Chapter 2

A survey of complementarities in growth and location theories

2.1 Introduction

In our ever-changing economy, few trends last so long that they may be used to characterize the developments from the industrial revolution until today. Yet over the centuries, two phenomena seem to have stood the test of time: every year, on average, economic output grows by a few percentage points (Romer 1986). And, through the years, economic activity has always agglomerated into small areas, instead of spreading out evenly (Krugman 1991a).

As an example of continuing growth, consider figure 2.1 on page 10 which shows Dutch GDP per capita over more than a century. On average, growth is about 1.5% per year and apart from the period 1930–1945, the crisis and war years, economic growth is a regular phenomenon. In figure 2.2 on page 11, observe the year-2000 production per hectare[1] for each of the 12 Dutch provinces. There is an astounding eleven-fold difference between the most and least producing province. The differences in total production reflect different populations as well as differences in productivity, but are not necessarily the results of exogenous differences in natural endowments of the provinces.

This thesis is concerned with the things that economic science has to say about these two matters. What exactly brings about the growth of an economy, and why is it that production is so unevenly distributed? Using economic theory and models, we try to answer these questions and evaluate how policy affects growth and concentration.

In terms of models, the treatment of both growth and agglomeration has been rather upside-down. As for growth, the Solow (1956) model explains transitory adjustment processes, but the persistence of growth is an

[1] One hectare is 10,000 m^2 and approximately 2.47 acres.

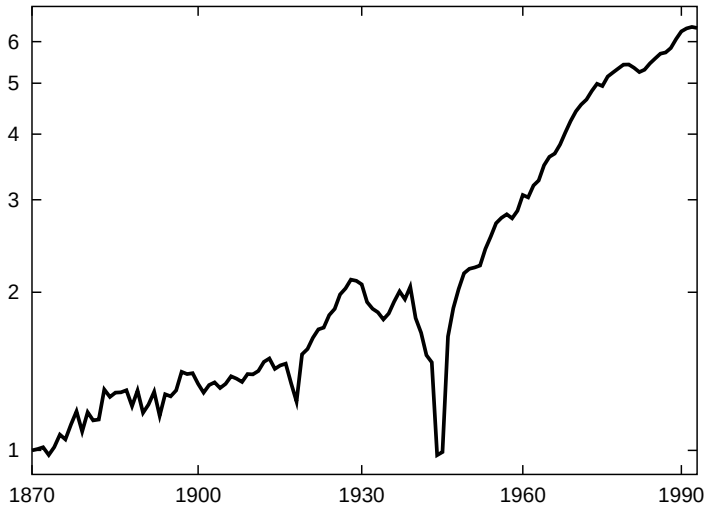

Figure 2.1: GDP per capita in the Netherlands, index numbers ($1870 = 1$), log scale. Source: Maddison (1995)

assumption rather than an outcome. Agglomerations can be studied using land-rent models based on von Thünen (1842) or using the central-place theory of Lösch (1967). The former shows how the existence of a center affects the hinterland, while the latter constructs the efficient placement of centers on a featureless plain. In both theories however, the center is assumed rather than derived.

There exists an interesting connection between the deficiencies of these two approaches, and it is this connection that will be the theme of this survey. The inability of both models to generate the phenomena that seem so characteristic of real life is caused by the market form that is used. Both assume that economic activity is exclusively conducted by firms that are in full competition. This market form is in accordance with the firms' technical specification, namely, it is assumed that all firms are subject to constant returns to scale. In the context of both growth and location theory, we will contrast models which use CRS[2] with alternatives that feature increasing returns to scale and monopolistic competition. These models use the Dixit and Stiglitz (1977) MC framework to construct theories where growth and agglomeration are a consequence of the model, rather than an assumption.

[2]When no ambiguity may arise, I use the expressions 'constant returns to scale,' 'constant returns,' and the acronym CRS interchangeably. The same holds for MC as an acronym for Monopolistic Competition.

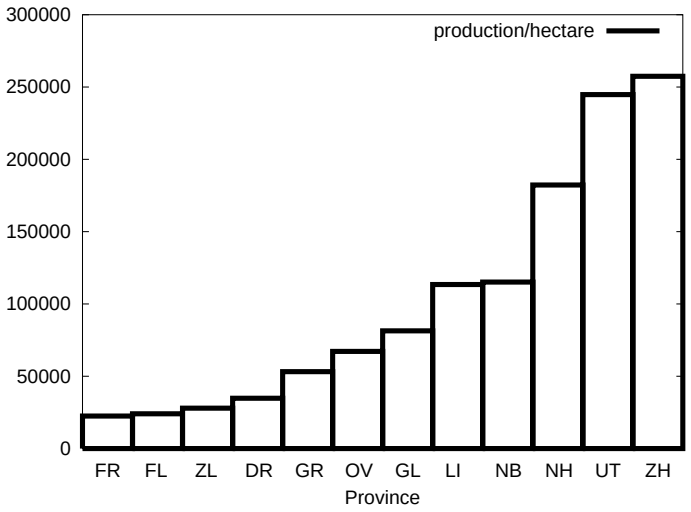

Figure 2.2: Gross regional product in Euros per hectare for twelve Dutch provinces, 2000. Source: CBS (2002)

Growth becomes endogenous growth, and the city center becomes an endogenous agglomeration.

The MC model has caused quite an upheaval in many areas of economics. The different fields that have profited from this innovation are discussed in Buchanan and Yoon (1994) and Brakman and Heijdra (2003), for instance.

As for the CRS assumption, it is easily seen that constant returns to scale severely limit the possible outcomes. If production is conducted under CRS, each separate factor in production faces decreasing returns. When growth is based on the accumulation of a subset of factors, this means that the economy cannot grow without bounds, and ends up in a steady state with zero growth.[3]

In location theory, the spatial impossibility theorem of Starrett (1978) (which can be found in Fujita 1986) states that a model with mobile agents on a closed, homogeneous space, facing a CRS production technology, can never explain the occurrence of agglomerations. Land rent will disperse economic activity without any countervailing force, because dividing up production over many locations leads to no loss in efficiency.[4]

Given these shortcomings of the CRS framework, it would seem tempting to use a wider class of firms, including those with increasing returns to

[3]For a detailed analysis, see section 2.4.

[4]The clash between economic geography models and the need to specify the market structure is discussed at length in Krugman (1995a).

scale. However, with the relinquishing of CRS, the assumption of full competition becomes untenable. The occurrence of other market forms greatly complicates the analysis, and allows for few analytical results. Fortunately, as we indicated above, a concise model of monopolistic competition introduced by Dixit and Stiglitz (1977) can be used to circumvent these problems.

This paper will briefly look at the monopolistic competition framework, and surveys the endogenous growth theory and economic geography in the light of it. It turns out that many interesting results in the two branches of literature can be attributed to the same fundamental properties of the monopolistic competition framework. The interplay between growth and geography is therefore not purely coincidental. While the models that show this were only recently made rigorous, their conclusions have been anticipated decades ago by such economists as Kaldor (1970) and Myrdal (1957):

> "[...] the movements of labour, capital, goods and services do not by themselves counteract the natural tendency to regional inequality. By themselves, migration, capital movements and trade are rather the media through which the cumulative process evolves—upwards in the lucky regions and downwards in the unlucky ones." Myrdal (1957, p. 27)

The antiquated notion of 'cumulative causation' is revived today as a process caused by complementarities in the model.

I will first look into the nature of monopolistic competition and the complementarities that characterize it. This is done in section 2.2. The findings are then used to provide a selective survey of economic geography (section 2.3) and endogenous growth theory (section 2.4). The two strands of literature are brought together in section 2.5, and section 2.6 concludes.

2.2 Complementarities and the monopolistic competition framework

In the classical framework of economics, many important results are obtained under a broad set of assumptions. For instance, the propositions of welfare economics as they may be found in Arrow and Hahn (1971) or in Takayama (1985, p. 185), guarantee that in general, decentralized market outcomes are socially optimal.

The theory assumes, among others, that all producers of goods are in full competition. This assumption implies a number of important simplifications: under full competition, one producer's pricing decision does not influence the market price. Also, no producer makes a profit, and prices should equal marginal and average costs. These simplifications produce,

in general, a single, unique, welfare-maximizing solution. They also allow for simple pricing rules in the absence of strategic considerations. In this environment, a great number of analytical results may be derived.

As noted by Dixit and Stiglitz (1977, p. 297), the existence of a unique and optimal market equilibrium can challenged for at least three reasons, one of which is a failure of the model to reflect economies of scale that are observed on the level of a firm.[5] Allowing for these economies of scale, however, means letting go of the CRS assumption. This alters the behavioral assumptions that are appropriate for the firms (Helpman 1984). Increasing returns imply, for instance, that the largest firm has the lowest average costs, and is able to push the smaller competitors off the market. Even if this may seem realistic for some sectors, it makes it much harder to derive analytical results.

There is a case for abandoning CRS however. The assumption tends to bend reality, and paints a world in which economic transactions are basically a zero-sum game. In a CRS economy, it is of no consequence whether all people divide their time over the same range of activities, or whether each person specializes in a single activity and trades with the others. Clearly, this outcome is unsatisfactory as a reflection of real economic activity. It goes as much against common sense as it goes against the founding words of economics as a science, dedicated to the productivity gains from dividing labor (Smith 1776, p.13).

The issue whether to assume CRS thus turned out to be rather crucial for a coherent model of general equilibrium, but unrealistic in practice. This left economists divided for a long time:

> '... there seem to be two traditions, which persist. On the one hand there are those who are so impressed by what has been done by the CRS method that they have come to live with it; on the other, those for whom scale economies are so important that they cannot bring themselves to leave them aside.' (Hicks 1989, p. 12)

Among the efforts to bridge the gap was the work by Chamberlin (1933) and Robinson (1933), who sketched an alternative market form to the full competition implied by CRS. Their framework, monopolistic competition, held the promise of reconciling the two camps, but was rejected by most economists because of supposed inconsistency (Heijdra 1997). A severe blow to the MC model was dealt by Stigler (1968), who considered the model a failure and argued that economists should restrict themselves to the analysis of perfect competition and monopolies.

Despite these problems, the attractions of the MC model remained such that economists kept searching for a formulation that would be both math-

[5] The other two are distributive justice and external effects.

ematically consistent and useful in practice. This would be a version where strategic interactions between the different firms would have to be somehow 'sanitized' from the model, for it was clear that such considerations would severely cloud the search for equilibrium. A formulation that allowed exactly that was finally found by Spence (1976) and Dixit and Stiglitz (1977). Their breakthrough articles set a chain of events in motion in which the MC-alternative to CRS became widely used, especially in industrial economics, trade theory, growth theory and economic geography. Although theirs is 'a very restrictive, indeed in some respects, a silly model' (Krugman 1998, p. 164), allowing the economist to focus on the effects of increasing returns without worrying about strategic interactions between firms made it an instant classic. The apparent arbitrariness of the model is not denied, but taken for granted, hoping that insights will extend beyond the model:

> "Unfortunately, there are no general or even plausible tractable models of imperfect competition. The tractable models always involve some set of arbitrary assumptions about tastes, technology, behavior, or all three. This means that [...] one must have the courage to be silly, writing down models that are implausible in the details in order to arrive at convincing higher-level insights." (Krugman 1995a, pp. 14-15)

It is important to realize that the monopolistic-competition approach is not the only available route into increasing returns, and that some insights are sacrificed when it is chosen. As Dixit and Norman (1980) write,

> For descriptive purposes, one must [...] choose among the numerous alternative ways in which imperfect competition can be modelled; and the conclusions one arrives at will in general depend on the particular specification chosen. [...] The best one can hope for is a catalogue of special models. (p. 265)

Neary (2003) argues that the MC model has nothing to say, for instance, about the effects of globalization on market structure. In that case, a model of strategic oligopolistic interaction is needed.

This section provides a short introduction to the Dixit-Stiglitz monopolistic competition framework. Before looking at the model itself, we will briefly discuss the problems that surround returns to scale in general, and the notion of externalities.

2.2.1 Returns to scale

A firm's production possibilities are summarized in its production function. If for an amount A of a certain product a firm uses inputs, whose

quantities are summarized in a vector $\mathbf{B}$, the correspondence between different values of A and $\mathbf{B}$ defines the production function $f(\mathbf{B})$. For any $\mathbf{B}$, we can evaluate the returns to scale of the firm by looking at the point elasticity

$$\varepsilon_B = \left. \frac{\partial f(\lambda \mathbf{B})}{\partial \lambda} \frac{\lambda}{f(\mathbf{B})} \right|_{\lambda=1}.$$

When ε_B is larger than one, there are increasing returns to scale. Note that ε_B is a function of the inputs $\mathbf{B}$. A firm can have increasing returns for all possible $\mathbf{B}$, but also for a limited set of values of $\mathbf{B}$.

On the level of the entire economy, increasing returns to scale are fairly undisputed. In this case, we can think of f as a nation's production function, with $\mathbf{B}$ indicating the supply of labor and capital. Increasing returns have been attributed to the division of labor (Smith 1776), splitting up complex production methods into multiple simple steps (Young 1928, Stigler 1951), and the fact that technological knowledge, once produced, is nonrival and nonexcludable (Romer 1990). It would be a positive quality of any economic model to have the possibility of including increasing returns on the macro level.

Much of today's macroeconomic theory is derived explicitly from microeconomic foundations (see, for instance, Romer 1993). The occurrence of increasing returns at the micro-level spells trouble. Helpman (1984) shows that the modeler needs to specify a host of parameters to even start working: the conditions of firm entry, the heterogeneity of the good, and the type of market are just a few among them. The outcome of the model is highly dependent on these assumptions, for instance, do firms compete in a Bertrand– or a Cournot–market?

The simplest of these assumptions is that every sector is dominated by a single monopolist, who fully exploits the increasing returns. Apart from the question of realism, the presence of monopolists causes problems in a general-equilibrium model. One source of problems is the occurrence of monopoly rents: the model needs to specify how these rents are spent by the monopolist. In full competition, profits are zero by definition.

To avoid these issues altogether, one can assume that part of the returns to scale are external to the firm. The idea, originally from Marshall (1920), separates internal economies ('those dependent on the resources of the individual houses of business engaged in it', p. 266) from external economies ('those dependent on the general development of the industry', p. 266). External economies, or externalities, do not affect the firm's optimization; thus, they can be incorporated in a consistent profit-maximizing framework, where firms perceive their situation as one of full competition. Between externalities, we can find two types (Scitovsky 1954): pecuniary externalities, those which are mediated by markets, and the rest, non-pecuniary externalities.

Non-pecuniary externalities use a production function, at the firm level, like $f(\mathbf{B}) = \tilde{f}(\mathbf{B}, X)$. Here, $\mathbf{B}$ again are the inputs and X is industry output (Helpman 1984). Every single producer considers X as given, and controls only $\mathbf{B}$. But f may have increasing returns in $\mathbf{B}$ and X together.

Using non-pecuniary externalities, it is possible to construct a model of general equilibrium that features increasing returns. Although this has indeed been done (Chipman 1970), such models have not been used extensively. By their nature, non-pecuniary externalities are not observed so that the economist can assume anything about them. Any possible outcome can thus be 'doctored' into the model.

Pecuniary externalities are more subtle. It could be possible that a producer, by entering a market, increases the consumers' utility because of the increased variety that he/she provides. Although profit opportunities were the firm's original motive for entering, the variety effect may influence the perceived price level faced by the consumer, and alter the allocation of goods. Another example would be the entry of a firm that, because of its demand for an input, affects the price that input for all other firms.

However, the methodological problems outlined by Helpman (1984) still need to be solved. A particular model that knits together increasing returns at the firm and macro level in a consistent way, and thus solves these problems, is the Monopolistic Competition model of Dixit and Stiglitz (1977). The introduction of this model, in which pecuniary externalities drive the equilibrium, for the first time allowed the analysis of increasing returns and caused what Brakman and Heijdra (2003) call the 'second[6] monopolistic revolution.' We will introduce the MC model in the following section.

2.2.2 Monopolistic competition

The key difference between full competition and monopolistic competition[7] is in the nature of the traded good. With full competition, the good is assumed to be homogeneous, and its price the only criterion of selection. With MC, consumers discern different varieties, and products from different producers are imperfect substitutes.[8] Even if each individual producer faces increasing returns to scale in production, the largest producer is not always able to push smaller competitors out of the markets because substitution between products is limited.

In most applications of MC, consumer preferences are modelled as in

[6] The first monopolistic revolution was the idea of MC being formulated by Chamberlin (1956).

[7] We will use the acronym MC for 'monopolistic competition' from now on.

[8] Chamberlin (1956, p. 56) suggests that such elements as 'the conditions surrounding its sale', trade marks and the seller's reputation 'may be regarded as [being purchased] along with the commodity itself.'

Dixit and Stiglitz (1977)[9]. The quantities of goods x_i consumed are aggregated in a CES function,

$$U(x_1, \ldots, x_n) = \left(\sum_{i=1}^{n} x_i^{\theta} \right)^{1/\theta}. \tag{2.1}$$

with $0 < \theta < 1$. By choosing suitable units of measurement for the different goods, we can abstain from adding scale parameters to the different x_i. It is clear that for each of the goods, an increase in the amount consumed will increase total utility. If we maximize (2.1) with respect to a budget constraint $\sum x_i p_i = E$, we find that

$$x_i = \frac{E}{q} \left(\frac{p_i}{q} \right)^{-\sigma} \tag{2.2}$$

where $\sigma = 1/(1-\theta) > 1$, and we have used the associated (ideal) price index $q = \left(\sum p_j^{1-\sigma} \right)^{1/(1-\sigma)}$. We assume a large number of producers n so that the effect that one producer's price has on q is vanishingly small. So, each producer takes the price index as given and faces a demand elasticity σ for his product. Also, he does not need to take the behavior of other producers into account when deciding on price and quantity. Strategic motives are absent, and this makes the model tractable and easy to solve.

If every variety sells for the same price p, all are purchased in the same amount. In this case, formula (2.1) shows that utility is $n^{1/(\sigma-1)} E/p$. That is, an increase in variety brings an increase in utility even if the nominal budget remains the same. Helpman and Krugman (1985, p. 117) call this the 'love-of-variety effect.'

The more varieties (n) there are, the less influence a single producer's price exerts on the consumer's real income. To completely eliminate every producer's market power, it is often assumed that the range of goods $[0 \ldots n]$ is continuous, and each producer is infinitely small. Though awkward, this assumption can be given some rigor. This is done in appendix 2.A.

Producers are usually assumed to face a fixed cost for setting up production and a variable cost per item produced. This implies the average cost per product declines with total production, so that producers are subject to increasing returns technology. This encourages firms to expand their output as much as possible; however, they also face a downward sloping demand curve as we saw in formula (2.2). Thus, producers maximize profits by setting marginal benefit equal to marginal costs, which given (2.2)

[9] Weitzman (1994) shows that this model is much related to the Lancaster (1979) 'spatial competition' model, where each consumer has an ideal product and picks the one closest to it.

results in a mark-up over marginal costs of size $1/\theta$. In equilibrium, all producers set the same price. The number of active producers adjusts so that discounted profits are just enough to recoup the initial investment F. With free entry, this means that n adjusts to drive profits to zero.

The constant elasticity of demand, faced by a producer, is at once an advantage and a disadvantage of the model (Dixit 2000). It allows us to get a simple form for the pricing equation, which gives the model much of its appeal. However, as the number of varieties increases, we would expect the products to become more similar and the elasticity of demand to increase. This way, there would be a competitive limit to the model. In the current formulation, this is not the case. We should recognize this flaw when we discuss models where n grows *ad infinitum*.[10]

In an alternative interpretation of the same model, Ethier (1982) used the aggregator function in (2.1) as a production function. Output U is made with inputs x_i; each input is produced by a single intermediate goods producer. The production function belongs to a class of firms that convert the intermediate goods into a final consumer good. These firms face constant returns to scale, as may be checked from (2.1), and are in full competition. The 'love-of-variety effect' from above has now become quite another thing: when entrance is free, there are increasing returns to scale at the economy's macro level. We will return to this interpretation below, as well as in the following chapters.

Now that increasing returns to can be modelled consistently, we are able to construct a general equilibrium theory where the actions of one firm affect the conditions of other firms, though not intentionally. We will find that many equilibria in MC models, for their stability, depend on the fact that the actions of several firms complement each other. Complementarity is the subject of the next section.

2.2.3 Complementarities

Matsuyama (1993, 1995) discusses complementarities, the notion that "two phenomena (or two actions, two activities) reinforce each other." (1995, p. 702). Complementarities often arise in the MC framework.

As a specific example, assume that in an economy, people consume a single final product that is made out of several intermediate goods with production function (2.1). That is, there are n different intermediate goods, and total production is U. This is the Ethier-setup from above. Assume also that intermediate-goods producers face fixed costs F and variable costs θx_i

[10] An extension of the model that goes into this direction is introduced by Heijdra and Yang (1993)

which are both incurred in labor, that is,

$$L_i = F + \theta x_i \tag{2.3}$$
$$\pi_i = p_i x_i - w(F + \theta x_i) \tag{2.4}$$

where x_i is the output of firm i (the double use of parameter θ is here and in formula (2.1) is for mathematical convenience), π_i is firm i's profit and w is the wage rate. Remembering that price, in this model, is a markup $1/\theta$ over marginal costs, we can use the elasticity of substitution in the price equation, writing it as

$$p_i = \frac{\sigma}{\sigma - 1} \cdot \theta w = w \tag{2.5}$$

where σ is defined as above. From this and (2.4) it follows that a firm that makes zero profits employs $L_i^* = \sigma F$ workers. When there are L workers in the economy and there is free entry in the intermediate sector, it follows that the number of producers in that sector will be

$$n^* = \frac{L}{\sigma F} \tag{2.6}$$

The production of the final good, per capita, is increasing in n^*, because of increasing returns to scale on the macro level. In fact, per capita production is $(n^*)^{1/(\sigma-1)}$.

Now if there exist two of these economies, with different intermediate goods, and they open up for trade, both economies will see the range of available intermediate goods increase. Because of this, both economies will experience an increase in production per capita. When the two economies interact, they are complementary to each other. This principle has been the basis for a large class of trade models, for instance in Helpman and Krugman (1985).

Hirschman (1958) discusses a related issue in the context of economic development. In his terminology, there exist *linkages* between different firms in a region. These linkages concern the input-output relations among the firms. Hirschman distinguishes *backward* linkages when a firm demands inputs from other firms, and *forward* linkages when a firm produces inputs for other firms. The conjecture is that with positive costs of transport for intermediate goods, linkages between firms can make an agglomeration stable.

In fact, the conjecture requires that linked firms are complementary to each other. It is true that in general, the arrival of a downstream firm can induce an upstream firm to expand. However, when this happens in a constant-returns world, the expansion has no effects on the original activities of the individual upstream firm, and merely leads to entry of upstream firms. The linkage is rather weak in this case. But should the upstream firm exhibit increasing returns to scale, expansion means that it can now operate at a higher level of efficiency. In that case, the two firms are complementary.

2.2.4 Review, and a look ahead

To study a complex phenomenon, it can be necessary to make a number of assumptions that simplify the problem. We have argued that the CRS assumption fulfilled such a role in economics, as it allowed the derivation of a simple rule of conduct for firms, namely, marginal cost pricing. It also solved the problem of which market form would prevail, in favor of full competition.

We have also introduced an alternative framework, based on a different assumption: the MC setup. This setup is not any more general than full competition, the number of assumptions has even increased. Yet it is an interesting alternative because it allows for complementarities and increasing returns to scale.

The short introduction above does not do justice to all the intricacies of MC, but that is not the point of this survey. Rather, we now want to look at the application of this framework to two fields, economic geography and growth theory. The application of MC to these fields has allowed a large number of innovations. Those in economic geography are discussed in the following section, while those in growth theory are the subject of section 2.4. The two strands of literature are brought together in section 2.5.

2.3 Economic geography

Ironically, economic geography or location theory has been a rather peripheral field of study within economics. In part, the small amount of attention for issues of location can be attributed to the institutional, geographical and sociological factors that play such an important role in the problem. Yet over the years, many interesting results have been obtained using methods of economics. We look at the foundations of location theory in section 2.3.1. Then we turn to a new class of models that involve monopolistic competition and increasing returns in section 2.3.2.

2.3.1 Foundations of location theory (1): Exogenous agglomeration

The earliest theory of location can be divided in two branches (Greenhut 1956): least-cost theory, oriented on the supply side, and spatial competition theory, oriented on the demand side of the economy. The striking characteristic of least-cost theories is that they start by assuming a form of agglomeration; they do not explain why the agglomeration came about in the first place. This problem is tackled to some extent by spatial competition theories, as well as by the theories based on externalities and those that use increasing returns. We look at the different theories in chronological order.

Least-cost or land use theory starts with assuming that all demand in the economy is located at a single point. This can be a mining town demanding agricultural produce as in von Thünen (1842), or a central business district in which all trade is conducted, as in Fujita (1986). Transportation is costly, and costs increase with distance from the center, r. From their production function and the costs of transport, suppliers can compute how much rent they want to pay as a function of r. This information is aggregated in a rent gradient, according to which the suppliers settle. The approach is refined by Weber (1909) to account for the location of raw materials, and Alonso (1964) adds, among other, endogenous lot size. Many models of urban structure still use this setup.

Spatial competition or locational interdependence theory, on the other hand, does not assume the existence of a center. Rather, (consumer) demand is distributed over locations and (zero-size) producers are looking for the optimal spot. With land rent out of the model, this approach clearly deals with questions of attraction and repulsion among different firms, which places it in the realm of game theory. The founding paper of this field is Hotelling (1929), who shows that two producers of a homogeneous good will locate next to each other halfway a line with evenly spread consumers. This is not the socially optimal situation. Chamberlin (1956, pp. 260-265) shows that increasing the number of sellers in this problem will cause their dispersion, converging to the optimal dispersion as the number of sellers goes to infinity. Gabszewicz and Thisse (1986) provide a survey of this method.

The two approaches above may be combined. Lösch (1967) and Greenhut (1952) introduce profit-maximization as the relevant criterion. Given that demand and supply conditions may vary with location, this tends to make the problem less tractable. There exist fewer general rules on spatial dispersion than in the above, simplified, analysis. An important limitation of both these approaches is the assumption that consumers do not change their location in response to the suppliers' whereabouts.

2.3.2 Foundations of location theory (2): Endogenous agglomeration

The location theory in the preceding section has said very little about the causes of agglomeration. We can think of at least three types of forces that drive people and firms to the same location. Firstly, there are the autonomous characteristics of the landscape. Some places may be more pleasant as a place of residence, or productive as a place of business than others. Natural harbors or strategic points fall in this category.

Secondly, it is often thought that nonmarket externalities are an important factor in the creation of agglomerations. Such hard-to-measure concepts as informational and technical spillovers between firms, or in general

informational exchanges between agents (Fujita and Thisse 1996, p.347) cause people to cluster together. The reason for clustering is the fact that the amount of spillovers between two firms is assumed to decline rapidly with distance. The spillovers are embodied in such acts as face-to-face talks and casual inspection of the other firm's production site. Nonmarket externalities are emphasized in Jacobs (1969).

The problem with the above two conjectures about the causes of agglomeration is that they are difficult to verify. Saying that agglomerations are caused by agglomeration economies is close to a tautology. Designating the spots where people have clustered 'attractive' is not much better. The predictive power of these theories is small. It is therefore preferable to have a model where agglomerations are a result of more fundamental properties like the way people consume and produce.

Recently, economists working on the theory of (urban) agglomeration have put forward a number of such micro-foundations. Duranton and Puga (2003) classify them using three motives for agglomeration: *sharing*, *matching* and *learning*. In each of these classes, models are formulated featuring pecuniary externalities. In the class of models in which *sharing* is the driving force behind agglomeration, the monopolistic competition framework of section 2.2.2 is the preferred vehicle of analysis. We briefly discuss each class of motives here, before concentrating on a particular kind.

- **Sharing**. The agglomeration of a large number of people into a city may be explained by the presence of an indivisible service to consumers such as a stadium, or by production-side indivisibilities such as a large factory. However, even with smaller sized services and firms, the *variety* they constitute when gathered into one place may be a force of attraction. Firms are complementary to each other in this respect, each one constituting a small part of the total supply of variety. We will spend most of this section on a model that shows how such sharing of variety may lead to equilibrium agglomeration. Other models in the *sharing* category use the returns to specialization and the sharing of (labor market) risk, a subject that goes back to Marshall (1920).

- **Matching**. Another way to bring out the advantages of a large labor pool is by looking at the heterogeneity of labor. Both the probability and the quality of labor market matches are larger when the number of firms and laborers is big. Thick labor markets in cities can thus play a role in the stability of the agglomeration.

- **Learning**. Finally, the role of cities as repositories of knowledge can be used to explain their success. Non-pecuniary effects in knowledge generation as emphasized by Jacobs (1969) fall into this category, as do models of skill transmission. Finally, models of learning-by-doing

> with specific urban features exist. These will be discussed in section 2.5.

In the rest of this section, we will concentrate on a model in which the variety offered by different producers leads to agglomeration. Because of their differences, producers are complementary to each other and when different producers are located at the same spot, their combined presence is an equilibrium. The principle, complementarity-induced agglomeration, was recognized by Krugman (1979) in a paper about monopolistic competition and international trade. In his model, where the MC setup was slightly different from above[11], when trade was prohibited but factors were mobile,

> [...] there will be an incentive for workers to move to the region that already has the larger labor force. [...] In equilibrium all workers will have concentrated in one region or another. (p. 20)

It is not very difficult to see the agglomerative tendencies using the model from section 2.2.2. Suppose that there are two regions in which economies with an MC structure exist, and that trade is prohibited. The number of firms in each region is linear in the number of inhabitants (section 2.2.3); the aggregate price index faced by each inhabitant is (section 2.2.2)

$$q = \left(\sum_{j=1}^{n} p_j^{1-\sigma} \right)^{1/(1-\sigma)} = \left(np^{1-\sigma} \right)^{1/(1-\sigma)}$$

which is decreasing in n (this is due to the 'love-of-variety-effect'). If inhabitants are given the choice where to live, they will move to the more populated region. Hence, agglomeration results naturally.

The first to design explicit models of location based on MC were Fujita (1988) and Rivera-Batiz (1988). These models featured agglomeration economies as well as land rents based on a least-cost framework (see Section 2.3.1 above). Despite the countervailing force of the rents, Rivera-Batiz shows that for some parameters, "[t]he economy's population [...] ends up completely in city m" (1988, p. 148).

Since so much of this book will be based on this type of models, we take the time to specify an economic geography model in the next section, and characterize the different varieties that are known.

[11] Specifically, the subutility-function x^θ is replaced by a function $v(x)$. The elasticity of demand is now $-v'/v''x$, and it is assumed that the elasticity decreases in x (this does not happen with the x^θ form). The assumption leads to the result that wages are higher in the most populated region; with the x^θ form this is not true, unless the wages are corrected for the local price index.

2.3.3 The Core-Periphery model and the home market effect

Endogenous agglomeration is the hallmark property of class of models known as 'new economic geography.' The complementarities that cause agglomeration arise because of the MC framework, but they may travel through different markets. In this section, we will look at the model of Krugman (1991b) where complementarities go through the labor market. We follow the exposition of Neary (2001), which is a little more concise than the original.

There are two types of production, agricultural and industrial. The former is homogeneous and operates under CRS, the latter is MC. Consumers in either region maximize utility,

$$\bar{U} = A^{1-\mu} \cdot U^{\mu} \tag{2.7}$$

where A is consumption of agricultural products and U is the aggregate utility obtained from the consumption of different varieties of the industrial product. U is defined in equation (2.1) above. We deduce that μ is the share of income spent on manufactures. Because of the MC assumption, producers face a demand schedule as in (2.2), where they take the price index q as given. The budget E is the amount of money that consumers spend on industrial goods, or μ times their income. This form of the demand function leads to mill pricing, where the price is a markup over marginal costs.

The production function uses only labor and is given in (2.3). This leads to the result that production per firm is $y_i = \sigma F$, and thus does not vary with any endogenous variable. This result is rather remarkable, and rather special. It requires that increases in costs because of rising wages, for instance, are exactly balanced against increases in revenue because of rising prices. A number of these special properties makes the MC model tractable, but very special. Because of the fixed output per firm, shifts in the level of total production are caused by changes in the number of firms, n^*, which is defined in (2.6).

Up until now, our model is a standard full equilibrium MC model of a closed economy. This changes when we assume that there are two regions for which the above specification holds, and that there are positive transport costs between them. These costs of transport take the *iceberg* form, where they are incurred in the shipped product itself. We assume that of what is shipped between the two regions, only a fraction τ actually arrives. The analogy of this type of transport costs to a melting iceberg is due to Paul Samuelson.[12] Agricultural goods are shipped without incurring transport costs and serve as the numéraire.[13]

[12] A similar analogy recently occured at a diner with a number of economists, one of whom refused to pass a bottle of wine without pouring some in his own glass.

[13] The assumption about the absence of transport costs for agricultural goods is not in-

The demand for industrial goods from the other region is different from the home demand (formula 2.2) for two reasons. Firstly, the price that is faced by foreign consumers is higher because of transport costs. Secondly, for each unit that is to be received in the other region, $1/\tau$ units must be shipped. This leads to the following form for foreign demand:

$$\begin{aligned} x_i^* &= \frac{E^*}{q^*}\left(\frac{p_i/\tau}{q^*}\right)^{-\sigma}\frac{1}{\tau} \\ &= E^*(q^*)^{\sigma-1}p_i^{-\sigma}\tau^{\sigma-1} \end{aligned} \tag{2.8}$$

In these equations, foreign variables are indicated with an asterisk. From this, we can immediately see the use of the iceberg-assumption. Even though the foreign price is increased by transport costs, the elasticity of foreign demand with respect to p_i remains equal to σ. Thus, the maximization problem for the producer remains the same, as does the optimal price.

Of course, the price indices q and q^* change when trade between the regions is allowed. The home price index now takes the form

$$q = \left[np^{1-\sigma} + n^*\left(\frac{p^*}{\tau}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \tag{2.9}$$

We can write a similar expression for q^*.

Notice the mathematical form of (2.9), which has a sum inside a power expression. This form cannot be simplified through manipulation, and remains at the heart of many variants of the Core-Periphery model. This makes it impossible, in general, to solve the model analytically, which leads to the fact that many results have to be derived through numerical simulation. Krugman (1998) calls the need for reliance on numerical results one of the hallmark properties of economic geography models.

We pause for a moment to summarize what we have constructed thus far. We have two regions which produce agricultural and industrial products, both of which can be traded. Prices are set the same as in autarky, courtesy of our special assumption about the form of transport costs. The number of farms and industrial producers is governed by the zero-profit assumption. We have not made any special assumptions yet about the labor force and whether it can change sectors, but so far we have assumed that there is no interregional migration.

Consider now a situation where both regions are exactly equal. What will happen when the home region faces an increase in demand? We know that, since per-firm production and prices are fixed at their optimal level, the demand shock leads to the entry of new industrial firms. These firms

nocuous. Davis (1998) shows that the introduction of positive costs of transport may negate several results, including the Home Market effect (see below and appendix 2.B).

produce new varieties of the industrial product, which will lead to a change in home's price index q. This in itself leads to a change in the equilibrium, which could result in a further bout of entry or exit. Will the demand shock lead to a more or less than proportional increase in the number of firms?

In appendix 2.B, we derive a result known as the *home market effect*, which answers this question. The result, originally by Krugman (1980), says that the region with higher demand will have a disproportionally *higher* number of firms. That is, the change in q leads to the entry of even more firms. From this, we can conclude that regions with a larger home market will be net exporters of manufactures. This result has been the basis of empirical tests of the MC model, which are discussed in section 5.2.1 below.

The home market effect tells us that an increase in demand leads to an even larger increase in the number of firms. If it were true that an increase in the number of firms, through some channel, caused another increase in demand, we would have a closed causal loop that could explain spontaneous agglomeration and persistence of regional differences as discussed in the quote by Myrdal on page 12.

There are a number of different possible channels through which the causal loop can be closed, and the choice of channel defines the type of CP model. Ottaviano and Puga (1997) classify three different types of models according to these media. They discern migration linkages, input-output (or intermediate good-) linkages and intertemporal linkages. We discuss the three types of models in turn. In each of them, the basic setup is as described above: there are two regions and each region has two sectors, agriculture and manufacturing, which are competitive and MC, respectively.

2.3.4 Three channels in the Core-Periphery model

The migration-based CP model

The first CP model, constructed by Krugman (1991a, 1991b), is based on migration linkages. In it, manufacturing workers decide where they want to live based on their real wage. From (2.7) and the fact that the price of agricultural goods is normalized at one, we find that

$$\omega = \frac{w}{q^{\mu}} \tag{2.10}$$

for real wage ω.

Agricultural workers are supposed to stay put. To see how this changes the model, remember the hypothetical shock in demand of the previous paragraph: it led to a (disproportionally large) change in the number of firms, but there the causality stopped. Now, the changing number of firms will also, because of the variety effect, alter the price index of industrial

goods. This in turn affects real wages and will lead to migration. And the channel may not end there: we will have to trace the effects of this migration, to find out where, or whether, the process ends.

Neary (2001, p. 542) finds that there are three effects of a change in the number of firms in a region. The first is the competition effect: when there are more firms, each gets a smaller piece of total revenue. This effect does not depend on labor mobility and was already present in the model above. It is a stabilizing force, in the sense that it limits the number of firms that can profitably enter after a shock in demand.

The second and third effect work through the mobile labor force. As the number of firms in a region increases, the increased scarcity of labor will drive up real wages, inducing migration into the region. This has two effects. Firstly, the new workers will demand manufactures from local producers. This *demand linkage* leads to an increase in profitability. Neary (2001) shows that, assuming wages return to their prior levels, the balance of the first and the second effect depends on the relative sizes of μ, the share of manufactures in demand, and Z, the index of transport costs defined in appendix 2.B. If μ is larger, the (destabilizing) demand linkage dominates, while a larger Z implies that the (stabilizing) competition effect is stronger.[14]

However, we must also take into account that the decline in price index q caused by the increasing number of firms, leads to a lower cost of living in the region where the demand shock took place. Since real wage ω must be equal in both regions, the assumption that wages return to their previous level must be false. Nominal wages can fall, leading to a further increase in profitability. This is the third effect.

The balance of (stabilizing) effect 1 versus (destabilizing) effects 2 and 3 determines whether a symmetric equilibrium, in which both regions have the same number of firms, can be stable. If a small demand shock in one region leads to a cumulative process of migration and firm entrance, the equilibrium is unstable; if instead it fails to lead to a cumulative process the equilibrium is stable. Using the properties of the model, we can derive a condition on the parameters that tells us whether the symmetric equilibrium is stable or not. The level of transport costs at which stability changes is called the *break point* τ^B.

Similarly, we can ask whether complete agglomeration is stable. That is, when all manufacturing is concentrated in one region, and there is a demand shock in the 'empty' region, does a cumulative process ensue which leads to a symmetric equilibrium? If not, the agglomeration is stable and

[14]This result is appealing: a large Z corresponds to low costs of trade and diminishes the market power that producers exercise over local demand. With low costs of transport these consumers can easily substitute imported goods. A large value of μ indicates that consumers spend a big share of their income on manufactured goods, making their arrival more interesting to producers, but only to the extent that buying local goods is attractive.

the model returns to its previous state after the shock. There exist values of τ for which this is indeed the case, so that the model can explain endogenous agglomeration, as promised. However, for transport costs that are too high the equilibrium is not stable. Hence, a level of transport costs may be derived at which the stability of the concentrated equilibrium changes. This level is called the *sustain point* τ^S.

Both points are derived in appendix 2.C. It turns out that the sustain point and the break point are generally not the same, and that $\tau^B > \tau^S$. This means that there exist transport costs τ^+, with $\tau^B > \tau^+ > \tau^S$, where both the agglomerated and the symmetric equilibrium are stable. Which equilibrium actually occurs depends on the initial conditions: if the model starts off close to symmetric, the symmetric equilibrium will be attained. If the model starts out with all industry concentrated in one region, it stays that way. The CP model with transport costs τ^+ has a path-dependent solution.

The intermediate goods-based CP model

Venables (1996a), in a model where labor is not mobile, shows that it is possible that input-output linkages between firms fulfill the same role as a mobile workforce. Using a monopolistic competition setup for both an upstream and a downstream sector, Venables shows that it is possible that an increase in the size of one industry brings the other industry to a higher level of efficiency. The model's conclusions remain the same in Krugman and Venables (1995), who extend the framework by collapsing the upstream and downstream industries into one layer. The monopolistic competitive market structure is preserved by a specific form of the final demand function. Amiti (1997) shows that a similar outcome may be obtained without the use of an MC framework. In her model, a scale effect arises because of a pricing game that is played between firms in a sector. An increase in the number of firms has a negative effect on collusion and ups the sector's efficiency.

Later in this book, we will make good use of the model where intermediate goods transmit the complementarity between firms. A detailed introduction to this model can be found in chapters 3 and 4.

Intertemporal linkages and the CP model

Aspects of factor accumulation can also serve as a medium for agglomerative tendencies. Baldwin and Martin (2003) survey the interdependencies between agglomeration and growth; they divide the subject into two classes: in the first class, growth influences agglomeration but there is no causality going the other way. We will discuss this class in the current paragraph, as it illustrates how the accumulation of capital can lead to agglom-

eration. The other class, in which technological spillovers are only local, will be discussed after our introduction to growth theory below. In models of the second class, agglomeration can affect the rate of growth, and vice versa. They are the subject of section 2.5 on dynamic economic geography.

Assume that there exist knowledge spillovers, and that they are global. That is, the cost of capital investment declines as a function of the world stock of capital. In that case,

$$\dot{K}_i = \gamma \cdot L_i^I \cdot K_{\text{world}} \tag{2.11}$$

where the growth of the capital (or 'knowledge') stock of region i depends on the number of people working in the innovation sector, L^I and the world stock of capital K_{world}.

Whether capital accumulation of this kind can lead to full agglomeration depends on the mobility of capital. If we assume that inhabitants of one region can own and operate capital in another region, while spending the proceeds at home, capital is mobile. If instead we assume that most capital takes the form of human knowledge, which cannot be separated from its owner, capital is immobile.

Baldwin and Martin (2003) show that with perfect mobility, the initial distribution of firms and capital between regions is stable. Both regions save and accumulate capital, deploying it where it is most productive. With zero capital mobility, however, agglomeration in one region can occur. This happens when trade costs are sufficiently low.

The reasoning behind agglomeration is the following: agents can only invest in capital that is used in their own region. The incentive to invest depends on the profitability of operating a firm; the firm's profitability in turn depends on the demand for its products. Now if trade costs are high, local demand can be enough to sustain firms in either region. But with low trade costs, it is possible that one region enters into a downward spiral: if the number of firms declines, the income from capital declines (all capital is owned locally due to the immobility) which drives down local demand. Meanwhile, imports from the other region substitute for products that are no longer available locally. This further decreases the incentive to invest in local capital. Ultimately then, all investments are made in the other region.

2.3.5 Conclusion

Even though the mechanics, as well as the economic rationale of the above models are substantially different, there are some common characteristics that are worth spelling out. The most important outcome is that in all three models, the combination increasing returns - transport costs can in principle lead to agglomeration. Both are a necessary factor. If there are no increasing returns, firms may as well split up and be spread out over space

without any loss in efficiency. If transport costs are zero, then the whole concept of location does not matter in economic decisions (this is the spatial impossibility theorem referred to in Section 2.1).

The relation between transport costs and agglomeration tendencies is often found to be an 'inverted U' (Junius 1996, Ottaviano and Puga 1997, Venables 1996a). At very high transport costs each region is self-sufficient and no interaction takes place. At intermediate transport costs the above agglomeration effects are stronger, and at very low transport costs the 'centrifugal' forces congestion and factor market competition take over and firms spread out again.

A natural question that arises when these theoretical models are presented is whether there is any empirical relevance to them. We will discuss the empirical literature that complements this theory in chapter 5, which deals with estimation. The survey is in section 5.2.

2.4 Endogenous growth theory

In the introduction, we spoke briefly about the inability of traditional theories of growth to explain lasting growth as an economic phenomenon. I will now substantiate these claims and introduce several alternatives that fall under the header of 'new' endogenous growth theory.

The MC framework that was introduced above does not play a pivotal role throughout endogenous growth theory. The new growth models were erected for a number of reasons, summarized by Romer (1994). Besides dissatisfaction with the inability of classical models to explain lasting growth, Romer identifies two other causes. One is the so-called convergence controversy: the (perceived) neoclassical prediction that poor countries must catch up with rich countries was disputed by data that became available around that time (Maddison 1982, Summers and Heston 1988). The other cause is the fact that the neoclassical model is at odds with a number of easily observable facts, facts which can only be explained if imperfect competition is incorporated.[15] As the MC framework was the first to allow imperfect competition to be modelled in a concise way, it has been the framework of choice for a lot of endogenous growth models.

There are basically two 'waves' of models within the theory; the first wave (started by Romer 1986) describes growth as a process of ceaseless

[15]The facts are: 1. There are many firms in the economy, not one monopolist. 2. Discoveries are nonrival. This makes them different from other inputs. 3. Physical activities can be replicated; therefore production functions should be homogeneous of degree one. 4. Technological advance comes from things that people do. It does not occur by itself. 5. Many individuals have market power and earn monopoly rent on discoveries even though they are nonrival: informaton can be excludable.

Classical growth models are at odds with facts 4 and 5. Not all endogenous growth theories accomodate all these facts.

accumulation of factors. It is possible to retain the assumption of perfect competition in these models, using externalities. We will look at a sample model that employs the MC framework, though. The second wave (started by Romer 1990) explains growth by organized technological progress, and uses the MC framework together with an explicit sector for R&D.

We first briefly look at the exogenous (Solow-) growth model and compare it with some first-wave endogenous growth models. We then look at the second-wave models in Section 2.4.2.

2.4.1 Neoclassical and endogenous models of accumulation

The macro level

The neoclassical growth model was developed independently in Solow (1956) and Swan (1956), and the setup can be summarized quite concisely. The economy of a country uses two factors, L and K, and produces a single output. A proportion $(1-s)$ is consumed[16], the rest is used to increase K:

$$Y_t = F(K_t, L_t) \tag{2.12}$$
$$\dot{K}_t = sY_t \tag{2.13}$$

The aggregate production function F exhibits constant returns to scale, and the population of laborers L grows exponentially at rate n. To each factor L and K taken alone, the function has decreasing returns to scale. We may thus assume that the aggregate production function is a representation of an indeterminate number of firms that are in full competition.

The qualitative results of the model of course depend on the shape of F. Solow considers quite a number of different possibilities, but the one best remembered and usually quoted is when F has the Inada properties ($F_x \to \infty, 0$ as $x \to 0, \infty$ and $F(0,c) = F(c,0) = 0$). Because of the CRS assumption, we may write this model in *per capita* terms by dividing both sides of (2.12) by L and substituting (2.13) in. This leads to the differential equation

$$\dot{k} = sf(k) - nk$$

where lowercase variables are per capita, and $f(k) = F(K/L, 1)$. By the Inada assumption, f exhibits decreasing returns to scale, so that the equation has a single solution k^* to which all time-paths must converge. This implies that there exists a level K/L at which the extra capital only just compensates the increase in population. This is the steady state to which the economy converges, and in which the growth in production per capita

[16]The assumption of a fixed rate of saving can be relaxed without altering the basic results of the model. A model of intertemporal optimization was built by Cass (1965) and Koopmans (1965); the result may also be found in Barro and Sala-i-Martin (1995) and Rensman (1996).

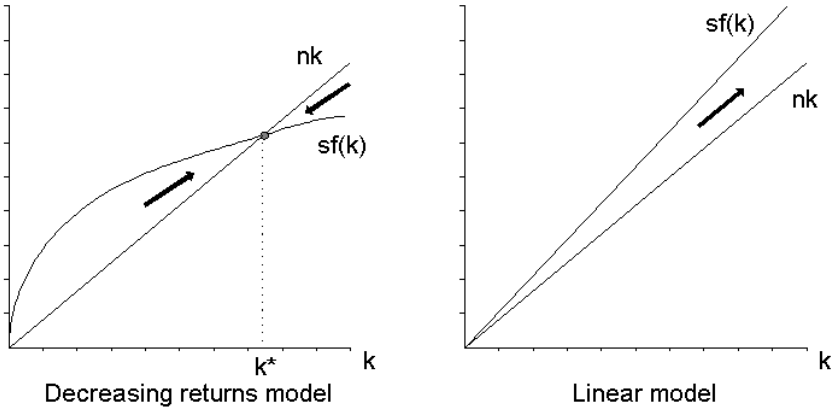

Figure 2.3: Direction of motion of k in two models of growth

stops. The model is depicted in the left-hand panel of Figure 2.3. Capital per worker converges to the steady state level k^* from every initial level k_0.

To stay in line with the empirical fact that the economy keeps growing, the neoclassical model is usually amended with exogenous technological growth. This growth is necessarily Harrod-neutral (for a proof, see Barro and Sala-i-Martin 1995, p. 54) and can be incorporated by substituting $\hat{L}_t$ for L_t in (2.12), with $\hat{L}_t = A_t L_t$. Regular increases in A then result in a growing income per capita, even if the economy is in the steady state. If the rate of growth of A is assumed constant it is possible to estimate values for it for different countries using time series data. In another paper, I estimated exogenous growth for the U.S. to be 0.0180 $[.0009]$ and for the Netherlands 0.0149 $[.0021]$ (standard errors in brackets, Knaap 1997).

The neoclassical model highlights the process of capital accumulation in a closed economy and does not consider the interactions between several economies. It does make a prediction about the dispersion of capital per head over several closed economies, if these economies can all be described by the same production and investment functions: regardless of the initial level of capital, the economies will converge to the same equilibrium, and thus to the same level of K/L. This property of the model is known as the convergence property.

The temporary nature of growth in this model has to do with the fact that the factors that can be accumulated together face decreasing returns to scale. The more of these accumulable factors are around, the less their added productivity is. This is an assumption of the model, and not necessarily a fact of life. The assumption was made because the neoclassical model also considers the factor labor, which cannot be accumulated by sheer economic means, and together the factors must exhibit CRS. For, if they do not exhibit CRS, the assumption of perfect competition is inappro-

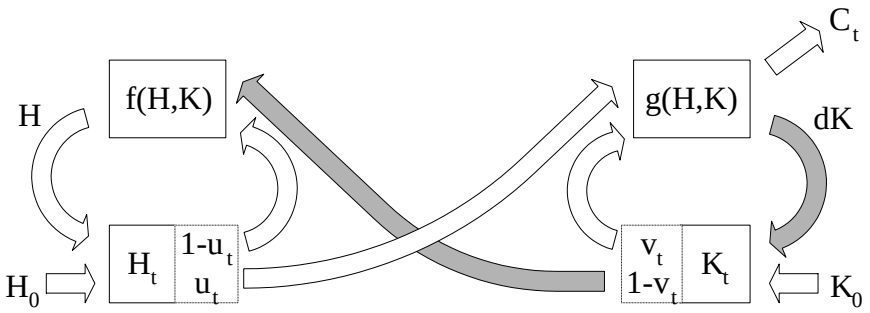

Figure 2.4: A box-arrow sketch of the two-sector model

priate.

On the premise that we will discuss the appropriate market structure below, let us now explore what would happen on a macro-level if *all* factors of production could be accumulated. This implies a return to the models proposed by Harrod (1939) and Domar (1946), who supposed that every addition to the stock of capital per worker allows production to be increased proportionally. Then the per capita stock of capital can never be too high, in the sense that additions to it are relatively unproductive. This can be seen when we substitute $F(K_t, L_t) = AK_t$ in formula (2.12) above. The accumulable resources in this case must be understood to include human capital and other production factors as well, besides capital in the narrow sense.

A graphical analysis of this linear model of production is in the right-hand panel of Figure 2.3. It is clear that if all factors can be accumulated, while the CRS condition still holds, we have specified a model of endogenous, ever-lasting growth.

An important point made by Rebelo (1991, p. 502) is that to achieve this result, not every part of the economy needs to have constant returns. It is sufficient that there exist a sector that uses a core of accumulable factors with a constant returns technology. This sector then becomes the economy's "engine of growth" as it pulls the rest of the economy.

We will illustrate this and other issues by considering the following two-sector model of an economy, taken from Barro and Sala-i-Martin (1995, p. 198):

$$C_t + \dot{K}_t + \delta_K K_t = A\left(v_t K_t\right)^{\alpha_1}\left(u_t H_t\right)^{\alpha_2} \tag{2.14}$$

$$\dot{H}_t + \delta_H H_t = B\left(\left(1-v_t\right)K_t\right)^{\eta_1}\left(\left(1-u_t\right)H_t\right)^{\eta_2} \tag{2.15}$$

A box-arrow sketch of this model is in Figure 2.4. The different colors of the arrows are used later; for now consider them all equal.

We see that there are two sectors, one with production function f (formula 2.15) and one with production function g (formula 2.14). Both sectors

use two factors, K and H. In principle, both factors K and H can be accumulated. The variables v, u and C are control variables. The sectors differ in parameters α_i, η_i, A, and B and in the fact that the sector that produces K also produces consumption, C. Consumers solve the dynamic problem

$$\max_{u_t, v_t} \int_0^\infty \frac{C_t^{1-\sigma}}{1-\sigma} e^{-\rho t} \mathrm{d}t$$

given H_0 and K_0 and the parameters.

The complete model (2.14)-(2.15) is analyzed by Mulligan and Sala-i-Martin (1993). They derive the conditions under which this system can generate a steady state growth path, that is, a solution path where all variables grow at a constant rate. It turns out that this is only possible under the following condition:

$$(1-\alpha_1)(1-\eta_2) = \alpha_2 \eta_1 \tag{2.16}$$

A model whose parameters do not obey this condition either comes to rest at equilibrium levels of H and K or 'explodes', which means that it generates infinitely large state variables in finite time, and the objective integral becomes improper. This knife-edge condition on the parameters bothered Solow (1994) who discusses the value of α_1 in the AK model (see below). If that parameter is only slightly different than assumed, condition (2.16) is not satisfied and the endogenous growth results vanish. It causes him to call this type of theory "unpromising on theoretical grounds" (p. 51).

The model (2.14)-(2.15) has a number of well known special cases. We briefly list them below.

Example 2.4.1 *The AK model. For this model, the sector on the left in Figure 2.4 is taken out. The other sector is assumed to have constant returns: $\alpha_2 = 0$, $\alpha_1 = v_t = 1$. Notice condition (2.16) is satisfied. This is a limiting case of the neoclassical Solow-Cass-Koopmans model with $f(K) = AK$, hence the name. The steady state solution is $\dot{C}/C = \dot{K}/K = (A - \delta_K - \rho)/\sigma$. The model does not have any transitional dynamics. The growth rate of C always remains positive under suitable parameters.*

Example 2.4.2 *The engine of growth. The two grey arrows in figure 2.4 are taken out. Both sectors have constant returns: $\eta_1 = 0$, $\eta_2 = 1$, $\alpha_1 = 1 - \alpha_2$, $\delta_K = \dot{K}_t = 0$. Here, K represents the invariant stock of non-reproducible, non-depreciating capital goods (think of land, for instance) and H is the stock of factors that can be accumulated. Again, the model only has a steady state solution and lacks transitional dynamics. Rebelo (1991) shows that the solution is $\dot{C}/C = \alpha_1 \dot{H}/H$ which is equal to $\alpha_1 (B - \delta_H - \rho) / (1 - \alpha_1 (1 - \sigma))$. It is natural to designate the sector producing H as the engine of growth, as it is the constant returns accumulation of H that causes C to grow.*

Example 2.4.3 *The Lucas model. This is a slightly more general version of the 'engine' model from Example 2.4.2, analyzed in Lucas (1988). This time we take out only the middle grey arrow. The parameters are* $\eta_1 = 0$, $\eta_2 = 1$, $\delta_{H,K} = 0$, $\alpha_1 + \alpha_2 > 1$. *H is understood to be human capital and K is conventional capital. Thus capital goods play no role in the (constant returns) creation of human capital. The goods sector shows increasing returns. In fact, Lucas assumes constant returns plus an external effect of the average stock of human capital, so that a competitive equilibrium exists (more on this below). The optimal steady state growth rate of consumption (with zero population growth) is* $\dot{C}/C = \dot{K}/K = (\frac{1-\alpha_1+\gamma}{1-\alpha_1}B - \rho)/\sigma$. *Here,* $\gamma = \alpha_1 + \alpha_2 - 1$, *the size of the external effect. This shows that increasing returns are not essential for the resulting endogenous growth, as* $\gamma = 0$ *still permits a positive value for* γ_C.

These models can be classified as to their stability. Because there are no transitional dynamics in the first two models, a small perturbation of the initial value has lasting effects. Because the growth rate of the accumulable factor is constant, the difference between the solutions starting in F and $F+\varepsilon$ grows exponentially (F is the initial value of the relevant state variable, K and H for the AK and the 'engine' model, respectively). A similar result holds for the Lucas model, although derivation of this result is not trivial. See Barro and Sala-i-Martin (1995, p. 184) and Mulligan and Sala-i-Martin (1993, p. 758).

The micro level

The models presented above pose a difficulty additional to the knife-edge condition on the parameters. If they include increasing returns to an accumulable factor, the usual fully competitive environment is no longer feasible; in other words, the set of supporting prices does not exist. We look at two approaches that have been used to circumvent this problem. One is to introduce increasing returns only at the level of the sector, and not of the firm. The sectorial returns take the shape of externalities. The other approach is to explicitly model the imperfect competition that arises because of the increasing returns.

Externalities We discussed externalities in Section 2.2.1 as a means to reconcile CRS and increasing returns. Some endogenous growth models use non-pecuniary externalities to do just this. We have already mentioned the use of externalities in the Lucas (1988) model, and we now look at the approach in Romer (1986). Because of a careful specification of the externality setup, the model does not suffer from the knife-edge condition (2.16).

The production function for a representative firm is $F(k_i, K, \mathbf{x}_i)$ with k_i the state of knowledge available to firm i and $\mathbf{x}_i$ a vector of additional factors (capital, labor). The variable K is the aggregate level of knowledge

$\sum_{i=1}^{N} k_i$ which can be used by all firms to some extent because knowledge is partly non-rival and non-excludable. It is assumed that F has constant returns to the factors k_i and $\mathbf{x}_i$, and increasing returns to all three factors. However, each firm takes the value of K as given when making its decisions. Output can be consumed or invested in k_i ($\mathbf{x}_i$ is constant). The latter goes through the knowledge production function: $\dot{k}/k = g(I/k)$. The function g is increasing and bounded from above by a finite constant M. These conditions on g prevent the 'explosion' that the models above suffered from: a firm can never let its stock of knowledge grow at a faster rate than M so that k_i and K cannot reach infinity in finite time. Note that the g-functions above were usually linear in the state variable.

Romer finds that the socially optimal solution is different from the competitive solution because the latter does not take the external effects into account. Both solutions do generate endogenous growth, albeit that the rate of growth is larger in the optimal solution. The competitive solution is properly defined in all models that satisfy the above specification.

Monopolistic Competition As an alternative to the use of externalities above, Romer (1987) explicitly introduces markets that are monopolistically competitive; the model is very similar to that in Section 2.2.3. There exists an all-purpose capital good Z, which is transformed into a continuum of n^* intermediate goods; this is done by a continuum of firms (see appendix 2.A). These intermediate goods are then used as inputs for the final good. The final good can again be added to Z or can be consumed. Consumers maximize utility (a function of consumption) intertemporally. The production function in the final goods sector is as described in section 2.2.2. An increasing number of intermediate inputs (n^*) increases output as in the example in Section 2.2.3. Varieties $x(i)$ are produced using an increasing returns production function.

The most important characteristic of this model is that output Y turns out to be a linear function of the stock Z. This is because the efficient scale of the intermediate producers does not change as Z changes, so n^* is linear in Z. As Y is linear in n^* this means that the model behaves much as though it were the AK model above, and generates stable endogenous growth. It also suffers from the above-mentioned drawbacks, notably the fact that it is parameter-unstable. However, constant returns of Y to Z seem a little less "luck" (cf. Solow 1994, p. 51) than above, as they can be defended on economic grounds rather than just being mathematically convenient. Also, this model became the backbone of more advanced growth models. We will come across those models in the next section.

2.4.2 Growth through innovation

Above, economic growth was mostly brought about by an ever increasing supply of factors. In Romer's (1987) model, an increase in the number of varieties played a role, but this increase was 'free,' *i.e.* no sacrifices needed to be made to discover the new varieties; the increase was a matter of efficient scale. Yet stylized fact #4 (footnote 15) specified that 'technological advance comes from things that people do.' The second wave of growth models thus concentrated on a situation where R&D absorbs resources and new varieties are discovered in return.

New varieties can be substitutes or complements to older ones. In growth theory parlance one thus distinguishes *horizontal* and *vertical* innovation. The term 'horizontal innovation' is from Grossman and Helpman (1991), and the first model in this direction was drafted by Judd (1985). It is replicated here.

It is assumed that consumers maximize an intertemporal CES utility function

$$U = \int_0^\infty e^{-\beta t} \left(\int_0^{V(t)} x\left(v,t\right)^\theta \mathrm{d}v \right) \mathrm{d}t. \tag{2.17}$$

The only factors of production are labor, which is constant at L, and the known range of varieties $V\left(t\right)$. For each variety there holds that one unit can be produced using one unit of labor. The range of varieties $V(t)$ grows through R&D, whose only input also is labor. It is assumed that $\dot{V} = L_{\text{R\&D}}/k$.[17]

There holds that $\theta < 1$, so that in equilibrium the quantities $x\left(v\right)$ are the same across varieties. Call this quantity y. The problem may then be written as

$$\begin{aligned} &\max_{0 \leq y \leq LV^{-1}} \int_0^\infty e^{-\beta t} y^\theta V \mathrm{d}t \\ &\text{subject to } k\dot{V} = L - yV. \end{aligned}$$

The solution (see Judd 1985) is that the economy converges to a stationary state where both y and V are constant. That is, there exists an optimal variety of goods, and once this variety has been attained innovation comes to a halt.

It is possible to see why innovation stops if we compare the problem to the basic monopolistic competition model of section 2.2.2. In that setup, an increase in that number of firms lowers each firm's profit margin. Profit is used to repay a fixed cost that is associated with entry. A situation of too many producers leads to profits that are too low to recoup the initial investment. Hence there exists an optimum number of producers. In this

[17] Note that there is no uncertainty involved in research. This rather quaint assumption is maintained through much of the growth-through-innovation literature.

model, the fixed cost associated with entry is the labor that must be hired to conduct R&D. If that cost cannot be repaid because profit margins are too low, innovation stops.

Note that the MC market form is essential in this model because, as opposed it full competition, it allows producers to make a profit. Those profits can be used to pay off the initial R&D expenses. Without the possibility to price higher than marginal costs, innovation would never occur.

Horizontal innovation, endogenous growth

One way to keep the economy growing in the model above, is by lowering the costs of innovation as the number of varieties increases. If the outcome of the model should be a constant growth rate g of the number of varieties, and we know that $\dot{V} = L_{\text{R\&D}}/k$, then we can deduce

$$\begin{aligned} \frac{\dot{V}}{V} &= g = \frac{L_{\text{R\&D}}}{kV} \\ g \text{ constant} &\Rightarrow kV \text{ constant} \end{aligned}$$

So, if the R&D productivity parameter k^{-1} is proportional to the number of varieties V, we can have everlasting growth.

Romer (1990) presents an adapted version of his model in Romer (1991) that "emphasizes the importance of human capital in the research process" (p. S78). Like above, it features three sectors: R&D, intermediate and final goods. Knowledge has a rival component H and a non-rival component A; the latter can be interpreted as the 'state of technology' and is allowed to grow without bounds.

In the production of final output Y, human capital H plays a role next to labor L and a continuum of intermediate goods $x(i)$:

$$Y(H_Y, L, x) = H_Y^{\alpha} \cdot L^{\beta} \cdot \int_0^A x(i)^{1-\alpha-\beta} \mathrm{d}i \tag{2.18}$$

(notice the similarity to formula 2.17 above). The stock of H is split up in a part H_Y that works in the final goods sector, and a part H_A that works in the R&D sector. The interval over which $x(i)$ is positive has size A, the level of technology. An increase in A, that is, a rise in the level of technology, does not render older types of the intermediate good obsolete. This is due to the additively separable form of (2.18).

In line with the derivation above, the technology used in the R&D sector is such that A changes according to

$$\dot{A} = \delta \cdot H_A \cdot A \tag{2.19}$$

This form is justified by claiming that a larger stock of knowledge will enhance current research possibilities. The model is closed by specifying

that the stock of intermediate goods $K = \int_0^A x(i)\mathrm{d}i$ evolves according to $\dot{K}_t = Y_t - C_t$.

Romer's analysis shows that the model, specified above, yields unbounded endogenous growth. This is caused by the assumption of constant returns to scale in equation (2.19) above. With respect to this assumption, Romer writes:

> "...in this sense, unbounded growth is more like an assumption than a result of the model. [...] Whether opportunities in research are actually petering out, or will eventually do so, is an empirical question that this kind of theory cannot resolve."

Vertical innovation, endogenous growth

Aghion and Howitt (1992) consider a model of growth that features vertical innovation. Newer types of intermediates replace the older types, and therefore the model represents the concept of Creative Destruction introduced by Schumpeter (1942).

The economy consists of three sectors: the R&D sector, the intermediate goods sector and the sector that produces consumption goods. The tradeoff in the economy is the decision how many workers are allotted to work in R&D instead of the intermediate goods sector. This number depends on the expected profitability of innovations.

A new intermediate good completely replaces the older type. The inventor is the only producer of the goods, and is thus allowed to earn some monopoly rents until the next innovation takes place. The time until the next innovation is random and exponentially distributed, and depends negatively on the number of people working in R&D. The marginal product of an extra R&D worker is decreasing, so that there exists an optimal number of people engaged in research and development.

There is only one kind of uncertainty in the model, namely the time of arrival of a new technology. The increase in the level of technology, caused by the invention, is fixed. By defining a 'period' as the elapsed time between two innovations, the authors in effect make the monopoly rent earned off the inventions the random variable in the model.

Without being explicit about such things as the aggregate production function, Aghion and Howitt (1992) examine the motives for investing in R&D and find that, depending on the 'arrival function' of new technologies, there may exist a fluctuating or steady (possibly zero) number of researchers in the economy. Endogenous growth is implied as soon as there is a positive number of researchers active, and its rate is determined by both endogenous and exogenous variables.

2.4.3 Empirical tests

In this section we will mostly look at empirical tests of the implications of the above models. As some of the results came out negatively, interest in the neoclassical Solow model was revived in the early 1990s. The results of such interest can be found in Mankiw et al. (1992) and Nonneman and Vanhoudt (1996).

As Pack (1994) notices, much early empirical research on endogenous growth models is conducted in the neoclassical framework. Thus, instead of testing the new growth theory directly, it is only used as a possible alternative when the Solow model fails. A first direct test of the theory is performed by Jones (1995b), who tests the time-series predictions of new growth theory. The evidence is collected in two rounds.

The models of this section have the property that a permanent increase in investment causes a permanent increase in the economy's rate of growth. Or, even stronger, the two variables are linearly related. This is easily seen with the AK and the Lucas model, as the rate of growth of capital and the rate of growth of consumption are the same. The result does not hold for the engine-of-growth model. Jones (1995b, p. 500) shows that the growth rates of selected OECD countries are stationary variables, whereas a unit root in the OECD investment rates can only be rejected in four out of 15 cases. Almost all investment rates show a positive trend. This contradicts the (supposed) linear relationship between investment rates and growth rates. Further time series estimations show that the effects of an increase in the investment rate can only be observed for eight years after the shock, much less than the proposed everlasting effect.

The testable proposition of the R&D-based models of section 2.4.2 is that the growth rate of an economy is linearly related to the number of people active in the R&D sector. Using data on the number of researchers in the U.S., Germany, Japan and France, Jones (1995b, p. 517) again shows a strong upward trend in these explanatory variables, whereas the rate of growth of their respective countries remains stationary. These two results can be seen as a rejection of the testable propositions that came out of the endogenous growth models. Jones (1995a) proposes to 'fix' the R&D-based model by writing equation (2.19) in Romer's (1990) model as

$$\dot{A} = \delta \cdot H_A^{\lambda} \cdot A^{\phi}$$

If λ, $\phi < 1$ then the model will no longer exhibit endogenous growth but instead settle down in an equilibrium. As Jones (1995a, p. 766) puts it, "$\ldots \phi = 1$ represents a completely arbitrary degree of increasing returns and [...] is inconsistent with a broad range of time series data on R&D and TFP growth " (see also Romer's quote in section 2.4.2). The model proposed by Jones (1995a) can best be seen as an extended version of the Solow (1956) setup, with all its asymptotic characteristics.

2.4.4 Review

We have seen that classical growth models that use the CRS paradigm explain growth through accumulation, but this growth cannot last forever without exogenous propelling. Accumulation-based models can explain lasting growth if they have constant returns to all accumulable factors. The micro-foundations for these models use externalities or an MC-setup.

The second wave of endogenous growth models explains growth not by accumulation of factors, but by technological progress. Virtually all these models use the MC framework.

Some critical notes can be placed about endogenous growth models. The scale-effects that they predict are not observed, and they are parameter-unstable. Despite the critical notes above, at the time of writing, endogenous growth theory is still very much alive. It turns out that the spirit of the models can be maintained while accommodating empirical facts (see Aghion and Howitt 1998, chapter 12). And the ability of the models to handle a number of questions that exogenous growth theory cannot answer (questions concerning the long run growth rate, for instance) has made them popular with empirical researchers.

2.5 Dynamic economic geography

This review chapter has shown how the monopolistic competition model (section 2.2) made possible new ways of modelling economic geography (section 2.3) and economic growth (section 2.4). So far however, we have only discussed growth models without an explicit geographical dimension, and mostly static geography models.[18] It is only natural to combine the two strands of the literature, which are based on the same framework.

Recently, a large amount of research was done in this direction. A good survey of these efforts is in Baldwin and Martin (2003). They discern two classes of models, one of which we have already discussed above: models in which growth affects geography, but not the other way around. The other class of models allows for an interplay between growth and agglomeration; we survey it in this section.

The key assumption that determines whether growth and geography will interact concerns the (spatial) range of knowledge spillovers. As we have seen above, spillovers are crucial to endogenous growth theory. They imply that the stock of productive knowledge that is already available in the economy helps to reduce the costs of acquiring extra knowledge. An assumption about spillovers is implicit in capital-accumulation equations

[18] The exception being the CP-model in which capital accumulation is the medium for complementarities on page 29.

such as (2.19) above. There, the size of the available non-rival knowledge stock A determines the effort that is needed to increase it.

It is intuitive that whatever we assume about the geographical reach of this effect will determine the growth performances of different regions. If local knowledge only spills over to R&D efforts in the same region, we can imagine one region growing, while the other stagnates. When spillovers are worldwide, as we assumed above in formula (2.11), different regions can pool their knowledge.

In the next paragraph, we discuss the results when spillovers are local. We further survey a number of other approaches with different agglomeration links. Paragraph 2.5.2 discusses a model by Martin and Ottaviano (1996b) where the linkage runs through the R&D sector, in a way reminiscent of Krugman and Venables (1995). Paragraph 2.5.3 shows that the linkage can go through the (research-) labor market as well. We end with some other models in paragraph 2.5.4.

2.5.1 Agglomeration through local knowledge spillovers

Baldwin and Martin (2003) replace the capital accumulation equation in formula (2.11) with a slightly different version that involves transport costs for knowledge.

$$\dot{K}_i = \gamma \cdot L_i^I \cdot (K_i + \lambda \cdot (K_{\text{world}} - K_i)) \tag{2.20}$$

The parameter that measures knowledge transport costs, λ, lies between zero (only local spillovers) and one (global spillovers). Its value determines the rate of accumulation of region i's capital, $\dot{K}_i$. For now, we assume that $\lambda = 0$.

As before, we discern two cases with respect to capital mobility. Perfect mobility means that agents can own capital in every region; this equalizes the rate of return to capital in all regions. Capital immobility means that agents can only accumulate capital in their own region. This assumption applies to human capital, for instance, if people are not mobile.

Perfect capital mobility

When there are no restrictions on owning foreign capital, everybody will want to produce capital where the costs of production are lowest. According to (2.20), the region with the highest initial capital stock is where all capital production will take place. The lagging region does not have an innovation sector of its own, but is perfectly able to accumulate capital using the other region's knowledge pool. This means that there is no self-enforcing agglomeration in this model, as an initial disadvantage does not translate into lower income for the lagging region.

There exists an interesting trade-off in this model. Assume for a moment that region N has a larger share of total income, due to a higher initial share of world capital. The share of firms that locate in N will be higher than N's share of income, due to the home market effect (see appendix 2.B). This concentration of firms leads to a higher rate of growth, because of the larger local spillovers that now occur in region N.

How does this impact the people in the S-region? On the one hand, a larger share of income going to their neighbors in N is bad, and more firms locating in N means higher transport costs for the people in S, who will have to import more. But on the other hand a higher rate of growth benefits inhabitants of all regions. The balance of these two effects is determined by the costs of transport for knowledge and goods. Baldwin and Martin show that for small λ (when spillovers are mostly local) and small costs of transport for goods, the S region may *gain* from an extra concentration of industry in N. This is a positive and surprising outcome, that contrasts with the earlier result that agglomeration in one region is generally welfare-reducing for the other region.

Immobile capital

When capital is immobile, agents can only invest in their own economy. If the returns to capital in one region drop, this has an immediate effect on the income of the (capital owning) inhabitants of that region, leading to a feedback that may cause the forming of a periphery and an agglomeration. Earlier, on page 29, we stated that a model with global spillovers and immobile capital could generate a core-periphery outcome. It should not be surprising that the same holds for a model with immobile capital and local spillovers.

The extent to which an inequality between the rate of growth in two regions arises, depends on the level of transport costs. Baldwin and Martin (2003) show that there exist a threshold level of transport costs below which a process of agglomeration starts. When transport costs are high enough, local demand makes investment worthwhile in either region and a symmetric equilibrium obtains. However, when transport costs drop below the threshold level, one region completely stops investing while the other experiences a 'growth takeoff' (p. 28). The region that agglomerates sees the costs of investment fall more quickly due to local spillovers, and enters a period of high growth. The other region gets stuck in a situation where local demand is too low to justify investment in new firms.

> Hence, growth affects geography which itself affects growth and agglomeration is driven by the appearance of growth poles and sinks. (Baldwin and Martin 2003, p.28)

The question once again arises whether the region where innovation stops

is worse off; the higher rate of growth is beneficial to both regions, after all. As it turns out, the welfare of the region that was left behind depends on the share of differentiated goods in expenditures. For a high value of this share, the region may actually gain from the new, agglomerated, equilibrium. For low values, it certainly loses.

2.5.2 Agglomeration through the R&D sector

Localized spillovers are not the only way in which growth and geography may interact. The model in this section has a feedback between growth and agglomeration that is the result from vertical linkages between the R&D sector and the differentiated goods sector.

Martin and Ottaviano (1996b) present a model with two regions and three sectors: a full-competition agricultural sector, an MC industrial sector and a sector for R&D. It is the latter sector that is most interesting.

The R&D sector is fully competitive. The output of the sector is patents; each patent can be used to manufacture a variety in the industrial sector, the total number of varieties is n. The productivity of the R&D sector increases as n gets larger. These qualities are similar to the Romer-Grossman-Helpman models of Section 2.4.2. The only input to the R&D sector is the composite good D that is the output of the industrial sector. This creates a linkage between the two sectors akin to the linkages in Krugman and Venables (1995) and Venables (1996a). Wherever firms from the industrial sector are abundant, the costs of R&D are low. And wherever R&D is conducted, the demand for industrial goods is higher. The linkage causes agglomeration of industrial and R&D firms in the same location.

Consumers in this model maximize an intertemporal utility function that depends on the consumption of the agricultural and the industrial good. The model has two types of solutions. In one solution, both locations have exactly the same number of industrial producers. R&D is conducted in both locations. This solution is unstable. The other solution has all R&D taking place in one location, where also the majority of the industrial producers are active.

In the second, unbalanced, solution the rate of growth is higher. This is intuitive: if industrial producers are spread evenly the industrial composite costs are the same in both locations, say, c. In case of an imbalance, there always is a location in which the composite is cheaper than c. Because R&D uses only the composite, an even spread of the industrial producers maximizes production costs and minimizes growth.

An important conclusion of the model by Martin and Ottaviano (1996b) is that the rate of growth influences the location decision, and the location decision influences the rate of growth. This puts models in which both are treated separately at a disadvantage. The fact that the interaction causes agglomeration of industrial activity is in line with the quote by Myrdal on

page 12.

2.5.3 Agglomeration through the labor market

Migration between the two regions is the cornerstone of the model by Baldwin and Forslid (1997), just as it is in Krugman (1991a, 1991b).

The assumptions are roughly the same as above, except that the R&D sector now uses only labor as an input. Again, the input requirement decreases as the stock of knowledge gets larger. However, the spillovers are only regional; the stock of knowledge consists of the number of firms in ones own location only.

In the long run, the linkage now works as follows: wherever the most firms are is where the consumer price index is lowest. Personnel has an incentive to move to this location. So do all firms in the (competitive) R&D sector, because the costs of R&D depend negatively on the available pool of knowledge (in this case, the number of firms). On the other hand, where most people are is where firms like to be because of the demand that people exercise, and because of the larger labor market that the firms can draw from.

Again, there are two types of equilibrium in this model. In one, all activity is evenly divided between locations, and both locations grow at the same speed. The other equilibrium has all R&D and most labor and industrial firms in one location, the other deprived of most activity.

It turns out that the first equilibrium (the even spread) is very unstable, even at prohibitive trade costs. This was not the case in the Krugman (1991a, 1991b) models. Contrary to the static economy, the dynamic economy will agglomerate into one location for all possible parameters.

In this model, the R&D sector does not constitute a part of the linkages, as it did above. However, it does react to the outcome. In the long term, all R&D is concentrated in the agglomeration, because it is the cheaper place to work. This does not necessarily affect the location of the industrial firms developed by the R&D sector, as the patents are valid in both locations. Thus, the R&D sector reacts to the linkages, but is not a part of it.

The interplay between growth and location shows up in this model as well. When all R&D is done in the same location, all R&D firms add to the same stock of knowledge. This leads to faster rates of growth than if the advances are divided over two separate stocks of knowledge, because the efficiency of the R&D sector increases with K.

2.5.4 Other models of growth and geography

We discuss a number of other models where growth and location theories are integrated. The degree of interaction between the two is more limited than in the theories that were discussed above.

Quah (2002) proposes a highly theoretical model where the extent to which spillovers between different regions exist is a function of the distance between those regions. The advantage of his approach is that space is no longer limited to two regions, but can consist of a continuous plane, or a globe. The model shows that if there are adjustment costs for capital and spillovers are local, there exists an interesting transition path to the long run equilibrium, where all regions are equal. During the transition, an agglomeration force creates growth 'peaks' and 'troughs' in the space that is studied. These poles disappear when the long run steady state is attained.

Martin and Ottaviano (1996a) develop a model where migration does not occur. There are three sectors, agriculture, industrial and R&D. The MC industrial sector uses patents as in Section 2.4.2. The competitive R&D sector that develops the patents uses labor and the pool of knowledge. Patents can be used in any location and are not subject to transport costs. If the R&D sector has access to all knowledge in the economy (global spillovers), then R&D is conducted in both locations. If there are only local spillovers, the R&D sector agglomerates. The developed firms will be set up in both locations, though.

The model is a first attempt to merge theories of growth and location. The structure of the economy (industrial and agricultural production at the two locations) is so rigid that it does not change much under the different growth regimes, so that the interaction is limited to the location of the R&D sector.

Englmann and Walz (1995) construct a model with two locations without transport costs. The geographic structure plays a role however, because the knowledge pool is different between the two regions. This leads to a situation with nontraded inputs, where each location has its own intermediates. The initially larger region becomes the industrial center, whereas the other becomes a peripheral region. If there are interregional knowledge spillovers, so that inputs still are not traded but R&D *can* use them, many solutions become possible.

In this model, devoid of transport costs, it is the size of the knowledge pools that steers the regional development. Knowledge pools contain nontraded inputs, so that the factor that causes agglomeration is not traded itself. Though interesting, this is fundamentally different from the models of Section 2.3 and is the subject of another branch of literature (see, for instance Rivera-Batiz and Romer 1991) .

Duranton and Puga (2003) include a section on dynamic externalities that lead to growth as well as agglomeration. The payoff to investment in human capital is thought to be a positive function of the human capital stock in the immediate vicinity. External effects of other people's human capital fuel growth, as they are a particular variety of the endogenous growth models from section 2.4. At the same time, they are an agglomerat-

ing force.

Redding and Schott (2003) connect geography and growth indirectly as they look at the effect of remoteness on the accumulation of human capital. As we have seen in section 2.4, the accumulation of human capital is thought to be a mechanism for economic growth. The authors find that, under plausible assumptions, remoteness depresses the skill premium and reduces incentives to accumulate human capital. Though their model is static and yields no direct results pertaining to the rate of growth, this indirect evidence points to a negative relation between the latter and the geographic position of a country.

2.6 Conclusions

In this survey paper, we introduced the monopolistic competition framework as the foundation of two new strands of literature, on the one hand endogenous growth theory, and on the other hand economic geography. Both theories use the fact that MC allows scale economies to be used in a model of general equilibrium.

In our survey of endogenous growth, we showed that early models were based on the endless accumulation of resources, as are exogenous growth models. Later versions stressed technological progress as the source of growth. Progress can take the form of horizontal innovations and vertical innovations.

In the literature on economic geography, linkages between firms and consumers, and between firms themselves, play an important role. The different models can be classified as to the type of linkage they use. Most models predict a dramatic agglomeration at certain parameter values.

Because both strands of literature rest on the same foundation, and describe related phenomena, it is only logical to incorporate the two. We surveyed several attempts to that end. It turns out that the interplay between growth and location upsets the predictions of either literature by itself. Stable equilibria in static geography models turn out to be unstable in a dynamic context; the rate of growth again is influenced by the location pattern, which depends on initial values.

Studies that investigate the empirical value of both literatures are not overly enthusiastic. Whereas CRS-based theory stands up to the data in a reasonable way, many effects predicted by MC are not measured at all. However, this may be due to a lack of testing methodology capable of dealing with the nonlinear nature of the models. Tests can only be conducted on specific linear predictions of the model. It is unclear to which extent a refutation of such a prediction constitutes a problem for the whole body of theory.

It seems that the combination of endogenous growth theory and eco-

nomic geography is a promising field of research. The scattered results available so far indicate that more work needs to be done before any swaying conclusions can be drawn.

2.A A continuum of goods

The derivation of the equilibrium in the monopolistic competition framework holds in general 'when n is large.' This can be an awkward assumption; do we really need, in economic terms, an endless array of goods to work with this model?

The usual interpretation is that really, all we need is to be able to refine and differentiate goods enough. The range can remain the same, but we ought to be able to divide goods into as many different subtypes as we need. Mathematically, this means that we look at a continuum of goods $x(j)$ defined on a real interval $[0, n]$. In principle, each good $x(j)$ with $j \in [0, n]$ can be identified as a different variety. Quantities of goods, however, are only defined over intervals of j. The quantity $x(3) = 1$ is meaningless, but $x(j) = 1$ for all $j \in [0.1, 0.2]$ is a positive quantity.

How do our maximand U and the budget restriction change when we work with a continuum of goods? They can be derived as limiting cases of their discrete versions.

Suppose we call all the goods $x(j)$ with $0 \leq j < n_1$ good 1, all the goods with $n_1 \leq j < n_2$ good 2, and introduce a set of numbers $\mathcal{S} = \{n_0, n_1, n_2, \ldots, n_Q\}$ like this, with $n_0 = 0$ and $n_Q = n$. If two goods belong to the same interval, they are purchased in the same amount and priced the same.[19] With this set, we are back in the discrete goods setup. There holds

$$\begin{aligned} U &= \left[\sum_{i=1}^{Q} (n_i - n_{i-1})\, x(i)^{\theta}\right]^{\frac{1}{\theta}} \\ E &\geq \sum_{i=1}^{Q} (n_i - n_{i-1})\, x_i p_i. \end{aligned}$$

For any properly defined set $\mathcal{S}$, these formulae can be rewritten as

$$U = \left[\int_0^n x(i)^{\theta} \mathrm{d}i\right]^{\frac{1}{\theta}} \tag{2.21}$$

$$E \geq \int_0^n x(i)\, p(i)\, \mathrm{d}i \tag{2.22}$$

We see that there are two ways in which the number of goods can increase. By picking a larger set $\mathcal{S}$, we refine the definition of the goods, and allow for

[19] That is, we have $x(i)$ and $x(j)$ with $n_{k-1} \leq i < n_k$ and $n_{k-1} \leq j < n_k$, and both are purchased in the amount x_k.

more price and quantity differentiation. By increasing n, the range of goods is increased with the introduction of new varieties that can be purchased *instead of* the older set.

The monopolistic competition setup is usually introduced as in formulas (2.21) and (2.22), without a specific set $\mathcal{S}$ defined. To retrieve the results that hold in the integer case, however, we need to imagine such a set ourselves.

Suppose we want to maximize function U from formula (2.21) under the restriction (2.22). The problem can be written as a Lagrangian,[20]

$$\max_{\{x(i)|i\in[0,n]\}} \left[\int_0^n x(i)^\theta \mathrm{d}i\right] - \lambda\left[\int_0^n x\left(i\right)p\left(i\right)\mathrm{d}i - E\right]$$

The problem is hard to solve when we stick with the integral notation, but we can imagine that the differentiation between goods only goes as far as a set $\mathcal{S}$, which we do not specify. We may then write the maximand as

$$\mathcal{L} = \left[\sum_{i=1}^{Q}\left(n_i - n_{i-1}\right)x(i)^\theta\right] - \lambda\left[\sum_{i=1}^{Q}\left(n_i - n_{i-1}\right)x\left(i\right)p\left(i\right) - E\right].$$

Differentiate with respect to $x(i)$ and set equal to zero to find

$$\begin{aligned}\beta\left(n_i - n_{i-1}\right)x(i)^{\theta-1} - \lambda\left(n_i - n_{i-1}\right)p\left(i\right) &= 0 \Rightarrow \\ \theta x(i)^{\theta-1} &= \lambda p\left(i\right).\end{aligned}$$

Note that we may divide by $\left(n_i - n_{i-1}\right)$ because the requirements for $\mathcal{S}$ have it greater than zero. Because λk does not vary with i, we may write that for all i,

$$x(i)p\left(i\right)^{\frac{1}{\theta-1}} = \text{constant}.$$

If we substitute this into formula (2.22) we get that

$$x\left(i\right) = \frac{Ep\left(i\right)^{-\sigma}}{\int_0^n p(j)^{1-\sigma}\mathrm{d}j}$$

where $\sigma = 1/\left(1-\theta\right) > 1$.

2.B The home market effect

We follow the derivation in Neary (2001) in this section. Assume that the two regions are exactly equal and consider an equal and opposite change in the environment. That is,

$$\begin{array}{llll} p = p^*, & q = q^*, & n = n^*, & E = E^* \\ \hat{p} = -\hat{p^*}, & \hat{q} = -\hat{q^*}, & \hat{n} = -\hat{n^*}, & \hat{E} = -\hat{E^*} \end{array} \tag{2.23}$$

[20] We momentarily omit the exponent $1/\theta$, which does not change the outcome of the maximization.

where the *foreign* region has asterisks on its variables and a 'hatted' variable denotes a rate of change, *i.e.* $\hat{x} = \mathrm{d}x/x$. We will use equations (2.8) and (2.9) to derive the result, starting with the latter which we replicate here for convenience.

$$q^{1-\sigma} = np^{1-\sigma} + n^* \left(\frac{p^*}{\tau}\right)^{1-\sigma} \tag{2.9}$$

We totally differentiate equation (2.9), which gives us

$$(1-\sigma)q^{-\sigma}\mathrm{d}q = p^{1-\sigma}\mathrm{d}n + (1-\sigma)p^{-\sigma}\mathrm{d}p + \left(\frac{p^*}{\tau}\right)^{1-\sigma}\mathrm{d}n^* + n^*(1-\sigma)\left(\frac{p^*}{\tau}\right)^{-\sigma}\frac{\mathrm{d}p^*}{\tau}$$

Using the 'hat'-notation and the equalities from (2.23), we can write this as

$$(1-\sigma)q^{1-\sigma}\hat{q} = (\hat{n} + (1-\sigma)\hat{p})(1-\tau^{\sigma-1})np^{1-\sigma}$$

while we can rewrite (2.9) as

$$q^{1-\sigma} = np^{1-\sigma}(1+\tau^{\sigma-1})$$

which combines into the first result,

$$\hat{q} = Z\left(\frac{1}{1-\sigma}\hat{n} + \hat{p}\right) \tag{R1}$$

where $Z = \frac{1-\tau^{\sigma-1}}{1+\tau^{\sigma-1}}$.

Next, we use equation (2.8) to write the total demand for a firm, which is simply the sum of home and foreign demand:

$$x_i = p_i^{-\sigma}\left(Eq^{\sigma-1} + E^*(q^*)^{\sigma-1}\tau^{\sigma-1}\right) \tag{2.24}$$

which, after total differentiation and use of hats, gives

$$\begin{aligned}
\hat{x_i} &= -\sigma\hat{p_i} + \frac{p_i^{-\sigma}}{x_i}\left(\hat{E} + (\sigma-1)\hat{q}\right)Eq^{\sigma-1} + \\
&\quad \frac{p_i^{-\sigma}}{x_i}\tau^{\sigma-1}\left(\hat{E^*} + (\sigma-1)\hat{q^*}\right)E^*(q^*)^{\sigma-1} \\
&= -\sigma\hat{p_i} + (1-\tau^{\sigma-1})(\hat{E} + (\sigma-1)\hat{q})\frac{p_i^{-\sigma}Eq^{\sigma-1}}{x_i}
\end{aligned}$$

where we invoked the equalities from (2.23) in the second step. We now rewrite formula (2.24) as

$$x_i = p_i^{-\sigma}(1+\tau^{\sigma-1})Eq^{\sigma-1}$$

which, combined with the above, gives the second result

$$\hat{x_i} = -\sigma\hat{p_i} + Z(\hat{E} + (\sigma-1)\hat{q}). \tag{R2}$$

Now suppose that from the symmetric equilibrium, for some reason, the home region faces an increase in demand. The MC model tells us that per-firm output is fixed at its optimal level, as is the price of both types of goods. This means that an increase in demand leads to the entry of new firms into the home market. Because these new firms produce new varieties, this changes the price index of industrial goods which may lead to additional entry or exit of firms. The results derived above allow us to quantify the effect of a demand shock on the number of firms.

In the above notation, the increased demand means that $\hat{E} > 0$. We know prices and per-firm output do not change, which gives $\hat{p_i} = \hat{x_i} = 0$. From (R2) we find that $\hat{E} = (1 - \sigma)\hat{q}$. Using this to substitute $\hat{q}$ from (R1), we find that

$$\hat{E} = \frac{1}{Z}\hat{n}.$$

From the definition of Z above, we know that $0 < Z < 1$ if transport costs are positive. This allows us to interpret the above formula as the *home market effect* (Krugman 1980): the region with a higher demand has a proportionately higher share of manufacturing.

2.C The break and sustain point

To find the break point, the level of transport costs at which the symmetric equilibrium becomes unstable, we start at the same point as the previous paragraph. Suppose there are two regions and the symmetric equilibrium has been attained. That is, equation (2.23) is in force. We again use the notation $\hat{x} = \mathrm{d}x/x$.

We assume that workers are mobile between regions and move to equalize the level of real wage $\omega = wq^{-\mu}$. Firms enter and exit to drive profits to zero, but in this derivation we assume that this proces of adjustment is much slower than that of the (mobile) workers. This assumption allows us to assume that the equalization of real wages holds at all times. We then study the properties of the equilibrium by looking at the direction of change in firm profits, in response to a change in the number of firms. If we had reversed the assumption about the speed of adjustment, we would hold profits at zero and look at the change in real wages. Puga (1999) shows that the results of either assumption are the same.

We gather some relationships between rates of change in the model. From (2.5), the pricing rule, we find that $\hat{p} = \hat{w}$. Condition (2.6) implies that $\hat{L} = \hat{n}$. Finally, the real-wage condition (2.10) renders us $\hat{\omega} = \hat{w} - \mu\hat{q}$.

Total expenditure in a region is the sum of the total wages of the industrial and agricultural workers, or $E = wL + L_A$, where wages for the agricultural sector are normalized at one. The number of industrial workers is denoted by L, and there are L_A agricultural workers. Because changes in

E only arise through changes in the number of (mobile) industrial workers or through changes in their wage rate, we have

$$\begin{aligned} \hat{E} &= \frac{1}{E}\left(L\mathrm{d}w + w\mathrm{d}L\right) \\ &= \mu(\hat{w} + \hat{L}) \end{aligned} \tag{2.25}$$

where we have used the fact that in the symmetric equilibrium, the share of manufacturing wages in total output must be equal to the share of manufacturing in consumption, μ.

We totally differentiate the expression for a firm's profit (2.4) to find

$$\begin{aligned} \mathrm{d}\pi &= p\mathrm{d}x - p\theta\mathrm{d}x \\ &= \frac{p}{\sigma}\mathrm{d}x \end{aligned}$$

where we use the definition of θ defined earlier. From the fact that prices p and the elasticity σ are positive, we deduce that changes in profit $\mathrm{d}\pi$ and changes in production $\mathrm{d}x$ have the same sign. This is intuitive: a firm, at equilibrium, produces and sells just enough to cover its total costs. The price has been set higher than marginal costs in this monopolistic competition framework, so any increase in sales beyond the equilibrium quanitity will make the firm profitable. We will look at the sign of $\mathrm{d}x/\mathrm{d}n$ as an indication of the sign of $\mathrm{d}\pi/\mathrm{d}n$. The latter derivative determines the stability of the symmetric equilibrium: if an increase in the number of firms in a region leads to lower profits, some firms will exit and the equilibrium will be restored. This is the case if $\mathrm{d}\pi/\mathrm{d}n$ is negative. For positive values of the derivative, a small change in the number of firms will make the other firms more profitable, causing more entry and a runaway process of agglomeration.

One point is worth expanding upon before we embark on our calculations. We can check the stability of the equilibrium quite easily in the current setup, as 'a small perturbation in the number of firms' is well defined in this case. As there is only one type of industrial firm, any change in its number must take place along the same dimension. In the next chapter, we will introduce several types of firms, which are active in different sectors. In that model, a change in the number of firms in one sector will affect the profitability of the other firms in that sector, but also the profitability of firms in the other sectors. Also, a perturbation in the number of firms can take place in any of the sectors, or in multiple sectors at once, and to find out about the stability of an equilibrium we will have to check all possible perturbations and their associated changes in profits. We will develop a method of doing this efficiently in section 4.2.

We start with formula (R2) from the previous section. Replacing $\hat{p}$ with $\hat{w}$ and $\hat{L}$ with $\hat{n}$, using (2.25) and the relation $\hat{p} = \hat{w} = \mu\hat{q}$, which follows

from real wage equalization, we have that

$$\begin{aligned}\hat{x} &= -\sigma\hat{w} + Z\left(\mu(\hat{w}+\hat{n}) + (\sigma-1)\frac{\hat{w}}{\mu}\right) \\ &= \left(Z\mu - \sigma + \frac{Z}{\mu}(\sigma-1)\right)\hat{w} + Z\mu\hat{n} \qquad (2.26)\end{aligned}$$

We can eliminate the $\hat{w}$ from this expression by using formula (R1) from the previous section. Once again using $\hat{p} = \hat{w} = \mu\hat{q}$, it can be written as

$$\hat{w} = \frac{-\mu Z}{(\sigma-1)(1-Z\mu)}\hat{n}$$

Using this value for $\hat{w}$ in (2.26), we find

$$\hat{x} = \frac{(2\sigma-1)\mu - [\sigma(1+\mu^2)-1]Z}{(\sigma-1)(1-Z\mu)}Z\hat{n}$$

The sign of $\hat{x}/\hat{n}$ determines the stability of the symmetric equilibrium. The definition of Z from the previous section renders a value of τ that lies on the border between stability and instability. That value is

$$\tau_{\text{break}} = \left[\frac{(\sigma(1+\mu)-1)(1+\mu)}{(\sigma(1-\mu)-1)(1-\mu)}\right]^{\frac{1}{\sigma-1}}.$$

We look next at the sustain point, the value of transport costs at which the asymmetric equilibrium in which all firms have agglomerated in one region, is only just stable. Once again, we use the special form of the equilibrium (total agglomeration) to simplify certain relations in the model, and consider the fate of a breakaway firm. Meanwhile, we retain the assumption that firms enter and exit much slower than workers switch regions. This ensures the equality of real wages in both regions at all times.

When all firms have agglomerated in the *home* region, we know that the expenditures of each region, equal to the total wages paid to its inhabitants, are related by

$$E^* = \frac{1-\mu}{1+\mu}E \qquad (2.27)$$

Also, we can simplify the relationship between the two price indices to

$$q^* = q/\tau. \qquad (2.28)$$

This is a simplification of formula (2.9) above.

We now look at the demand that a firm receives when it is part of the agglomerated region, and compare it to the demand that it would receive if the firm decided to break away from the agglomeration and move to the

peripheral region. For the first case, we add demand from the home region (formula 2.2) and from the foreign region (formula 2.8) to find the demand that a firm in the agglomerated region can expect:

$$x_{\text{core}} = \mu \left[Eq^{\sigma-1} + E^*(q^*)^{\sigma-1}\tau^{\sigma-1} \right] p^{-\sigma} \tag{2.29}$$

We can use the simplifications in (2.27) and (2.28) to write this as

$$x_{\text{core}} = \frac{2\mu}{1+\mu} Eq^{\sigma-1}p^{-\sigma} \tag{2.30}$$

This is the demand that a representative firm faces when it is located in the core, the agglomerated *home* region. For the demand that the same firm would get if it were to move to the *foreign* region, we use a version of (2.29) where the transport costs work the other way:

$$\begin{aligned} x_{\text{periphery}} &= \mu \left[Eq^{\sigma-1}\tau^{\sigma-1} + E^*(q^*)^{\sigma-1} \right] (p^*)^{-\sigma} \\ &= \mu q^{\sigma-1} E \left[\tau^{\sigma-1} + \tau^{1-\sigma} \frac{1-\mu}{1+\mu} \right] (p^*)^{-\sigma} \end{aligned} \tag{2.31}$$

where we used the same simplifications. We now turn our attention to the ratio $x_{\text{core}}/x_{\text{periphery}}$. This ratio tells us if the agglomerated equilibrium is stable, for if it is smaller than one, a firm can expect more demand in the peripheral region than in the agglomerated region. As above, profits are proportional to demand, so when the ratio is below one we know that profits must be larger in the *foreign* region than they are in the *home* region. This means that the agglomerated equilibrium is unstable. Reversely, a value of the ratio larger than one indicates that the agglomeration is stable. We write

$$\begin{aligned} \frac{x_{\text{core}}}{x_{\text{periphery}}} &= \left(\frac{p}{p^*} \right)^{-\sigma} \left[(1+\mu)\tau^{\sigma-1} + (1-\mu)\tau^{1-\sigma} \right] /2 \\ &= \tau^{\sigma\mu} \left[(1+\mu)\tau^{\sigma-1} + (1-\mu)\tau^{1-\sigma} \right] /2. \end{aligned} \tag{2.32}$$

From this expression, we can calculate the value of τ for which the agglomerated equilibrium is just stable: in that case, the ratio is equal to one. This holds trivially for $\tau = 1$; if there are no transport costs, location is irrelevant and firms will receive equal demand wherever they locate. If we differentiate (2.32) with respect to τ at $\tau = 1$, we find that the derivative is equal to $(3\sigma - 2)(\mu - 1)$, which is negative. This means that a small decrease in τ (or, the introduction of transport costs) leads to a situation where the ratio in (2.32) becomes larger than one. In that case, agglomeration is stable.

Neary (2001) shows that, for most practical values of μ, there exists a second, fractional value of τ for which the expression in (2.32) is equal to one. He also proves that it must be higher than τ_{break} derived above, in which case there exist values of transport costs for which the agglomerated equilibrium is stable, as is the symmetric equilibrium. In that case, history decides which equilibrium obtains.