no increasing returns to city size since

$$Q = \phi N^{\gamma}$$
$$\gamma = 1 + (1 - \beta)(\mu - 1) = 1$$

A 10 point critique of the CES production function.

Neary(2001) gives an elegant and readable account based on the closely related geographical economics theory. His basic points of criticism can be summarized thus:

1) We would expect scale economies to depend on the existence of a fixed labour input above zero so that as production expands less labour is needed to produce a unit of output, in other words the average labour per unit of output falls as output increases. This is when we have internal economies of scale. However, under Dixit-Stiglitz monopolistic competition, at equilibrium scale economies depend on μ, there is no mention of the fixed labour input. This is because of the way μ is linked to marginal labour requirement to simplify things. However if we change μ we cannot distinguish the effect of a change in the elasticity of substitution (or price elasticity of demand) from a higher ratio of marginal to fixed labour requirement.

2) there is free entry of service firms, in other words a perfectly elastic supply. There is no role for strategic interaction between firms. In other words the assumed market structure is that firms take each other's pricing behaviour as given and this important assumption underpins the derivations above. Firms cannot create artificial barriers to entry. Firms are myopic, and not able to engage in industrial strategies to shore up their position and stop other firms entering to erode their profits.

3) there is no discussion of the policy implications, or

4) of factors that might influence key parameters such as μ.

5) there is only a limited consideration of the externalities that are present in the real world

6) As Neary(2001) concludes, the problem with this kind of model is that it is too simple. The focus is on monopolistic competition as the cause of agglomeration. Now it is true there is some benefit to be derived from focussing on a single feature, this type of model is essential for understanding the world. But 'no monocausal model can hope to capture the complexity of any applied problem' certainly not one where

7) firms are all identical,
8) infinitesimal in size
9) the CES function is so important to the outcome and
10) externalities are dealt with in such a limited way.

3. Introducing Externalities

External economies, or externalities, are becoming an increasingly important dimension of our understanding of economic development. Glaeser et al. (1992) make this point in their paper on the growth of cities, where they observe that recent theories of economic growth have stressed the role of technological spillovers, particularly in cities where close communication between people greatly facilitates knowledge spillovers. Being within a city provides external economies that are beneficial for economic activity. The following quotations, taken from Glaeser(1999), highlight the limited nature of the type of theory outlined so far. Glaeser notes that

'Even the seminal work of Krugman (1991) bases its triumph, in part, on its ability to explain economic agglomerations without resorting to ad hoc external effects.'

However, he is not convinced that this is the best way to proceed to get a proper understanding of agglomeration forces, arguing that

'Urban economics needs to specialize in non-market interactions, because these interactions are (I believe) central to understanding the causes and effects of cities. Krugman (1991) shows that a brilliant theorist can explain cities without non-market interactions. But it is less obvious to me why one would want to do so. The flow of ideas and values that occurs through face-to-face interaction may be the most interesting feature of city.'

In this section, we show how important, but thus far omitted, determinants of urban and regional agglomeration can be incorporated into our model. The issue is one of identifying and incorporating various externalities that have been assumed away. The most general definition of an externality is that it exists when an economic activity affects people not directly participating in it. In fact the existence of positive externalities is another reason, apart from the 'love of variety' effect we have thus far concentrated on, why we should see agglomeration of economic activity. Positive externalities promote increasing rather than diminishing returns. For a long time, economists accepted the possibility of increasing returns only as an exception occurring in rare instances. The assumption was that mainly the economy operates in a range of decreasing returns. However, this view has been overturned to a degree by the changing structure of the economy, which has become more oriented towards goods and services in which knowledge is a significantly part of delivering the service, and knowledge entails high fixed costs because workers have to be trained before they can use their knowledge. Thus, so far we have seen that internal increasing returns to scale in the service sector amounts to an externality for final goods production. These will be highest when production is clustered together, in other words when it is in a city, the larger the better. The other positive externalities which we have not yet considered enhance this tendency to cluster.

A nomenclature of external and internal economies of scale

The terminology used in the literature is somewhat unclear, although Brackman et. al. (2001, p. 27-8) provide some illumination. Therefore we set out a nomenclature which should help to introduce the clarity that is required.

- Increasing returns

While much current emphasis is on increasing returns, in fact we should acknowledge that with very strong negative externalities due say to urban congestion, we could see diminishing returns to city size. However, the evidence is that, given the existence of agglomerations, the dominant mechanism is one of increasing returns. This is divided into two, internal and external economies of scale.

- Internal economies of scale (also sometimes referred to simply as increasing returns)

This is the case where the average costs decrease as a result of the increase in the level of production by the firm itself. Higher output brings higher profits and a cost advantage to larger firms compared to smaller firms. This market structure must be imperfect competition, as internal economies of scale imply market power.

- External economies of scale (also sometimes referred to as spillovers)

With external economies, the decrease in average costs per unit is due to an output increase at the level of the industry or city and not the firm.

The focus from now on is external economies rather than on internal increasing returns. It is useful to divide these into two categories (as is widely acknowledged in the literature), namely technological and pecuniary externalities. A pecuniary externality affects the firm's demand function. A technological externality affects the production function. Pecuniary externalities do not cause inefficiency, technological externalities may.

- Pecuniary externalities (sometimes referred to as market interdependence)

The reason why they are sometimes referred to as market interdependence is that the effects operate via a market, involving prices. As Fujita and Thisse(1996) observe, pecuniary externalities are the benefits of economic interactions 'which take place through usual market mechanisms via the mediation of prices'. In contrast, technological externalities are outside the market and are commonly associated with market failure. Since pecuniary externalities do not imply market failure, they are often considered to be not true externalities.

This kind of externality will occur when, for instance, firms create an externality in the labour market by training workers who then leave to work for another firm, so that the initial firm fails to appropriate all the benefits flowing from their investment in training. This spills over to other firms who benefit as a result of labour market transactions.

Unlike a true externality, a pecuniary externality only affects prices (ie the demand function); it doesn't affect (consumers' utilities or) the firm's production function directly. For example, a large firm moves into an area and bids up local wages, increasing labour costs for the other firms in the area. Similarly, if the size of the city

influences the price of inputs to a firm, then it is market mediated and a pecuniary externality. Likewise, with a cluster of manufacturers as in a city, the existence of the large local market for specialized services promotes service variety, enhances final goods productivity, and thus helps to maintain the manufactures market. There is market interdependence.

Pecuniary externalities are transmitted by the market through price effects for the individual firm, which may alter its output as a result. They do not affect the level of technology in the production function, and thus the technical relationship between the inputs and outputs. The price effect requires imperfect competition. In recent theory such as that described above, it is pecuniary externalities, which can be placed in a market context, rather than technological externalities, which cannot, which have been most amenable to formal analysis. This is the theory captured by the work of Abdel –Rahman and Fujita(1990) and Rivera-Batiz(1988). Pecuniary externalities are captured by the 'love of variety' effect obtain as a result of using the CES production function and monopolistic competition as the market structure in the service sector. Greater service variety per se is relevant to the level of output of final goods firms.

- Pure or technological (also sometimes called spillovers)

It is clear then that we have already captured pecuniary externalities in the model outlined thus far, since they are an external effect enjoyed by final goods producers as a result of the internal economies of scale in the producer services operating under monopolistic competition. However there are no technological externalities present.

A technological externality is present whenever the well-being of a consumer or the production possibilities of a firm are directly affected by the action of another agent in the economy. The use of the word 'directly' means that we exclude any effects that are mediated by prices. For example, assume we have a river with two activities, a fishery and an oil refinery. A technological externality is present if the fishery's productivity is reduced as a direct result of water pollution from emissions from the upstream the oil refinery. In contrast, if the fishery's profits were affected by the price of oil which itself depended on the oil refinery's output, that would count as a pecuniary externality, because it involved the operation of the price system in the market place.

Technological externalities exist as arguments in the firm's production function (or individual's utility function), so that the level of output is partly a result of the actions of other agents over which the firm has no control. They are called technological because they affect the technological relationship between the firm's inputs and outputs. For example, the level of technology may change due to the spillover of information and this will mean more output per unit of input. In other words research and development carried out by one firm may increase the rate of technical progress of other firms in an industry who have not paid for it, as a result of knowledge spillover. This kind of idea dates back to Marshall (see Appendix). It is associated with the development of clusters of productive activity. However Marshall envisaged technological spillovers existing under perfect competition, whereas we now recognize that they can occur even when the market structure involves imperfect competition. The development of ideas about externalities under imperfect markets can be traced to Scitovsky(1954).

With regard to technological externalities in general, there has been much recent emphasis (Quigley, 1988, Glaeser 1999) on the role of crime and victimization and how these increase with urban scale. Quigley focuses on the concentration of poverty and segregation in the housing market and how this can be an external diseconomy for employment prospects.

In what follows we develop the model to include two specific types of technological externality

- congestion
- knowledge spillovers

Congestion effects

Congestion externalities are largely considered to be unpriced (Anas, Arnot and Small 1998), meaning that the market has failed and therefore there is no market so we treat them as technological as opposed to pecuniary externalities. They are concerned with the technical relationship between inputs and outputs, with the structure of the production function rather than with prices in a market.
On the production side congestion involves interaction between firms, who 'get in each others' way' or 'step on each others' toes' (Cameron, 1996) way, and this affects their costs. Congestion arises when firms use common, but unpriced inputs in short supply, for instance there may be inadequate physical space, or infrastructural inadequacies relating to power supplies, water (for cleaning, cooling etc), road and other communications etc. However, there is a wider sense in which congestion occurs, it is when firms innovate and the innovations tend to be substitutes rather than complements. For instance, if R&D within different firms is substantially the same, there will be congestion externalities and over-investment in R&D. On the other hand, if the R&D is complementary, there will be positive network externalities. These produce external increasing returns. By network externalities we mean that a product's value to a consumer changes as the number of users of the product changes. The most obvious example is the mobile phone. For example, firms researching new mobile phone design will be complemented by firms researching new mobile phone technology. This will be of mutual benefit.
With regard to traffic, it is well known that the congestion externality arises because the vehicle user does not pay for its marginal contribution to congestion. Therefore the private cost of travel falls short of the social cost. Travel is misallocated across mode, route and time and may be excessive. This externality can be internalized[2] by a congestion toll but these are rarely to be seen and congested travel is under priced almost everywhere.

[2] The private ownership of roads would not result in the congestion externality since users would be charged a competitive price by the owners. However internalizing congestion externalities in general within a city is difficult because the number of economic agents involved is large and the costs of transactions between them is costly to internalize.

We have argued that congestion is not simply road congestion but is a wider concept. Congestion comes from various sources which make production more costly on a restricted space. Measuring the impact of each is likely to be difficult, and as Gordon and McCann(2000) argue that we can only observe the net realized effects of diverse simultaneous externality mechanisms, rather than individual sources.

The net effect of congestion is external diseconomies of scale. Since the separate effects going to make up the net effect are difficult to measure, and since we are more concerned with the consequences of congestion and not the precise individual causes (see Brackman et. al. 2001) , we chose to resolve the totality of congestion effects as a parameter of the firm's production function. We do this in two alternative ways, for both producer service firms and for final goods firms, to illustrate possible ways of handling the issue of incorporating congestion externalities into our model structure.

First, let us look at how we might model congestion in the service sector. We commence with the service firm's production function, hence

Equation 24

$$L = s + ai(t)$$

The idea now is to modify this so that costs rise as the size of the city rises, in other words as the number of service firms x increases.

Equation 25

$$L = x^{f}(s + a(i)t)$$

Raising x to a power f changes the costs according to whether f is positive or negative. If f is positive, then increasing x means increasing costs for the firm. This is what we might expect, with congestion taking its toll. It may turn out however though that costs fall with increasing number of firms, which would be what would occur if x was raised to a negative power, although this could not be described as a negative congestion externality, so we assume that the power is positive to model the congestion effect. If f is exactly zero, then we see that the multiplying factor is exactly 1, so there is no externality effect in that case.

We can carry out a similar modification to the production function of the final goods sector (for simplicity assume there are now no congestion effects in the service sector). We assume that the production function facing each firm is the same, and proceed as if the total output Q in the final goods sector in the city is produced by a single representative firm behaving competitively (as in Abdel Rahman and Fujita, 1990).

- The initial specification is as before, hence

Equation 26

$$Q = M^{\beta} I^{1-\beta}$$

but we now adjust this by also including a coefficient α which lies between 0 and 1 depending on the strength of congestion effects (see Ciccone and Hall, 1996). To show this consider that output depends on the number of units of labour and the amount of land that is used. Thus far we have ignored land as an input. The production function in this case is

Equation 27

$$Q = \left[M^{\beta} I^{1-\beta}\right]^{\alpha} L^{1-\alpha}$$

in which the term in the square brackets refers to labour of all kinds, either final goods or services, and *L* refers to the amount of land. The value assigned to the coefficient α determines the relative importance of these two inputs. So output will be greater if there is more labour or more land input. Now so far we have in effect assumed that α = 1, in other words the amount of land is irrelevant to the amount of final goods produced. In other words, *L* to the power 0 is equal to 1, we simply multiply by one. So we could ignore *L* and ignore α.

Assume now that *L* = 1, in other words we are calculating output for a unit of land. Assume also $0 < \alpha < 1$, in other words the amount of land is relevant to the amount produced. Since $1^{1-\alpha} = 1$, we still can leave *L* out of the production function completely, since we are simply multiplying by 1, but since $\alpha < 1$ we have to include α. The resulting equation, allowing for congestion effects, then becomes

Equation 28

$$Q = (M^{\beta} I^{1-\beta})^{\alpha} = (\phi' N^{\gamma'})^{\alpha}$$

Note that I have written *Q* in terms of the original production function, and also as a function of *N*, the total workforce, which we derived earlier. However, in this version of the production function I have replaced the coefficient ϕ which was used in the initial equation (equation 5) by ϕ' and replaced the original γ by γ', I want to reserve ϕ and γ to use later where they have a slightly different definition to their original one.

To summarize , α controls the strength of the impact of congestion effects on output. If we set α close to its lower limit 0 then congestion effects greatly inhibit output. As α approaches the upper limit of 1, congestion effects have less and less impact on *Q*. We see this in the following diagram. As α approaches 0 we need more workers and more services to reach the same level of final goods output : hence the isoquants move upwards (each line is for the same *Q*).

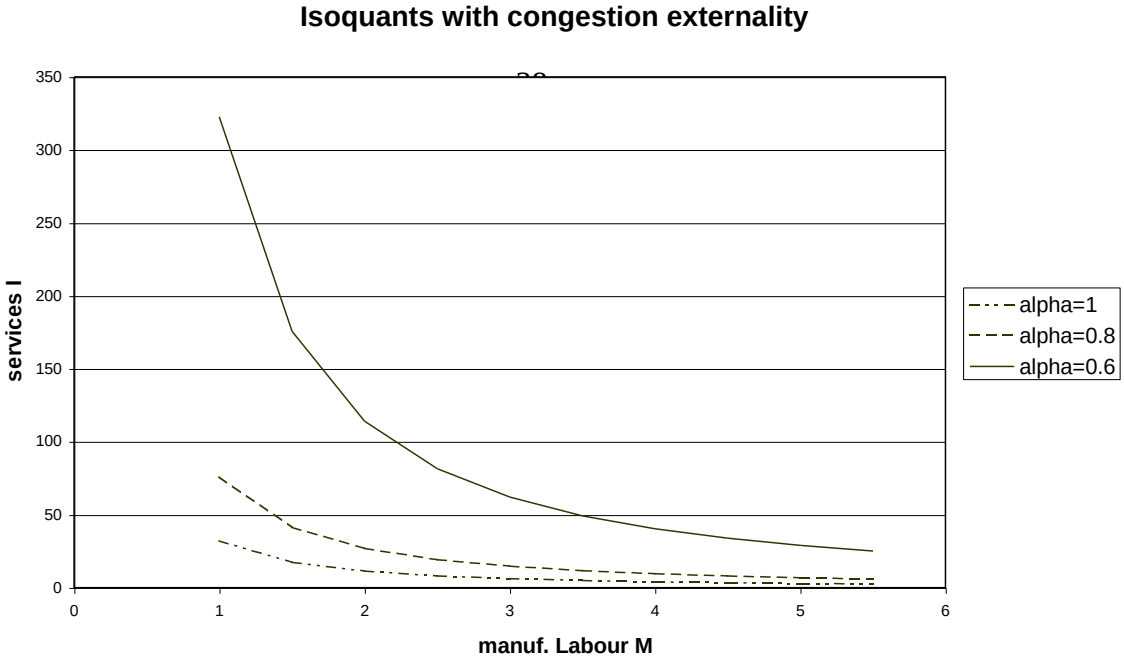


Figure 12

What does including congestion imply for the new definitions of the parameters ϕ and γ. Let us remind ourselves of the old definition of the more important of the two which is now given by γ'. We saw earlier (see equation 5) that the value of this was determined by two other parameters. One is the value of μ, which is the parameter of the services CES sub-production function, and which reflects the strength of monopoly power, the elasticity of substitution, and the amount of internal increasing returns. The second determinant of γ' was β, the value of which defines the relative importance of workers and composite services in the final goods producer's Cobb-Douglas production function, hence $\gamma' = 1 + (1-\beta)(\mu - 1)$. However, when we also include congestion in the form of α, then on expanding equation (28) we find that

Equation 29

$$\gamma = \alpha[1 + (1-\beta)(\mu - 1)]$$

As equation (30) shows, ϕ is also a function of the original ϕ'.

Equation 30

$$Q = (\phi' N^{\gamma'})^{\alpha}$$
$$\gamma' = 1 + (1-\beta)(\mu - 1)$$
$$Q = \phi'^{\alpha} N^{\alpha\gamma'}$$
$$\alpha\gamma' = \alpha(1 + (1-\beta)(\mu - 1)) = \gamma$$
$$\phi'^{\alpha} = \phi$$
$$Q = \phi N^{\gamma}$$

With the earlier definition $\gamma' = 1 + (1-\beta)(\mu - 1)$, if services were irrelevant ($\beta = 1$) or if there were no internal increasing returns for service firms ($\mu = 1$) then there are no increasing returns for final producers with density ($\gamma' = 1$). Otherwise, there will be

increasing returns for final producers. However with the new definition $\gamma' = \alpha[1+(1-\beta)(\mu-1)]$ there is a range of outcomes, depending on the respective values of μ, β and α,

- we see either increasing returns ($\gamma > 1$) so that increasing density is increasingly rewarded, as shown by the line with increasing slope in Figure 13.
- or diminishing returns ($\gamma < 1$). In the latter case, the effect of congestion is so severe that it completely overturns any tendency to increasing returns. Increasing density is not accompanied by a commensurate increase in output, as shown by the diminishing slope of the line in Figure 13.

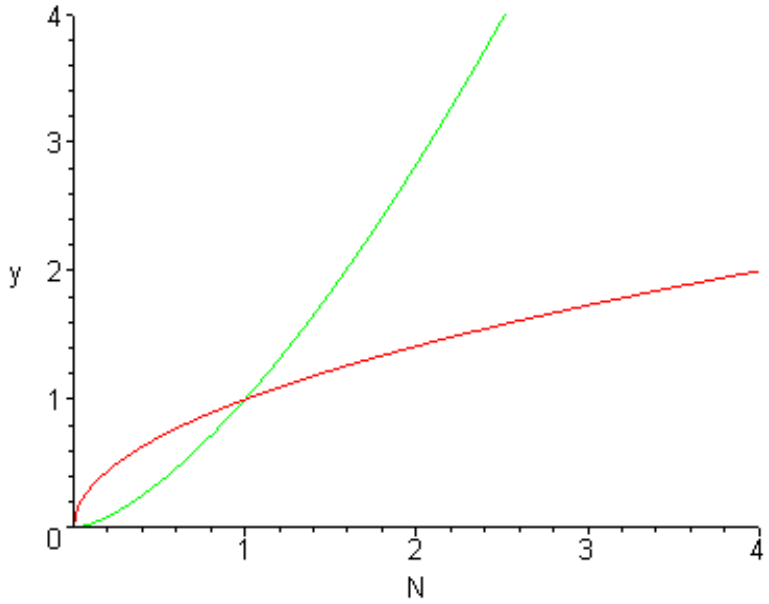


Figure 13

Usually we do not know the values of α, β and μ but we can obtain an estimated value of γ if we have data on Q and N. For example we may have a number of different cities where we know the level of final goods output and the size of the working population N (which equals the sum of service and final goods workers). Alternatively, we may have data on a single city over time. At each time point we have final goods output Q and city size N. Then we can get an estimate of the value of the coefficient γ from the loglinear regression equation

Equation 31

$$\ln(Q) = \ln(\phi) + \gamma \ln(N)$$

and test the null hypothesis that $\gamma = 1$.

If we do reject the null, then there is evidence for increasing returns, although by incorporating congestion effects there is more of a chance that we will see constant ($\gamma = 1$) or diminishing ($\gamma < 1$) returns to scale.
Note that we have no estimate of the contribution of the separate fundamental parameters μ, β and α that determine the value γ. Being able to estimate these fundamental parameters would be a significant help in being able to say more about determinants of agglomeration mechanisms (See Fingleton, Journal of Economic Geography, 2001 where this is discussed and evidence given showing variations in these 'exogenous' parameters across time and space).

Some published work (see Fingleton, 2001) estimates the value of γ by the relationship between final goods (ie manufacturing) productivity and final goods output. In fact it is easy to show that this relationship is equivalent to the relationship between Q and N described above. Hence

Equation 32

$$\ln(Q) = \ln(\phi) + \gamma \ln(N)$$
$$\gamma \ln(N) = \ln(Q) - \ln(\phi)$$
$$\ln(N) = \frac{\ln(Q) - \ln(\phi)}{\gamma}$$
$$\ln(Q) - \ln(N) = \ln(Q) - \frac{\ln(Q)}{\gamma} + \frac{\ln(\phi)}{\gamma}$$
$$\ln(Q/N) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q)$$

This is turned into a relationship between the level of final goods productivity (Q/M) and the level of final goods output(Q), since we know that $M = \beta N$. Hence

Equation 33

$$\ln(Q/M) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) - \ln(\beta)$$

The reason we know $M = \beta N$ derives from the production function for final goods output, which we previously extended to explicitly include land (L) as a factor, so that $Q = [M^{\beta} I^{1-\beta}]^{\alpha} L^{1-\alpha}$. Acknowledging that the factor L should also be present tells us that the total output Q is divided up between what is paid to land (L) and what is paid to labour (N), either directly to final goods workers (M) or indirectly to labour producing services (N-M).

In what follows we use the following theory. If we assume the economy is at a competitive equilibrium, then economic theory tells us that a factor is paid an amount that is equal to the value of its marginal product.
Let us look first at what is paid to land, and then work out what is paid to labour and finally what is paid to final goods labour. We know that M and I jointly comprise labour (the only input, which is either direct final goods labour or labour used to produce composite services). Hence the level of output is a function of total labour N,

so we write *f(N)* in place of $M^{\beta}I^{1-\beta}$. Equation 33 shows that differentiating *Q* with respect to *L* gives the marginal product of land, which competitive equilibrium theory tells us equals the rent paid to land, which is denoted by *r*. So the share of final product (*Q*) being paid to land is equal to the rent per unit of land *r* times the number of units of land (*L*) divided by *Q*, and this is equal to the coefficient in the Cobb-Douglas production function relating to *L*, 1-α.

Equation 34

$$Q = [f(N)]^{\alpha} L^{1-\alpha}$$
$$\partial Q / \partial L = f(N)^{\alpha} L^{1-\alpha}(1-\alpha)/L = (1-\alpha)Q/L$$
$$r = (1-\alpha)Q/L$$
$$rL/Q = 1-\alpha$$

Turning next to labour, since there are only two factors of production, land and labour, it must then be the case that the share of *Q* going to labour of both types (*N*) must be what is left over, in other words α. As with the example of land, this is equal to the wage rate per worker *w* times the number of workers (both final goods and services) divided by final product Q, in other words

Equation 35

$$wN/Q = \alpha$$

Turning next to labour directly employed in the production of final goods and services, which we represent by *M*. Using our competitive equilibrium theory, we see that the marginal product of final goods labour, the derivative $\delta Q/\delta M$, which is equal to $Q\alpha\beta/M$, is the wage rate *w*. Combining the result in equation (35) and $w = Q\alpha\beta/M$ gives the relationship $M = \beta N$.

Hence

Equation 36

$$Q = [M^{\beta}I^{1-\beta}]^{\alpha} L^{1-\alpha}$$
$$\partial Q/\partial M = \frac{[M^{\beta}I^{1-\beta}]^{\alpha}\alpha\beta L^{1-\alpha}}{M} = Q\alpha\beta/M = w$$
$$wM = Q\alpha\beta$$
$$wN = Q\alpha$$
$$\frac{wM}{wN} = \frac{Q\alpha\beta}{Q\alpha}$$
$$\frac{M}{N} = \beta$$
$$M = \beta N$$

We have used this relationship between *M* and *N* throughout, the above gives the reasoning on which it is based.

Knowledge spillovers

So far we have emphasized the role played by congestion as a form of technological externality, which has the potential to weaken the production gains that final goods producers obtain in dense cities. Other forms of technological externality exist, and in this section we focus on knowledge spillovers within and between cities and regions that also have an effect on production. In the modern economy, technological externalities often appear as benefits or costs due to transfers of information or knowledge. Knowledge generated by one agent for its own benefit is not exhausted by use but persists and spreads, affecting other economic agents. Following Glaeser et al (1992), It is useful to divide knowledge spillovers into three types, what we refer to below as MAR, Porter and Jacobs externalities.

First, with MAR there is knowledge that spills over between firms within a sector, but is essentially confined to that sector or industry. The idea here comes basically from Marshall(1890), hence the name Marshall-Arrow-Romer externalities, or MAR for short, with the acronym highlighting the leading proponents of this kind of spillover. The idea is that we see a boost to production due to non-pecuniary, and hence technological, externalities that involve firms picking up or somehow acquiring, at less than market cost to themselves, innovations and ideas generated by other firms within their industry. The use of the existing knowledge base has been referred to as 'the standing on shoulders effect' (Cameron, 1996). Not only does each firm benefit from its own R&D, but it also benefits from the research results of other firms, the domestic science base and research carried out by foreign governments and firms.

How is knowledge transferred? The main mechanisms are via patents, scientific literature, technology licences, technology embodied in capital and intermediate inputs, and personal contacts. One argument is that knowledge transfer is likely to be easier in a dense city since the firms will be effectively clustered together within a restricted area. This means that social networks will develop and be channels for information flow. Information flows more easily locally than at a distance. This suggests that personal contacts, for instance seminars, conferences, trade fairs, and sales meetings are a significant transmission mechanism.

However there are more remote spillovers of knowledge that are not so much affected by distance. For example international trade promotes personal contacts across national boundaries. Without these personal contacts, it will be more difficult to decode foreign R&D to adapt it to domestic use, involving substantial and costly R&D.

Knowledge can be transferred by formal channels, for example as a result of subcontracts and technology agreements between firms. Or, transfer may be by informal channels as workers change jobs, or as a result of industrial espionage, or via informal vertical linkages between firms. As Glaeser et al(1992) observe, there are various methods of acquisition and transmission; for example 'through spying, imitation, and rapid interfirm movement of highly skilled labour, ideas are quickly disseminated among neighbouring firms'.

However, regardless of mode of transmission and means of acquisition, the basic mechanism in all this is the inability of innovating firms to appropriate all the knowledge they create. This new knowledge can be acquired or imitated by other firms who do not make the same level of investment in R&D as the innovating firm. This has a consequence, it may not be worthwhile for a firm to carry out its own R&D if it can be acquired from others at lower cost.

Glaeser et al(1992) argue that

"In MAR models of externalities, innovators realize that some of their ideas will be imitated or improved on by their neighbours without compensation. This lack of property rights to ideas causes innovators to slow down their investment in externality-generating activities, such as research and development. If innovators had a monopoly on their ideas, or at least if they had fewer neighbours who imitated them immediately, the pace of innovation and growth would rise. The MAR models tend to imply that whereas local competition is bad for growth, local concentration is good for growth because innovators internalize the externalities" (Glaeser et al, 1992)

In other words while it may benefit firms on the receiving end of these knowledge flows, there may be dis-benefits for the innovating firms. It may be better from their point of view if knowledge does not travel so fast in the city, so that barriers can be erected which restricts the flow of ideas to others. In fact if a firm lost all of its ideas as soon as they were hatched, there would be no point in innovating and it would give up. Hence, MAR theory implies that growth is maximized in a situation where there is a local monopoly for innovations rather than local competition, since firms can hang onto their ideas, in other words internalize what would otherwise be externalities, allowing innovators to receive the full benefits of their effort. However, as we shall explain below, others would disagree.

Note that MAR externalities have here been related to growth, while so far in our theorizing we have been focusing on the different levels of output and productivity, which are the realm of static externalities. However, MAR externalities do also explain the location of activity, or why certain areas are specialized in certain sectors, as well as growth rates. The reason is that, irrespective of growth rates, by clustering together firms in the same sector maximize knowledge spillovers. As indicated by Glaeser et al(1992), while MAR simultaneously account for both locational specialization and differentiated growth rates, we might best refer to them as essentially dynamic externalities, since this differentiates them from purely static externalities that solely explain spatial concentration. As we shall see below, we can easily translate our model into a growth model.

While some would argue that growth is maximized when there is some form of monopoly over ideas so that firms can reap the benefits of their own R&D, other such as Porter(1990,1998) argue that local competition is better, since it causes firms to be better innovators or faster adopters of others' innovations than they otherwise would in order to survive, and that enhances the growth rate. This has been referred to as a 'creative destruction' effect where new knowledge makes old processes and products obsolescent and induces knowledge transfer. In this model, the presence of imperfect competition would be lead to the consumption, say in the form of excess wages and perks, of the incidental benefits arising from monopoly power within the sector. In contrast if the sector is highly competitive, it will tend to maximize its R&D effort and grow at the fastest possible rate. While intense competition may mean that fully capturing the returns on innovation may be more difficult than in the protected environment of less than perfect competition, this is more than offset by the enhanced rate of innovation and imitation within the sector. Information transmission is all-important, and so to ensure full access to any spillovers from competitors within the sector, it is a good strategy for firms to locate near to each other in clusters. Porter uses the example of the intensely competitive clusters in the jewelry industry and the ceramics industries in Italy as evidence supporting this clustering process.

The third type of external economy, namely Jacobs externalities, also is associated with a competitive environment. The name comes from the work of Jane Jacobs(1969, 1984). The essence of Jacobs externalities is the existence of spillovers between sectors, in contrast to MAR and Porter externalities which are essentially within sector. Jacobs externalities are external to the sector but internal to the city, in other words there are benefits to economic growth of a sector from the activities of other sectors within a city due to the ease of transmission of knowledge. Many others, for instance Saxenian(1994), have highlighted ‘face to face’ communication in the creative process as a mechanism that made it rational for firms to cluster together in the same location. As Glaeser et al (1992) explain, 'Jacob's idea is that the crucial externality in cities is the cross-fertilization of ideas across different lines of work'. The notion is that the variety of activity in a city adds to technological progress, so that 'the diversity of urban activities quite naturally encourages attempts to apply or adopt in one sector (or in one specific problem area) technological solutions adopted in another sector' (Bairoch, 1988). Jacobs envisages this kind of spillover to be at a maximum when competition between firms is maximized, arguing that innovation is stimulated rather than suppressed by monopolies that are unlikely to favour the creation of alternative technologies, methods products and services that would weaken their power.

Modelling varying levels and rates of growth of technology

Neoclassical growth theory assumes that technical progress is exogenous and proceeds at a steady rate, for some reason. This has been called the 'manna from heaven' view of technology. However, once we accept that the level of technology and its rate of growth (the technical progress rate) within a city or region will be strongly influenced by knowledge spillovers, then it has to be recognized that the rate of technical progress can no longer be considered to be a constant. For instance cities will be the source of more external economies than small towns, due to Jacobs externalities, and regions with concentrations of firms in the same sector will benefit from greater external economies than regions where the concentration is lower,

according to the arguments embodied in the theory of MAR and Porter externalities. The idea that technical progress rates depend on other factors rather than being an unexplained, exogenous constant is somewhat at variance with traditional neoclassical growth theory, which has nothing to say about causes of technical progress even though the equilibrium rate of growth depends on its rate of growth. In this book, while the theory that has been described so far is one step removed from neoclassical growth theory, it is nevertheless also the case that no account has been taken of the impact of differentiated levels and rates of growth of technology. This section is dedicated to extending our theory to allow this.

One of the reasons why there has hitherto been only limited attention given to knowledge spillovers is that it is very difficult to measure and model such phenomena, so that capturing localized knowledge pooling externalities has proved somewhat elusive and led to some negative remarks. For instance Krugman(1991) considers Marshall's technological externality resulting from knowledge spillovers between firms to be 'invisible' and impossible to model in a formal way. He argues that knowledge flows 'leave no paper trail by which they can be measured or tracked, and there is nothing to prevent the theorist from assuming anything about them that she likes'. Gordon and McCann(2000) argue that we can only observe the net realized effects of diverse simultaneous externality mechanisms, rather than individual sources.

However Quigley(1988) argues that while we cannot observe knowledge as it spills out among the buildings and streets of a city, some of this spillout does leave a paper trail, contrary to what Krugman argues. Hence, despite the difficulties, some attempts to explicitly model knowledge externalities have been made. According to Cameron(1996), since they do not know exactly where and to what extent spillovers are occurring, researchers typically use some proxy for the flow of spillovers; hence knowledge flow proxies take four main forms, input-output tables, patent concordances, innovation concordances, and proximity analysis. Jaffe, Trajtenberg and Henderson(1993) traced knowledge spillovers using patent citations. They looked at the geographic locations of successful patent applicants and the locations of the intellectual and commercial forebears of these innovations. The location of the forebears in practice means the location of earlier patent owners who are cited by the new successful patent applicant. They find that there is stronger geographical association between new patents and their forebears than would be expected from the pre-existing concentration of economic activity. Patents and their ancestors tend to be clustered in the same metropolitan area of the USA.

The approach adopted here is not to attempt the difficult task of trying to explicitly measure and model technological externalities per se, but to focus on what determines the level of technology. We start with a given technology level, and then look at the rate of technical progress, which changes the level of technology. We then look at what determines the rate of technical progress in different cities and regions, so that rather than being exogenous, the technical progress rate is endogenous, explained by factors that vary across different cities, or vary across time in the same city.

The assumption is that technological externalities, or spillovers, determine the rate of technical progress and therefore the level of technology. We then link this to our model structure, which was erected earlier, arguing that the level of technology at any point in time determines the efficiency of final goods workers at that point in time. In order to see this, we change the meaning of M so that it now becomes measured not in

terms of the number of workers, but in efficiency units, which equals the number of workers (E_t) multiplied by the level of technology they employ (A_t), with the subscript t signifying the point in time. Remember that up to now, we have reserved t to mean typical, so that *i*(t) means the output of the typical producer service firm. So, as time progresses, and the level of technology improves, we have more labour efficiency units even in the number of workers (*E*) remains constant. Hence the number of labour efficiency units is

Equation 37

$$M_t = E_t A_t$$

Let us also denote the rate of technical progress by the variable λ (anticipating the fact that technical progress rates will vary rather than remain constant). Assume that there is an exponential growth of technology, so that the level of technology at any point in time, equal to A_t. So the level of technology A_t depends on the initial technology level at time t = 0, which is equal to A_0, the time that has elapsed (given by t), and also depends on the rate of technical progress per unit of time (or proportional rate of growth of the level of technology) λ. We can see that λ represents the rate of change by the following piece of mathematics[3]. Since

Equation 38

$$A_t = A_0 e^{\lambda t}$$
$$\partial A / \partial t = \lambda A_0 e^{\lambda t}$$
$$\partial A / (A_t \partial t) = \lambda$$

It is therefore the case that the number of labour efficiency units is given by

Equation 39

$$M_t = E_t A_t = E_t A_0 e^{\lambda t}$$

With this revised definition of *M*, we can substitute for *M* in the model developed thus far and write our model in terms of the level of output per worker. The following sequence of equations shows this.

[3] Another way to see the same thing uses the fact that differentiating the log with respect to time give the (proportional) growth rate, since $\partial \ln(y)/\partial t = (\partial y/\partial t)(1/y)$, which is the proportional growth rate of y. We use this result, differentiating $\ln(A_t) = \ln(A_0) + \lambda t$ with respect to time. The result is $\lambda = \partial \ln(A_t) / \partial t$ which is therefore the growth rate.

Equation 40

$$\ln(Q/M) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) - \ln(\beta)$$

$$\ln(Q) - \ln(M) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) - \ln(\beta)$$

$$\ln(M) = \ln(E) + \ln(A_0) + \lambda t$$

$$\ln(Q) - \ln(E) - \ln(A_0) - \lambda t = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) - \ln(\beta)$$

and therefore

Equation 41

$$\ln(Q/E) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) - \ln(\beta) + \ln(A_0) + \lambda t$$

(See Fingleton, 2001).

Up to this point the rate of technical progress λ is exogenous. The analysis from this point onwards assumes that λ is a variable the value of which is determined endogenously, thus we focus on what we assume determines λ. The following analysis depends on creating a submodel in which λ is a function of some causal variables representing the effects of technological externalities. However, we take a rather general view of the operation of spillovers. While we believe there may be externalities that are internal to a sector (As in MAR or Porter externalities), or internal to a city (as in Jacobs externalities), we also believe that spillovers can be even more wide ranging, and cross formal city boundaries. Part of this relates to the difficulty of deciding what the limits to a city are. Given this problem, it seems appropriate to allow spillovers to occur which cross the boundaries that define the formal statistical units used for the purposes of empirical modeling.

What then might be the determining variables that control the facility with which information is spread within the city or region? One important factor considered by many analysts is educational attainment, which can be taken as a measure of the level of human capital and therefore we denote this variable by the letter *H*. The assumption is that higher educational attainment rates will boost the adoption and spread of innovations within the city. In other words new knowledge creation and technological externalities in the form of knowledge spillovers will be higher whenever the level of human capital (*H*) is higher. This is a reasonable assumption; the creation of new knowledge in the form of patenting rates is highest in cities with a high density of highly educated people engaged in research and development. Also, adopting and adapting foreign knowledge in particular requires R&D by the firm doing the copying. This requires human capital.

A second reason why technical progress rates might differ between cities and regions is differential rates of adoption of innovations, regardless of the human capital attributes of the area. An important reason which two cities with the same educational attainment rates should have different rates of technical progress is the level of

technology that already exists within the city. In other words the rate of technical progress is assumed to be a function of the size of the technology gap (*G*). This has to be defined at a point in time, so we take time zero, the start of the period of analysis. The technology gap is assumed to be the level of final goods technology in the city in relation to the highest possible level at time zero. The thesis is that if you are in a city with a big technology gap, then new knowledge will be very valuable as an enhancement to subsequent productivity growth. Your technical progress rate will be fast. It is for this reason that we have seen many third world cities catch-up to, or in some cases surpass, the level of final goods productivity seen in the developed world. On the other hand, if your city is a technological leader, then the technology gap will be zero or very small. The available knowledge on the world scene will not make much of an impact, you already know it. You are using state of the art production techniques, which cannot be improved. The small or zero technology gap means that technical progress does not come from the diffusion of knowledge from research and development in other cities and regions at a superior level of technology.

We have argued that knowledge spread best within the city, but nevertheless it is also true that for some kinds of spillover, spatial proximity within a city or region, and closeness to an innovating city, is irrelevant. Global spillovers unrestricted by distance decay may have been occurring for a long time, but the advent of the Internet and the growth of English (or American!) as the language of business, commerce and science, has made it even easier and we cannot ignore this kind of spillover if we wish to create a realistic model. Much knowledge is available instantaneously, as soon as it is created, on the Internet or other fairly instantaneous worldwide communications media. Moreover, many new ideas are published in Journals that have worldwide circulation. That is the type of knowledge diffusion referred to above. Up to this point we have highlighted key factors determining the adoption rate of new technologies which are essentially related to conditions within the city or region, be they educational attainment (*H*) or the level of technology gap (*G*), and which are independent of location.

However, despite modern communications, it remains true that the spread of some types of knowledge will be spatially impeded, although not restricted to within city boundaries. It will tend to spill across formal district boundaries into surrounding suburban and rural areas, since the functional city is often much more extensive than the formal, administrative city. The growth of edge cities in the USA (Garreau 1991, Anas et al 1998) has on the whole not been accompanied by an extension of political and administrative territories to encompass these new peripheral developments. Also when knowledge spreads it will not be confined even within the functional city, although there will tend to be spatial impedance so that it spreads first to local and well-connected cities, and last to remote and isolated ones. Or example, there is good evidence that it takes longer for innovations to diffuse from the USA to Europe than between domestic US firms. Therefore the third factor controlling the rate of technical progress is geographical proximity. New knowledge will diffuse more readily if you share a border with an innovating city or region. Physical contiguity is not necessary, but it helps. However if the city has good or regular communications links with another remote city, then we might see knowledge being transmitted quite long distances. For instance the good air communication between London and New York and the low cultural and linguistic barriers between the two cities means that we will see knowledge spillovers between the two even though they are on either side of the

Atlantic. We refer to this effect as the spillover (S) of knowledge from other regions and cities.

Some localized information transmissions may be embodied within workers moving jobs from one company to another. This is likely to be spatially impeded, particularly as job changes are likely to be more prevalent that house moves, and therefore commuting distance is a major restricting factor. Even when workers move residence in order to cut commuting costs, it is likely that the new place of abode will not be far from the previous one. Also, there will be spillover between firms directly because if they are linked in chains of production, then these will tend to be channels along which information about optimal production technologies will flow. More often than not production chains will be localized, for instance to facilitate just-in-time methods (McCann and Fingleton, 1996). Even if they are not cooperating as producers and suppliers, but are competitors, Porter reminds us that we still will see spillovers between firms within a locality.

We gather together these effects determining the rate of technical progress in the linear function

Equation 42

$$\lambda = b_0 + b_1 H + b_2 G + b_3 S$$

Next, put this back in our larger model, of which the rate of technical progress is but one cause of the level of final goods productivity, hence

Equation 43

$$\ln(Q/E) = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma - 1}{\gamma}\right]\ln(Q) - \ln(\beta) + \ln(A_0) + (b_0 + b_1 H + b_2 G + b_3 S)t$$

Equation 43 is an expression for the level of final goods productivity at point t in time. While we are very interested in productivity level differences, we are also very much concerned with the dynamics of the economy, with differences in productivity growth rates. Looking at growth rates rather than levels also eliminates some terms from the equation since they are assumed to be constant over time. This is shown if we turn our levels equation into a growth equation by differentiating with respect to time. In general for variable *y*, the growth rate is $\partial \ln(y)/\partial t$. Differentiating the natural log of the level of final goods productivity (Q/E or P) with respect to time gives the rate of growth of final goods productivity which we denote by p. Likewise differentiating $\ln(Q)$ with respect to time gives the growth of final goods and services output q (see Appendix).

I work with the discrete time analogue of the (proportional) growth rate[4] $\partial \ln(y)/\partial t$, which is the difference in logs $\ln(y_{t+1}) - \ln(y_t)$. Let us apply this to our model in

[4] We see that the difference in logs is the growth rate also using the economist's favourite approximation which says that for small h, $\ln(1+h) \approx h$. For example, say we have a proportional growth rate of 105/100 = 1.05 = 1+0.05 = 1 + h, then $\ln(105/100) = \ln(1+0.05) = \ln(1+h) = 0.049 \approx h$. The same result is given by the difference in logs since $\ln(105/100) = \ln(105) - \ln(100)$.

levels, looking at the level of output per worker or productivity at two times t=1 and t=2. If we subtract t=1 ln(productivity level) from t=2 ln(productivity level), then the result is a model in which the growth of productivity depends on the growth of output since the difference in logs is a growth rate. Hence,

Equation 44

$$\ln(Q/E)_t = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma-1}{\gamma}\right]\ln(Q)_t - \ln(\beta) + \ln(A_0) + (b_0 + b_1H + b_2G + b_3S)t$$

$$\ln(Q/E)_1 = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma-1}{\gamma}\right]\ln(Q)_1 - \ln(\beta) + \ln(A_0) + (b_0 + b_1H + b_2G + b_3S)1$$

$$\ln(Q/E)_2 = \frac{\ln(\phi)}{\gamma} + \left[\frac{\gamma-1}{\gamma}\right]\ln(Q)_2 - \ln(\beta) + \ln(A_0) + (b_0 + b_1H + b_2G + b_3S)2$$

$$\ln(Q/E)_2 - \ln(Q/E)_1 = \left[\frac{\gamma-1}{\gamma}\right](\ln(Q)_2 - \ln(Q)_1) + (b_0 + b_1H + b_2G + b_3S)$$

$$\ln(Q/E)_2 - \ln(Q/E)_1 = \ln(P_2/P_1) \quad since \quad \ln(a) - \ln(b) = \ln(a/b)$$

the result is the growth rate since

$$\ln(1+h) \approx h \quad when \quad |h| \quad small$$

$$p = \left[\frac{\gamma-1}{\gamma}\right]q + b_0 + b_1H + b_2G + b_3S$$

hence

Equation 45

$$p = \frac{\gamma-1}{\gamma}q + b_0 + b_1H + b_2G + b_3S$$

and the estimation of this type of equation this is the object of recent empirical work.

APPENDICES

The urban economics model

Endogenous variables

1. manufacturing labour (workers):

Equation 29

$$M = N\beta$$

manufacturing labour (workers) (M) equals total labour (N) times β which is the equilibrium allocation of labour to manufacturing under competitive conditions

2. manufacturing output:

Equation 30

$$Q = M^{\beta} I^{1-\beta}$$

this is a Cobb-Douglas production function
output (Q) equals workers (M) raised to the power β, multiplied by the level of composite services (I) to the power (1- β).

3. composite services:

Equation 31

$$I = \left[\int_{t=0}^{t=x} i(t)^{1/\mu} dt \right]^{\mu}$$

Equation 32

$$I = [xi(t)^{1/\mu}]^{\mu} = x^{\mu} i(t)$$

this is the CES (constant elasticity of substitution) (sub) production function for I, which is a function of the output of the typical services firm ($i(t)$), the number of services firms(x) and the elasticity of substitution, which diminishes with increasing μ. As μ approaches 1, then the services level approaches the number of firms times their output, as $\mu >> 1$ it is more than this due to the effect of the number of varieties(x), so that increasing firms results in a proportionately larger I

4. equilibrium output level of typical service firm:

Equation 33

$$i(t) = \frac{s}{a(\mu - 1)}$$

when firms are at equilibrium, so that (marginal) costs equal (marginal) revenues and profits are driven to zero, the output per firm can be shown to equal the fixed labour requirement (s) divided by the marginal labour requirement (a) times μ-1.

5. cost :

Equation 34

$$c = w(ai(t) + s)$$

cost of production equals wage rate (w) times amount of labour ($ai(t) + s$)

marginal cost:

Equation 35

$$mc = wa$$

equals wage rate(w) times marginal labour requirement(a)

6. revenue:

Equation 36

$$r = wa\mu i(t)$$

equals wage rate (w) times marginal labour requirement (a) times markup on costs (μ) ($wa\mu = p$ = price) times equilibrium output ($i(t)$)

marginal revenue:

Equation 37

$$mr = \frac{wa\mu}{\mu} = wa$$

equals price($p = wa\mu$) times (1-1/E) where E is the constant (subjective) price elasticity of demand (which can be shown to equal 1/(1-1/μ)), thus (1-1/E) = 1/μ.

Hence mr = p times 1/μ = p/μ. Note, here we are talking about imperfect competition so that price is unequal to marginal revenue. In fact

Equation 38

$$p = wa\mu$$

price (p) = wage rate (w) times marginal labour requirement(a) times markup (μ). If μ = 1 we have perfect competition so then mr = p.

7. the number of service firms(varieties):

Equation 39

$$x = \frac{(1-\beta)N}{ai(t)+s}$$

the number of firms (x) equals the total services labour force((1-β)N divided by the labour force per firm ($L = ai(t)+s$) at equilibrium

8. labour requirement:

Equation 40

$$L = s + ai(t)$$

the labour requirement equal to fixed labour requirement(s) plus marginal labour requirement(a) times firm's output ($i(t)$)

Exogenous variables

1. marginal labour requirement(a):

this is the exogenously determined increase in labour needed by the firm per unit increment of output (note that since output can be measured in any units, this can be left as 1).

2. fixed labour requirement(s>0):

this is the fixed cost in terms of service labour that must be incurred to produce any variety. It implies that increasing returns to scale exist in the service sector.

3. monopoly power/elasticity of substitution (μ):

as μ increases, the elasticity of substitution diminishes, as μ approaches 1, the services approach being perfect substitutes and variety diminishes in importance as a determinant of I.
Note that the elasticity of substitution is

Equation 41

$$\frac{\mu}{\mu-1}$$

4. total labour force (N):

Note how total manufacturing output(Q) is a nonlinear function of N, showing increasing returns with city size. However the latter is not modeled here and we treat N as exogenously determined.

5. The relative importance of workers versus services (β)

Equilibrium

Occurs when the level of output is such that marginal revenue(mr) equals marginal cost (mc), firms have entered shifting the demand curve to the left, driving down profits to zero, at which point entry stops. This is the equilibrium, when total revenue equals total costs and there are zero profits. This determines the equilibrium output level $i(t)$.

Hence
At equilibrium, profits are zero and costs equal revenues,

Equation 42

$$c = w(ai(t) + s) = r = wa\mu i(t)$$

Hence,

Equation 43

$$ai(t) + s = a\mu i(t)$$

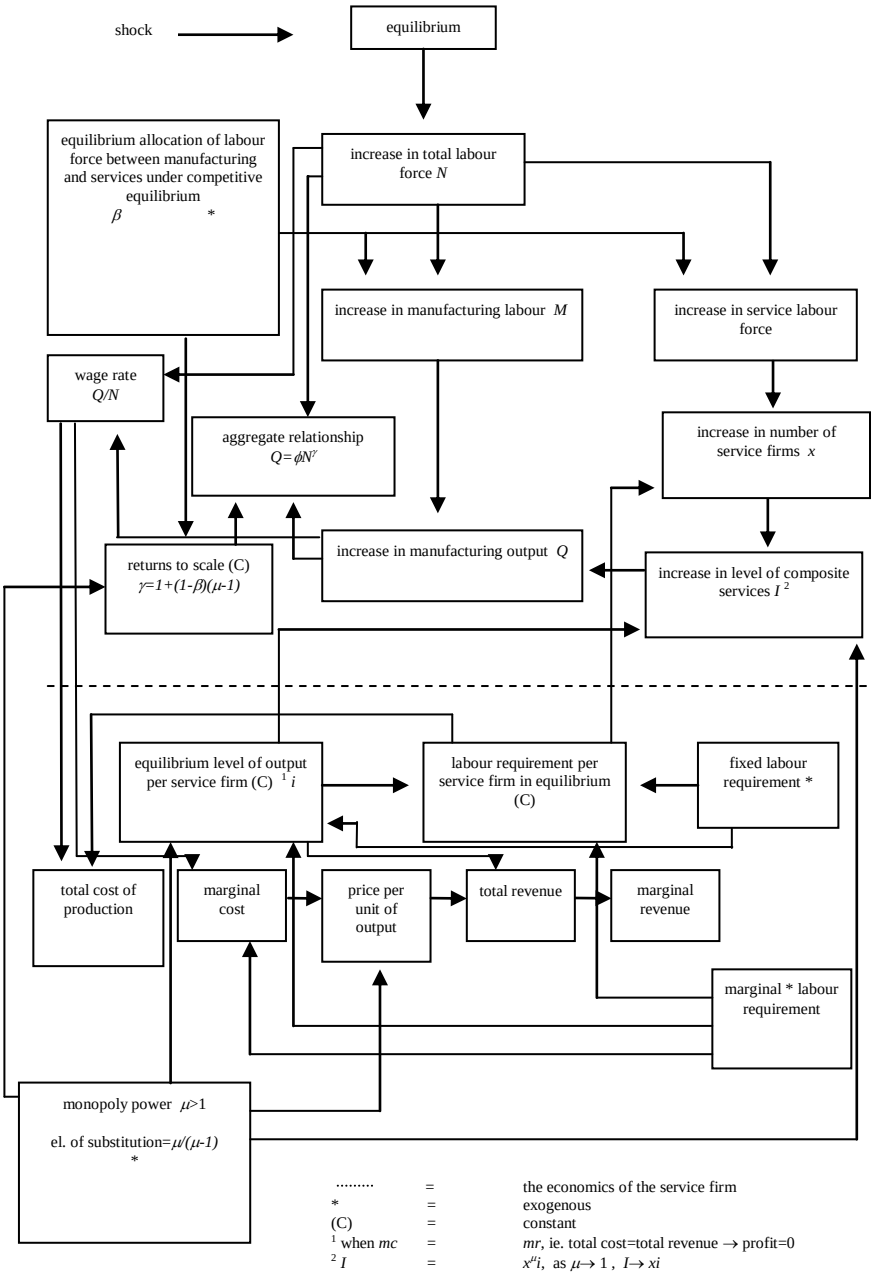
$$i(t) = \frac{s}{a(\mu - 1)}$$

We can choose units of output to be anything we want, which means we can choose them so that the marginal labour requirement $a = 1$. This gives the simplified version

Equation 44

$$i(t) = \frac{s}{\mu - 1}$$

Relations in the basic model



The elasticity of substitution

Recall that our model makes use of a constant elasticity of substitution production (CES) function. Define the elasticity of substitution as the percentage change in the quantities for a 1% change in prices. Take two firms t =1,2, then this is equal to the relative change in quantities i(t) divided by the relative change in prices p(t). Don't forget the negative sign, as prices rise, quantity demanded falls.

$$rel.ch.in.i = \frac{\Delta[i(2)/i(1)]}{i(2)/i(1)}$$

$$relch.in.p = \frac{\Delta[p_2/p_1]}{p_2/p_1}$$

$$eos = e_s = \frac{-\dfrac{\Delta[i(2)/i(1)]}{i(2)/i(1)}}{\dfrac{\Delta[p_2/p_1]}{p_2/p_1}} = -\frac{\Delta[i(2)/i(1)]}{\Delta[p_2/p_1]}\frac{p_2/p_1}{i(2)/i(1)}$$

It is the ratio of changes to levels. This is a general statement of elasticity of substitution. Let us look at the elasticity of substitution for our particular model. Let us start with the demand functions for two varieties of services

$$i(1) = kp_1^{-(\mu/(\mu-1))}$$
$$i(2) = kp_2^{-(\mu/(\mu-1))}$$

Next, we form the ratio of these, simplifying by writing ε in place of $\mu/(\mu-1)$

$$\left[\frac{i(2)}{i(1)}\right] = \frac{p_2^{-\varepsilon}}{p_1^{-\varepsilon}} = \left(p_2/p_1\right)^{-\varepsilon}$$

Now we know the change in i(2)/i(1) for unit change in (p_1/p_2) is the derivative of i(2)/i(1) with respect to (p_1/p_2). This is obtained from the above equation, with the simplification in the first line below resulting from the fact that we know the exponentiated term in the numerator from the previous equation. Then, multiplying the derivative by the prices ratio over the quantities ratio gives the elasticity of substitution (e_s) .
That is

$$\frac{\partial[i(2)/i(1)]}{\partial(p_2/p_1)} = -\frac{\varepsilon(p_2/p_1)^{-\varepsilon}}{(p_2/p_1)} = -\frac{\varepsilon[i(2)/i(1)]}{(p_2/p_1)}$$

$$el. \quad of \quad substition = e_s = \frac{-\varepsilon[i(2)/i(1)]}{(p_2/p_1)} \frac{(p_2/p_1)}{[i(2)/i(1)]} = -\varepsilon = -\frac{\mu}{\mu - 1}$$

So as μ increases, elasticity of substitution tends to 1
As μ approaches 1, elasticity of substitution tends to infinity

Marshall's ideas

Alfred Marshall(1890) was the first to use the notion of external economies and he wrote specifically about how they were caused industrial districts, such as the specialized production of cutlery in Sheffield, to emerge. Marshallian externalities are usually grouped into 3:

- Specialist suppliers

The idea here is that the concentration of similar producers in a specific locale provided a sufficiently large market to support specialized local suppliers, while the presence of specialist suppliers is attractive to producers

- Thick markets

Thick market have numerous buyers and sellers and make it easy to find a buyer or seller and thus reduce the amount of time to consummate trade. Any market structure that allows for greater communication and creates a 'focal point' for traders will reduce transactions costs. Furthermore, goods and services that must be purchased frequently incur frequent transactions costs. Thus, in these types of markets people have strong incentives to reduce transactions costs. A city is ideal for reducing transactions costs.
The notion applies in particular to the buying and selling of labour services. Even if there are many buyers and sellers of labour, many trades require very specialized skills or job descriptions. Agglomeration allows firms and workers to quickly and efficiently search for that person with the precise skills for the job, and for highly skilled workers to find a job that suits. Firms and people located in a thick labour market have an obvious locational advantage. The local concentration of workers and jobs is self-evidently symbiotic. If there are thick markets for specialized labour adjustment costs can be presumed to be low, as labour can move easily and hiring and firing costs are low. In such an environment, workers tend to move more frequently between jobs, thus providing a readily accessible common labour market pool for existing and potential firms within the sector.

- Externalities due to the transmission of information and expertise

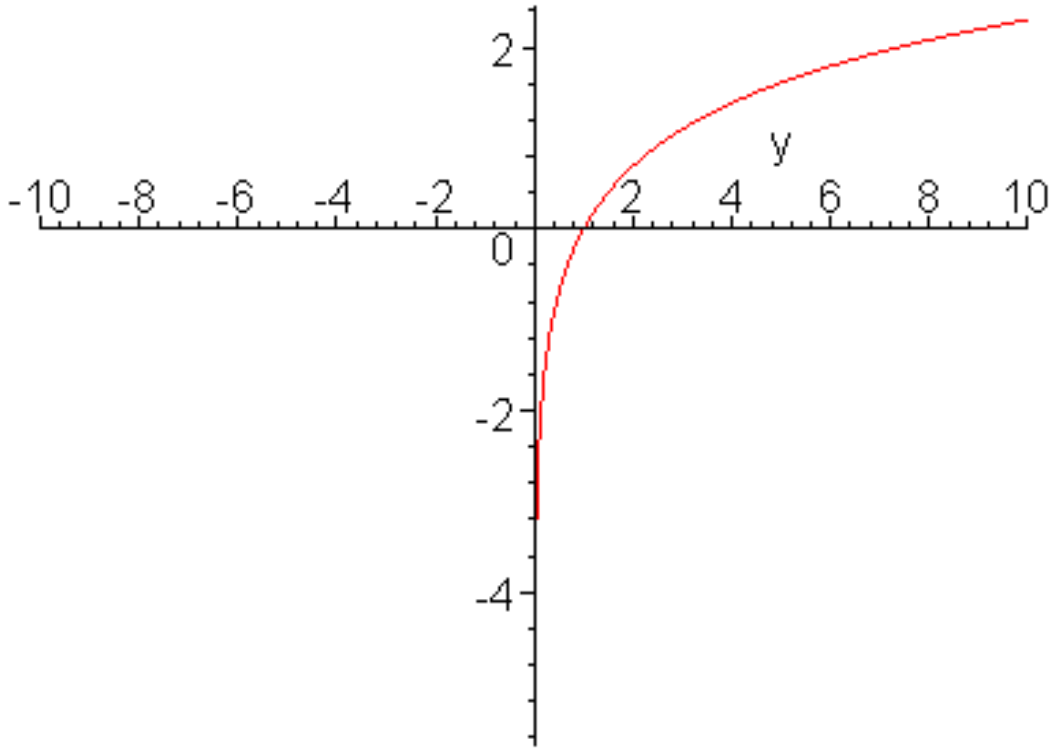
The fact that industry and labor tended to agglomerate concentrates know-how and skill within the labour market, thus reinforcing the agglomeration tendency and further deepening the pool of local knowledge

Marshall's considered his trio to be external effects acting on firms under perfect competition and subject to constant returns to scale. Thus economies of scale are

external not internal. This assumption of perfect competition might be viewed as a limitation, but nonetheless Marshall's theory does give useful insights, and external economies remain an important part of the contemporary spatial concentration of production story. Marshall's trio can be thought of as a mix of pecuniary and technological externalities. López-Bazo, Vayá and Artís (2001), for instance, treat input-output and thick-market effects as equivalent to pecuniary externalities, and knowledge spillovers as technological externalities.

Growth rates

The reason that differentiating the log with respect to time gives the growth rate is as follows. Since $\partial y/\partial t$ is the rate of change, therefore $1/y(\partial y/\partial t)$ is the proportional rate of growth. Now it is also the case that $1/y = \partial \ln(y)/\partial y$, a fact that can be obtained from any mathematics for economists textbook (eg Pemberton and Rau, 2001). Simply, on a graph of y versis ln(y), as y increases the slope (ie $\partial \ln(y)/\partial y$) becomes very small.



y versus ln(y)

Multiplying $1/y = \partial \ln(y)/\partial y$ by $(\partial y/\partial t)$ to get the proportional growth rate is the same as $\partial \ln(y)/\partial t$ since the ∂s cancel via the composite function rule of differentiation: in other words differentiating a log with respect to time gives the rate of growth. This is set out mathematically below.

$$y = f(t)$$
$$\partial y / \partial t = f'(t)$$
$$\frac{1}{y}\frac{\partial y}{\partial t} = \frac{1}{f(t)} f'(t) = \frac{f'(t)}{f(t)}$$
$$\frac{\partial \ln(y)}{\partial y} = \frac{1}{y}$$

$$\frac{\partial \ln(y)}{\partial y}\frac{\partial y}{\partial t} = \frac{1}{y}\frac{\partial y}{\partial t} = \frac{\partial \ln(y)}{\partial t}$$

The demand function for a variety

The demand function given as equation (10).

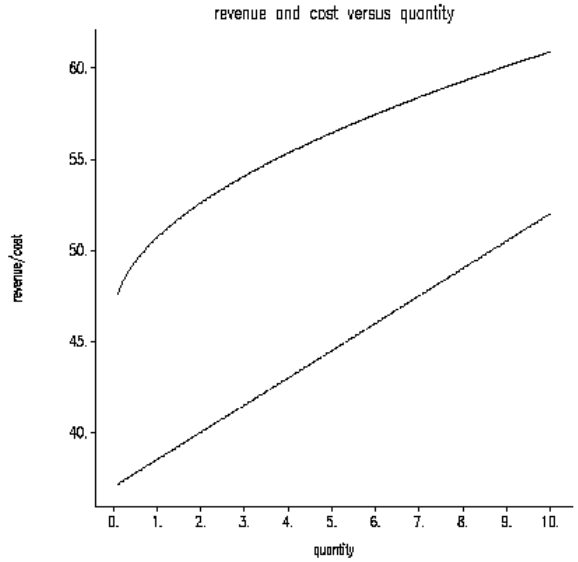
$$i(t) = kp_t^{-(\mu/(\mu-1))}$$

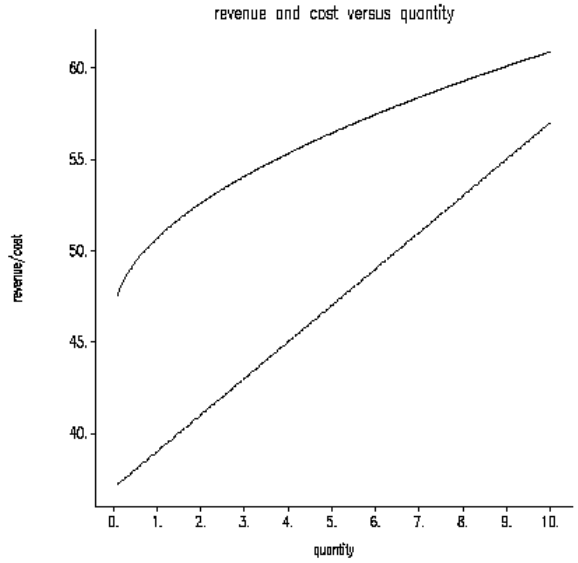
This shows the relationship between the quantity $i(t)$ and its price p_t . At any given price p_t there is an amount $i(t)$ which is demanded by the competitive sector that is consistent with profit maximization in that sector. We can see competitive sector profits Π rise to a maximum then fall as the quantity $i(t)$ increases, assuming a given price p_t . This is the outcome of profits equal to revenue minus costs, where revenue is a nonlinear function of $i(t)$ while costs are a linear function, as shown by equation (11)

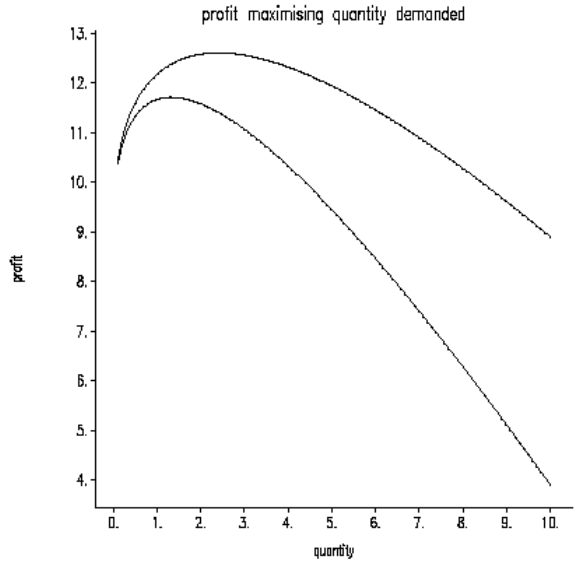
$$\Pi = PQ - [Mw + \sum_{t=1}^{x} p_t i(t)]$$

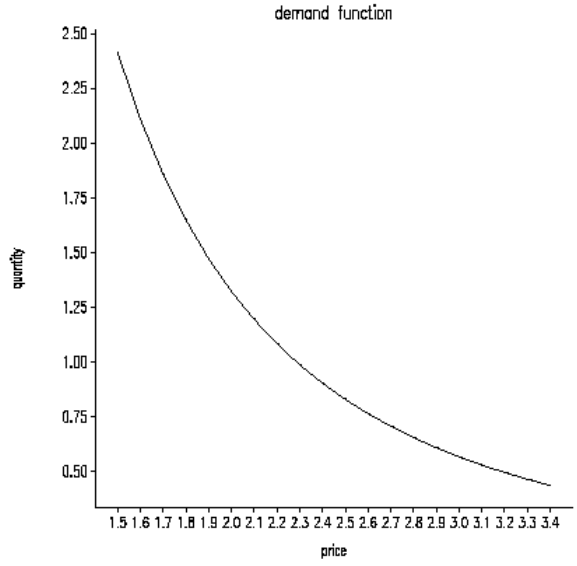
If we change the quantity $i(1)$ of a single variety at a given price, holding everything else constant, we can trace the path of revenue, costs and profit as in the Figure below. The nonlinear revenue path is because, while P is set by world market conditions and is treated as constant, $Q = M^{\beta} I^{1-\beta}$, so that there are diminishing returns to *I*, and *I* is determined by $i(1)$, so as $i(1)$ and therefore *I* increases, there is a less than proportionate increase in *Q*. On the other hand, costs are linear in $i(1)$, so as $i(1)$ increases, there is proportionate increase in $[Mw + \sum_{t=1}^{x} p_t i(t)]$ where Mw is constant and p_t is constant.

Now look (Figure below) at what happens if the price of variety 1 changes, hence we increase p_1 to obtain the resulting revenues, costs and profits at different levels of demand $i(1)$. The change in price p_1 is assumed not to affect the revenues at all, both P and I hence Q are the same as previously. However the cost line does change, since with higher price p_1 the costs are higher than they otherwise would be, which means that the profits are lower. Moreover, the quantity demanded $i(1)$ at which competitive sector profits are maximized changes, down from about 3 to about 1.5 on the figure. If we repeat the exercise, changing the price p_1 systematically to obtain the profit maximizing quantity demanded $i(1)$, we obtain the downward sloping curve relating demand to price. This can be shown mathematically to exactly equal equation (10)









References

Abdel-Rahman H, Fujita M (1990) Product variety, Marshallian externalities, and city sizes. *Journal of Regional Science,* 30: 165-183

Ciccone A, Hall RE (1996) Productivity and the density of economic activity. *American Economic Review,* 86: 54-70

Fingleton B (2000) Spatial econometrics, economic geography, dynamics and equilibrium : a third way? *Environment & Planning A*, 32: 1481-1498

Fingleton B (2001a) Theoretical economic geography and spatial econometrics: dynamic perspectives *Journal of Economic Geography*, 1: 201-225

Fingleton B (2001b) Equilibrium and economic growth: spatial econometric models and simulations. *Journal of Regional Science*, 41: 117-148

Fingleton B. (2003) 'Increasing returns: evidence from local wage rates in Great Britain', *Oxford Economic Papers*, 55, 716-739

Fingleton B (2006) 'The new economic geography versus urban economics : an evaluation using local wage rates in Great Britain', *Oxford Economic Papers* 58 501-530

Fujita M, Krugman P, Venables AJ (1999) *The Spatial Economy: Cities, Regions, and International Trade*. MIT Press, Cambridge and London

Fujita M, Thisse JF (1996) Economics of agglomeration. *Journal of the Japanese and International Economies,* 10: 339-78

Glaeser EL (1999) Learning in cities. *Journal of Urban Economics,* 46: 254-277

Glaeser EL et al (1992) "Growth of cities", Journal of Political Economy, 1992, vol.100, no. 6, 1126-1152

Jacobs J (1969) *The Economy of Cities*. Random House, New York

Neary J P (2001) 'Of Hype and Hyperbolas: Introducing the New Economic Geography' J*ournal of Economic Literature* XXXIX 536-561

Quigley J (1998) Urban diversity and economic growth**.** *Journal of Economic Perspectives,* 12: 127-138

Rivera-Batiz F (1988) Increasing returns, monopolistic competition, and agglomeration economies in consumption and production. *Regional Science and Urban Economics,* 18: 125-153