Spatial Organization in International Economics

by

Peter Berezin

A thesis submitted with the requirements for the degree of Doctor of Philosophy
Graduate Department of Economics
University of Toronto

© Copyright by Peter Berezin (2000)
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
Abstract

Spatial Organization in International Economics

Peter Berezin
Department of Economics
University of Toronto

This thesis consists of three essays that study key aspects of the spatial organization of economic activity. The first essay develops a theoretical framework that helps explain i) the considerable productivity dispersion among plants, ii) why exporters are more productive than non-exporters iii) why exporters tend to be larger than non-exporters, iv) why the fraction of plants that export is small, v) why even among those plants that do export, the fraction of revenues earned from exporting tends to be small and vi) why older plants tend to grow output at a slower rate than younger firms. It is also revealed that country size, as measured by the number of markets in a country, can have a significant effect on how quickly firms grow. The essay develops a novel way to solve dynamic optimization problems when "state-space" constraints are present. The solution strategy is applicable to a wide variety of problems in which agents are not constrained now but expect to be constrained in the future. The second essay develops a general equilibrium model in which oligopolistic competition leads to agglomeration. When trade costs fall sufficiently far, one region will become highly industrialized (the core) while the other region will become deindustrialized (the periphery). Higher demand for services and manufacturing workers will increase wages in the core but depress wages in the periphery. As a result, global inequality is seen to be an inevitable consequence of globalization and international trade. The third essay develops a model in which an increase in the volatility of firm-specific productivity shocks increases national income. Labour market pooling ensures that workers are always moving from less productive firms to more productive firms, thus raising the average productivity of the workforce. It is for this reason that firms tend to locate close to other firms, especially when they employ the same type of labour. It is shown that countries whose firms experience temporarily large firm-specific productivity shocks will run current account surpluses. The labour market pooling model is examined with data on job turnover. It is shown that the top 5 percent of most concentrated industries tend to have rates of labour turnover which are nearly twice the national average.
Acknowledgments

I would like to acknowledge the help of the many people who guided and assisted my research over the course of my degree. Daniel Trefler, my supervisor, as well Diego Puga and Nadia Soboleva were all instrumental in encouraging me and helping me overcome the obstacles that I encountered along the way. In addition, I would like to thank the Institute for Policy Analysis for providing facilities and support for my work. I am also grateful to the Social Science and Humanities Research Council of Canada for providing me with generous financial assistance, without which I could never have completed my work.
Contents

1 Trade, Productivity, and the Life Cycle of Firms 1

1.1 Introduction ................................................................. 2
1.2 A Formal Model ............................................................... 4
1.3 Comparative Static Results ............................................... 9
  1.3.1 Consumer’s Problem .................................................. 10
  1.3.2 Producer’s Problem ................................................... 12
  1.3.3 Solution: Step 1 ....................................................... 16
  1.3.4 Solution: Step 2 ....................................................... 22
  1.3.5 Shut-Down Condition .................................................. 26
1.4 Characterizing the Optimal Path ......................................... 28
  1.4.1 Firm Growth and the Profitability of Markets ....................... 29
  1.4.2 The Cost of Entering New markets and its Effect on Firm
    Growth ................................................................. 31
  1.4.3 Country Size and Firm Growth ...................................... 31
  1.4.4 The Superior Performance of Exporters ............................ 33
1.5 Conclusions ................................................................. 36
2 Why Isn’t the Whole World Rich? Trade and Inequality in a Model with Oligopolistic Competition

2.1 Introduction ............................................................. 41
2.2 The Basic Story ......................................................... 45
2.3 A Formal Model ......................................................... 47
  2.3.1 Agricultural Sector ................................................. 48
  2.3.2 Manufacturing Sector .............................................. 48
  2.3.3 Intermediate Services Sector .................................... 50
2.4 When does Agglomeration Occur? Stability of the Symmetric Equilibrium ... 54
2.5 Agglomeration and Global Inequality .................................. 61
2.6 Conclusions .............................................................. 63
2.7 Appendix A: Computing the Level of Trade Costs at which the Symmetric Equilibrium Ceases to be Stable ......................................................... 66
2.8 Appendix B: Three Possible non-symmetric equilibria ...................... 71
  2.8.1 North only Produces Manufactured goods; South Produces only Agriculture ......................................................... 71
  2.8.2 North only Produces Manufactured goods; South produces some Manufactured goods and all the food ......................................................... 73
  2.8.3 North produces manufacturing, Services, and Agriculture; South Produces only Agriculture ......................................................... 74
3 The Ins and Outs of Labour Market pooling

3.1 Introduction ................................................................. 78
3.2 A Macroeconomic Model of Labour Market Pooling .................. 81
  3.2.1 The Consumer's Problem ........................................ 82
  3.2.2 The Firm's Problem ............................................... 84
3.3 The Benefits of Labour Market Pooling ................................ 85
3.4 The Macroeconomic Effects of Labour Market Pooling .............. 87
3.5 Labour Market Pooling and Unemployment .............................. 92
  3.5.1 Aggregate labour demand never exceeds aggregate labour supply .... 92
  3.5.2 Aggregate labour demand may exceed aggregate labour supply ....... 93
3.6 An Empirical Test of the Labour Market Pooling Model .............. 97
  3.6.1 Data and Methodology ............................................ 99
  3.6.2 Results ............................................................. 101
3.7 Conclusions .................................................................. 104
Chapter 1

Trade, Productivity, and the Life Cycle of Firms
1.1 Introduction

Recent empirical evidence has documented the superior economic performance of exporters [Bernard and Jensen (1997 and 1999), Bernard and Wagner (1998), Bernard, Eaton, Jensen and Kortum (2000), Clerides, Lach, and Tybout (1996), Roberts and Tybout (1997), Tybout and Westbrook (1995)]. Exporters are larger, more productive, grow faster in the short run in terms of shipments and employment, pay higher wages, and are less likely to exit the industry than non-exporter [Bernard and Jensen, (1997)]. An important question is whether exporting makes firms more productive or whether good firms are more likely to become exporters. The bulk of the evidence supports the latter theory: high productivity firms self-select into the export market. Less productive firms confine themselves to domestic markets since they are not productive enough to compete abroad. Exporting does not make firms more productive. Rather, high productivity firms are more likely to become exporters.

To place these results within a theoretical framework, researchers have tended to employ models in which plant productivity levels are either exogenously determined or the outcome of random productivity shocks\(^1\). In these models, firms will only undertake the necessary investment to penetrate a foreign market if they can recoup the cost of entering the market. Since high productivity firms are more likely to fare well in foreign markets, only they will choose to become exporters. Thus, one finds a positive correlation between the propensity to export and firm productivity, with the latter determining the former. Although such models acknowledge the heterogeneity of plants, they do not explain what causes the heterogeneity.\(^2\) Furthermore, there is a growing consensus among researchers that most of the growth in total factor productivity is

\(^1\) See Bernard, Eaton, Jensen, and Kortum (2000) for an example of such a model.

\(^2\) Important exceptions are Hopenhayn (1992a, 1992b), Melitz (1999), McGuire and Pakes (1994) and Ericson and Pakes (1995). The papers by Pakes et al. shed light on the strong positive correlation between gross entry rates and gross exit rates. The amount of “turbulence” in an industry varies sharply across industries. Yet, in most industries, larger firms, since they tend to have a smaller chance of exiting [Dunne, Roberts, and Samuelson (1988)] tend to be unaffected by the churning of younger firms in the industry.
within-firm and not between-firm. [Bernard and Jensen (1999)].\(^3\) Yet, most models within this field, since they concentrate on productivity differences across firms at the expense of what happens within firms, overlook this vital source of total factor productivity growth. As this paper will argue, a model in which plant heterogeneity is endogenously determined sheds light on a variety of topical trade issues.

Any discussion of firm productivity and its relationship with exporting is bound to benefit from the vast literature on the life cycle of firms.\(^4\) As Baily, Hulten, and Campbell (1992) document, there is a positive correlation between both plant size and plant age and the likelihood that the plant will export.\(^5\) Furthermore, as Mansfield (1962) and Evans (1987a and 1987b) have shown, big firms tend to be old firms - firms with long histories and well-established relationships with customers and suppliers. Thus, to the extent that old firms tend to be big firms and big firms are more likely to export, there is a positive association between the age of a firm and its likelihood to export.\(^6\)

In addition, old firms are more likely to have exhausted all domestic markets. Many old firms, if they want to grow, must expand abroad. No such constraint applies to young firms who have yet to exhaust their domestic opportunities. Although there can be no doubt that heterogeneity in productivity levels across firms is what determines whether some firms export while others do not, that heterogeneity is often a function of where firms are in their life cycles. A model that exploits the life cycle characteristics of firms will yield a rich set of predictions that simpler models do not capture.

\(^3\) Bernard and Jensen (1999) find that 80 percent of productivity growth is within plant, 20 percent is between plant (due to more productive plants expanding output more quickly than less productive plants), and none of the growth is attributable to entry and exit (low productivity firms exiting and being replaced with high productivity firms).

\(^4\) See Sutton (1997) for a good survey of this literature.

\(^5\) Using 1987 productivity data on U.S. plants, Baily, Hulten, and Campbell (1992) find that plants born 10 to 15 years earlier were almost twice as likely to be in the top productivity quintile compared to plants born less than 5 years earlier. On the other hand, plants in the lowest productivity quintile were almost 3 times as likely to have been born less than 5 years earlier compared to plants born 10 to 15 years earlier.

\(^6\) Using a large sample of U.S. manufacturing firms from 1976 to 1982, Evans (1987b) showed that there is a strong correlation between firm size and firm age for firms less than 20 years old. Because young firms tend to be small, they grow more quickly than old firms. Firms less than 6 years old had a mean employee growth rate of 6.09 percent while firms between the ages of 7 and 20 years had a mean employee growth rate of 1.38 percent. For older firms (older than 20 years), the relationship between age and size and age and employee growth is weakly negative.
This paper constructs a life cycle model of the firm that captures all five of the "basic facts" that Bernard, Eaton, Jensen and Kortrum (2000) note about U.S. manufacturing plants: (i) the productivity dispersion across plants, (ii) the productivity advantage of exporters, (iii) the small fraction who export, (iv) the small fraction of revenues from exporting among those that do, and (v) the much larger size of exporters. The model captures all five facts without relying on the assumption of exogenous productivity differences across firms. To the extent that there are productivity differences across firms, these differences arise because they are at different stages of their life cycles, not because some firms have intrinsic productivity advantages over other firms.

Furthermore, the model developed in this chapter predicts that older firms will grow less quickly than younger firms, a result that has been empirically confirmed by Evans (1987a and 1987b). The model generates the prediction that trade liberalization will lead firms to grow more quickly, even if a firm is still in the stage of its life where it has yet to become an exporter. This prediction is important because many studies of trade liberalization focus exclusively on the performance of exporters and import-competing firms. This model suggests that trade liberalization will have a direct effect on all firms, even those without a foreign presence. The paper also offers an explanation for the "border-effect" phenomenon that McCallum (1995) first documented. Most people are startled to hear that the mere existence of national borders chokes a huge percentage of trade. Why should national borders impede so much trade when trade costs are so low? This paper argues that a life cycle model of a firm, in which firms incur expansion costs upon entering new markets, can help resolve the border effect puzzle.

1.2 The Basic Story

Imagine an industry in which firms produce differentiated manufactured varieties and compete in a monopolistically competitive marketplace. Firms face an exogenous probability that in any period, their variety will become obsolete and they will have to
exit the industry. In every period, a firm incurs three types of costs. First, a firm must cover its fixed cost before it can produce anything. Second, a firm must cover its variable costs. Third, a firm must incur an expansion cost if it wishes to enter a new market. Once a market is entered, a firm can stay there until its variety has become obsolete. As one might imagine, entering new markets is costly. Let us assume that the cost of entering new markets increases more than proportionally with the number of new markets entered (so that it is more than twice as costly to enter two markets as to enter one market).

This simple set-up yields a number of interesting predictions. First, since it is costly to enter new markets, no firm will choose to enter every market instantaneously. Rather, firms will be selective as to which market they enter, concentrating first on the more profitable markets. Let us assume that foreign markets are marginally less profitable than domestic markets (perhaps because extra transport costs must be incurred when selling goods abroad). This implies that firms will concentrate on domestic markets first and expand abroad only when all domestic markets are exhausted.

Although this prediction may sound obvious, it gets to the heart of the “border effect” and “missing trade” puzzles that McCallum (1995), Trefler (1995), Engles and Rogers (1996), and Helliwell (1997) have documented. Most Canadians, for instance, assume that trade in North America flows north-south: that is, Canadian provinces trade a lot more with U.S. states than they do with other Canadian provinces. In fact, that is false. For instance, consider three North American cities: Toronto, Seattle and Vancouver. Seattle and Vancouver are roughly the same size and roughly the same

---

7 Bernard and Wagner (1998) find “substantial sunk costs in export entry” for a sample of German firms. Likewise, Roberts and Tybout (1997) document that Columbian firms generally bear large sunk costs when penetrating foreign markets. In this paper, I assume that the cost of entering a new market does not depend on whether the new market is located at home or abroad. This assumption is made to keep the math less daunting. If we assume that foreign markets are more costly to enter (a very plausible assumption), this would only strengthen the qualitative predictions of the model.

8 This knife-edge result stems from the symmetry assumption that foreign and domestic markets are identical. Suppose, however, that some foreign markets were more profitable than some domestic markets despite the fact that trade costs must be incurred only when servicing foreign markets. The domestic firm would enter those foreign markets before it entered the relatively less profitable domestic markets. This implies that there would be an ordering of markets, with a bias towards domestic markets for any positive trade cost.
distance from Toronto. Yet, firms in Toronto sell more than ten times as much in Vancouver as they do in Seattle. In other words, the Canada-U.S. border chokes off more than 90 percent of all trade, a huge number given the relative ease with which goods and can be traded between the two countries.⁹

As long as firms incur expansion costs, border effects will arise simply because firms are forced to pick and choose which markets they first enter. Even if Seattle is slightly less profitable than Vancouver, because it would be too expensive for a firm in Toronto to enter both markets simultaneously, the firm will concentrate on Vancouver, and indeed, concentrate only on Canadian markets until all Canadian markets have been entered. Only at that stage will the firm expand abroad. However, if the firm faces a high probability that its variety will become obsolete and if the number of domestic markets is large, that time may never come. Thus, a simple life cycle model where firms face expansion costs can explain the border effect phenomenon.

The model developed in this paper makes two important additional predictions. The first prediction is that the number of markets that firms can potentially enter affects how quickly firms grow. In the absence of trade, firms in large countries will grow more quickly than firms in small countries. When there is free trade, what matters is not country size, but rather the size of the global market--the bigger the global market, the more quickly will firms grow. If this statement sounds familiar, it is because most politicians and pundits simply assume that it is true. Yet, to my knowledge, such a relationship has never been formally modeled. The reason is that most trade models are

⁹ Helliwell (1998) finds that "trade among Canadian provinces with that between Canadian provinces and U.S. states show interprovincial trade in 1988-90 to have been more than twenty times as dense as that between provinces and states, with some evidence of a downward trend since then, due to the post-FTA growth in trade between Canada and the United States." Wei (1996) finds evidence of border effects among OECD countries that, although sizable, are far smaller than Helliwell's estimates. The discrepancy, however, appears to be largely due to methodological issues. Helliwell concludes that the border effect for unrelated OECD countries lies in the range of eleven to fifteen (meaning that OECD borders choke off more than 90 percent of trade). Head and Reis (1998) attempt to analyze the link between international migration and the border effect phenomenon. Using Canadian trade data with 136 partners from 1980 to 1992, they find that a 10% increase in immigrants is associated with a 1% increase in Canadian exports to the immigrant's home country, and a 3% increase in imports.
based on a static framework. A dynamic model, however, can deeply enrich our understanding of firm evolution and its effect on a variety of key trade issues.

Why should the number of markets that a firm can enter have a bearing on how quickly the firm grows? The answer has to do with the opportunity cost of not entering new markets. The more markets there are to enter, the more a firm will lose by not having a presence in those markets. To see why, consider the following analogy: imagine a simple Robinson Crusoe economy in which Crusoe grows oranges and sells these oranges at the local market. For simplicity, assume that an orange tree will produce oranges from the very first day it is planted and will continue to produce oranges forever. Because planting orange trees is strenuous work, Crusoe finds it optimal to plant only one tree a day. Due to the small size of his island, Crusoe can plant no more than 20 trees in total. On the morning of day 20, Crusoe has planted 19 trees. Consider what happens if he takes the day off and neglects to plant the twentieth tree, his final tree, on the twentieth day and instead plants the final tree on the twenty-first day. Since he cannot plant more than 20 trees anyway, all he loses is the oranges that the twentieth tree would have yielded on the twentieth day, a rather insignificant loss. Now consider what would happen if there were no limit to how many orange trees Crusoe could plant. Crusoe would resume his optimal plan of planting one tree every day. However, because he neglected to plant a tree on the twentieth day, he would forever be missing the fruit produced by that tree, a rather substantial loss. The cost to Crusoe of slowing down the rate at which he plants trees is much greater when there is no upper bound on the number of trees that he can plant. Thus, he would find it optimal to spend more time planting trees if he could plant more than 20 trees in total. The higher is the maximum number of trees that Crusoe can plant, the higher will be the opportunity cost of not planting new trees, and thus, the more incentive he will have to increase the rate at which he plants new orange trees.

Similarly, when there are an unlimited number of potential markets that a firm can enter, the firm loses more by not having a presence in those markets. Imagine, for instance, what would happen if a firm erred by entering one less market than was optimal
at time $t$ and then resumed its optimal path in the next period. What is the cost of making this mistake? If there is no maximum to the number of markets that the firm can enter, then after time $t$ the firm will forever be in one less market than if it had maintained its optimal path all along. Thus, after time $t$, the firm will forever be losing the profits that it could have earned from that additional market. If the number of markets is constrained, however, we can no longer make this argument. Eventually, the firm would enter every possible market. Since the firm made a mistake by entering one less market at time $t$ than what was optimal, it will lose the profits from that market only until all markets are entered, and not forever as was the case when the maximum number of markets was unbounded. Thus, the opportunity cost of not entering a market is higher when there is no maximum to the number of markets that the firm can enter. Similarly, when the maximum number of markets that a firm can enter rises, the opportunity cost of not entering a market also rises. This leads the firm to grow more quickly, entering new markets at a faster pace.

The second significant prediction of the model is that trade liberalization, by making foreign markets more profitable, will lead firms to grow more quickly both at home and abroad. The fact trade liberalization will lead domestic firms to penetrate foreign markets more quickly is not surprising. If foreign markets become more profitable due to, for example, a drop in tariffs, it stands to reason that firms will enter foreign markets at a faster rate. What is less obvious is that trade liberalization should lead firms that have no presence abroad to grow more quickly at home. Even if trade liberalization does not make domestic markets more profitable, firms will still grow more quickly at home because they will be eager to exhaust all their domestic markets (which by assumption are still more profitable than foreign markets) so that they can become exporters.

To summarize, this model will explain the following facts: i) there is considerable productivity dispersion among plants, ii) exporters are more productive than non-

\footnote{Note how the second prediction is different from the first prediction. The first prediction had to do with the number of markets while the second prediction has to do with the profitability of those markets.}
exporters, iii) exporters are much larger than non-exporters, iv) the fraction of firms that export is small, v) even among those that do export, the fraction of revenues earned from exporting tends to be small vi), older firms grow more slowly than younger firms. Facts i) through v) correspond to the key stylized facts as delineated by Bernard, Eaton, Jensen and Kortrum's (2000). Facts iv) and v) correspond to the "border effects" and "missing trade" puzzles of Helliwell (1998) and Trefler (1995). Fact vi) corresponds to the empirical observations of Evans (1987a and 1987b) and Dunne, Roberts, and Samuelson (1988). Lastly, the model makes two key predictions. First, firms that have access to more markets should grow more quickly. This implies that small countries, if they are cut off from world trade, will tend have firms that grow more slowly than firms in countries that are open to trade. Second, the rate at which firms grow domestically will increase if foreign markets become more profitable (for instance, as a result of trade liberalization), even if the profitability of domestic markets stays the same. Thus, firms will grow more quickly if they have access to more markets and firms will grow more quickly if those markets become more profitable.

1.3 A Formal Model

Consider the following partial equilibrium model. Imagine a world in which there are two identical regions, home and foreign. In each region, there are \( m \) identical markets. Let us concentrate on the economy in home (since both regions are identical, the analysis would be the same for foreign). The price index for manufactured varieties in a home market, \( \hat{P}_{h,m} \), includes all the varieties of manufactured goods available in that market. Since there is trade between home and foreign, the manufacturing price index will include both domestic and foreign made varieties.

\[ \text{footnote}{Berezin (2000 - forthcoming) develops a general equilibrium version of this model in which the profits that a firm earns from every market that it services is endogenous. Unfortunately, the greater complexity of the general equilibrium set-up makes it impossible to derive analytical results. Instead, numerical simulations are used to illustrate the model’s qualitative predictions. Given the general equilibrium nature of his model, one can track the comparative static effects on real wages and firm profitability. For instance, real wages rise as the size of the market expands while the profitability of each market falls. However, since firms now grow faster and service more markets, their discounted stream of profits does not change.} \]
The manufacturing sector is assumed to be monopolistically competitive, à la Dixit and Stiglitz (1978). Firms have access to \( m \) home markets and \( m \) foreign markets (we could let the number of foreign markets differ from the number of domestic markets, but this would not change the essence of the story). At any given point in time, a firm faces a probability \( \rho \) that its variety will become obsolete. If a firm's variety does become obsolete, it must exit the industry. As is consistent with the empirical findings of Clerides, Lach and Tybout (1996) and Bernard and Wagner (1998), firms face "expansion" costs when they enter new markets. I assume that these costs are quadratic and that they are the same for home and foreign markets (allowing for expansion costs to be higher in foreign markets, as one might expect to observe, would make the qualitative results of the model more robust, but make the model less analytically tractable).

1.3.1 Consumer’s Problem

There are \( L \) workers in each market. Workers are assumed to devote one unit of income on manufactured varieties. Monopolistically competitive firms produce differentiated manufactured varieties over which workers have CES preferences. Define \( n_{h,h} \) as the number of home varieties available in each home market. The first subscript denotes where the good is produced and the second subscript denotes where the good is consumed. In addition, define \( n_{f,h} \) as the number of foreign varieties available in each home market, \( n_{f,f} \) as the number of foreign varieties available in each foreign market, \( n_{h,f} \) as the number of home varieties available in each foreign market and \( c_{h,h} \) and \( c_{f,h} \) as the quantities of a home and foreign variety, respectively, purchased by a worker in home.

Lastly, define \( \mu \geq 1 \) is a trade cost that must be borne when a home consumer purchases a foreign variety. To have one unit of the foreign variety for consumption, a home worker must purchase \( \mu \) units of the variety. The trade cost is broadly defined and can include such things as freight costs, tariffs, or the "patriotic" loss of utility that one
suffers from buying a good that was not produced at home. Workers solve a “second stage” maximization problem of the form:

$$\begin{align*}
\text{Max} & \left\{ c_{h,h}(i), c_{f,h}(i) \right\} \\
& \int_{0}^{n_{h,h}} c_{h,h}(i) \frac{\sigma-1}{\sigma} d(i) + \int_{0}^{n_{f,h}} c_{f,h}(i) \frac{\sigma-1}{\mu} d(i) \\
& \frac{\sigma}{\sigma-1}
\end{align*}$$

subject to

$$\int_{0}^{n_{h,h}} p_{h,h}(i) c_{h,h}(i) d(i) + \int_{0}^{n_{f,h}} p_{f,h}(i) c_{f,h}(i) d(i) = 1.$$ 

Solving the maximization problem in equation (1) produces the following demand equations:

$$c_{h,h}(i) = \frac{p_{h}(i)^{-\sigma}}{\hat{p}_{h,m}},$$

where the price index for manufacturing varieties in home, $\hat{p}_{h,m}$, is defined as

$$\hat{p}_{h,m} = \int_{0}^{n_{h,h}} p_{h}(i)^{1-\sigma} d(i) + \int_{0}^{n_{f,h}} p_{f}(i)^{1-\sigma} d(i).$$

Similarly, the quantity of each foreign variety $i$ purchased by a worker in home is
1.3.2 Producer's Problem

As mentioned before, at every time $t$, firm $i$ faces a constant probability $\rho$ of having its variety become obsolete. If the variety becomes obsolete the firm will earn no revenues and have no costs. If the variety does not become obsolete, firm $i$ will earn revenue,

\begin{equation}
\begin{aligned}
\left(5\right) & \quad p_h(i)[x_{h,h}(i,t) y_h(i,t) + x_{h,f}(i,t) y_f(i,t)],
\end{aligned}
\end{equation}

where $y_h(i,t)$ is the number of home markets in which home firm $i$ sells its variety, $y_f(i,t)$ is the number of foreign markets in which the home firm sells its variety, $x_{h,h}(i)$ is the quantity of a home firm’s variety sold in each home market and $x_{h,f}(i)$ is the quantity of a home firm’s variety sold in each foreign market. Additionally, as long as the variety does not become obsolete, home firm $i$ will incur the following costs at every time $t$:

\begin{equation}
\begin{aligned}
\left(6\right) & \quad \beta [x_{h,h}(i,t) y_h(i,t) + x_{h,f}(i,t) y_f(i,t)] + \frac{1}{2} \alpha u(i,t)^2 - F,
\end{aligned}
\end{equation}

where $\beta$ is the marginal cost of producing one unit of the variety, $u(i,t)$ is the number of new markets entered at time $t$, and $\alpha$ is a parameter signifying how large the cost of entering new markets is.

The first term in equation (6) captures the firm’s variable cost, the second term captures the firm’s expansion cost, and the third term captures the firm’s fixed cost. Note that the firm must pay a fixed cost every period in which it operates and not just at time
zero, when it commences operations. At time \( t \), firm \( i \) operates in \( y_h(i,t) \) home markets and \( y_f(i,t) \) foreign markets, selling \( x_{h,h} \) and \( x_{h,f} \) units of the variety in each market, respectively, with each unit costing \( \beta \) dollars to produce. Note that \( y_h(i,t) \) and \( y_f(i,t) \) play the role of "state" or "stock" variables in the problem while \( u(i,t) \) plays the role of the "control" or "flow" variable.\(^{12}\)

In the tradition of "costly capital adjustment models"\(^{13}\), I assume that entry costs are quadratic. Thus, the cost of entering new markets increases more than proportionately with the number of markets entered at time \( t \). Expansion costs are sunk costs: once the firm has paid to establish a new market, it can operate freely in that market until the firm’s product becomes obsolete. The quadratic cost structure is designed to capture the fact that excessive growth of a company’s market base can lead the firm to overextend itself and as a result, incur burdensome expansion costs. We can think of \( \alpha \) as a parameter that measures how expensive it is to penetrate new markets. A high \( \alpha \) implies that it is very costly to establish a retail network in the new market capable of selling the variety that the firm produces. The trade cost, \( \mu \), on the other hand, represents the extra expense incurred in buying foreign varieties. This cost can be considered of the "iceberg" form, first introduced by Samuelson (1954). To consume one unit of a foreign variety, a consumer must buy \( \mu \) units since \( \mu - 1 \) units of the variety will "melt" in transit. The firm incurs a fixed cost, \( F \), which might represent the cost of overhead (such as plant and equipment).\(^{14}\) Finally, the firm discounts future cash flow at rate \( r \).

\(^{12}\) The notation, \( y(t) \) for state variables, and \( u(t) \) for co-state variables, is commonly used in dynamic optimization problems. See Chaing (1992), for instance.

\(^{13}\) See Obstfeld (1996, p. 105) for an example of such a model.

\(^{14}\) The fixed cost, \( F \), is paid every period and not just at time zero. One could assume that \( F \) is paid only once, when the firm commences operations, without altering any of the qualitative predictions of the model. The assumption that the firm incurs a fixed cost in every period is useful for the analysis of how firm age interacts with productivity. Since a firm produces more as it ages, it will spread its fixed cost over an increasingly greater quantity of output. This will in turn reduce average cost, thus raising productivity. This provides an additional (but not the only) channel through which older firms become more productive.
As long as $\mu > 1$, home markets will always be more profitable than foreign markets. Thus, firms will first concentrate on domestic markets and enter foreign markets only when all domestic markets have been exhausted. Since the cost of entering new markets is quadratic, it will turn out to be optimal for firms to expand into new markets slowly. Let us denote $\tau_h$ as the time when the home firm exhausts all home markets and $\tau_f$ as the time when it exhausts all foreign markets, after which time there are no new markets in which the firm can enter.

Let us denote $V_h$ as the expected present value of profits that a home firm earns from home markets and $V_f$ as the expected present value of profits that a home firm earns from foreign markets. The firm runs the risk that at every time $t$ its variety will become obsolete with probability $\rho$. Since the firm comes into existence at time zero, the probability that the firm is still around at time $t$ is $(1 - \rho)^t$, which in continuous time is equal to $e^{-\rho t}$. Thus, we must scale the firm's profit function by $e^{-\rho t}$ at every time $t$ to reflect the fact that its variety may have already become obsolete by then. The home firm's optimization problem can then be expressed as:

\begin{equation}
\max_{\{x_h, h(i,t), u(i,t), y_h(i,t)\}} \int_0^\infty \left[ e^{-\rho t} \left[ p_h(i,t) - \beta \right] x_h(i,t) y_h(i,t) - \frac{1}{2} \alpha u(i,t)^2 - F \right] dt,
\end{equation}

subject to

\begin{equation}
\dot{y}_h(i, t) = u(i, t) \quad \dot{y}_h(i, t) \geq 0 \quad y_h(i, 0) = 0 \quad y_h(i, \tau_h) = m \quad y_h(i, t) \leq m,
\end{equation}

and subject to
subject to

\[ \max_{\{x_h, f(i,t), u(i,t), y_f(i,t)\}} V_f = \int_{\tau_h}^{\infty} e^{-(\tau+P)x} \left[ \left( p_h(i,t) - \beta \right) x_h, f(t) - \frac{1}{2} \alpha u(t)^2 - F \right] dt \]

Notice that equation (9) is itself a maximization problem. Thus, we first have to solve the maximization problem posed in equations (9) and (10) before we can go back and solve the maximization problem in equation (7) and (8). Although it is tempting to try to solve the problem all in one shot, such an approach will yield incorrect results. The problem stems from the fact that in most dynamic optimization problems the solution is predicated upon the continuity of both the state and costate variables. If a state-space constraint (a constraint that puts a bound on the value of the state variable) is present, however, the costate variable (i.e. the multiplier on the state-space constraint) will take a discrete jump at the point where the constraint becomes binding. Failure to take into account the discrete jump of the costate variable will result in a solution that is "myopic"; that is, the firm, when it is unconstrained, will act as if there is no constraint and will change its behaviour only when the constraint is reached. Obviously, what we want is for the firm to alter its behaviour in anticipation that the constraint will eventually be reached. In general, this is not a trivial matter. A proper solution to the foregoing problem would require the firm to foresee the constraint and adjust its behaviour well before the constraint becomes binding. This implies that the optimal path in the constrained case will differ from the optimal path in the unconstrained case even before the constraint is reached. Mathematically, this implies that the optimal \( u(t) \) will involve
Fortunately, by formulating the problem as I have done in equations (7) to (10), we can analytically derive solutions that are rationally forward looking. This procedure works because it endogenizes the state-space constraint directly within the objective function. This solution strategy can be verbosely described as "dynamic optimization subject to state-space constraints via backward induction". Essentially, this means we work backwards to piece together the optimal course of action for the firm to follow. As the first step in solving the problem, we ask "given that a home firm exhausts all foreign markets at some time $\tau_f$, what is the optimal rate of market expansion between the time it exhausts all home markets, $\tau_h$, and $\tau_f$? In the second step, knowing what the firm will do after time $\tau_h$, we go back and determine the optimal rate of expansion between time zero and time $\tau_h$ and then calculate the optimal values of $\tau_h$ and $\tau_f$. Optimal market expansion is only one application of this methodology. In general, this technology can be applied to a wide array of economic problems in which agents must decide how to behave when they are not presently constrained but expect to be constrained in the future.

1.3.3 Solution: Step 1

As the first step, let us solve the problem posed in equation (9) and (10). To do this, we first write the Hamiltonian for this problem,

\begin{equation}
H_f(i,t) = e^{-(r+p)t} \left[ p_h(i,t) - \beta \right] \kappa_{h,f}(i,t) y_f(i,t) - \frac{1}{2} \sigma u(i,t)^2 + \Lambda_f(i,t) u(i,t) - \Theta_f(i,t) y_f(i,t)
\end{equation}

for $\tau_h \leq t \leq \infty$.

The first order conditions are:
\begin{align}
(12) \quad \frac{\partial H f(i,t)}{\partial x_h f(i,t)} &= e^{-(r+\rho)t} \left[ p_h(i,t) \left( 1 + \frac{\partial p_h(i,t)}{\partial x_h f(i,t)} \frac{x_h f(i,t)}{p_h(i,t)} - \beta \right) \right] y_f(i,t) = 0 \\
&\text{for } \tau_h \leq t \leq \infty,
\end{align}

\begin{align}
(13) \quad \frac{\partial H f(i,t)}{\partial u(i,t)} &= -e^{-(r+\rho)t} \alpha u(i,t) + \Lambda_f(i,t) = 0 \quad \text{for } \tau_h \leq t \leq \infty,
\end{align}

and

\begin{align}
(14) \quad \dot{\Lambda}_f(i,t) &= -\frac{\partial H f(i,t)}{\partial y_f(i,t)} = -e^{-(r+\rho)t} [p_h(i,t) - \beta] x_h f(i,t) + \Theta_f(i,t) \\
&\text{for } \tau_h \leq t \leq \infty,
\end{align}

\begin{align}
(15) \quad y_h(t) &\leq m,
\end{align}

and

\begin{align}
(16) \quad \Theta_f(i,t)[m - y_f(i,t)] &= 0 \quad (\text{complementary slack}).
\end{align}

By definition, the firm cannot enter any new markets after time $\tau_f$, implying $u(i,t) = 0$ after time $\tau_f$. From equation (13), this implies $\Lambda_f(i,t) = 0$ for $\tau_f \leq t \leq \infty$. Equation (12) implies that

\begin{align}
(17) \quad p_h(i,t) &= \frac{e_d(i,t)}{e_d(i,t) - 1} \beta
\end{align}
where \( e_d(i,t) = \frac{\partial x_{h,h}(i,t)}{\partial p_h(i,t)} \frac{p_h(i,t)}{x_{h,h}(i,t)} \), the own-price elasticity of demand for firm \( i \)'s variety at time \( t \). Let us now make some symmetry assumptions. First, let us as assume that all firms, both in home and foreign, have the same technology. Second, let us assume that the demand for their varieties enter symmetrically into a consumer's utility function. Let us also assume that home firms do not differ from foreign firms so that the price of a variety is the same at home and abroad. Also, assume that whatever trade costs consumers incur, that these costs are the same in both countries (so that the trade cost for someone in home to buy a variety in foreign is the same as the trade cost for someone in foreign to buy a variety in home). Finally, assume that the two countries are the same size (both have \( m \) markets with \( L \) workers in each market). These assumptions imply that the home price of variety will equal the foreign price of a variety. Moreover, since all firms are identical, the price of all varieties will be the same. Thus, we can drop the \( i \) subscript from our equations. Furthermore, these assumptions imply that the price and quantity indices for manufactured varieties will be the same in home and foreign. In addition, these symmetry assumptions yield a familiar and useful result: \( e \), the own-price elasticity of demand, will equal \( \sigma \), the elasticity of substitution across varieties from the CES utility function in equation (1).

Let us confine the analysis to stationary equilibria where the total number of firms and the number of varieties offered to consumers are the same in all markets and do not change over time. This assumption implies that \( x_{h,h}(t) \) and \( x_{h,f}(t) \) do not change over time, regardless of the age of the firm. Let us define the profit net of expansion costs and fixed costs that a home firm earns in each foreign market that it services as

\[
\Pi_f = (p - \beta) x_{h,f}.
\]

Similarly, the net profit that a home firm earns in each home market is

\[
\Pi_h = (p - \beta) x_{h,h}.
\]
Since there are \( L \) consumers in each market, goods market clearing implies
\[
x_{h, h} = Lc_{h, h} \quad \text{and} \quad x_{h, f} = Lc_{h, f}.
\]
Thus, demand for a variety in each market where it is sold is
\[
(20) \quad x_{h, h} = \frac{p^{-\sigma} L}{\hat{p}}
\]
and
\[
(21) \quad x_{h, f} = \frac{\mu^{1-\sigma} p^{-\sigma} L}{\hat{p}}.
\]
Equations (18), (19), (20), and (21) imply
\[
(22) \quad \frac{\Pi f}{\Pi h} = \frac{x_{h, f}}{x_{h, h}} = \mu^{1-\sigma}.
\]
For convenience, define \( \Phi = \mu^{1-\sigma} \). Thus, when iceberg trade costs are zero \( (\mu = 1) \), then \( \Phi = 1 \). When trade costs are infinite \( (\mu = \infty) \), then \( \Phi = 0 \). Using equation (22), equation (14) can be simply re-written as
\[
(23) \quad \dot{\Lambda}_f(t) = -e^{-(r+\rho)t} \Pi f = \Theta f(t) - \Phi \Pi_h e^{-(r+\rho)t} \quad \text{for} \quad \tau_h \leq t \leq \infty.
\]
Differentiating equation (13) with respect to \( t \) we have,
\[
(24) \quad \dot{\Lambda}_f(t) = \alpha e^{-(r+\rho)t} \left[ \dot{u}(t) - (r + \rho)u(t) \right] \quad \text{for} \quad \tau_h \leq t \leq \infty.
\]
Equating equation (23) with equation (24), we see,

\[ e^{-(r+\rho)t} [\dot{u}(t) - (r + \rho)u(t)] = \frac{1}{\alpha} \left[ \Theta_f(t) - \Phi \Pi_k e^{-(r+\rho)t} \right] \text{ for } \tau_h \leq t \leq \infty. \]  

Notice that the left-hand side of equation (25) is simply \( \frac{d}{dt} \left( e^{-(r+\rho)t} u(t) \right) \).

Integrating both sides of equation (25) between \( t \) and \( \infty \), we get,

\[ \int_t^\infty \frac{d}{dt} \left( e^{-(r+\rho)t} u(t) \right) dt = \lim_{T \to \infty} e^{-(r+\rho)T} u(T) - e^{-(r+\rho)t} u(t) = \frac{1}{\alpha} \int_t^\infty \Theta_f(t) dt - \frac{\Phi \Pi_k e^{-(r+\rho)t}}{\alpha(r+\rho)}. \]

The transversality condition for this problem is \( \lim_{T \to \infty} e^{-(r+\rho)T} u(T) = 0 \). This is very intuitive for it rules out paths in which the firm keeps increasing the rate at which it enters new markets, which would be impossible, since the total number of markets is bounded by \( 2m \).

From the complementary slackness condition in equation (15), when \( y_f(t) \leq m \), it must be the case that \( \Theta_f(t) = 0 \). Since \( y_f(t) \leq m \) from time zero to time \( \tau_f \), it must be true that \( \Theta_f(t) = 0 \) from time zero to time \( \tau_f \) as well. Thus,

\[ \int_t^{\tau_f} \Theta_f(t) dt = 0 \text{ for } 0 \leq t \leq \tau_f. \]

Let \( \eta_f = \int_{\tau_f}^{\infty} \Theta(t) dt \). Thus, \( u(t) \) is
(28) \[ u(t) = \frac{-1}{\alpha} \eta_f e^{(r+\rho)t} + \frac{\Phi \Pi_h}{\alpha(r+\rho)} \text{ for } \tau_h \leq t \leq \tau_f. \]

To calculate \( \eta_f \), note \( u(t) = 0 \) for \( t \geq \tau_f \) (since after \( \tau_f \) there are no more markets left into which the firm can enter). From equation (13), this implies

(29) \[ \Lambda_f(t) = 0 \quad \text{for } \tau_f \leq t \leq \infty. \]

Differentiating equation (29) with respect to \( t \) we see,

(30) \[ \dot{\Lambda}_f(t) = 0 \quad \text{for } \tau_f \leq t \leq \infty. \]

Equations (23) and (30) together imply

(31) \[ \Theta_f(t) = \Phi \Pi_h e^{-(r+\rho)t} \quad \text{for } \tau_f \leq t \leq \infty. \]

Thus, \( \eta_f \) must be

(32) \[ \eta_f = \int_{\tau_f}^{\infty} \Theta(t) dt = \int_{\tau_f}^{\infty} \Phi \Pi_h e^{-(r+\rho)t} dt = \frac{\Phi \Pi_h e^{-(r+\rho)\tau_f}}{(r+\rho)}. \]

Substituting equation (32) into equation (28) we thus have,

(33) \[ u(t) = \frac{\Phi \Pi_h}{\alpha(r+\rho)} \left( 1 - e^{-(r+\rho)(\tau_f-t)} \right) \quad \text{for } \tau_h \leq t \leq \tau_f. \]

We can compute \( \Lambda_f(t) \) by substituting equation (33) into equation (13) to get:
Differentiating equation (34) with respect to \( t \) yields:

\[
(35) \quad \dot{\Lambda}_f(t) = -\Phi \Pi_h e^{-(r+\rho)t} \quad \text{for } \tau_h \leq t \leq \tau_f.
\]

By integrating equation (33), \( y_f(t) \) can calculated as,

\[
(36) \quad y_f(t) = \int_{\tau_h}^{t} u(t) dt = \frac{\Phi \Pi_h}{\alpha(r+\rho)} \left( t - \tau_h + \frac{1}{(r+\rho)} \left[ e^{-(r+\rho)(\tau_f - \tau_h)} - e^{-(r+\rho)(\tau_f - t)} \right] \right)
\]

for \( \tau_h \leq t \leq \tau_f \).

1.3.4 Solution: Step 2

Let us now use the results we derived in Step 1 to solve the maximization problem posed in equation (7) and (8). The Hamiltonian for this problem is:\(^{15}\)

\[
(37) \quad H_h(t) = e^{-(r+\rho)t} \left( \Pi_h y_h(t) - \frac{1}{2} \alpha u(t)^2 \right) + \Lambda_h(t) u(t) - \Theta_h(t) y_h(t) \quad \text{for } 0 \leq t \leq \infty.
\]

The first order conditions are:

\[
(38) \quad \frac{\partial H_h(t)}{\partial u(t)} = -e^{-(r+\rho)t} \alpha u(t) + \Lambda_h(t) = 0 \quad \text{for } 0 \leq t \leq \infty,
\]

and

---

\(^{15}\) In writing the problem this way, I have assumed that the firm is already optimally choosing how much to sell in each market that it services; as was shown in Step 1, this assumption is valid only if \( \Pi_h \) does not vary with time. This will be the case in a stationary equilibrium in which the number of firms is constant and the price and quantity indices do not change over time.
\( \dot{\lambda}_h(t) = - \frac{\partial H_h(t)}{\partial y_h(t)} = \Theta_h(t) - e^{-(r + \rho)t} \Pi_h \) for \( 0 \leq t \leq \infty \).

\( y_h(t) \leq m \),

and

\( \Theta_h(t)[m - y_h(t)] = 0 \) \( \text{(complementary slack)} \).

As before, differentiating equation (38) with respect to \( t \) we have,

\( \dot{\lambda}_h(t) = \alpha e^{-(r + \rho)t} \left[ \dot{u}(t) - (r + \rho)u(t) \right] \) for \( 0 \leq t \leq \infty \).

Equating equation (39) with equation (42), we see,

\( e^{-(r + \rho)t} \left[ \dot{u}(t) - (r + \rho)u(t) \right] = \frac{1}{\alpha} \left[ \Theta_h(t) - \Pi_h e^{-(r + \rho)t} \right] \) for \( 0 \leq t \leq \infty \).

Similar to what we did in Step 1, integrate both sides of equation (43) between \( t \) and \( \infty \) to get,

\( \int_t^\infty \frac{e^{-(r + \rho)t}}{\alpha} \left[ \Theta_h(t) - \Pi_h e^{-(r + \rho)t} \right] dt = \lim_{T \to \infty} e^{-(r + \rho)T} u(T) - e^{-(r + \rho)t} u(t) = \frac{1}{\alpha} \int_t^\infty \Theta(t) dt - \frac{\Pi_h e^{-(r + \rho)t}}{\alpha(r + \rho)} \).

As in Step 1, the transversality condition implies \( \lim_{T \to \infty} e^{-(r + \rho)T} u(T) = 0 \).
From the complementary slackness condition (equation (33)), when \( y_h(t) \leq m \), it must be the case that \( \Theta_h(t) = 0 \). Since \( y_f(t) \leq m \) from time zero to time \( \tau_h \), it must be true that \( \Theta_h(t) = 0 \) from time zero to time \( \tau_h \) as well. Thus, we get

\[
\int_0^{\tau_h} \Theta_h(t) \, dt = \int_0^{\tau_h} \Theta_h(t) \, dt \quad \text{for } 0 \leq t \leq \tau_h.
\]

Let \( \eta_h = \int_0^{\tau_f} \Theta_h(t) \, dt = \int_0^{\tau_h} \Theta_h(t) \, dt + \int_\tau_h^{\tau_f} \Theta_h(t) \, dt \). Thus, \( u(t) \) is

\[
u(t) = -\frac{1}{\alpha} \eta_h e^{(r+\rho)t} + \frac{\Pi_h}{\alpha(r+\rho)} \quad \text{for } 0 \leq t \leq \tau_h.\]

As in Step 1, to calculate \( \eta_f \), note that \( u(t) = 0 \) after \( \tau_f \), and so equation (38) implies

\[
\Lambda_h(t) = 0 \quad \text{for } \tau_f \leq t \leq \infty.
\]

Differentiating equation (47) with respect to \( t \), we see,

\[
\dot{\Lambda}_h(t) = 0 \quad \text{for } \tau_f \leq t \leq \infty.
\]

Equations (20) and (40) together imply

\[
\Theta_h(t) = \Pi_h e^{-(r+\rho)t} \quad \text{for } \tau_f \leq t \leq \infty.
\]

Comparing equation (13) with equation (37), we see that
Differentiating both sides with respect to \( t \), we have

\[
\Lambda_h(t) = \Lambda_f(t) \quad \text{for } \tau_h \geq t \geq \infty.
\]

Using equation (51) and equating (35) with (39) yields

\[
\Theta_h(t) = (1 - \Phi)\Pi_h e^{-(r+\rho)t} \quad \text{for } \tau_h \leq t \leq \tau_f.
\]

Hence,

\[
\eta_h = \frac{\Pi_h}{r + \rho} \left[ (1 - \Phi) e^{-(r+\rho)\tau_h} + \Phi e^{-(r+\rho)\tau_f} \right].
\]

Plugging equation (53) into equation (46) gives,

\[
u(t) = \frac{\Pi_h}{\alpha(r + \rho)} \left( 1 - \left[ \Phi e^{-(r+\rho)(\tau_f - t)} + (1 - \Phi) e^{-(r+\rho)(\tau_h - t)} \right] \right) \quad \text{for } 0 \leq t \leq \tau_h.
\]

As before, we can integrate (54) to get

\[
y_h(t) = \int_0^t u(t) \, dt = \frac{\Pi_h}{\alpha(r + \rho)} \left( \frac{1 - \Phi}{r + \rho} \left[ e^{-(r+\rho)\tau_h} - e^{-(r+\rho)(\tau_h - t)} \right] + \frac{\Phi}{r + \rho} \left[ e^{-(r+\rho)\tau_f} - e^{-(r+\rho)(\tau_f - t)} \right] \right) \quad \text{for } 0 \leq t \leq \tau_h.
\]

We can compute \( \tau_h \) and \( \tau_f \) by simultaneously solving the following two equations:
and equation (57) can be more succinctly expressed as

\begin{equation}
(57') \quad (r + \rho)\tau_h + \left(1 - \Phi\left(e^{-(r + \rho)\tau_h} - 1\right) + \Phi\left(e^{-(r + \rho)\tau_f} - e^{-(r + \rho)(\tau_f - \tau_h)}\right)\right) = \frac{\alpha(r + \rho)^2 m}{\Phi \Pi_h}
\end{equation}

and equation (57) can be more succinctly expressed as

\begin{equation}
(57'') \quad (r + \rho)(\tau_f - \tau_h) + e^{-(r + \rho)(\tau_f - \tau_h)} = \frac{\alpha(r + \rho)^2 m}{\Phi \Pi_h}.
\end{equation}

1.3.5 Shut-down Condition

As the last step, we need to specify a shutdown condition for the firm. That is, we need to specify how low \( \Pi_h \) can fall before the expected present value of future profits that the firm earns becomes negative. Let us call this number \( \Pi_{min} \). For any \( \Pi_h \) below \( \Pi_{min} \), the firm will never begin operations since its expected present value of future profits will be negative. Expected profits are zero when

\begin{equation}
(58) \quad V_h^* + V_f^* = 0
\end{equation}
Using equations (7), (9), (18) and (19), we can write,

\[(59)\]

\[V_h + V_f = \int_0^\infty e^{-(r+\rho)t} \left\{ \Pi_h y_h(t) - \frac{1}{2} \alpha u(t)^2 \right\} dt + \int_0^\infty e^{-(r+\rho)t} \Phi \Pi_h m dt + \int_0^\infty e^{-(r+\rho)t} \Phi \Pi_h m dt = 0\]

Substituting equations (33), (36), (54), (55) into equation (59) and solving for \(\Pi_h\) yields\(^{16}\)

\[(60)\]

\[\Pi_h = \frac{-b + \sqrt{b^2 - 4ac}}{2a},\]

where

\[a = \frac{1}{\alpha(r+\rho)^3} \left( \frac{1}{2} - (r+\rho)\tau_h - \frac{1}{2} e^{-(r+\rho)\tau_h} \right) + \Phi^2 \left( \frac{1}{2} e^{-(r+\rho)(\tau_h - \tau_f)} e^{-(r+\rho)\tau_f} + \frac{1}{2} \left( e^{-(r+\rho)\tau_h} + e^{-(r+\rho)\tau_f} \right) - e^{-(r+\rho)(\tau_f + \tau_h)} - e^{-(r+\rho)(2\tau_f - \tau_h)} \right)\]

and

\[b = \frac{m}{r + \rho} \left( e^{-(r+\rho)\tau_h} + \Phi e^{-(r+\rho)\tau_f} \right)\]

\(^{16}\) Strictly speaking, we must also consider the other solution to this equation,

\[\Pi_h = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.\]

However, this solution implies that \(\Pi_h < 0\) which contradicts the constraint \(\dot{y}_h(t) \geq 0\). Thus, this solution can be discarded.
\[ c = -\frac{F}{r + \rho}. \]

### 1.4 Characterizing the Optimal Expansion Path

Three interesting features characterize the optimal expansion path for a firm. First, as long as there are trade costs, no matter how small, young firms will concentrate strictly on domestic markets and avoid all foreign markets until they have reached age \( \tau_h \). If a firm's variety becomes obsolete by then, the firm will never export anything.

Second, as a firm matures, it will grow more slowly, entering fewer markets in each period. In other words, \( u(t) \) falls over time. One can see this by differentiating equation (53) with respect to \( t \):

\[
(61) \quad \frac{\partial u(t)}{\partial t} = -\frac{\Pi_h}{\alpha} \left[ \Phi e^{-(r+\rho)(\tau_f - t)} + (1 - \Phi)e^{-(r+\rho)(\tau_h - t)} \right] < 0 \text{ for } 0 \leq t \leq \tau_h.
\]

For the range between time zero and time \( \tau_h \), firms will decrease the rate at which they enter domestic markets. Similarly, by differentiating equation (35) with respect to \( t \), one can show that firms will also decrease the rate at which they enter foreign markets:

\[
(62) \quad \frac{\partial u(t)}{\partial t} = -\frac{\Phi \Pi_h}{\alpha} e^{-(r+\rho)(\tau_f - t)} < 0 \text{ for } \tau_h \leq t \leq \tau_f.
\]

Evans (1987a) documents precisely this result. Using a sample of firms operating within 100 manufacturing industries, Evans concludes firm growth decreases at a diminishing rate with firm size. Similarly, Dunne, Roberts, and Samuelson (1988) note that although younger firms are more likely to exit than older firms, this effect is more than outweighed by the tendency of young firm to grow more quickly. Thus, on net,
even if one takes into account that a larger proportion of young firms will exit (and hence have growth rates of minus one), young firms, on average, still tend to increase output and employment faster than older firms. These results contradict "Gibrat's Law" which postulates that firm growth is independent of firm size.\textsuperscript{17}

Third, there is no discontinuity in \( u(t) \) when the firm passes \( \tau_h \) and begins exporting, regardless of the relative profitability of foreign markets. Thus, even if foreign markets are considerably less profitable than domestic markets, the rate of entry into new markets will not take a discrete fall when the firm begins to expand abroad. Rather, the firm's growth in the period before it exhausts all domestic markets is entirely a function of how quickly it will grow in the period after which it exhausts all domestic markets. This is due to the forward-looking nature of firms. The rate of entry into new markets depends exclusively on the profitability of markets that have yet to be entered. If all but a few domestic markets have been entered, then the speed at which the firm will enter those few remaining domestic markets will be almost entirely determined by the profitability of foreign markets. We can show this formally by setting \( t = \tau_h \) and dividing equation (53) by (35):

\[
\frac{u(\tau_h - \varepsilon)}{u(\tau_h + \varepsilon)} = 1 \tag{63}
\]

Thus, there is no discrete fall in the rate at which a firm enters foreign markets, regardless of the relative profitability of foreign markets.

\subsection*{1.4.1 Firm growth and the Profitability of Markets}

It is straightforward to prove that an increase in the profitability of foreign markets (i.e. an increase in \( \Phi \)) will cause the home firm to expand into foreign markets.

\textsuperscript{17} See Sutton (1997) for details about Gibrat's Law.
more quickly.\footnote{An increase in $\Phi$ (arising from a decrease in $\mu$) will increase the profitability of foreign markets only from the standpoint of a home firm. Since foreign consumers will incur lower trade costs when purchasing home goods, they will purchase more of each home variety. This will in turn increase the profitability of foreign markets.} Similarly, an increase in the profitability of \textit{domestic} markets (i.e. an increase in $\Pi_H$) will cause the home firm to expand into \textit{domestic} markets more quickly. This, of course, is not particularly surprising; it is only natural for firms to want to expand into new markets more expeditiously if those markets become more profitable.

What is more surprising is that the speed at which firms grow at \textit{home} is affected by the profitability of \textit{foreign} markets.\footnote{The converse, however, is not true. A change in the profitability of home markets has no effect on how quickly firms grow abroad.} That is, firms will enter domestic markets more quickly if foreign markets become more profitable, even if the profitability of domestic markets does not change. To see this, differentiate equation (57') with respect to $\tau_h$ and $\Phi$:

\begin{equation}
\frac{\partial \tau_h}{\partial \Phi} = \frac{1}{r + \rho} \left[ \frac{e^{-(r+\rho)(\tau_f - \tau_h)} - e^{-(r+\rho)(\tau_f - \tau_h)}}{1 - e^{-(r+\rho)(\tau_f - \tau_h)}} \right] < 0. \footnote{The reason this expression is necessarily negative is due to the convexity of the exponential function.}
\end{equation}

Thus, the time at which the firm exhausts all domestic markets decreases when foreign markets become more profitable (that is, a higher $\Phi$ leads to a lower $\tau_h$). This implies that the speed at which firms grow domestically depends not only on the profitability of domestic markets, but also on the profitability of foreign markets. As a policy experiment, imagine that the home government imposes an export tax or erects some other trade barrier that reduces the profitability of foreign markets. A decrease in the profitability of foreign markets will reduce the speed at which firms expand into \textit{both} domestic and foreign markets. The fact that higher trade barriers cause firms to expand...
into domestic markets more slowly is intriguing since the export tax applies only to goods sold abroad. Intuitively, the explanation hinges on the opportunity cost of slowing the rate of entry into domestic markets. When $\Phi$ is close to one, foreign markets are almost as profitable as domestic markets, and thus, firms will be more anxious to enter these relatively lucrative foreign markets. On the other hand, if foreign markets are very unprofitable, firms will be in no rush to exhaust all domestic markets since the payoff from "going global" will be small.

1.4.2 The Cost of Entering New Markets and its Effect on Firm Growth

The cost of entering new markets depends on the parameter $\alpha$. To determine the effect that a change in $\alpha$ has on the rate at which firms expand into foreign markets, we differentiate equation (57') with respect to $\tau_h$ and $\alpha$:

$$\frac{\partial \tau_f}{\partial \alpha} = \frac{r + \rho}{\Phi \Pi_h} \left[ \frac{1}{1 - e^{-(r+\rho)(\tau_f - \tau_h)}} \right] > 0.$$ (65)

As one might expect, an increase in $\alpha$ (holding $\tau_h$ constant) raises $\tau_f$, implying that it takes longer for the firm to exhaust all available foreign markets. A similar calculation can be made to show that a higher $\alpha$ decreases the rate at which a firm expands into home markets. Thus, as one might intuitively suspect, an increase in $\alpha$ implies that it becomes more costly to enter new markets and hence, firms will find it optimal to enter new markets more slowly.

1.4.3 Country Size and Firm Growth

One of the alleged benefits of trade liberalization is that firms will operate more
efficiently if they have access to larger markets (WTO, 2000). Unfortunately, in a static CES framework, how much a firm produces does not depend on how many competitors it has or on the size of the market area in which the firm operates. Within a dynamic monopolistic model with multiple markets, things are not so simple. The issue of “how big” a firm is depends both on how much it sells in each market and how many markets it services. As the number of markets, \(m\), within a country expands, firms grow more quickly, entering more markets at every point in time. That is, \(u(t)\) depends positively on \(m\). Since firms that grow more quickly will, on average, be bigger, this implies a positive relationship between firm size and country size.

Let us first show that for \(0 \leq t \leq \tau_h\), an increase in the number of markets, \(m\), leads firms to enter domestic markets more quickly (that is, leads a higher \(u(t)\)). To see this, differentiate equation (54) with respect to \(\tau_h\) and \(\tau_f\) and then implicitly differentiate equation (56'), first with respect to \(m\) and \(\tau_h\), and then with respect to \(m\) and \(\tau_f\). This yields:

\[
\frac{\partial u(t)}{\partial m} = \frac{\partial u(t)}{\partial \tau_h} \frac{\partial \tau_h}{\partial m} + \frac{\partial u(t)}{\partial \tau_f} \frac{\partial \tau_f}{\partial m}
\]

\[
(66)
\]

\[
= \frac{(r + \rho)(1 - \Phi)e^{-(r+\rho)(\tau_f - \tau_h)}}{1 - [e^{-(r+\rho)\tau_h} + \Phi e^{-(r+\rho)(\tau_f - \tau_h)}]} + \frac{e^{-(r+\rho)(\tau_f - \tau)}}{e^{-(r+\rho)(\tau_f - \tau_h)} - e^{-(r+\rho)\tau_f}} > 0
\]

for \(0 \leq t \leq \tau_h\).

---

21 On its homepage, the WTO matter-of-factly asserts, “Most firms recognize that the bigger the market the greater their potential — they can expand until they are at their most efficient size, and they can have access to large numbers of customers.” For instance, in the U.S., a large open economy, only 3 percent of firms have between 1 and 9 employees while 85 percent of firms have more than 49 employees. In India, a large but poor, ethnically fragmented, and relatively closed economy, 77 percent of firms have between 1 and 9 employees while only 16 percent of firms have more than 49 employees (Tybout, 2000).
Similarly, by differentiating equation (33) with respect to $\tau_f$ and then implicitly differentiate equation (57') with respect to $m$ and $\tau_f$, we can show that $u(t)$ rises when $m$ rises between $\tau_h \leq t \leq \tau_f$.

\[
\frac{du(t)}{dm} = \frac{du(t)}{d\tau_f} \frac{d\tau_f}{dm} = \frac{(r + \rho)e^{-\tau_f}(r + \rho)(\tau_f - t)}{1 - e^{-\tau_f}(r + \rho)(\tau_f - \tau_h)} > 0
\]

This is an important result for it predicts that firms in large countries should grow more quickly than firms in small countries. To understand this point, let us imagine that home and foreign are two regions of the same country. As a mental experiment, one might ask: what would happen if the number of markets within home and foreign doubled? As this section makes clear, a doubling of $m$ would cause firms to enter new markets more quickly. This implies that trade liberalization, by expanding the number of new markets that firms can enter, should lead firms to grow more rapidly. This is a novel result and one that can be derived only within the sort of dynamic framework that this model develops.

### 1.4.4 The Superior Performance of Exporters

As Bernard and Jensen (1997) note, exporters are more efficient than non-exporters. The bulk of the evidence, however, suggests that exporting does not make firms more efficient. Yet, if exporting does not make firms more efficient, what does? This model provides one plausible answer—firm age. Older firms have had time to establish a presence in many markets and hence, they tend to produce a lot [Kumar (1985), Evans (1987a)]. By calculating a firm's average cost, we see that there is a positive correlation between firm age and efficiency.

---

22 We do not have to worry about the indirect effect on $m$ through a change in $\tau_h$ since $\tau_f - \tau_h$ is independent of $\tau_h$ (i.e. $\tau_f$ changes one for one with $\tau_h$).
Equation (68) can be re-written as:

\[
AC(t) = \frac{TC(t)}{x_{h,h}y_h(t) + x_{h,f}y_f(t)} = \frac{\beta(x_{h,h}y_h(t) + x_{h,f}y_f(t)) + F + \frac{1}{2} \alpha u(t)^2}{x_{h,h}y_h(t) + x_{h,f}y_f(t)}.
\]

As firms age, they become more efficient to the extent that their average costs fall over time. This occurs for the following three reasons: first, big firms can spread out fixed costs over more output, thus lowering their average costs. This reduces the second term in equation (68'). Second, big firms benefit from the investment they previously made to capture new markets. Thus, unlike new firms that service only a few markets, big firms will earn revenue from a larger number of markets. This increases the denominator in the third term of equation (68'), thus decreasing average costs. Third, big firms grow more slowly (in the sense that they enter fewer new markets every period), and as result, incur lower expansion costs. This reduces the numerator in the third term. Thus, old firms are more efficient than young firms and since only old firms (older than age \( \tau_h \)) export, exporters will be more efficient than non-exporters.

### 1.4.5 The Border Effects Puzzle

A key stylized fact that has emerged in the empirical trade literature is the large degree to which national borders impede trade\(^{23}\) [Trefler (1995), McCallum (1995), Wei

---

\(^{23}\) Trefler (1995) argues that the "missing trade" is most likely the result of a home consumption bias coupled with international technology differences. This paper, in contrast, generates the prediction that border effects will arise even if there is no home consumption bias and all countries have the same technology. To the extent that countries do have domestic consumption biases, this would amplify the border effects that already exist.
(1996), Helliwell (1997), Head and Reis (1998). This result is especially surprising given the popular belief that we live in a global village where national borders have lost much of their significance. 24 The errant conviction is nurtured by the largely accurate belief that trade costs are fairly small across countries due to the telecommunications revolution, innovations in transportation, significant progress in trade liberalization, and reduced cultural and political frictions among developed nations. 25

This paper offers a story to help resolve the border effect puzzle. When firms face expansion costs, they will be selective about which markets they enter. If it proves too costly to enter many markets simultaneously, a firm will enter only the most profitable markets. If home markets are only marginally more profitable than foreign markets, this will lead firms to shun foreign markets altogether, thus causing a border effect. Thus, the border can choke off a large fraction of trade despite very low trade costs.

To quantify the border effect in this model, we need to measure the ratio of exports to total output that a typical firm will generate over the course of its life. Some firms, if their varieties become obsolete before time $\tau_h$, will have no exports while other firms, who are lucky enough to avoid going out of business before time $\tau_f$, will eventually earn half of their revenues from foreign markets. The probability that a firm will one day become an exporter is $e^{-\rho \tau_h}$. If the number of domestic markets is large, firms face a high probability of going out of business, and it is costly to enter new markets, this number may be very low. The probability that a firm will eventually grow so large that it exhausts both domestic and foreign markets is $e^{-\rho \tau_f}$. Again, if the number of foreign markets is large and it takes a long time to exhaust all domestic markets, this number likely will be very low.

24 See Greider (1996) for an example of how the popular press simply assumes the world is already tightly integrated.
25 Tariffs on industrial products have fallen steeply since the Second World War. By January 1999, tariffs in industrial countries averaged less than 4% WTO (2000).
When a firm begins operations at time zero, the expected value of exports (measured in terms of output sold) as a proportion of total output (from both domestic and foreign sales) is

\[
\frac{\tau_f}{\tau_f} \int e^{-\rho t} x_{h,f} y_f (t) dt + \int e^{-\rho t} x_{h,f} mdt \\
\int e^{-\rho t} x_{h,h} y_h (t) dt + \int e^{-\rho t} x_{h,h} mdt \\
0
\]

In a standard monopolistic trade model in which trade costs are infinitesimally small, consumers have no demand bias towards locally made varieties, and both countries are the same size, the fraction of exports to GDP would equal $\frac{1}{2}$. In the equation above this would only be true if either $\rho = 0$ (implying that the economy would be swamped with huge firms that have exhausted all home and foreign markets) or $\tau_h = 0$ (implying that firms instantaneously exhaust all home markets). As long as neither restriction holds, a large border effect will exist, even for infinitesimally small trade costs. Furthermore, note that border effects can still be sizable even if the elasticity of demand across varieties and trade costs are both extremely low. This result should be seen as complementary to Obstfeld and Rogoff (2000) who show within a theoretical model that reasonably low trade costs (in the order of 5 percent) and an elasticity of important demand of around 5 can generate the sort of border effects that we observe in the data.\(^26\)

1.5 Conclusions

The predictions of the model coincide nicely with Bernard, Eaton, Jensen and Kortrum's (2000) observation that there is i) considerable productivity dispersion among plants, that ii) exporters are more productive and iii) much larger than non-exporters and that iv) the fraction of firms that export is small and v) even among those that do export,
the fraction of revenues earned from exporting tends to be small as well. The model also theoretical underpinning for Evans' (1987a and 1987b) observation that larger firms tend to grow more slowly than smaller firms. Additionally, the model offers a useful re-interpretation of the "border effects" and "missing trade" puzzles. Lastly, the model makes two interesting predictions: That firms who have access to more markets will grow more quickly and that the rate at which firms grow domestically will increase if foreign markets become more profitable (for instance, as a result of trade liberalization), even if the profitability of domestic markets stays the same.

These observations result from the way firms evolve. When firms are young, they grow quickly, spending vast sums of money to establish a presence in new markets (Evans 1985a and 1985b). This leads to short-term losses. As a result, young firms tend to be less productive than old firms. As a firm gets older, it gets bigger. If technology is subject to increasing returns to scale, this will imply that older firms will produce more output over which to spread out their fixed costs, thus lowering their average cost and raising their productivity. Furthermore, older firms will benefit more from the prior investment they made to capture new markets and since older firms grow more slowly, they will also incur lower expansion costs. This leads to the productivity dispersion across firms: older firms are more productive than younger firms and since there will always be a mix of young firms and old firms in an economy, this will lead to a dispersion of productivity levels across firms.

The fact that exporters are more productive than non-exporters results from the fact that only older firms export. Young firms are too busy penetrating domestic markets to worry about foreign markets, even if foreign markets are only marginally less profitable than domestic markets. As long as firms must incur expansion costs when they enter new markets, this will make them selective about which markets they enter. If a foreign market is even slightly less profitable than a domestic market but it costs too much to enter both markets simultaneously, the firm will shun the foreign market. This

---

Trefler and Lai, using panel data from 28 industries in 36 countries, estimate an average elasticity of demand in the neighbourhood of 5 to 6.
creates a domestic bias since firms only service foreign markets if they have exhausted all their domestic markets; most of the goods that will be available to consumers will be domestically produced. This helps resolve the border effect puzzle by offering an explanation for why national borders choke off so much trade. The answer is that most foreign goods are simply not available to domestic residents and as a result, most of the goods consumed will be locally produced.

In big countries with many markets, there will be a high likelihood that most firms will never become exporters since their variety will become obsolete before the firm has a chance to exhaust all domestic markets. This explains why so few firms export; the vast majority of firms never outgrow the domestic economy. Even among those firms that do export, there will always be a large fraction that are just beginning their trek into foreign markets. Only the largest and oldest firms will exhaust all potential domestic and foreign markets. If the number of potential markets is large and $\rho$ is high, very few firms will ever evolve to the point that half of their revenues come from foreign markets.

The model's prediction that firms that have access to more markets will grow more quickly stems from the fact that the opportunity cost of slowing down the rate at which a firm grows is larger when the firm has access to many markets. If you find yourself mining for gold you will most likely mine more quickly if there is a lot of gold in the ground (even if you are not worried about competitors). If there is a lot of gold in the ground, then the slower you mine it, the more you lose by not taking it to market earlier. If there is only a small amount of gold in the ground, there is a greater incentive to relax, to not exhaust yourself, since it will not take you long to mine all the available gold. Thus, to the extent that policy makers believe that "growth" is good thing, trade liberalization, by giving firms access to more markets, will lead them to grow more quickly, potentially stimulating employment and heightening competition in the process.

The model's second prediction, that firms will expand more quickly into foreign and domestic markets if the profitability of foreign markets rises, stems from the forward-
looking nature of firms. If foreign markets become more lucrative, firms will want to exhaust domestic markets faster in order to begin expanding abroad. Thus, firms will grow more quickly if they have access to more markets and firms will grow more quickly if those markets become more profitable. Both effects serve to increase the average size of firms within the economy. Since large firms are more efficient firms, this implies that trade liberalization, by either making foreign markets more profitable or by giving domestic firms access to new markets, will increase the average level of productivity in the economy.\footnote{This result has been validated in a number of empirical studies. Trefler (1999), in his analysis of the Canada-U.S. Free Trade Agreement (covering 1988 to 1996), concludes that the FTA raised labour productivity by 2.1\% per year in industries that experienced the largest tariff reductions and by 0.6\% for the manufacturing sector as a whole. As is consistent with other studies [Bernard and Jensen (1999), for instance], most of the productivity gain (about 80\%) was within plant. Furthermore, Head and Ries (1999) find that the FTA and subsequent efforts to liberalize trade between Canada and the U.S. increased output per firm in Canada by 24\% between 1988-94. They conclude that trade liberalization induced "substantial rationalization" in Canadian manufacturing. Similarly, Tybout and Westbrook (1995), looking at Mexican plant-level data during the trade liberalization phase between 1984 and 1990, conclude that average costs fell by 1.1\% per year and the fall in average costs was most pronounced in relatively open sectors.}
Chapter 2

Why isn't the Whole World Rich? Trade and Inequality in a Model with Oligopolistic Competition
2.1 Introduction

The idea that location choice and market structure play a vital role in economic development was first articulated by Weber [1909], Young [1928], Christaller [1933], Lösch [1940], Rosenstein-Rodan [1943], and Harris [1954] and codified in the influential works of Myrdal [1957], who developed the idea of ‘circular causation’ and Hirschman [1958], who coined the terms “backward” and “forward” linkages. Yet, after the mid-1960s, economists virtually abandoned the study of location choice and spatial agglomeration and it was not until the late 1980s, led by the pioneering efforts of Krugman [1991a], that economic geography began to enjoy a renaissance of new interest.

Models of economic geography, almost without exception, rely on the presumption of increasing returns. To model increasing returns, trade economists have adopted a variety of technical tools from the industrial organization literature. Unfortunately, these new tools are not without their drawbacks. One of the key "modeling tricks", as Krugman [1998] calls them, of the new trade theory and the new economic geography is the use of the Dixit-Stiglitz [1977] monopolistic competition model. The Dixit-Stiglitz framework is extremely tractable, a desirable quality considering that even relatively simple economic geography models often require numerical simulations to produce any meaningful results. Unfortunately, as Krugman [1998, p. 164] puts it, "...Dixit-Stiglitz is a very restrictive, indeed in some respects, silly model. Above all, the assumed symmetry among varieties, and the resulting absence both of monopoly rents in equilibrium and of any strategic behavior by firms, mean that Dixit-Stiglitz analyses undoubtedly miss much of what really happens in imperfectly competitive industries."

The absence of strategic interaction in the Dixit-Stiglitz framework is lamentable, since as this chapter will argue, strategic interaction is not only ubiquitous, but is itself a vital source of agglomeration. Moreover, the original pioneers of economic geography
did not have the Dixit-Stiglitz story, which hinges on variety effects, in mind when discussing the reasons why firms agglomerate. Rather, what they had in mind was that strategic interaction among a large number of suppliers would reduce price-cost margins and attract enough new customers to a location that this would make the location an even more attraction place for firms to locate. Unfortunately, the Dixit-Stiglitz monopolistic framework is largely silent on the issue of price-cost margins since the assumption that the elasticity of substitution across varieties is constant implies that firms will charge a constant markup over marginal cost regardless of the scale at which they operate or how many competitors they have.

To a great extent, therefore, the aim of this chapter is to formally model the original story of agglomeration in which strategic interactions are at the fore of analysis. Yet, this chapter is much more than an attempt to mathematically express what was verbally articulated 40 years ago. The early models of economic geography suffered from two major shortcomings. First, they were expressed within a partial equilibrium framework, and in many cases, not even that. Thus, the effect on wages, prices, and market shares was often ignored and adding up constraints overlooked. Second, although the original authors were not explicit in pointing this out, the earlier works tended to describe planning solutions, not market outcomes. Thus, for instance, Lösch’s [1940]

---

28 Scotchmer and Thisse [1992] call this the ‘folk theorem of spatial economics’. In the absence of increasing returns, location choice ceases to be an issue; firms will simply build small enough plants to service every region, thereby eliminating transport costs.

29 One classic model in which strategic interaction figured heavily is Hotelling’s [1929] “linear city model”. Hotelling showed that if there are only two firms, they will locate side by side and evenly split the market between them, a result referred to as the Principle of Minimum Differentiation. Hotelling’s model, although extremely insightful, is not a true model of agglomeration. Agglomeration necessarily involves “cumulative causation”: the idea that if a critical mass of firms gather in a certain location, that this will entice more firms to move to this location, thereby creating an even greater incentive for other firms to move to this location, and so on. This notion of agglomeration is absent in Hotelling’s model. Furthermore, recent work has cast doubt on Hotelling’s results. Osborne and Pitchik [1987], by allowing firms to choose mixed strategies over prices, derive equilibria for the Hotelling game in which firms locate away from the center. D’Aspremont, Gabszewicz and Thisse [1983] show that a “side by side” equilibrium is impossible to sustain because this equilibrium entails zero profits (Bertrand competition will force the price down to marginal cost) and a firm will always be able to make positive profits by defecting away to an isolated location. Furthermore, D’Aspremont et al show that in a two-stage game in which there is quadratic transport costs and firms first choose a location and then compete on price, the best strategy is to move away from competitors in order to mitigate price competition.
argument that a hexagonal layout would represent the most efficient layout of a city core assumed that central planners could design the city as they saw fit. Whether such a layout would naturally result if firms and workers were allowed to pursue their own self-interest was left unanswered.

One of the advantages of using a general equilibrium framework is that we can deal directly with issues of economic welfare, and specifically the issue of what effect economic integration will have on wages. Back in the sixties and seventies, critics of the American establishment often argued that the wealth of the developed world had come at the expense of the less developed countries in Africa, Asia, and Latin America. We were rich because they were poor, so the argument went. It therefore followed that the only viable solution for Third World countries was to “de-link” themselves from the rich capitalist world and pursue economic development on their own terms.

By the late eighties, this argument was supplanted by a rather different, and seemingly contradictory, fear. Led by such populist leaders like Ross Perot, Richard Gephart, and Pat Buchanan, the new foes of free trade argued that workers in the United States were being impoverished by low wage competition from less developed countries. Suddenly, it was no longer the North that was getting rich at the expense of the South; rather, it was the South that was getting rich at the expense of the North.

To most economists, this may seem like yet another example of muddled thinking; evidence the most people have little grasp of basic economic issues, let alone such esoteric principles like comparative advantage. Yet, as this chapter will demonstrate, both sets of arguments may contain grains of truth. I will argue that it is possible to construct a fully specified general equilibrium model where globalization inevitably leads to a world of rich and poor countries. The first phase of globalization is marked by even development where all countries grow at the same rate. The next phase is characterized by divergence. Some countries become industrialized and rich while others become deindustrialized and poor. This occurs not because some countries have intrinsic advantages over other countries nor because there is some sort of coordination
failure in which rich countries are able to gravitate towards superior equilibria that poor countries, for whatever reason, fail to reach (as in Murphy, Shleifer, and Vishny [1989]; Matsuyama [1991, 1992], and Rodrik [1996]). Rather, in my model, global inequality is an unavoidable consequence of the world trading system.

When transport costs fall to a sufficiently low level, international trade will cause agglomeration of manufacturing and intermediate services within one region (the "core"), and that region will become rich at the expense of the other region (the "periphery"). Yet, as the world becomes more integrated, a third period of economic convergence ensues. The core is forced to compete more intensely with periphery. This depresses the wages paid to workers in the core, giving credence to Ross Perot's argument that the United States will be forced to cut wages or risk hearing the "giant sucking sound" of manufacturing jobs moving to Mexico.

An ancestor to the model presented in this chapter is Krugman and Venables [1995]. Although Krugman and Venables generate similar predictions to my model, the model presented here is more forthright in the way these predictions are derived. Since Krugman and Venables do not model strategic interactions among firms, they are forced to introduce a fairly complicated cost linkage, based on variety effects via intermediate goods, as a way to induce agglomeration. This chapter goes a step further by embedding strategic interactions within the model. This is not a trivial extension. Monopolistic competition, in and of itself, is incapable of generating agglomeration. Consequently, one cannot regard monopolistic competition as a "reason" for why agglomeration takes place. Strategic interaction, however, does constitute a reason for why we observe agglomeration. As a result, there is no need to introduce complicated cost linkages since the simple presence of oligopolistic competition is enough to generate agglomeration. This makes modeling agglomeration easier and more intuitively appealing.

One insightful model that does not overlook strategic interaction is Combes [1997]. Combes shows that when firms compete through Cournot competition, then agglomeration may ensue. There are two competing forces at work in Combes' model.
When more firms are added to a region, this increases price competition and also reduces
the market share of all firms, leading to lower profits. This force works against
agglomeration. However, the region with the higher number of firms will also have
more income and since firms produce more for the local market than they export abroad,
this will help raise expenditure on the products of firms in that region. This force works
to induce agglomeration.

Yet, even in Combes' model, wages are fixed, even in the long run. Furthermore,
there is no inter-industry labour mobility; manufacturing workers cannot become farmers
or vice-versa. The assumption of fixed wages is useful, especially in Combes' model
since his desire is to study the relatively rigid labour markets of the European
Community. However, the assumption of fixed wages makes it difficult to say much
about the welfare implications of asymmetric regional development. For instance, will
wage inequality increase or decrease when agglomeration occurs due to strategic
interaction amongst firms? Clearly, to answer this question, we need a model where
wages adjust in response to changes in labour demand.

2.2 The Basic Story

Before formally outlining the model, let me paint a basic picture of my story. I
imagine a world in which there are two regions and three sectors: an agricultural sector, a
manufacturing sector, and an intermediate services sector. Manufacturing firms use a
basket of intermediate services (such as accounting, information technology, marketing,
human resource management, etc.) as an input. I assume that intermediate service
providers are quantity-setting Cournot oligopolists. This may seem a little strange since
the pricing decision of a service provider is usually more involved than simply deciding
how much of the service to supply on the market. The assumption of quantity setting,
however, is an innocuous abstraction. Kreps and Scheinkman [1983] show that a more
complicated model where firms first decide the capacity at which they want to operate
and then decide what price to charge for the goods or services that they provide is
equivalent to a simple Cournot quantity setting game. Thus, for instance, a model where accounting firms choose the quantity of accounting services to provide on the market generates the same predictions as a more complicated model where accounting firms first decide how many accountants to employ (i.e. choose capacity) and then decide what price to charge for accounting services.

Suppose that both regions are initially identical, but then, for whatever reason, there is a slight relocation of manufacturing from one region to the other region. The region that ends up with the higher share of manufacturing will demand more intermediate services, causing new service providers to enter the market. This will make the service sector more competitive, causing the price of services to fall. This, in turn, will reduce the costs of manufacturing firms and as a result, entice more manufacturing firms to move to the region with the larger manufacturing base. This "cost effect", as I call it, serves to induce agglomeration.

Working against the cost effect is the "demand effect". As more manufacturing firms enter a region, they are forced to compete more intensely for a slice of the market, thereby decreasing revenues. This dampens the desire for new manufacturing firms to move to the region with the bigger manufacturing base. Hence, the demand effect serves to impede agglomeration. As it turns out, which effect dominates depends crucially on the level of transport costs. When transport costs are high, the demand effect will prevail and so, there will be "even development". Yet, in a global economy that is becoming increasingly more integrated, transport costs will eventually fall to a sufficiently low level and when this happens, the cost effect will dominate and agglomeration will occur. The world then splits into a core and a periphery. The living standard in the core will rise and the living standard in the periphery will fall. This phase of economic development is marked by economic divergence. As transport costs continue to fall, however, a new era is ushered in. Now, the living standard in the two regions begins to converge. Workers in the core must now compete more intensely with the low-wage workers in the periphery. This serves to decrease the relative wage of workers in the core.
2.3 A Formal Model

Let us formally develop a general equilibrium model in which all prices and wages are flexible and workers are free to move among different sectors of the economy. It will be shown that Cournot competition will induce agglomeration if trade costs are sufficiently low. Imagine a two-region world. Although both regions are identical the forces of agglomeration may dictate that one region will become highly industrialized (i.e. the core) while the other region becomes deindustrialized (i.e. the periphery). Let us call the region that becomes the core “North”, and the region that becomes the periphery “South”. There are three sectors: the first sector produces food; the second sector produces differentiated manufactured goods, and the third sector produces a homogenous intermediate service good. Food is produced under constant returns to scale using labour. Manufactured goods are produced by monopolistically competitive firms under increasing returns to scale. Each firm produces a different variety of a manufactured good. Manufacturing firms employ both labour and intermediate services using a Cobb-Douglas production function. Intermediate services are produced under increasing returns to scale using labour as the only input. I assume that service providers behave as quantity setting Cournot oligopolists. Firms in all three sectors are subject to a zero profit condition. I assume that food can be freely traded, manufactured goods can be traded but are subject to iceberg trade costs and intermediate services cannot be traded at all. 

Workers in both regions consume food and manufactured goods. Lastly, each worker inelastically supplies one unit of labour. I assume that labour is perfectly mobile between sectors within a region but not between regions. This implies that the wage

---

30 The assumption that services cannot be traded across regions, while intuitive in its own right, also helps to greatly simplify the model. If Northern service providers could sell their services in South, they would then have to consider not only the impact that their output choice would have on the domestic price of services, but also what impact this would have on Southern service providers. Even in a symmetric equilibrium, this would lead to reciprocal dumping where North would sell some services to South and South would sell some services to North. In one sense, this would be an inefficient waste of resources since transport costs would be incurred on the services imported (by assumption there is no difference between Northern and Southern services so there cannot be any benefits from increasing variety). On the other hand, the possibility of reciprocal dumping would force the oligopolists to compete more intensively, thereby reducing the cost of services and decreasing the average cost of service providers (see Brander and Krugman [1983] for an exposition of the reciprocal dumping model). In any event, the possibility of reciprocal dumping would great complicate the analytics of the model.
earned by all workers within a region must be the same although the wage may differ between regions. Let us now examine each industry in turn. I will focus on the economy in North, but since the regions have the same endowments and technology, the analysis is the same for South. Denote all South variables with an asterisk.

2.3.1 Agricultural Sector

Agriculture is produced under constant returns to scale using only labour. Without loss of generality, assume that one unit of labour is necessary to produce one unit of food. Let the price of food be the numeraire (i.e. \( p_a = 1 \)). Zero profits in the agriculture industry imply that the cost of producing one unit of food must equal the price of food. Hence, if North produces any food, it must be true that the wage, \( w \), is equal to one. If North produces no food and imports all its food from South, the wage in North can exceed one. Thus, in general,

\[
(2.1) \quad w \geq 1. 
\]

2.3.2 Manufacturing Sector

Firms in the manufacturing sector produce differentiated varieties using labour and intermediate services via a Cobb-Douglas production function. The price of each variety is \( p_m \). Let \( \mu \) be the share of total cost devoted to services and \( p_s \) be the price of services. Let \( x \) be the quantity of the variety sold in North and \( y \) be quantity of the variety exported to South. The firm requires \( \alpha \) units of the input to cover fixed costs and \( \beta \) units of the input to produce every additional unit of output thereafter. The profit function for a representative manufacturing firm is therefore,

\[
(2.2) \quad \Pi_m = p_m (x + y) - p_s^\mu w^{1-\mu} [\alpha + \beta (x + y)].
\]
Define $\sigma$ as the elasticity of substitution between varieties. Differentiating equation (2.2) with respect to either $x$ or $y$ yields the profit maximizing price for a variety:

\begin{equation}
    p_m = \frac{\sigma}{\sigma - 1} \beta \rho^\mu_x w^{1-\mu}.
\end{equation}

As is always the case in models of this form, the zero-profit condition (found by setting equation (2.2) equal to zero) implies that firm size is constant and equal to

\begin{equation}
    x + y = \frac{\alpha(\sigma - 1)}{\beta}.
\end{equation}

Consumers have Cobb-Douglas preferences over agriculture and a composite manufactured good and CES preferences over the various varieties that comprise the composite manufactured good. This entails that they follow a two-stage budgeting process. In the first stage, consumers decide what fraction of their incomes to allocate on manufacturing and in the second stage, having decided how much to spend on manufacturing, they then decide how to allocate this money on the various varieties of manufactured goods available. Denote $n_m$ and $n_m^*$ as the number of varieties produced in North and South respectively. Furthermore, following Samuelson [1954], define $\tau$ as the iceberg cost of transporting manufacturing goods between regions (i.e. $\tau$ units of the variety must leave the port so that one unit arrives at the destination).\(^{31}\)

---

\(^{31}\) The idea that some fraction of the good melts in transit (or alternatively, some fraction of the good must be used in its own transportation (i.e. to ship coal, you need to burn coal)) is a common feature in the new economic geography literature. The assumption of iceberg transport costs is convenient for it obviates the need to model a transportation sector. Unfortunately, the assumption is not entirely innocuous. Since a constant fraction of the good melts in transit, this implies that the monetary cost of transporting a good rises if its price rises. This tends to bias prices downward since exporters can artificially reduce transport costs simply by selling their goods at a lower price. To remedy this shortcoming, Ottaviano and Thisse [1999] propose a model in which transportation is treated as a separate sector that uses resources other than the transported good.
It is straightforward to show that the price of the composite manufactured good is\(^{32}\)

\[
Q_m = \left[ n_m p_m^{1-\sigma} + n_m^* (\varphi_m^*)^{1-\sigma} \right]^{1/(1-\sigma)}.
\]

Assume that workers devote a fraction \( \gamma \) of their budgets on the composite manufactured good and the remainder they spend on agriculture. The aggregate demand functions for a manufactured variety are then:

\[
x = \frac{p_m^{-\sigma} \gamma \nu L}{n_m p_m^{1-\sigma} + n_m^* (\varphi_m^*)^{1-\sigma}} \quad \text{and} \quad y = \frac{p_m^{-\sigma} \gamma \nu^* L \tau^{1-\sigma}}{n_m^* p_m^{1-\sigma} + n_m (\varphi_m)^{1-\sigma}}.
\]

### 2.3.3 Intermediate Services Sector

The intermediate services sector produces a homogenous service (think of it as some basket of accounting, marketing, IT, human resource, and financial services) and is composed of \( n_s \) Cournot quantity setting oligopolists. Each service provider takes her rival's output as given but knows that by changing her own output, she can affect the market price of services.

The profit function for a service provider is

\[
\Pi_s = p_s s - w(\theta + \delta s),
\]

where \( \theta \) is the fixed cost of providing a service provider, \( \delta \) is the marginal cost of

---

\(^{32}\) The price of the composite manufactured good is derived from the expenditure function in the second stage of the consumer's maximization problem. The price is equal to how much expenditure is necessary to yield one unit of utility. It might be helpful to think of a variety as some part of a computer (i.e. the hard drive, monitor, CPU) and the composite manufactured good as the computer when it has been fully assembled. The price of each component in the computer is \( p_m \) and the price of the computer itself is \( Q_m \).
providing each unit of services, $p_s$ is the price of the services and $s$ is the quantity of services produced by each firm. Note that the price of services depends on the total quantity of services produced, $S = \sum^n s$, where $n_s$ is the number of service providers.

Total expenditure, $E$, on services in North is

$$E = p_s S. \tag{2.8}$$

To make things more tractable, I assume that service providers take $n_m$ and $n_s$, the number of firms in the manufacturing and service sectors, and all prices (except the price of services, of course) as given. Substituting equation (2.8) into equation (2.7), differentiating with respect to $s$, and invoking the symmetry condition $S = n_s s$, we can solve for the output of each service provider:

$$s = \frac{E}{wS} \left( \frac{n_s - 1}{n_s^2} \right). \tag{2.9}$$

The total quantity of services produced is

$$S = n_s s = \frac{E}{wS} \left( \frac{n_s - 1}{n_s} \right). \tag{2.10}$$

Hence, the price of services is

$$p_s = \frac{E}{S} = \frac{wS n_s}{n_s - 1}. \tag{2.11}$$
Using the zero profit condition for service providers (obtained by setting equation (2.7) equal to zero) and equation (2.11), we can re-write equation (2.9) as a function only of the number of firms in the service sector,

\[(2.9')\]
\[s = \frac{\theta}{\delta}(n_s - 1).\]

Equation (2.9') tells us that the more service providers there are, the more services each firm will produce. Since marginal costs are constant and each firm incurs a fixed cost, this implies that the average cost curve for a service provider is downward sloping. Yet, the zero profit condition requires that price of services be equal to the average cost of producing services. This implies that, all things equal, service providers in large markets will operate at a more efficient scale, despite having access to the same technology as service providers in small markets. The fact that the price-cost markup of service providers decreases as the number of service providers grows can be readily seen by re-arranging and differentiating equation (2.11):

\[(2.13)\]
\[
\frac{\partial}{\partial n_s} \left( \frac{p_s}{w \delta} \right) = -\frac{1}{(n_s - 1)^2} < 0.
\]

Intuitively, what happens is that as the number of service providers grows, they compete more intensely with one another. Naturally, this reduces price-cost margins, which, in turn, forces each firm to expand output to cover its fixed costs. Note that this result stems from the assumption that service providers interact strategically via Cournot competition. We could not obtain this result within the Dixit-Stiglitz monopolistic competition framework because there, the assumption that the elasticity of substitution is constant, implies that the optimal price-cost markup does not vary with firm size or the number of firms in the industry.
A further feature is worth highlighting: The ratio $\frac{\theta}{\delta}$ in equation (2.9') can be interpreted as a measure of increasing returns since it measures the ratio of fixed to marginal costs of service providers. The higher is this ratio, the more pronounced are increasing returns. Thus, we see that equilibrium firm size increases when increasing returns become more pronounced (higher $\theta$) and when the number of firms in the industry expands.

Using equation (2.9), equation (2.11) and the zero profit condition, we can solve for the equilibrium number of service providers:

$$n_s = \sqrt{\frac{E}{w\theta}}.$$  \hspace{1cm} (2.12)

Equation (2.12) tells us that the number of service providers grows when there is an increase in expenditure on services. Moreover, since the number of service providers is a function of the square root of expenditure on services, a $k$ percent increase in expenditure will result in a less than $k$ percent increase in the number of firms. Thus, an increase in expenditure on services will result in an increase in the number of service providers and an increase in the scale at which each service provider operates. This, in fact, is the channel through which agglomeration occurs. If there is a slight relocation of manufacturing firms from the one region to another, the region with the larger share of manufacturing firms will require more intermediate services, thus encouraging more service providers to enter the market. This will bid down the price of services, thus reducing the cost of manufacturers. If trade costs are sufficiently low, this will, in turn, make manufacturing more profitable and will entice more manufacturing firms to move into the region with the bigger manufacturing base.

Examples of such real-world interactions are not hard to come by. For example, Anna Lee Saxenian, in an analysis of agglomeration patterns in Silicon Valley, interviewed a variety of senior managers to find out why their firms had decided to
relocate from the Route 128 area in Massachusetts to Silicon Valley. According to Jeffrey Kalb, an engineer who worked for DEC along Route 128 before moving to Silicon Valley to start the MasPar Computer Corporation, commented: 'It's hard for a small company to start in Route 128 because you can't get stuff like IC's and disk drives fast. [...] In Silicon Valley, you can get anything you want on the market. You can get all those things in Route 128 sooner or later, but the decisions are much faster if you're in Silicon Valley.' Furthermore, Tom Furlong, who headed a DEC workstation division in Maynard (Massachusetts) before moving west in 1985 remarked 'The same job of bringing a new workstation to market takes two times as long in the East coast and many more people than it does here. [...] It's easier and cheaper for me to rely on the little companies in Silicon Valley to take care of the things I need, and it forces them to compete and be more efficient.' Both managers stress a common theme: by locating close to your suppliers, you can purchase intermediate goods more quickly and cheaply. This makes relocation more profitable and as a result, helps to induce agglomeration.

2.4 When does agglomeration occur? Stability of the symmetric equilibrium

Let us consider the symmetric equilibrium in which North and South produce the same quantity of food, manufactured goods, and intermediate services and then determine the conditions under which this equilibrium is stable. If the symmetric equilibrium is stable, then a slight relocation of firms from South to North will decrease the profitability of Northern firms and increase the profitability of Southern firms, causing the initial imbalance to be corrected. If, on the other hand, the symmetric equilibrium is unstable, then a slight relocation of manufacturing firms from South to North will make manufacturing more profitable in North, causing even more firms to move to North,
thereby leading to the deindustrialization of South and the creation of a core-periphery.\footnote{Two technical, although admittedly esoteric, issues arise in this analysis: The first issue concerns the precise nature of the initial disturbance that moves the economy away from the symmetric equilibrium. The best way to think about this issue is to suppose that there is a "trembling hand" dynamic at work (see Kreps [1990, pp. 437-441] for details about this game-theoretic concept). That is, suppose that firms occasionally make mistakes by locating in the region that offers a lower profit. In this way, the symmetric equilibrium is perpetually being tested. If the symmetric equilibrium is stable, a slight errant reallocation of manufacturing firms from South to North will decrease profits of Northern firms and consequently, will not attract new firms to migrate from South to North. If the symmetric equilibrium is unstable, then an errant reallocation of firms from South to North will increase the profits of Northern firms, thereby attracting even more firms from South. The second issue concerns the neglected role of expectations in the analysis. This issue is somewhat problematic since there are no explicit dynamics in the model (see Krugman [1991b] for an example of a model where expectations play a crucial role in determining the economic evolution of a region). Fortunately, it is possible to regard the problem as one where agents pick locations as strategies. In this case, the problem becomes not one of modeling static expectations, but rather an exercise in evolutionary game theory. (See Osborne and Rubinstein [1994, pp. 48-51] for an exposition of this modeling technique)) As Krugman [1998, p. 165] cleverly puts it, "to middle-brow modelers like myself, it sometimes seems that the main contribution of evolutionary game theory has been to re-legitimize those little arrows we always wanted to draw on our diagrams."}

Let us denote symmetric equilibrium values with a circumflex. By definition, a symmetric equilibrium requires $\hat{\rho}_m = \hat{\rho}_m^*, \hat{\rho}_s = \hat{\rho}_s^*, \hat{n}_m = \hat{n}_m^*, \hat{n}_s = \hat{n}_s^*$. Food will be produced in both regions in a symmetric equilibrium, implying that $\hat{\omega} = \hat{\omega}^* = 1$. Furthermore, both regions will produce the same number of manufactured varieties. Thus, the total expenditure on manufacturing in North will equal the total revenue of all Northern manufacturing firms, which, through the zero profit condition, is equal to the total cost of all Northern manufacturing firms. Since manufacturing firms employ Cobb-Douglas technology, total expenditure on intermediate services must then be

$$E = \hat{\rho}_s \hat{s} = \mu \hat{n}_m \hat{\rho}_s^\mu \left[\alpha + \beta (\hat{x} + \hat{y})\right] = \mu \hat{n}_m \hat{\rho}_m (\hat{x} + \hat{y}) = \mu \gamma L.$$ \hspace{1cm} (2.14)

The number of service providers is therefore

$$\hat{n}_s = \sqrt{\frac{E}{\theta}} = \sqrt{\frac{\mu \gamma L}{\theta}}.$$ \hspace{1cm} (2.15)

The price of services is\footnote{...}
(2.16) \[ \hat{p}_s = \frac{\delta n_s}{\hat{n}_s - 1} = \frac{\delta}{1 - \sqrt[\gamma]{\mu \lambda}}. \]

The fact that the price of services rises when \( \theta \) and \( \delta \), the fixed and marginal cost respectively of service providers, increases, is not surprising. More revealing is the fact that the price of services falls when either \( \mu \), \( \gamma \), or \( L \) rise. All three parameters are related to the market potential of service providers. The idea of market potential has its roots in the writings of Harris [1954] and Pred [1966]. Harris argued that there is an obvious circularity in regional development. Firms want to be where there are many consumers. But who do consumers tend to be? They are the employers and employees of firms. Thus, firms want to be where there are other firms.

While Harris was more concerned with studying the economy within a static framework, Pred [1966] concentrated on the dynamics of regional growth. Central to Pred’s analysis was the notion of a “base-multiplier.” All first year undergraduates learn that imports constitute a “leakage” in the expenditure cycle. The higher the fraction of disposable income spent on imports, the lower is the spending multiplier. Pred argued that the same argument applies to regions. The more that people in a region spend on goods produced in other regions; the lower will be the “base-multiplier” for that region. Pred pointed out that the marginal propensity to buy goods from other regions is itself endogenous and depends on the size of the region. Big regions will produce a wide variety of goods and as a result, people will have less desire to buy goods from other regions. The fact that the base-multiplier rises as the region expands helps fuel agglomeration. All things equal, bigger regions will spend more internally and thus, a dollar of expenditure in a big region will help expand demand more than a dollar spent in

\[ ^{34} \text{Since the price of services is undefined when there is less than one service provider, we need to assume that the number of workers is sufficiently large so that } L \geq \frac{\theta}{\mu \gamma}. \]
a small region. Hence, firms will find that "market potential" rises when they locate in a larger region.

Looking at our model, we see that the notion of market potential is succinctly captured by $\mu$, $\gamma$ and $L$. A higher $\mu$ implies that manufactures will devote a larger share of their costs on services, thereby increasing the demand for services. A higher $\gamma$ implies that consumers will buy more manufactured goods and less food, which indirectly raises the demand for services (since services are an intermediate good in manufacturing production). Finally, $L$, the number of workers in the region, directly captures the size of the region. As $L$ grows, the region expands, thus increasing the demand for intermediate services. This helps make the service sector more competitive, which, in turn, puts downward pressure on the price of services. When the price of services falls, this decreases the price of manufactured goods and consequently, increases the real wage of workers. This, in turn, further increases the demand for manufactured goods and services.

Let us now complete the derivation of the model. By plugging equation (2.16) into equation (2.3), we can solve for the symmetric equilibrium price of manufactured varieties:

\[
\hat{p}_m = \frac{\sigma}{\sigma - 1} \beta \hat{p}_s^\mu = \frac{\sigma}{\sigma - 1} \beta \left[ \frac{\delta}{1 - \frac{\theta}{\mu \gamma L}} \right]^\mu.
\] (2.17)

Using equation (2.4), equation (2.17), and the fact that $n_m \hat{p}_m (\hat{x} + \hat{y}) = \gamma L$, we can calculate the number of manufactured varieties produced in each region:
Using equation (2.5) and equation (2.18), we see that the price index of manufactures is

\[
\hat{n}_m = \frac{\gamma \lambda}{\alpha \sigma} \left[ 1 - \sqrt{\frac{\theta}{\mu \gamma \lambda}} \right]^{\mu}
\]

(2.18)

Suppose the economy is initially at a symmetric equilibrium. There are two forces that affect the stability of this equilibrium. Suppose there is a slight relocation of manufacturing firms from South to North. Since North will now have more manufacturing firms than South, Northern firms will require more intermediate services. The number of service providers will thus rise in North and fall in South, causing the price of services in North to decrease and the price of services in South to increase. This “cost effect” will increase the profitability of Northern firms, leading to an outflow of manufacturing production from South to North. Hence, the cost effect works to destabilize the symmetric equilibrium and as such, is the driving force behind agglomeration.

Working against the “cost effect” is the “demand effect.” As the number of varieties produced in North rises, Northern consumers end up buying less of each Northern variety. This reduces the profitability of Northern firms and as a result, dampens the incentive to agglomerate. To see this, consider a small relocation of manufacturing firms from South to North. Refer back to equation (2.6); holding the price of manufactured varieties constant and evaluating everything around the symmetric equilibrium, we see that
Putting equation (2.20) and (2.21) together, we get,

\[
\frac{\partial x}{\partial n_m} \bigg|_{dn_m = -dn_m^*} + \frac{\partial y}{\partial n_m} \bigg|_{dn_m = -dn_m^*} = \frac{\alpha \sigma (\sigma - 1) \tau \left(1 - \frac{1}{\tau^2}\right)}{\beta \gamma L \left(1 + \frac{1}{\tau^2}\right)} < 0.
\]

When there is a slight relocation of manufacturing firms from North to South, Northern firms end up selling less of a variety domestically and more of the variety abroad. When trade costs are present (i.e. \(\tau > 0\)), Northern firms will sell the majority of their output domestically and so, the negative effect on domestic sales in equation (2.20) will outweigh the positive effect on exports in equation (2.21). The net result of the "demand effect" is to reduce the profitability of Northern firms, thereby helping to restore the symmetric equilibrium (equation (2.22)).

To illustrate this point, consider the case of New York City. New York City (to be more precise, Manhattan) is the single most important center for financial services in the United States, and probably the world. Why do financial firms want to locate in New York City? Partly because so many of their customers are there and partly because so many suppliers of key intermediate financial services (data vendors, financial
accountants, etc.) are also there. The latter point corresponds to the “cost effect.” Why don’t all financial firms move to New York City then? Largely because not all their customers are in New York City. Since it is important for firms that supply financial services to be close to their customers, some firms will want to locate in smaller markets, despite the fact that doing so would leave them farther away from their suppliers. In other words, if too many financial firms move to New York City, this would saturate the market and reduce each firm’s profits. This, of course, is precisely what the “demand” effect captures.

Let us now turn to the issue of how the relative strength of the demand effect varies with the level of transport costs. A decrease in $\tau$ weakens the demand effect, thereby making agglomeration more likely. At one extreme, when there is complete free trade, the demand effect disappears completely and so, the symmetric equilibrium must necessarily be unstable. This can be readily seen by noting that as $\tau$ falls to one, the derivatives in equation (2.20), (2.21) and (2.22) go to zero. Intuitively, if consumers do not incur trade costs when they purchase a variety produced in another region, they will purchase equal quantities of Northern and Southern varieties. Thus, the

---

35 Clearly, there are other relevant forces at work. Financial firms need highly skilled workers, which they are more likely to find in New York City than in a rural town. At the same time, workers want to be close to firms, both to facilitate the matching of skills with job types (See Helsley and Strange [1990] and Midelfart-Knarvik [1995] for an exposition of this point) and as a way to reduce labour risks. The notion that “labour market pooling” can reduce labour risks and as a result, lead to agglomeration was first articulated by Krugman [1991b] and extended and formalized within a general equilibrium model by Berezin [Chapter 3]. For the same reason that people find it advantageous to hold well diversified portfolios, workers will find it to their advantage to live in a region where there are many firms because then decreases in labour demand by some firms will tend to be counterbalanced by increases in demand by other firms, thus reducing the variability of wages and aggregate employment. Furthermore, the presence of fixed factors (such as land or labour) will dampen the desire to agglomerate for low trade costs (see Ottaviano and Puga [1997] for a discussion of this point). Puga [1999] develops a model in which labour mobility serves to facilitate agglomeration. His main point is that labour mobility serves to increase both local expenditure on goods or services and ease labour market competition in the region receiving the inflow of workers. Both effects make the region a better place for other firms to locate, thus fueling agglomeration.

36 Actually, another effect is present in the model that corresponds to the sort of “backward” linkage described by Hirschman [1958]. If North ends up completely specializing in manufacturing and services, then the wage will rise above one. This will increase the disposable income of workers, allowing them to buy even more manufactured goods. This, in turn, will make manufacturing more profitable, reinforcing the incentive for manufacturing firms to agglomerate in North. However, since the wage in both regions is necessarily equal to one at the symmetric equilibrium, this effect will not play a role until agglomeration has proceeded sufficiently far.
revenue earned from selling a variety does not depend on the region in which the firm is located. The cost of producing the variety, however, does depend on the region in which the firm is located. Since intermediate services are more expensive in the region with the smaller manufacturing base, any relocation of firms from South to North will give further impetuous for firms to leave South and move to North.

Hence, as trade barriers fall over time and the world becomes more closely integrated, the symmetric equilibrium will become less and less stable and eventually, once \( \tau \) has fallen below some critical value, the symmetric equilibrium will crumble and a core-periphery will emerge. Let us define this critical value as \( \bar{\tau} \). For values of \( \tau \) above \( \bar{\tau} \), the symmetric equilibrium is stable; for values of \( \tau \) below \( \bar{\tau} \), the symmetric equilibrium is unstable. Appendix A describes how to calculate \( \bar{\tau} \). Essentially, what is required is to take the equations that characterize the solution to the model, linearize them around the symmetric equilibrium, and then see whether manufacturing profits increase or decrease as we vary the share of manufacturing firms in North.

2.5 Agglomeration and Global Inequality

When economic integration has progressed sufficiently far so that \( \tau \) falls below \( \bar{\tau} \), the symmetric equilibrium will break down and the economy will form a core-periphery; North will become more industrialized while South will become deindustrialized. There are three possibilities for the resulting non-symmetric equilibrium. One, both regions will fully specialize; North will produce only manufacturing and services and South will produce only agriculture. Two, North will continue to produce only manufacturing and services but now, South will also produce some manufacturing and services. Three, North will produce some agriculture and all the manufacturing and services and South will produce only agriculture. Which scenario will occur depends on parameter values, most critically on \( \gamma \), the fraction of income that consumers spend on manufactured goods. If \( \gamma \) is high, both regions will end up having to produce manufactured goods, with North producing no food at all. If \( \gamma \) is low, both
regions will produce some food, with South producing only food. The case where both regions specialize fully in either manufacturing or agriculture occurs for intermediate levels of $\gamma$. Appendix B develops all three cases analytically.

What happens to the wage (measured in terms of the numeraire, agriculture) as trade costs fall over time? To illustrate what happens, let us assume that when the symmetric equilibrium breaks down and a core-periphery emerges, the resulting non-symmetric equilibrium will be one where North fully specializes in manufacturing but South also retains some manufacturing production. Looking at figure 2.1, when trade costs are high (higher than $\tau_{\text{max}}$), the symmetric equilibrium is stable and so the wage in North will equal one. Between $\tau_{\text{max}}$ and $\bar{\tau}$, both equilibriums can exist. If the economy is at a symmetric equilibrium, it will stay there. If the economy is at the non-symmetric equilibrium in which North produces no food, it will stay there too. Both equilibria are stable. Which equilibrium will prevail will depend on initial conditions. After trade costs fall below $\bar{\tau}$, the symmetric equilibrium will necessarily break down and North will specialize exclusively in manufacturing. Since North produces no food, the wage that Northern workers earn will exceed one. As figure 2.1 shows, the equilibrium wage depends on the level of trade barriers. As trade barriers continue to fall, South will demand more of each Northern variety, thus allowing the wage in North to rise. However, as trade barriers continue to fall, Northern manufactures will compete more intensively with Southern firms. Since Southern firms pay their workers a wage equal to one, this puts downward pressure on the wage that Northern firms can pay. Thus, as trade costs continue to fall, low wages in the South will eventually depress wages in the North.

Figures 2.2 and 2.3 depict the real wage (measured in terms of both agriculture and the composite manufactured good) in North and South as a function of trade costs. Again, imagine a world where trade costs are falling through time. When trade barriers fall below $\bar{\tau}$, the symmetric equilibrium will become unstable and the economy will gravitate towards the non-symmetric equilibrium. At this point, the real wage in North rises and the real wage in South falls, leading to what Baldwin, Martin, and Ottaviano
[1998] describe as the "virtuous cycle" of rising living standards in the core and the "vicious cycle" of falling living standards in the periphery.\textsuperscript{37} This divergence of living standards occurs because manufacturing shifts from South to North. Now consumers in South must purchase the bulk of their manufactured varieties from North and since they pay trade costs when they purchase Northern varieties, this means that they have to pay more for the manufactured goods that they buy. Furthermore, because the service sector shrinks in South, Southern made manufactured goods become more expensive, thereby further depressing the real wage in South. The exact opposite happens in North. Northern consumers now buy more manufactured varieties at home, thereby incurring lower trade costs. This helps to raise the real wage in North. In addition, because the service sector grows in North, services become less expensive to acquire and hence the price of Northern made manufacturing varieties falls. Both effects help to raise the real wage in North.

As trade barriers continue to fall below $\bar{r}$, the real wage in both regions begins to rise since lower transport costs are incurred on manufactured goods purchased abroad. The real wage in South continues to rise monotonically after $\bar{r}$. The real wage in North, depending on parameter values\textsuperscript{38}, may eventually hit a maximum after which further reductions in trade barriers cause the real wage in the North to fall. In figure 2.2, the real wage in North continues to rise even as transport costs fall to zero. In this case, the decline in the price of manufactured goods (both domestic and foreign) is sufficiently strong to counterbalance a falling wage. In figure 2.3, however, a falling wage in North

\textsuperscript{37} In an interesting attempt to synthesize the main lessons of the new economic geography literature with the endogenous growth literature and in the process, provide a micro foundation for Rostow's [1960] famous "Four Stages of Economic Growth" thesis, Baldwin, Martin, and Ottaviano [1998] show that the periphery may suffer not only from a shrinking manufacturing base, but also from a reduction in investment. The reduction in capital accumulation in the periphery (and the increase in capital accumulation in the core) leads to an even greater divergence in income than that posited by more conventional economic geography models.

\textsuperscript{38} The key parameter that determines whether the real wage will continue to rise monotonically in North (as in Figure 2) or eventually decline (as in Figure 3) is $\sigma$, the elasticity of substitution between varieties. A high value of $\sigma$ implies that consumers are very price sensitive. As transport costs fall, Northern manufactures will be put under strong pressure to reduce prices to ensure that price sensitive consumers keep buying the bulk of their manufactured goods from North. This will, in turn, drag down wages in North, leading to the non-monotonic relationship between transport costs and real wages shown in figure 2.3.
eventually outweighs the benefit from lower trade costs and as a consequence, the real wage in North actually decreases.

### 2.6 Conclusions

Economists in the 1950s, such Myrdal [1957] and Hirschman [1958], wrote extensively about such ideas as cumulative causation and backward and forward linkages. Crucial to their story was the view that an increase in the number of suppliers in a region would reduce price-cost margins and attract enough new customers to make further entry of firms profitable, thereby leading to agglomeration and a 'virtuous' cycle of industrialization. Recent work on economic geography has tried to tell the same story, but has done so in a roundabout way, focusing more on such things as variety effects than on shrinking price-cost margins resulting from increased competition among firms. Thus, the goal of this chapter was to present a model of agglomeration in which strategic interactions among firms take center stage. Interestingly, the predictions generated in this chapter arise by making the model more realistic than models that rely solely on monopolistic competition. Most firms operate in markets in which they know that their own actions will have an effect on the price at which they sell their products. Instead of taking prices as given (as in a perfectly competitive model) or by simply charging a constant markup over marginal cost (as in the monopolistically competitive model) firms interact strategically, perpetually asking themselves how their actions will affect the market in which they operate.

Furthermore, strategic interaction is itself a vital source of agglomeration. Whereas other economic geography models can produce agglomeration only by combining monopolistic competition with some other feature (such as variety effects via intermediate goods, as in Krugman and Venables [1995] and Venables [1996]), my model produces agglomeration simply as a byproduct of strategic interaction. This allows us to model agglomeration in a very tractable and intuitive way.
No doubt, this model, like so many models of the new trade theory, can provide ammunition to a wide assortment of protectionist arguments. Those who see the wealth of the North as the cause of poverty in the South can cite the model's prediction that closer economic integration will cause the deindustrialization of the South, thereby increasing global inequality. On the other hand, those who fear that competition from low wage workers in the South will depress wages in the North can find theoretical support from the model's prediction that eventually, when transport costs have fallen sufficiently far, the relative wage of Northern workers will drop as Northern manufacturing firms face increased price competition from their Southern competitors.

To regard this chapter as a plea for protectionism, however, would be a grave mistake. First, although a Northern tariff on Southern manufactured varieties may help raise the real wage of Northern workers when transport costs are low, the improved standard of living in the North can only come at the expense of Southern workers. The notion that a tariff can help both the North and the South is simply wrong in the framework of this model. Second, although it is tedious to prove analytically, it can be shown that the average real wage (the average of the real wage in the South and in the North) rises after the core-periphery is formed. This results from the simple fact that there are large efficiency gains from pooling manufacturing within a single region. If one region ends up with the lion's share of manufacturing, this will increase demand for services in that region, leading to a more competitive service sector. This heightened competition will lower margins and force service providers to operate farther along their average cost curves, allowing them to lower their unit costs. This resulting increase in efficiency implies that the world, as measured by aggregate global income, is better off from the formation of a core-periphery.
2.7 Appendix A: Computing the Level of Trade Costs at which the Symmetric Equilibrium Ceases to be Stable

To compute $\tau$, let us consider what happens at the symmetric equilibrium when there is a slight relocation of manufacturing firms from South to North (i.e. $dn_m = -dn_m^*$). If this relocation creates economic losses for Northern firms this will cause the initial imbalance to be corrected and hence, the symmetric equilibrium is stable. If, on the other hand, the relocation makes Northern firms more profitable, this will induce even more firms to enter the manufacturing sector in North, thereby causing the manufacturing sector in South to shrink. This will in turn destabilize the symmetric equilibrium, engendering a core-periphery. Since we are only considering a small relocation of firms from South to North, both regions will retain an agricultural sector, thus ensuring that the wage is equal to one in both regions. For convenience, define total production of a single variety as $z = x + y$. From equation (2.2) and (2.4), we see that a relocation of manufacturing firms from South to North will lead to positive economic profits if total production of the variety, $z$, rises. That is, $\Pi_m > 0 \Rightarrow z > \frac{\alpha(\sigma - 1)}{\beta}$. Thus, if the relocation of firms leads to higher total production of the variety in North, then the equilibrium will be unstable and the economy will form a core-periphery. If instead $z$ falls when there is an increase in the number of manufacturing firms in North, then the symmetric equilibrium will endure. The value of $\tau$ at which the economy is indifferent between staying at the symmetric equilibrium and forming a core periphery must then correspond to the case where $\frac{dz}{dn_m} = 0$.

Let us first totally differentiate equation (2.3) along with its Southern counterpart around the symmetric equilibrium:

\begin{equation}
(2.23) \quad dp_m = -\frac{\beta \delta \mu \sigma}{\sigma - 1} \frac{p_{S}^{\mu - 1}}{(\bar{a}_s - 1)^{2}} dn_s,
\end{equation}
\[ (2.24) \quad dp^*_m = - \frac{\beta \delta \mu \sigma}{\sigma - 1} \frac{\hat{p}_{s}^{-1}}{(\hat{n}_s - 1)^2} dn^*_s. \]

The number of Northern service providers is

\[ (2.25) \quad n_s = \sqrt{\frac{E}{\theta}} = \sqrt{\frac{\mu n_m p_s^\mu (\alpha + \beta z)}{\theta}}. \]

Totally differentiating this equation, along with its Southern counterpart, we get

\[ (2.26) \quad dn_s = \frac{1}{2} \left[ \frac{\theta}{\alpha \sigma \mu \hat{n}_m \hat{p}_s^\mu} \right]^{1/2} \left\{ \frac{\alpha \sigma \mu \hat{p}_s^\mu}{\theta} dn_m - \frac{\alpha \delta \mu^2 \sigma \hat{n}_m \hat{p}_s^{\mu - 1}}{\theta (\hat{n}_s - 1)^2} dn_k + \frac{\beta \mu \hat{n}_m \hat{p}_s^\mu}{\theta} dz \right\}, \]

and

\[ (2.27) \quad dn^*_s = \frac{1}{2} \left[ \frac{\theta}{\alpha \sigma \mu \hat{n}_m \hat{p}_s^\mu} \right]^{1/2} \left\{ -\frac{\alpha \sigma \mu \hat{p}_s^\mu}{\theta} dn_m - \frac{\alpha \delta \mu^2 \sigma \hat{n}_m \hat{p}_s^{\mu - 1}}{\theta (\hat{n}_s - 1)^2} dn_k + \frac{\beta \mu \hat{n}_m \hat{p}_s^\mu}{\theta} dz^* \right\}. \]

Finally, we can totally differentiate equation (2.6) and its Southern counterpart to get:

\[ (2.28) \quad dz = \frac{\gamma}{\hat{n}_m \hat{p}_m} \left[ \left( \sigma - 1 \right) \frac{1 + \tau^{2(1-\sigma)}}{(1 + \tau^{1-\sigma})^2} - \sigma \right] \frac{dp_m}{\hat{p}_m} + 2 \frac{(\sigma - 1) \tau^{1-\sigma}}{(1 + \tau^{1-\sigma})^2} \frac{dp^*_m}{\hat{p}_m^*} + \left\{ \frac{1 - \tau^{1-\sigma}}{1 + \tau^{1-\sigma}} \right\}^2 \frac{dn_m}{n_m}. \]
Let us re-write equations (2.23), (2.24), and (2.26) through (2.29) as follows

\[(2.23')\]
\[dp_m + Adn_s = 0,\]

\[(2.24')\]
\[dp_m^* + Adn_s^* = 0,\]

where
\[A = \frac{\beta \delta \mu \sigma}{\sigma - 1} \frac{\hat{p}_s^{\mu-1}}{(\hat{n}_s - 1)^2} = -\frac{\beta \delta \mu \mu \sigma}{\sigma - 1}\left(\frac{\mu \theta}{\theta}\right)^{1/2} \left[1 + \frac{1}{2} \left(\frac{\mu \theta}{\theta}\right)\right]\]

\[(2.26')\]
\[dn_s + Bdp_m + Cdz = Ddn_m,\]

\[(2.27')\]
\[dn_s^* + Bdp_m^* + Cdz^* = -Ddn_m,\]

where
\[B = 1 + \frac{1}{2} \left[\frac{\theta}{\alpha \sigma \mu \hat{n}_m \hat{p}_s^\mu} \left(\frac{\theta}{\hat{n}_s - 1}\right)^2\right] = 1 + \frac{1}{2} \left(\frac{\mu \theta}{\mu \theta\gamma}\right)\left(1 - \left(\frac{\theta}{\mu \theta}\right)^{1/2}\right)\]
\[ C = -\frac{1}{2} \left[ \frac{\theta}{\alpha \mu \hat{m}} \hat{p}_s^\mu \right]^{1/2} \left[ \frac{\beta \mu \hat{m} \hat{p}_s^\mu}{\theta} \right] \frac{1}{2} \left( \frac{\beta}{\alpha \sigma} \right) \left( \frac{L}{\theta} \right)^{1/2} \left\{ \frac{1}{2} \left( \frac{\theta}{\mu \gamma \lambda} \right) \right\}^{1/2} (1-\mu) \]

\[ D = \frac{1}{2} \left[ \frac{\theta}{\alpha \mu \hat{m}} \hat{p}_s^\mu \right]^{1/2} \left[ \frac{\alpha \mu \hat{p}_s^\mu}{\theta} \right] = \frac{1}{2} \alpha \sigma \left( \frac{\mu}{\gamma \theta \lambda} \right)^{1/2} \left( \frac{\delta}{1 - \left( \frac{\theta}{\gamma \mu} \right)^2} \right) \]

(2.28') \[ Fp_m + Fdp_m^* + dz = Gdn_m \]

(2.29') \[ Fp_m + Edp_m^* + dz^* = -Gdn_m \]

where \[ E = \frac{\beta}{\hat{m}} \left[ \frac{(\sigma - 1) \left( \frac{1 + \tau}{1 + \tau^{1-\sigma}} \right)^2}{\sigma} \right] = \frac{\alpha (\sigma - 1)^2}{\beta} \left\{ \frac{1}{2} \left( \frac{\mu \lambda}{\theta} \right)^{1/2} \right\} \left\{ \frac{1 - \sigma}{\theta} \frac{(1 + \tau^{2(1-\sigma)})}{(1 + \tau^{1-\sigma})^2} \right\} \]

\[ F = -\frac{2 \gamma L (\sigma - 1)^{1-\sigma}}{\hat{m} \hat{p}_m^2 (1 + \tau^{1-\sigma})^2} = -\frac{2 \alpha (\sigma - 1)^3}{\beta} \left\{ \frac{1}{2} \left( \frac{\mu \lambda}{\theta} \right)^{1/2} \right\} \left\{ \frac{\tau^{1-\sigma}}{(1 + \tau^{1-\sigma})^2} \right\} \]
We are interested in the value of $\tau$ that sets the derivative $\frac{dz}{d\tau} = 0$. Using Cramer's rule, this implies that the parameters in the matrix above must set the following determinant equal to zero.

\[
\begin{vmatrix}
1 & 0 & A & 0 & 0 & 0 \\
0 & 1 & 0 & A & 0 & 0 \\
0 & 0 & B & 0 & C & 0 \\
0 & 0 & 0 & B & 0 & C \\
E & F & 0 & 0 & 1 & 0 \\
F & E & 0 & 0 & 0 & 1
\end{vmatrix} = 0
\]
(2.32)
\[ f(\bar{r}; \alpha, \beta, \delta, \gamma, \mu, \theta, \sigma) = B^2G - ABDF + ACFG + ABCEG + ABDE + A^2CDE^2 - A^2CDF^2 = 0 \]

Since this is a fairly complex expression, the value of \( \bar{r} \) must be computed numerically.

## 2.8 Appendix B: Three possible non-symmetric equilibria

This appendix analyzes the three possible non-symmetric equilibria.

### 2.8.1 North only Produces Manufactured goods; South produces only Agriculture

Since North completely specializes in manufacturing, the wage in North may exceed one. The wage in South, however, must equal one to satisfy the zero profit condition in agriculture. Total expenditure on Northern manufactured goods is \( \gamma wL + \gamma L \) (the first term is expenditure by Northern workers and the second term is expenditure by Southern workers). Due to the zero profit condition, total expenditure on Northern manufactures must equal the total cost of northern manufactures, a fraction \( \mu \) of which is allocated to the purchase of intermediate services. Thus, total expenditure on services in North is

\[
E = \gamma \mu L(w + 1)
\]

Balanced trade requires that Northern imports of agriculture, \((1 - \gamma)wL\), equal Northern exports of manufactured goods, \(\gamma L\). Thus, the wage must be

\[
w = \frac{\gamma}{1 - \gamma}
\]
Since we know \( w \geq 1 \), this equilibrium can only exist if \( \gamma \geq 0.5 \). Thus, total expenditure on services is

\[
E = \frac{\eta \mu L}{1 - \gamma}.
\]

This implies that the number of service providers is

\[
n_s = \sqrt{\frac{E}{w \theta}} = \sqrt{\frac{\mu L}{\theta}}.
\]

Hence, the price of services is

\[
p_s = \frac{\gamma}{1 - \gamma} \left( \frac{\delta}{1 - \frac{\theta}{\sqrt{\mu L}}} \right).
\]

Thus, the price of a manufactured good is

\[
p_m = \beta p_s \omega^1 w^{1 - \mu} = \beta \left( \frac{\gamma}{1 - \gamma} \right)^{2 - \mu} \left( \frac{\delta}{1 - \frac{\theta}{\sqrt{\mu L}}} \right) \mu.
\]

Total expenditure on manufactured goods, \( \gamma \omega L + \gamma L \), must equal the total revenue of manufacturing firms, \( p_m n_m(x + y) \). Using equation (2.4), this implies that the number of manufacturing firms is
2.8.2 North only Produces Manufactured Goods; South produces some Manufactured Goods and all the Food.

Since South produces food, \( w^* = 1 \), in order to satisfy the zero profit condition in South's agricultural sector. North must import all its food from South. Hence, we know that North's imports of food must be \((1 - \gamma)wL\). Since trade must be balanced, North's imports of food must equal North's net exports of manufacturing goods. Thus,

\[
(1 - \gamma)wL = n_m p_m y - n_m p_m y^* = \frac{n_m p_m^{1-\sigma} y L^{1-\sigma}}{n_m p_m^{1-\sigma} + n_m (\varphi m)^{1-\sigma}} - \frac{n_m^* p_m^{1-\sigma} \gamma w L^{1-\sigma}}{n_m p_m^{1-\sigma} + n_m (\varphi m)^{1-\sigma}}.
\]

Using equations (2.4) and (2.6), we can write

\[
x + y = \frac{\alpha(\sigma - 1)}{\beta} = \frac{p_m^{-\sigma} \gamma w L}{n_m p_m^{1-\sigma} + n_m (\varphi m)^{1-\sigma}} + \frac{p_m^{-\sigma} \gamma L^{1-\sigma}}{n_m p_m^{1-\sigma} + n_m (\varphi m)^{1-\sigma}}.
\]

\[
x^* + y^* = \frac{\alpha(\sigma - 1)}{\beta} = \frac{p_m^{*\sigma} \gamma L}{n_m^* p_m^{1-\sigma} + n_m^*(\varphi m)^{1-\sigma}} + \frac{p_m^{*\sigma} \gamma w L^{1-\sigma}}{n_m^* p_m^{1-\sigma} + n_m^*(\varphi m)^{1-\sigma}}.
\]

Total expenditure on services in North is

\[
n_m = \frac{L}{\alpha(\sigma - 1)} \left( \frac{1-\gamma}{\gamma} \right)^{1-\mu} \left( \frac{\delta}{\sqrt{\theta}} \right)^{\mu}.
\]
Similarly, total expenditure on services in South is

\[ E^* = \mu p^*_{m,n_m}(x + y) = \frac{p_m n_m \alpha \mu (\sigma - 1)}{\beta} \]  

This implies that the number of service providers in North and South are

\[ n_s = \sqrt{\frac{p_m n_m \alpha \mu (\sigma - 1)}{w \beta \theta}} \quad \text{and} \quad n_s^* = \sqrt{\frac{p_m n_m \alpha \mu (\sigma - 1)}{\beta \theta}}. \]

Thus, the price of services in North and South is

\[ p_s = \frac{w \delta n_s}{n_s - 1} = \frac{\delta}{1 - \sqrt{\frac{w \beta \theta}{p_m n_m \alpha \mu (\sigma - 1)}}} \quad \text{and} \quad p_s^* = \frac{\delta n_s^*}{n_s^* - 1} = \frac{\delta}{1 - \sqrt{\frac{\beta \theta}{p_m n_m \alpha \mu (\sigma - 1)}}}. \]

Equations (2.40), (2.41), (2.42), (2.45) and (2.46) provide a system of seven equations and seven unknowns \((p_s, p_s^*, w, n_m, n_m^*, p_m, p_m^*)\). The solution to this system of equations can be used to solve for the other endogenous variables in the model.

### 2.8.3 North Produces Manufacturing, Services and Agriculture; South Produces only Agriculture

Since agriculture is produced in both regions, \(w = w^* = 1\). Since South produces no manufacturing, it must import all its manufacturing from North. Thus, South spends
\(\gamma L\) on Northern manufacturing exports, implying that North buys \(\gamma L\) worth of food from South. Since Northern workers wish to consume \((1 - \gamma)L\) worth food, expenditure on food produced in North must equal \((1 - \gamma)L - \gamma L = L(1 - 2\gamma)\). Hence, a necessary condition for this equilibrium to exist is \(\gamma \leq 0.5\).

Workers in both North and South spend \(\gamma L\) on manufacturing. Thus, total expenditure on services in North is

\[(2.47) \quad E = 2\mu\gamma L\]

Thus, the number of service providers in North is

\[(2.48) \quad n_s = \sqrt{\frac{2\mu\gamma L}{\theta}}\]

Hence, the price of services is

\[(2.49) \quad p_s = \frac{\delta}{1 - \frac{\theta}{\sqrt{2\mu\gamma L}}}\]

This implies that the price of manufactured goods is

\[(2.50) \quad p_m = \frac{\sigma\beta}{\sigma - 1}\left\{\frac{\delta}{\theta} - \frac{1}{1 - \sqrt{2\mu\gamma L}}\right\}^{\mu}\]

Using equation (2.4), equation (2.40), and the fact that \(p_m n_m (x + y) = 2\gamma L\) we can solve for the number of manufactured varieties that are produced in equilibrium:
\( n_m = \frac{2\gamma L}{\alpha \sigma} \left\{ \frac{1 - \sqrt{\frac{\theta}{2\mu \gamma L}}}{\delta} \right\}^\mu \)
Chapter 3

The Ins and Outs of Labour Market Pooling
3.1 Introduction

Why do firms in the same industries often choose to locate close to one another? It is a very simple question, but unfortunately, one that is difficult to answer within the confines of a model that relies on perfect competition. In a perfectly competitive world, one would expect economic activity to be spatially dispersed since firms will face less demand for the goods that they produce and less demand for the factors of production that they employ if they are located far from other firms. Of course, this is not what we observe. Economic activity tends to be clustered in cities. Furthermore, cities themselves often tend to specialize in particularly industries. Dalton, Georgia is the "carpet capital" of the United States. Flint, Michigan is the home of numerous automobile parts manufacturers. Milan boasts one of the world's most vibrant fashion industries (Krugman [1991b]).

Although a variety of stories have been proposed to explain the propensity for firms to agglomerate39, detailed empirical work by Dumais, Ellison, and Glaeser [1997] suggests that one story is most compelling.40 The story is based on the idea of "labour market pooling".41 Imagine that a firm is located in an isolated area and all of a sudden it receives a major contract that requires it to temporarily double production. To accomplish this, the firm will require more workers, but since the firm is located in an isolated area, finding more workers will likely prove to be difficult. Hence, it makes sense for the firm to locate in a city where there are many firms and workers in the same industry. Then, if a firm experiences good times and needs to expand employment, chances are that there will be firms that are experiencing bad times and are laying off workers. The successful firm can draw on these workers to meet its employment needs.

40 Dumais, Ellison, and Glaeser [1997, p. 30] conclude that "the presence of input suppliers and customers is relatively unimportant in explaining why firms in different industries locate near one another. Intellectual spillovers appear to be somewhat more important, but the location process appears to be dominated by the labor mix of a particular area: plants do seem to locate near other industries when they share the same type of labor. This effect is large and suggests that labor market pooling is a dominant force in explaining the agglomeration of industry."
41 Marshall [1920] was the first to outline this story.
Thus, labour market pooling will increase the expected profits of firms. Furthermore, workers will benefit by living in regions where there are many firms because then decreases in labour demand by some firms will tend to be counterbalanced by increases in labour demand by other firms, thus reducing the variability of wages and aggregate employment.

An important question arises from this analysis: Does labour market pooling provide any additional benefits beyond insulating workers from firm-specific shocks and giving firms the ability to hire more workers when they experience good times? For instance, can increased labour demand shocks actually boost productivity? As this chapter will argue, the answer is yes. To motivate the intuition for this result, consider the following dilemma: suppose that you have a career in academic research and you have two options in life. The first option allows you to be moderately intelligent every day. For the second option, God flips a coin to determine whether you will be a simpleton or a sage; hence you have a fifty percent chance of either spending the day dumb as an ox or as smart as (insert the name of your favorite economist here). I hazard to guess that most people would choose the second option, at least if their prime goal were to maximize the quantity and quality of their academic work. On the days that you are a sage you will be able to do brilliant academic research and on the days that you are a simpleton you will be able to do all the mundane work that you would normally have to do anyway: household chores, washing the car, etc. By concentrating on research on days when you are most productive and shunning research on days when you are least productive, you will be able to increase your average productivity. On a more abstract level, the benefit from choosing the second option arises from the positive correlation between research output and productivity. Even if your productivity rises by the same amount on good days as if falls on bad days, aggregate output will increase as long as you work more on good days.

It is the positive relationship between how much a firm produces and its productivity level that drives the labour market pooling story. Suppose that firms experience idiosyncratic productivity shocks. If a firm experiences a positive
productivity shock, it will hire more workers. If a firm experiences a negative productivity shock, it will lay off workers. Labour will continually be drawn away from low productivity firms to high productivity firms. This will raise the average productivity of labour in the workforce. And the stronger are firm-specific shocks, the bigger will be the benefits from labour market pooling. In this sense, increased volatility in productivity levels across firms, instead of being a source of trouble, can actually serve to raise the average standard of living in an economy.

Incidentally, the same intuition holds true in a Ricardian trade model. Suppose the world consists of two countries that produce wine and cloth. It is a truism that if both countries have the same opportunity cost for wine, that there will be no gains from international trade. Suppose, however, that both countries experience productivity shocks. When one country is particularly good at wine production, it will export wine and import cloth. As long as the productivity shocks are not perfectly correlated between the two countries, there will be benefits from trade, and the whole world will be better off.

---

42 Whether this is so is, of course, an empirical matter. Higher labour productivity will allow firms to increase output without employing more workers but will also allow firms to produce the same amount of output with fewer workers. As a practical matter, the evidence suggests that favorable technological shocks tend to increase labour use, at least initially. Shea (1998) finds that increases in R&D and patents tend to increase the demand for labour in the short run. However, since technological improvements encourage substitution towards more capital intensive techniques, they decrease the demand for labour in the long run. Brainard and Cutler (1993) find the excessive positive stock returns (which proxy for unexpected favorable productivity shocks) predict subsequent increases in employment within industries. They also find a "positive and statistically significant correlation between cross-section volatility and subsequent employment". This implies that labour market pooling can shield workers from some of that volatility by allowing them to find jobs at firms that are experiencing good times if indeed their current employer is experiencing bad times. Furthermore, Baily, Hulten and Campbell (1992) and Foster, Haltiwanger, and Krizan (1998) both provide evidence that productivity is positively correlated with firm size; larger firms tend to be more productive than smaller firms. There is also substantial evidence that low productivity casts a "shadow of death" on struggling firms and as a result, helps predict exit [Baily, Hulten and Campbell (1992), Olley and Pakes (1996), Dwyer (1995)]. While established firms have higher average productivity levels than new firms, new firms tend to be more productive than firms on the brink of exiting the industry [Foster, Haltiwanger, and Krizan (1998)]. Furthermore, contrary to the convention wisdom that downsizing led to rising productivity levels in the U.S. in the 1980s, Baily, Bartelsman, and Haltiwanger (1994) find that "plants that increased employment as well as productivity contribute almost as much to overall productivity growth in the 1980s as the plants that increased productivity at the expense of employment."
The outline of this chapter is as follows. Section 3.2 constructs a formal macroeconomic model of labour market pooling with flexible wages. Section 3.3 shows that labour market pooling increases the expected profits that firms earn and section 3.4 describes what effect labour market pooling has on some key macroeconomic variables, such as national income and the current account. Section 3.5 considers the effects of labour market pooling when wages are fixed. It is shown that if there are times when aggregate labour demand exceeds aggregate labour supply, labour market pooling has the potential to reduce not only the variability of the unemployment rate but also reduce the expected value of the unemployment rate. Finally, section 3.6 tests one of the key predictions of the labour pooling model, namely that agglomeration is more likely to take place in industries where firms experience strong firm-specific shocks.

3.2 A Macroeconomic model of labour market pooling

The basic structure of this model is drawn from Krugman’s [1991b] partial equilibrium model of labour market pooling. By embedding the model within a general equilibrium framework, however, it is possible to highlight some key issues that go unresolved within a partial equilibrium setting. Specifically, we can explore such ideas as how labour market pooling affects aggregate consumption, production, and the current account. A number of interesting results emerge from this analysis. It is shown that countries whose firms experience temporarily large firm-specific productivity shocks will run current account surpluses. Furthermore, both current and future increases in population will raise per capita consumption. Also, it is demonstrated that an increase in uncertainty about future productivity will serve to increase consumption both in the present and in the future.

Imagine a small open economy that produces a single homogeneous good. Suppose the economy is comprised of \( n \) worker-entrepreneurs. Each worker-entrepreneur inelastically supplies 1 unit of labour, which pays a real wage, \( w \). Wages are perfectly flexible and adjust to ensure that all \( n \) worker-entrepreneurs are fully
employed. Each worker-entrepreneur owns his own firm, which earns a real profit $\pi_i$, where $i$ denotes the $i$th individual. Note that this implies that the number of workers is always equal to the number of firms in this economy. Thus, each worker-entrepreneur has two sources of income, the first from the labour he supplies and the second from the firm that he owns. The worker-entrepreneur may, but does not have to, work at her own firm (it might be helpful to imagine a yeoman farmer who owns her own plot of land but also sells her labour to other yeoman farmers). Since wages are perfectly flexible and there is perfect labour mobility across firms, each firm will pay the same wage. Thus, it will not matter to the individual at which firm she is employed. I assume that individuals maximize utility over an infinite horizon. Lastly, all individuals can purchase risk-free bonds that earn a real interest rate, $r$.

3.2.1 The consumer's problem

Consumer $i$'s problem is:

\[ \text{Max } E_0 \sum_{t=0}^{\infty} \beta^t U(C_{i,t}), \]

subject to

\[ B_{i,t+1} = (1 + r)B_{i,t} - C_{i,t} + w_t + \pi_{i,t}. \]

$B_{i,t}$ is the stock of bonds owned by individual $i$ at time $t$. $C_{i,t}$ is his consumption at time $t$, $w_t$ is the wage that he receives for the one unit of labour that he inelastically supplies and $\pi_{i,t}$ is the profit that he earns from the firm that he owns. Notice that $w_t$ does not have an $i$ subscript because labour is perfectly mobile across firms, thus guaranteeing that all firms pay the same wage. The solution to this problem yields the familiar Euler equation.
(3.3) \[ U'(C_{i,t}) = \beta (1 + r) E_i U'(C_{i,t+1}) . \]

The accumulation constraint (3.2) can be solved forward to yield:

(3.4) \[ \sum_{j=0}^{T} \left( \frac{1}{1 + r} \right)^j C_{i,t+j} = (1 + r) B_{i,t} + \sum_{j=0}^{T} \left( \frac{1}{1 + r} \right)^j (w_{i,t+j} + \pi_{i,t+j}) . \]

Notice that in going from (3.2) to (3.4) we have implicitly imposed the transversality condition,

(3.5) \[ \lim_{\tau \to \infty} \left( \frac{1}{1 + r} \right)^{T-\tau} B_{i,t+\tau} = 0 . \]

The transversality condition precludes individuals from saving so much that their stock of bonds becomes infinite or borrowing so much that they become infinitely indebted to the rest of the world.

In order to derive a closed-form solution for optimal consumption, let us assume that utility is quadratic. Thus,

(3.6) \[ U(C_{i,t}) = a C_{i,t} - b C_{i,t}^2 . \]

Also, for convenience, assume that \( \beta = \frac{1}{1 + r} \). Equation (3.3) can now be re-written

---

43 The assumption that \( \beta = \frac{1}{1 + r} \) implies that, on average, the country will neither be a net foreign lender nor a net foreign borrower. If all countries are the same, this is a necessary assumption since it is impossible for every country to run either a current account deficit or a current account surplus. Even if countries do vary in their growth prospects and their propensity to discount future income, one can still argue that the assumption is reasonable since there is such a strong domestic investment bias (resulting in a the strong correlation between national rates of savings and investment). See Feldstein and Horioka (1980) and French and Poterba (1991) for details.
simply as:

\[(3.7) \quad C_{i,t} = E_t C_{i,t+1} .\]

Solving forward and using the law of iterated expectations, we see that equation (3.7) implies:

\[(3.8) \quad C_{i,t} = E_t C_{i,t+j} .\]

Thus, the best estimate of consumption at time \(t+j\) is how much is consumed today.

### 3.2.2 The firm's problem

Let us now turn to the firm's problem. The number of firms is fixed and is equal to \(n\), the number of worker-entrepreneurs in the economy. Firms can not enter or exit. Assume that each firm is a price taker in both the goods market and in the labour market. To keep the notation simple, let us temporarily drop the \(t\) subscript. Each firm has a production function,

\[(3.9) \quad Q_i = (1 + \varepsilon_i) L_i , \quad \text{where} \quad \varepsilon_i \sim iid f(\mu, \sigma^2) \text{ over the support } \{0, \infty\} .\]

\(\varepsilon_i\) is an exogenous idiosyncratic firm-specific productivity shock and \(L_i\) is the quantity of labour that firm \(i\) employs. Without loss of generality, let the price of output be the numeraire and hence equal one. In addition, assume that each firm incurs quadratic costs of \(\frac{1}{2} L_i^2 + wL_i\). The profit function is therefore\(^{44}\):

\[^{44}\text{The resulting profit function is simplified version of the one found in Krugman (1991) with } \alpha = 0 \text{ and } \beta = \gamma = 1 .\]
Differentiating equation (3.10) with respect to $L_i$, we see that the first order condition implies:

$$L_i = (1 + \varepsilon_i - w).$$

Thus, firms will want to employ more workers when they experience positive productivity shocks. Substituting (3.11) into (3.10), we get,

$$\Pi_i = \frac{1}{2} L_i^2 = \frac{1}{2} (1 + \varepsilon_i - w)^2.$$

### 3.3 The Benefits of Labour Market Pooling

Labour market clearing implies that aggregate labour demand must equal aggregate labour supply. Thus,

$$L_D = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} (1 + \varepsilon_i - w) = n + \sum_{i=1}^{n} \varepsilon_i - nw = L_s = n.$$

Hence, the wage rate is simply,

$$w = \frac{\sum_{i=1}^{n} \varepsilon_i}{n}.$$

The expected wage is therefore:
(3.15) \[ E(w) = \mu. \]

Since each \( \epsilon_i \) must by assumption assume a value between zero and infinity, \( \mu \) must necessarily be greater than zero, implying that the expected wage will always be positive.

Substituting (3.14) into (3.12), we have:

(3.16) \[
\Pi_i = \frac{1}{2} \left( 1 + \epsilon_i - \frac{\sum_{i=1}^{n} \epsilon_i}{n} \right)^2 
\]

\[
= \frac{1}{2} \left[ 1 + \epsilon_i^2 + \left( \frac{\sum_{i=1}^{n} \epsilon_i}{n} \right)^2 - 2 \frac{\sum_{i=1}^{n} \epsilon_i}{n} \epsilon_i + 2 \epsilon_i - \frac{2}{n} \sum_{i=1}^{n} \epsilon_i \sum_{i=1}^{n} \epsilon_i \right].
\]

Making use of the fact that \( E(\epsilon_i^2) = \mu^2 + \sigma^2 \), \( E\left( \sum_{i=1}^{n} \epsilon_i \right)^2 = n(\mu^2 + \sigma^2) \), and

\[ E\left( \epsilon_i \sum_{i=1}^{n} \epsilon_i \right) = n\mu^2 + \sigma^2, \]

it is straightforward to show that,

(3.17) \[ E(\Pi_i) = \frac{1}{2} \left( 1 + \frac{n-1}{n} \sigma^2 \right). \]

Note that \( \frac{\partial E(\Pi_i)}{\partial n} = \frac{\sigma^2}{2n^2} > 0 \) and \( \frac{\partial E(\Pi_i)}{\partial \sigma^2} = \frac{n-1}{2n} > 0 \).

Thus, expected profits are increasing in both the number of firms, \( n \), and the uncertainty that firms face, \( \sigma^2 \). The former effect is a direct consequence of labour

\[ ^{45} \text{Recall the assumption that since each worker is also an entrepreneur who owns his own firm, the number of workers must equal the number of firms.} \]
market pooling. When there are many firms in a region, high labour demand by some firms will tend to be counterbalanced by low labour demand by other firms, thereby keeping wages relatively stable. We can see this by computing the variance of the wage rate in equation (3.14):

\begin{equation}
Var(w) = Var \left( \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \right) = \frac{\sigma^2}{n}
\end{equation}

Equation (3.18) tells us that the variance of the wage rate falls as \( n \) rises. If there were only a single firm, then the wage that it paid would equal the realized value of its productivity shock (see equation (3.14)). Thus, times when the firm enjoyed high productivity would coincide with times when the firm had to pay high wages. This is extremely undesirable from the point of view of the firm. Paying a high wage is most costly to the firm when it wants to employ many workers. Yet, if a firm is big enough relative to the market in which it operates, it will drive up wages precisely when it wants to employ many workers and drive down wages when it wants to employ few workers. However, if the firm is small relative to the market, it will be able to increase employment without putting much upward pressure on wages. This is beneficial to the firm since it cares more about what wage it pays when it wants to employ many workers. Thus, the firm can increase its expected profits if it can avoid paying a high wage when it wants to employ many workers (which is very costly) and a low wage when it employs few workers (which is only slightly helpful) and instead, pay a wage that does not depend on whatever idiosyncratic productivity shocks it experiences.\(^{46}\)

\(^{46}\) In principle, this is the same theory that underlies the popular investment strategy of "dollar cost averaging." The strategy advises the investor to spend a constant sum of money on stocks (or other volatile asset). The more volatile the price of the asset, the lower will be the average purchase price. For instance, if you spend $150 every month on a stock whose price is always $10, you average purchase price per share will be $10. If the share price has a 50% percent chance of being $5 and a 50% chance of being $15 in any given month, then on average, you will be able to purchase \( \frac{1}{2} \left( \frac{150}{5} + \frac{150}{15} \right) = 20 \) shares for expected price of \( \frac{150}{20} = $7.5 \). Thus, you can reduce your average purchase price (and hence increase your expected return) by spending a constant sum of money on the more volatile asset.
Notice that that expected profits also rise when uncertainty increases (that is, when $\sigma^2$ increases). When labour at firm $i$ is extremely productive (high $\varepsilon_i$) relative to labour at other firms, firm $i$ will tend to employ a lot of labour (recall; when you are a sage, you do a lot of academic research). When labour at firm $i$ is not very productive, labour will flow to other firms who happen to be experiencing higher productivity. This raises the average productivity of all firms, thereby increasing their expected profits. Of course, this is possible only if there are many firms in the locality. Thus, as we see from equation (3.17), an increase in $\sigma^2$ is beneficial only if the number of firms exceeds one.

### 3.4 The Macroeconomic effects of Labour Market Pooling

Now, let us turn to the aggregate effects of labour market pooling. Since equation (3.4) must hold in every state of the world, we can take its expectation at time $t$ and using equations (3.8), (3.15) and (3.17), we can then solve for the consumption of person $i$ at time $t$. Thus,

$$C_{i,t} = rB_{i,t} + \mu + \frac{1}{2} \left(1 + \sigma^2 \frac{r}{1+r} \sum_{j=0}^{\infty} \left(\frac{1}{1+r}\right)^j \frac{n_{i+j} - 1}{n_{i+j}}\right).$$

Consumption of person $i$ rises when there is an increase in expected productivity (higher $\mu$), an increase in uncertainty about future productivity shocks (higher $\sigma^2$) or an increase in either the current or future number of firms. Thus, if people expect the population of the region in which they live to grow, they should consume more today than people living in a region whose population is expected to remain stable. This stems from the fact that people want to smooth consumption over time. If the population is expected to increase, then people will also expect to benefit more from labour market pooling in the future. Since their incomes will be higher in the future, they can afford to consume more in the present.
As Krugman [1991b] shows, this sort of framework generates the prediction that labour mobility across regions has the potential to cause agglomeration of all labour within a single region. Imagine, for instance, that there are two regions and each region has an economy that can be described by this model. People will naturally want to live in the region where they can have the highest consumption over time. This implies that regions with high initial population levels, high expected population growth, and productivity that is high on average and extremely variable across firms, will tend to attract workers and firms from surrounding regions. This process, of course, is self-sustaining. A region that attracts workers and firms will benefit more from labour market pooling and this will, in turn, attract even more workers. As a result, some regions will become depopulated while others will become highly urbanized.

Let us now simplify things a bit. When the population is constant so that \( n_{i+1} = n \), equation (3.19) becomes,

\[
C_{i,t} = rB_{i,t} + \mu + \frac{1}{2} \left( 1 + \frac{n-1}{n} \sigma^2 \right).
\]

What are the consumption benefits of labour market pooling when the population is constant over time? To answer this question, denote \( \bar{C}_{i,t} \) as the consumption of individual \( i \) when there is only a single firm. Therefore, the increase in per capita consumption from living in a region where there are \( n \) firms is

\[
C_{i,t} - \bar{C}_{i,t} = \frac{n-1}{n} \sigma^2.
\]

As we increase the number of firms, the benefits from pooling increase, but do so at a decreasing rate. Note that when \( n \to \infty \), the gain in per capita consumption approaches \( \sigma^2 \). Thus, the gain in consumption from living in a huge city \( (n \to \infty) \) over living on your own \( (n = 0) \) is simply equal to the variance of firm-specific productivity shocks.
To determine aggregate consumption, sum over $n$ and denote aggregate consumption as $C_i = \sum_{i=1}^{n} C_{i,n}$ and the aggregate stock of bond holdings as $B_i = \sum_{i=1}^{n} B_{i,n}$.

Hence,

$$C_i = rB_i + n\mu + \frac{1}{2}[n + (n-1)\sigma^2].$$ (3.21)

Next, let us look at the dynamics of the current account. The current account is the difference between what a country earns abroad and what it spends abroad and is defined as,

$$CA_i = rB_i + Y_i - C_i.$$ (3.22)

Aggregate income, $Y_i$, is

$$Y_i = \sum_{i=1}^{n} (w + \Pi_i) = n \left( \sum_{i=1}^{n} \frac{\varepsilon_i}{n} \right) + \sum_{i=1}^{n} \Pi_i = \sum_{i=1}^{n} \varepsilon_i + \sum_{i=1}^{n} \Pi_i.$$ (3.23)

Notice that equation (3.16) can be re-written as

$$\Pi_i = \frac{1}{2} \left[ 1 + \left( \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \right)^2 + 2 \left( \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \right) \right].$$ (3.16')

Summing over $n$, we see that
where $s^2$ is the realized variance of $c_i$, which is not to be confused with its population counterpart, $\sigma^2$. Combining equations (3.21), (3.22), (3.23), and (3.24), we see that the current account must equal:

\[ CA_t = n(\bar{\varepsilon}_t - \mu) + \frac{n-1}{2} (s^2_t - \sigma^2), \]

where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\sum_{i=1}^{n} \varepsilon_i - \frac{\sum_{i=1}^{n} \varepsilon_i}{n})^2$ is the realized variance of $\varepsilon_t$, which is not to be confused with its population counterpart, $\sigma^2$.

Since $E(s^2_t) = \sigma^2$ and $E(\bar{\varepsilon}_t) = \mu$, on average, the current account is balanced. Shocks to the current account arise from deviations between the sample realizations of the mean and variance of productivity shocks and their population counterparts. An exceptionally high $\bar{\varepsilon}_t$, for instance, implies that firms on average experienced abnormally high productivity, implying that national income is higher than usual. Since people want to smooth consumption over time, they will save most of this extra income, thereby producing a current account surplus. Similarly, when $s^2_t$ is higher than usual, national income rises since the economy benefits by shifting workers to firms that are experiencing exceptionally high productivity. Since this deviation is temporary, most of the extra income is saved and as a result, the country will run a current account surplus.
3.5 Labour Market Pooling and Unemployment

The previous section examined the benefits that accrue from labour market pooling in an economy with flexible wages. This section considers some of the ramifications of labour market pooling when wages are fixed. When wages are fixed, fluctuations in labour demand will result in fluctuations in employment as opposed to fluctuations in the wage rate.

Since wages are fixed, let us assume, without loss of generality, that $w = 1$. The profit function in equation (3.10) then becomes

$\Pi_i = (1 + \varepsilon_i) L_i - \frac{1}{2} L_i^2 - L_i$.

The first order condition implies,

$L_i = \varepsilon_i$.

The unemployment rate, $u$, is therefore:

$u = \frac{L_s - L_D}{L_s} = \frac{n - \sum^{\infty}_{i=1} L_i}{n} = 1 - \frac{\sum^{\infty}_{i=1} \varepsilon_i}{n}$

Notice that the unemployment rate depends on the distribution from which the $\varepsilon_i$'s are drawn. As we shall see, the choice of support is crucial in determining the effects of labour market pooling on unemployment.

3.5.1 Aggregate labour demand never exceeds aggregate labour supply
Suppose, for instance, that \( \varepsilon_i \sim iid f(\mu, \sigma^2) \) over the support \( \{0,1\} \). Since each

\[
\sum_{i=1}^{n} \varepsilon_i
\]

\( \varepsilon, \) will be between zero and one, it must be the case that \( 0 \leq \frac{i=1}{n} \leq 1 \). Hence, aggregate labour demand will never exceed aggregate labour supply and thus, there will always be some unemployed workers. On average, the unemployment rate will be,

\[
E(u) = 1 - \mu
\]

with variance equal to

\[
Var(u) = \frac{\sigma^2}{n}
\]

Thus, when \( 0 \leq \varepsilon_i \leq 1 \), increasing the number of firms in the region reduces the variance of unemployment rate but does not change its expected value.

Using equation (3.27) and (3.28), we see that each firm will earn an expected profit of:

\[
E(\Pi_i) = \frac{1}{2} E(\varepsilon_i^2) = \frac{1}{2} (\mu^2 + \sigma^2)
\]

Notice that unlike the previous section, expected profits are now independent of the number of firms in the region.

### 3.5.2 Aggregate labour demand may exceed aggregate labour supply

Now, let us make a slight adjustment to the distribution from which the \( \varepsilon, \)'s are drawn. As before, continue to assume that the \( \varepsilon, \)'s are iid. This time, however, suppose the support from which the \( \varepsilon, \)'s are drawn is \( \{0,2\} \). Now, the \( \varepsilon, \)'s can assume values
greater than one, thus implying that there will be times when everyone who wants work will be able to find work. If total labour demand, $\sum_{i=1}^{n} e_i$, turns out to be greater than total labour supply, $n$, then there will be a shortage of workers and so, the unemployment rate will be zero. Thus, in this case, the unemployment rate is,

$$(3.31) \quad u = \max(0, 1 - \bar{e}),$$

where $\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n}$. The expected unemployment rate is now:

$$(3.32) \quad E(u) = \int_{0}^{1} (1 - \bar{e})f(\bar{e})d\bar{e},$$

where $f(\bar{e})$ is the probability distribution of $\bar{e}$ over the support $\{0, 2\}$ (since each $e$ is drawn from the support $\{0, 2\}$, $\bar{e}$ must also fall between $\{0, 2\}$).

Now it will turn out that increasing the number of firms in the region will indeed decrease the expected unemployment rate. To see this, let us first consider the case where there is only one firm. In this case $\bar{e} = e_i$ and so equation (3.32) becomes:

$$(3.33) \quad E(u) = \int_{0}^{1} (1 - e_i) f(e_i) d\bar{e} > 0.$$  

Thus, the expected unemployment rate when there is only one firm is always positive.

Using the fact that $\mu = \int_{0}^{1} e f(e) d\bar{e} = \int_{0}^{1} e f(e) d\bar{e} + \int_{1}^{2} e f(e) d\bar{e}$ and

$1 = \int_{0}^{2} f(e) d\bar{e} = \int_{0}^{1} f(e) d\bar{e} + \int_{1}^{2} f(e) d\bar{e}$, we can re-write equation (3.33) as:

$$(3.33') \quad E(u) = 1 - \mu + \int_{1}^{2}(e - 1)f(e) d\bar{e} > 1 - \mu.$$
Hence, the expected unemployment rate when there is only one firm *always* exceeds $1 - \mu$.

Now, let's go to the opposite extreme and consider what happens to the expected unemployment rate when there are an infinite number of firms. Proceeding in precisely the same way as in the case of only one firm, it is straightforward to show that,

$$E(u) = 1 - \mu + \int_1^2 (\bar{E} - 1) f(\bar{E}) d\bar{E}.$$  

(3.34)

First consider the case where $0 \leq \mu < 1$. As $n$ approaches infinity, the distribution of $\bar{E}$ becomes a spike centered around $\mu$. By the Central Limit Theorem, the third term in equation (3.34) goes to zero since it no longer contains any probability mass. Hence,

$$\lim_{n \to \infty} E(u) = 1 - \mu.$$  

(3.38)

From equation (3.33'), we see that when $0 \leq \mu < 1$, the expected unemployment rate when there is pooling is less than the expected unemployment rate when there is no pooling (i.e. only one firm).

Now consider the case where $1 \geq \mu \geq 2$. First, re-write equation (3.34) as

$$E(u) = 1 - \mu - \int_1^2 f(\bar{E}) d\bar{E} + \int_1^2 \bar{E} f(\bar{E}) d\bar{E}.$$  

(3.34')

As before, when $n$ goes to infinity, $f(\bar{E})$ degenerates into a spike centered at $\mu$. Since all the probability mass is in this spike and $\mu$ is between one and two, the third term in equation (3.34') is equal to one. This implies that the fourth term is simply equal to

$$\lim_{n \to \infty} \int_1^2 \bar{E} f(\bar{E}) d\bar{E} = \lim_{n \to \infty} \int_0^2 \bar{E} f(\bar{E}) d\bar{E} = \mu.$$  

Thus, in the case where $1 \geq \mu \geq 2,$
From equation (3.33), this implies that the expected unemployment rate when there is pooling is lower than when there is a single firm. Thus, labour market pooling decreases the expected unemployment rate only when there are times when aggregate labour demand exceeds aggregate labour supply. This is the case because pooling decreases the variance of the average quantity of labour demanded by firms. To see this, note that,

\[
\lim_{n \to \infty} E(u) = 0. 
\]

\[
(3.39)
\]

\[
Var\left( \frac{L_D}{n} \right) = Var\left( \frac{\sum_{i=1}^{n} L_i}{n} \right) = Var\left( \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \right) = \frac{\sigma^2}{n},
\]

and \( E\left( \frac{L_D}{n} \right) = \mu \). As the number of firms approaches infinity, the variance of the average quantity of labour demanded by firms approaches zero.

When \( 1 \geq \mu \geq 2 \), each firm will, on average, want to demand more than one unit of labour. Since each worker supplies only one unit of labour and there are just as many firms as workers, then as \( n \to \infty \), aggregate labour demand will always be greater than aggregate labour supply, implying that the unemployment rate will always be zero. On the other hand, if there is only a single firm, then even though on average that firm will demand more than one unit of labour, there will be times when it demands less than that. When that happens, there will be unemployment. Thus, when there is only a single firm in a region, the expected unemployment rate must be positive.

When \( 0 \leq \mu < 1 \), as \( n \) approaches infinity, each firm will, on average, demand less than one unit of labour. Thus, there will always be some unemployment. On the other hand, if there is only a single firm, then most of the time that firm will demand less than
one unit of labour, but occasionally it will demand more than that. On those few occasions, there will be no unemployment. Yet, because the firm cannot employ more than one unit of labour, it will not be able to realize the full benefit of exceptionally large productivity shocks. Thus, when labour demand exceeds labour supply, the large productivity shock gets partly squandered. Yet, as we have noted, when there are an infinite number of firms, then aggregate labour demand will never exceed aggregate labour supply when $0 \leq \mu < 1$, and hence, positive productivity shocks experienced by firms are never squandered. Thus the average unemployment rate will be lower in regions with many firms if there are times when aggregate labour demand exceeds aggregate labour supply.

### 3.6 An Empirical Test of the Labour Pooling Model

As this chapter has shown, labour market pooling can induce agglomeration in a variety of ways. When wages are flexible, labour market pooling can reduce the variability of wages. This is beneficial to risk averse workers who prefer stable wages. Furthermore, stable wages are beneficial to firms since firms can hire more workers when they experience productivity shocks without driving up the wage rate. As equation (3.17) demonstrates, this results in higher expected profits. When wages are fixed, labour market pooling will decrease the expected unemployment rate if there are times when labour demand exceeds labour supply.

A key assumption in the labour pooling models presented in this chapter is that productivity shocks are firm-specific. If productivity shocks occurred at the industry level, then much of the benefits of labour market pooling would disappear. For instance, if all software firms decided to vary employment by the same degree and at the same time, then there would be no benefit for software firms to locate in a single region. On the other hand, if some software firms were inclined to reduce employment while others increased employment, then the aggregate demand for software engineers would remain stable, bringing benefits for software firms and their employees who locate in a single
region. More formally, this can be shown by extending the model of section 3.2 by allowing shocks to be correlated across firms. Suppose that each firm experiences a productivity shock \( \epsilon_i \), with mean \( \mu \), variance \( \sigma^2 \), and covariance \( \sigma_{i,j} \) that captures the correlation between the productivity shocks of firm \( i \) and \( j \). It is straightforward to show that equation (3.17) then becomes:

\[
E(\Pi_i) = \frac{1}{2} \left[ 1 + \frac{n-1}{n} \left( \sigma^2 - \sigma_{i,j}^2 \right) \right].
\]

Defining the coefficient of correlation, \( r \), as \( \frac{\sigma_{i,j}}{\sigma^2} \), equation (3.55) now becomes

\[
E(\Pi_i) = \frac{1}{2} \left[ 1 + \frac{n-1}{n} (1 - r) \sigma^2 \right].
\]

By Slutsky's inequality, \(-1 \leq r \leq 1\). If productivity shocks are independent of one another, \( r = 0 \), then equation (3.55) reduces to equation (3.17). Clearly, as \( r \) rises, the benefits from labour market pooling decrease. At one extreme, when productivity shocks are perfectly negatively correlated \( (r = -1) \), then the gains from pooling are maximized. At the other extreme, when shocks are perfectly positively correlated \( (r = 1) \), there are no gains from pooling.

This brings forth a testable proposition: are firms in industries that experience strong firm-specific shocks more likely to agglomerate? \(^{47}\) One way to test this hypothesis is to look at the rates of labour turnover in narrowly defined industries. If the labour market pooling model is accurate, we should see high rates of labour turnover in highly concentrated industries. There is, of course, an issue of causality in doing this.

\(^{47}\) Diamond and Simon (1990) argue that specialized cities actually suffer a greater unemployment risk. They provide evidence that workers in less specialized U.S. cities receive higher wages than workers in cities with a wide mix of industries. Using the theory of compensating differentials, they conclude that workers in less specialized cities are compensated for the heightened risk of losing their jobs by receiving higher wages. Firms are willing to pay higher wages because the benefit from pooling more than outweighs the increased labour costs.
Does high labour turnover induce firms to locate close to one another or does high concentration simply facilitate labour turnover? For instance, suppose that firms locate at random, and by chance, an industry ends up highly concentrated. Now, if we observe much labour turnover in this industry, this could simply mean that it is easy for workers to move across firms when the firms are located close to one another.

With this important caveat in mind, there does seem to be some support, albeit weak, for the prediction that industries with high labour turnover will tend to be more concentrated than the average industry. Using 4-digit plant level data on agglomeration and labour turnover, I show that industries that are spatially highly concentrated tend to have high rates of job creation, job destruction, and job reallocation. Over a period of twenty years, the top ten most concentrated industries, on average, had a rate of job reallocation that was 45 percent higher than the rate for the 10 least concentrated industries and more than 25 higher than the average rate of job reallocation for all industries in the sample. This suggests that firms that exhibit strong firm-specific labour demand shocks will have an incentive to cluster together to allow workers to shift from plants that are either dying or shrinking to plants that are either just opening or expanding.\(^{48}\)

### 3.6.1 Data and Methodology

The data on agglomeration is from Ellison and Glaeser (1994). Ellison and Glaeser construct an index of agglomeration for 459 manufacturing industries defined by the 4-digit classifications from the Census Bureau’s 1987 S.I.C. system. There are 51 regions in their dataset. Each state corresponds to a region, with the 51\(^{st}\) region being the District of Columbia. Ellison and Glaeser construct a variable “gamma” which provides

---

\(^{48}\) This finding is consistent with Henderson (1983) who finds that for U.S. manufacturing export industries, economies of scale are based on localization (firms in the same industry clustering together) rather than urbanization (firms locate in big cities regardless of whether they are near other firms in the same industry). Thus, in choosing to locate in a city, it is more important for the firm to know that it is locating close to other firms in its industry than to know that it is locating close to either its customers or suppliers from other industries.
a measure of agglomeration for each 4-digit industry. There are a number of reasons why their approach is superior to that of Krugman (1991), who constructs "gini coefficients" for 3-digit industries. First, and most obvious, a 4-digit classification provides greater detail about what happens in the industry. This extra detail is especially important since most labor reallocation takes place within rather than between fairly narrow industry classifications [(Foster, Haltiwanger, and Krizan, 1998) and (Haltiwanger, 1997)] Thus, detailed data at what happens at the 4-digit level, where most labor reallocation takes place, is vital. Second, Ellison’s and Glaeser’s index measures agglomeration beyond what would be observed if firms chose their locations at random. Their "dartboard" approach thus acknowledges the fact that clusters of firms may simply arise by chance, and not as a result of agglomeration economies. Third, Ellison and Glaeser’s index takes into account the discreteness of plants. To understand why this is important, consider the vacuum cleaner industry (S.I.C. 3635). Seventy-five percent of employees in this industry work in one of the four largest plants. Although this fact certainly reflects economies of scale at the plant level, it says nothing about what agglomeration economies might be present in the vacuum cleaner industry. Yet, since 75 percent of the employment in this industry is necessarily contained in four states, a

---

49 Consider a given industry. Suppose there are \( N \) firms within this industry and there are \( M \) regions in which firms can locate. Gamma is defined as

\[
\gamma = \frac{G - H}{1 - H} \quad \text{where} \quad G = \frac{\sum_i (s_i - x_i)^2}{1 - \sum_i x_i^2}
\]

\[
H = \sum_j z_j^2.
\]

The variable \( H \) is a Herfindahl index, with \( z_j \) representing industry \( j \)’s share of aggregate employment. The variable \( x_i \) represents region \( i \)’s share of aggregate employment. Suppose the regions in which firms locate are independent identically distributed random variables \( v_1, v_2, \ldots, v_N \) with probabilities \( p_1, p_2, \ldots, p_N \) and \( u_{ji} \) is a Bernoulli random variable equal to one if and only if \( v_j = i \). The variable \( s_i = \sum_{j=1}^{N} z_j u_{ji} \) is then the share of employment in region \( i \). If plant location is determined entirely by chance, then the model predicts that \( E(G) = H p \). For details, see Ellison and Glaeser (1994).

50 Haltiwanger (Table 1, 1997) shows that 4-digit industry effects account for less than 10 percent of the cross-sectional heterogeneity in output, employment, capital equipment, capital structures, and productivity growth rates across establishments, with the remaining 90 percent of the variation stemming from idiosyncratic factors at the firm level.
Krugman (1991) style gini coefficient would imply a high degree of agglomeration, a potentially misleading result.

The plant-level data on labour flows comes from Davis, Haltiwanger, and Schuh (1996). They provide annual 4-digit level data on job creation (increases in employment at new and existing plants) and job destruction (decreases in employment at dying and existing plants). If we subtract the job destruction rate from the job creation rate, we get the net job creation rate, which tells us the change in net employment in the industry. Industries that are growing have positive net job creation rates, while industries that are contracting show negative net job creation rates. If, on the other hand, we add the job creation rate (the number of jobs created at new plants and existing plants as a percentage of total employment in the industry) and the job destruction rate (the number of jobs eliminated as a result of plant closure or downsizing as a percentage of total employment) we get what Davis, Haltiwanger and Schuh call the employment reallocation rate.

3.6.2 Results

The main finding of this section is that industries that are extremely spatially agglomerated tend to have significantly higher rates of job creation, job destruction, and job reallocation than the average industry. This result applies only to the top 5 percent of agglomerated industries. For all but the most concentrated industries, there is little evidence that agglomeration is correlated with job turnover. This suggests that labour market pooling is an important source of agglomeration for only the most concentrated industries.

Before proceeding to examine the link between labour turnover and agglomeration more closely, it is worthwhile to review the summary statistics of sectoral job flows. Figure 3.1 displays the mean annual rates of job creation, job destruction, net job creation, and job reallocation.51 For the period 1973 to 1993, the average rates of job destruction, job creation, net job creation, and job reallocation were 8.9%, 10.7%, -1.8%,

51 The mean rates are unweighted averages for the 459 4-digit industries in my sample.
and 19.6%, respectively. The large magnitude of these numbers reflects the economy's ability to adjust to allocative shocks by shifting employment from dying plants, or plants that are doing poorly, to new plants or plants that are doing well. Since plants that are expanding tend to be more productive than plants that are contracting, this tends to raise the average level of productivity in the economy. Surprisingly, the rate of job destruction, although cyclically volatile, has not trended upwards, contrary to the popular presumption that previous generations of workers enjoyed more job security than the current generation.\textsuperscript{52} Moreover, there is no trend in job creation either. In most years, more jobs were destroyed than created. This reflects the net reduction in employment in U.S. manufacturing over the past 30 years.

Table 3.1 lists the top ten most concentrated industries and the ten least concentrated industries. Among the top ten most concentrated industries, what is most striking is that there is no obvious geographic reason why most of these industries should have such high levels of agglomeration.\textsuperscript{53} The most concentrated industry, fur goods (S.I.C. 2371), is heavily concentrated in New York State. Upon closer inspection, one would discover that most fur good plants are actually located in New York City. And upon even further inspection, one would discover that most fur plants are concentrated on the east side of Manhattan. The dominance of Manhattan in the fur industry is, like so

\textsuperscript{52} This is somewhat surprising given that the shares of publicly traded companies have become more volatile. While market and industry variances have remained fairly stable since the early sixties, the average variance of individual shares more than doubled from 1962 to 1997 (Campbell, Lettau, Malkiel, and Xu, 2000). As a result, the mean pair-wise correlation in share prices across companies fell by more than 60 percent between the 1960s and the 1990s. The presumption is that this reflects stronger idiosyncratic shocks at the firm level. Although it is natural to assume that if companies experience stronger idiosyncratic shocks they will also have more volatile hiring and firing patterns, the labour evidence suggests that this has not happened. One might argue that individual shares are more volatile because, for whatever reason, investors have become less rational. If this were true, "contrarian investment strategies" which buy shares of undervalued companies (low price to book and price to earnings ratio) and short sell shares of overvalued companies should have become more profitable. The evidence, however, suggests that this has not happened (Lakonishok, Shleifer and Vishny, 1994). A more plausible reason is that the importance of conglomerates (companies that produce a wide spectrum of often unrelated goods) has fallen since the 1960's. As companies have become more single-focused, the risk of holding their shares has subsequently increased. Furthermore, there has been a growing trend towards issuing equity in companies that have yet to show solid revenue growth or produce positive earnings. Since the share prices of "start-up" companies will depend on future earnings, as opposed current earnings, their shares will be more volatile.

\textsuperscript{53} The exceptions are Wines, Brandy, and Brandy Spirits (S.I.C. 2084) which is concentrated in the wine growing regions of California and Oil and Gas Field Machinery (S.I.C. 3533) which is concentrated in the oil producing areas of Texas.
many stories of agglomeration, based on past historical accidents. In this case, the accident happened more than 400 years ago, when the early Dutch settlers built trading posts in New York City (at that time, called New Amsterdam) to trade guns and brandy for furs with Native Americans.

Figure 3.2 plots the gamma (the measure of agglomeration) for all 459 4-digit industries in my sample. The gammas are ordered from the most concentrated to the least concentrated industry. The resulting curve resembles a "power curve". Although 446 out of 459 industries are more concentrated than what one would expect if locations were chosen randomly, there are only twenty industries with gammas over 0.2. It is these twenty industries that have relatively high rates of job reallocation (the rate of job creation plus the rate of job destruction).

Figure 3.3 plots the average industry employment level per decile (the first decile represents the 10 percent most concentrated industries and the 10th decile represents the 10 percent least concentrated industries). There is a perceptible upward trend, suggesting that less concentrated industries tend to higher employment levels. The average level of employment in the 20 percent least concentrated industries is 60 percent larger than the average level of employment in the 20 percent most concentrated industries. The most likely explanation for this result hinges on the fact that there are likely to be congestion costs from concentrating firms within too dense an area. A large industry in which there are many firms might find the congestion costs too great to permit most firms to locate in a single region. This will cause multiple pockets of firms to spring up in different regions, thus reducing the industry's level of agglomeration. A good example is the telecommunications industry. There are large pockets of telecom firms in at least five separate regions of the United States: Silicon Valley, Boston's Route 128, the Virginia Technology Triangle, and Northern Texas.

Figure 3.4 plots the rates of job creation, job destruction, and job reallocation from 1973 to 1993 for the ten most concentrated industries as a percentage of the mean rate for all industries for the corresponding year. In every year, the average annual rate of
job reallocation for the ten most concentrated industries exceeded the average annual rate of job reallocation for all industries. On average, over the twenty-year span of the sample, the annual rate of job reallocation was 26 percent higher in the ten most concentrated industries. Furthermore, as figure 3.4 also demonstrates, the higher rate of job reallocation in the ten most concentrated industries is the result of both high rates of job creation and job destruction. The average rate of job creation among the most highly concentrated industries was 20 percent higher than the industry mean and the average rate of job destruction was 35 percent higher. This suggests that the most highly concentrated industries were shedding employment at a faster rate than the typical industry.

Once we move away from the top ten most concentrated industries, the stark results we previously obtained begin to fade. It seems as though the link between labour turnover and agglomeration is strong only for the most concentrated industries. For the remaining industries, there seems to be no obvious relationship between agglomeration and labour turnover. Regression analysis bares this out. Table 3.2 provides results from 3 different years: 1991, 1992, and 1993. The 4-digit industry gamma is regressed on the rate of job creation, job destruction, and the product of job creation and job destruction, for the corresponding industry. The $R^2$ in both years is very low and the estimated coefficients are statistically insignificant.

### 3.7 Conclusions

This chapter has developed a general equilibrium model of labour market pooling. It was shown that firm-specific productivity shocks benefit an economy by drawing workers from less productive firms to more productive firms. The benefits from pooling labour within a single locality depend on two things: one, the total number of workers that can be pooled and two, the variance of productivity shocks that firms experience. It was shown that average productivity in the economy depends positively on both variables. This stark result underscores some of the limitations of the model. First, the model ignores congestion effects. One might easily imagine that as more workers crowd
into a region, this will put upward pressure on land prices and as result, will make living in the region more expensive. Presumably, this is why the entire population of the United States does not reside in one city and instead, is clustered in hundreds of different cities.

Second, the assumption that productivity shocks are uncorrelated across firms is, of course, overly simplistic. There is a fair amount of serial correlation of productivity shocks across firms in most industries. At the extreme, if productivity shocks are perfectly correlated across firms, there will be no gains from pooling labour in the same location. After all, if my firm is doing poorly, I would like to be able to move to a firm that is doing well. But if all firms are doing poorly, then there will be nowhere to turn.54

Third, the assumption that workers can freely and instantaneously move across firms is unrealistic. One can easily imagine a model in which there are costs associated in transplanting workers from one firm to another. In such a model, the benefit from relocating a worker from a low productivity firm to a high productivity firm would have to be weighed against the adjustment cost of doing so. The greater is the productivity gap between firms and the longer the gap is expected to persist, then the more incentive there will be for workers to change jobs. Nevertheless, with these qualifications in mind, the model still yields a number of fruitful results, the most important of which is the notion that productivity shocks can be a source of good if they occur in countries that have flexible labour markets and in which workers can move relatively easily across firms.

The vast majority of industries in my sample were more concentrated than what one would expect if location were determined by chance. Even so, only a small percentage (about 3 percent) had levels of concentration considerably above the national average. The empirical evidence presented in this chapter reveals that the top ten most concentrated industries had rates of job creation and job destruction that were 20 percent and 35 percent higher than the industry average, respectively. For the remaining

---

54 Diamond and Simon (1990) argue that specialized cities actually suffer a greater unemployment risk. Using the theory of compensating differentials, they argue that workers in U.S. cities get compensated for the additional risk of losing their jobs by receiving higher wages. Firms are willing to pay higher wages because the benefit from pooling more than outweighs the increased labour costs.
industries, there is little correlation between labour turnover and agglomeration. This suggests that high labour turnover, as a predictor of agglomeration, is important only for extremely concentrated industries.
Figures and Tables
Table 3.1
Most and Least Localized Industries

<table>
<thead>
<tr>
<th>10 Most Localized Industries</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S.I.C. Code</td>
<td>Industry Name</td>
<td>Gamma</td>
</tr>
<tr>
<td>2371</td>
<td>Fur Goods</td>
<td>0.63</td>
</tr>
<tr>
<td>2084</td>
<td>Wines, Brandy, Brandy Spirits</td>
<td>0.48</td>
</tr>
<tr>
<td>2252</td>
<td>Hosiery, n.e.c.</td>
<td>0.44</td>
</tr>
<tr>
<td>3533</td>
<td>Oil and Gas Field Machinery</td>
<td>0.43</td>
</tr>
<tr>
<td>2273</td>
<td>Carpets and Rugs</td>
<td>0.38</td>
</tr>
<tr>
<td>2429</td>
<td>Special Product Sawmills, n.e.c.</td>
<td>0.37</td>
</tr>
<tr>
<td>3961</td>
<td>Costume Jewelry</td>
<td>0.32</td>
</tr>
<tr>
<td>2895</td>
<td>Carbon Black</td>
<td>0.30</td>
</tr>
<tr>
<td>3915</td>
<td>Jewelers' Materials, Lapidary</td>
<td>0.30</td>
</tr>
<tr>
<td>2874</td>
<td>Phosphatic Fertilizers</td>
<td>0.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10 Least Localized Industries</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S.I.C. Code</td>
<td>Industry Name</td>
<td>Gamma</td>
</tr>
<tr>
<td>2371</td>
<td>Rubber and Plastics Footwear</td>
<td>-0.013</td>
</tr>
<tr>
<td>2084</td>
<td>Canned Specialties</td>
<td>-0.012</td>
</tr>
<tr>
<td>2252</td>
<td>Malt Beverages</td>
<td>-0.010</td>
</tr>
<tr>
<td>3533</td>
<td>Household Vacuum Cleaners</td>
<td>-0.009</td>
</tr>
<tr>
<td>2273</td>
<td>Prerecorded Records and Tapes</td>
<td>-0.008</td>
</tr>
<tr>
<td>2429</td>
<td>Small Arms Ammunition</td>
<td>-0.004</td>
</tr>
<tr>
<td>3961</td>
<td>Steel Investment Foundries</td>
<td>-0.003</td>
</tr>
<tr>
<td>2895</td>
<td>Elevators and Moving Stairways</td>
<td>-0.001</td>
</tr>
<tr>
<td>3915</td>
<td>Cookies and Crackers</td>
<td>-0.0009</td>
</tr>
<tr>
<td>2874</td>
<td>Macaroni and Spaghetti</td>
<td>-0.0008</td>
</tr>
</tbody>
</table>
Table 3.2
Regression Results – Dependent Variable: Gamma (Measure of Concentration)

<table>
<thead>
<tr>
<th></th>
<th>1991</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>Creation</td>
<td>Destruction</td>
<td>Creation*Destruction</td>
</tr>
<tr>
<td>Estimated Coefficient</td>
<td>0.04805</td>
<td>-0.00191</td>
<td>0.00182</td>
<td>0.00026</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.01203</td>
<td>0.00130</td>
<td>0.00095</td>
<td>0.00087</td>
</tr>
<tr>
<td>t-ratio</td>
<td>3.99518</td>
<td>-1.46627</td>
<td>1.91871</td>
<td>0.29917</td>
</tr>
</tbody>
</table>

R²=0.03170

<table>
<thead>
<tr>
<th></th>
<th>1992</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>Creation</td>
<td>Destruction</td>
<td>Creation*Destruction</td>
</tr>
<tr>
<td>Estimated Coefficient</td>
<td>0.04726</td>
<td>-0.00130</td>
<td>0.00033</td>
<td>0.00012</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.01116</td>
<td>0.00122</td>
<td>0.00087</td>
<td>0.00009</td>
</tr>
<tr>
<td>t-ratio</td>
<td>4.23339</td>
<td>-1.06610</td>
<td>0.38585</td>
<td>1.33469</td>
</tr>
</tbody>
</table>

R²=0.01950

<table>
<thead>
<tr>
<th></th>
<th>1993</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>Creation</td>
<td>Destruction</td>
<td>Creation*Destruction</td>
</tr>
<tr>
<td>Estimated Coefficient</td>
<td>0.04764</td>
<td>-0.00186</td>
<td>0.00180</td>
<td>0.00003</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.01202</td>
<td>0.00130</td>
<td>0.00095</td>
<td>0.00009</td>
</tr>
<tr>
<td>t-ratio</td>
<td>3.96243</td>
<td>-1.42932</td>
<td>1.89489</td>
<td>0.30471</td>
</tr>
</tbody>
</table>

R²=0.03085

* The independent variable is the 4 digit SIC industry gamma from Ellison and Glaeser (1994) and the dependent variables are i) the rate of job creation, ii) the rate of job destruction and iii) the product of the rate of job creation and job destruction. All the job flow data is from Davis and Haltiwanger (1996).
Figure 2.1

$\alpha = 10, \beta = 1, \gamma = 0.8, \delta = 1, \mu = 0.8, \sigma = 2, \theta = 2, L = 100$

Wage (in terms of the numeraire, food)

Wage in North

Wage in South

Symmetric Wage

Iceberg Trade Cost

$\tilde{\tau}$

$\tau_{\text{max}}$
Figure 2.2

\[ \alpha = 10, \beta = 1, \gamma = 0.8, \delta = 1, \mu = 0.5, \sigma = 2, \theta = 2, L = 100 \]
Figure 2.3
\[ \alpha=10, \beta=1, \gamma=0.8, \delta=1, \mu=0.8, \sigma=2, \theta=2, L=100 \]
Figure 3.1
Mean Annual Rates of Job Creation, Job Destruction, Net Job Creation, and Job Reallocation (1973-1993)
Figure 3.2
4-Digit Industry Gamma

Industry Rank by Gamma (From Most Concentrated to Least Concentrated)
Figure 3.3
Average Industry Employment Level by Decile

Decile (From Most Concentrated to Least Concentrated Industry)
Figure 3.4
Rate of Job Creation and Job Destruction for the 10 Most Concentrated Industries as a Percentage of the Respective Mean for All Industries
Bibliography


Lösch, A. (1940), *The Economics of Location* (Jena: Fischer).


