

NBER WORKING PAPER SERIES

HETEROGENEOUS FIRMS, AGGLOMERATION
AND ECONOMIC GEOGRAPHY: SPATIAL
SELECTION AND SORTING

Richard Baldwin
Toshihiro Okubo

Working Paper 11650
<http://www.nber.org/papers/w11650>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2005

We would like to thank Marc Melitz, Tony Venables and Jan Haaland for suggestions, Virginia di Nino for excellent proof reading and the 4, 5 or 6 referees that looked at this for the Journal of Economic Geography. We especially thank Diego Puga, the editor, for his valuable inputs that have substantially improved our paper. Funding was provided by the Swiss National Science Foundation (100012-105675/1). The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

©2005 by Richard Baldwin and Toshihiro Okubo. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Heterogeneous Firms, Agglomeration and Economic Geography: Spatial Selection and Sorting
Richard Baldwin and Toshihiro Okubo
NBER Working Paper No. 11650
September 2005
JEL No. H32, P16

ABSTRACT

A Melitz-style model of monopolistic competition with heterogeneous firms is integrated into a simple New Economic Geography model to show that the standard assumption of identical firms is neither necessary nor innocuous. We show that re-locating to the big region is most attractive for the most productive firms; this implies interesting results for empirical work and policy analysis. A 'selection effect' means standard empirical measures overestimate agglomeration economies. A 'sorting effect' means that a regional policy induces the highest productivity firms to move to the core while the lowest productivity firms to move to the periphery. We also show that heterogeneity dampens the home market effect.

Richard Baldwin
Graduate Institute of International Studies
11A, avenue de la Paix
CH-1202 Geneva
SWITZERLAND
and NBER
baldwin@hei.unige.ch

Toshihiro Okubo
Graduate Institute of International Studies
11A, avenue de la Paix
CH-1202 Geneva
SWITZERLAND
okubo3@hei.unige.ch

Heterogeneous firms, agglomeration and economic geography: Spatial selection and sorting

Richard E. Baldwin and Toshihiro Okubo¹

Graduate Institute of International Studies, Geneva

September 2005

ABSTRACT

A Melitz-style model of monopolistic competition with heterogeneous firms is integrated into a simple New Economic Geography model to show that the standard assumption of identical firms is neither necessary nor innocuous. We show that re-locating to the big region is most attractive for the most productive firms; this implies interesting results for empirical work and policy analysis. A ‘selection effect’ means standard empirical measures overestimate agglomeration economies. A ‘sorting effect’ means that a regional policy induces the highest productivity firms to move to the core while the lowest productivity firms to move to the periphery. We also show that heterogeneity dampens the home market effect.

JEL H32, P16.

Keywords: heterogeneous firms, economic geography, estimation of agglomeration economies, home market effect.

1. INTRODUCTION

One of the great contributions of the new economic geography (NEG) is to explicitly model “the self-reinforcing character of spatial concentration” (Fujita, Krugman and Venables 1999 p.4). The early work in this literature, e.g. Krugman (1991), and Venables (1996), achieved this with a modelling approach – Dixit-Stiglitz monopolistic competition – that ignored many important aspects of locational economics. An intense effort by theorists over the past decade has broadened the modelling to allow for many important effects, much of this relies on the monopolistic competition framework of Ottaviano, Tabuchi and Thisse (2002). See Fujita and Thisse (2002) for a succinct synthesis of this work.

One of the most convenient, but least realistic assumptions in the new economic geography (NEG) literature is that of identical firms. An extensive empirical literature shows that firms vary enormously in terms of size (Cabral and Mata 2003) as well as in terms of productivity and trade behaviour (Bernard, Jensen and Schott 2003, Helpman, Melitz and Yeaple 2004). Our paper argues that this ‘assumption of convenience’ is neither necessary nor innocuous in NEG models. A more recent empirical literature suggests that big plants are more likely to be

¹ We would like to thank Marc Melitz, Tony Venables and Jan Haaland for suggestions, Virginia di Nino for excellent proof reading and the 4, 5 or 6 referees that looked at this for the Journal of Economic Geography. We especially thank Diego Puga, the editor, for his valuable inputs that have substantially improved our paper. Funding was provided by the Swiss National Science Foundation (100012-105675/1).

found in clustered in areas that are specialised in a particular sector (Lafourcadey and Mionz 2003, Alsleben 2005).

We show how a Melitz (2003) style model of monopolistic competition with heterogeneous firms can be integrated into a simple NEG setting. We use the model to demonstrate that relaxing the standard assumption of homogenous firms has several important implications – all of which turn on the fact that re-locating to the big region is most attractive for the most productive firms.

Intuition for spatial selection

Before turning to the implications, we provide intuition for why spatial selection occurs in our model, i.e. why firms that move to large markets tend to have above-average productivity. In most NEG models, the spatial equilibrium occurs at a degree of spatial concentration where the agglomeration forces balance the dispersion forces. In simple NEG models, the agglomeration forces consist of backward and forward linkages, while the dispersion force consists of local competition (also called market crowding). Highly productive firms are systematically subject to greater agglomeration forces and weaker dispersion forces than are less productive firms. Because more productive firms have lower marginal costs, they tend to sell more so the backward and forward linkages operating in the bigger market are systematically more attractive to the most efficient firms. Likewise these firms' high productivity also means that they are systematically less harmed by the higher degree of local competition in the big market. Plainly, then the delocation of firms from a small region to a large region will involve spatial selection as far as firm-level productivity is concerned. The most productive firms will move to the big market first.

Implications

The first implication is a cautionary tale for empirical researchers. Since relocation is a non-random process, a 'selection effect' plagues standard empirical techniques for measuring agglomeration economies. We sign the bias, showing that standard techniques will overestimate the importance of agglomeration economies since firms that move to the agglomerated region have above average firm-level productivity independently of any agglomeration economies.

The second concerns the impact of regional policy. Most regional policies aim to increase the share of industry in periphery regions. Taking production subsidies as an example, we show that regional policies tend to attract the least productive firms since they have the lowest opportunity cost of leaving the agglomerated region (or not moving there in the first place). The result is a 'sorting effect'. A policy that succeeds in increasing the periphery region's share of industry will induce the highest productivity firms to move to the core and the lowest productivity firms to move to the periphery. This sorting has several implications for policy. For example, it may explain why modest production subsidies have very little impact on regional welfare. Small subsidies attract few firms and all of these are intrinsically inefficient.

1.1. *Previous literature*

Our paper is not the first to consider selection effects in a model with heterogeneous firms. For example, Melitz and Ottaviano (2003) and Melitz (2003), among others consider selection in the sense of the elimination of the least efficient firms within a nation. These papers, however, ruled out the possibility of spatial relocation by fiat. Our paper concerns a very different aspect of selection, namely the spatial dimension of selection.

This distinction between national selection and spatial matters when it comes to our primary contribution – the idea that standard econometric techniques for measuring agglomeration economies overestimate the strength of these forces. One observes a correlation between geographical clustering of firms and high average productivity. But which causes which? On one hand, it may be that the big market makes firms more productive (agglomeration economies); on the other hand, it may be that the geographical gathering of the most productive firms raises big market's average productivity (spatial selection). Here is an example of why this distinction matters. If the cluster-productivity correlation is due to agglomeration economies, a nation can raise the total output of its productive factors by encouraging spatial clustering. However, if the correlation is due to spatial selection, a pro-clustering policy merely fosters spatial inequality.

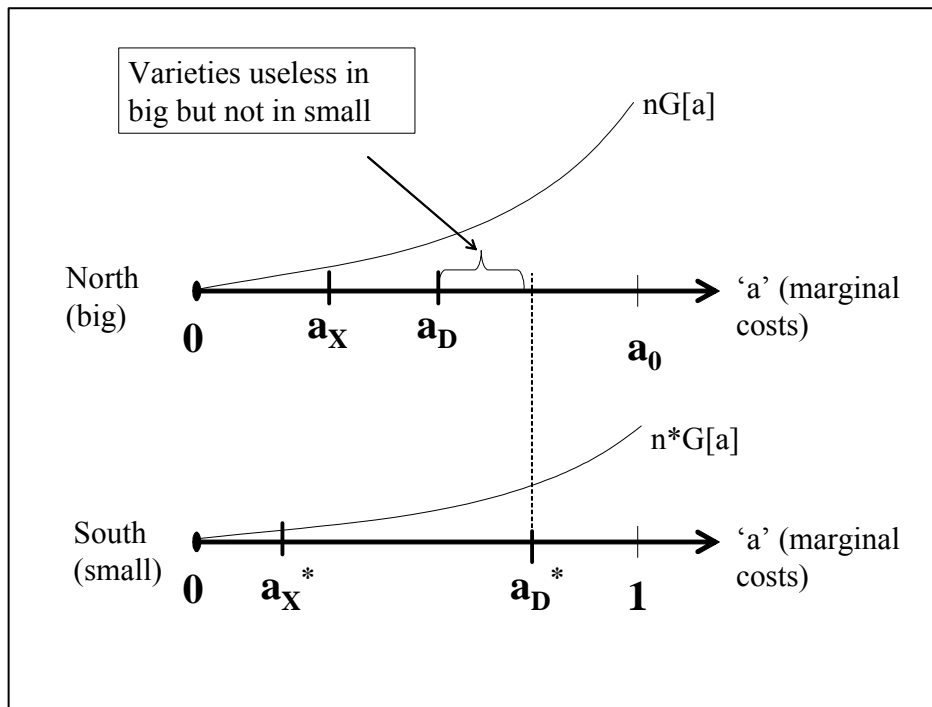
1.1.1 Melitz and Ottaviano (2003)

This paper marries the Ottaviano, Tabuchi and Thisse (2002) framework (OTT for short) of monopolistic competition in a linear demand system (assuming zero inter-regional mobility of factors) with the Melitz-Hopenhayn mechanism for the development of firms, each associated with a particular labour input coefficient (i.e. marginal cost). The Melitz-Hopenhayn mechanism assumes a continuous drawing of new firms (since firms are continuously dying according to a Poisson process) from the underlying distribution $G[a]$, with support $[0..a_0]$; this yields a mass of potential firms $nG[a]$, but only the most efficient find it worthwhile to actually producing.

Specifically, firms that have marginal costs above the demand curve intersection will not sell/produce anything (firms are atomistic and so take the intercept as given even though it is endogenous in the aggregate; greater competition lowers the intercept). In short, competition truncates the distribution of marginal costs at some point, call it a_D , so only firms with marginal costs less than a_D actually produce; the unlucky innovators who drew a 's greater than a_D let bygones be bygones, realizing ex post that they have wasted their fixed innovation costs. The mass of active firms rises to the point where the ex ante expected profits from getting an ' a ' less than a_D just balances the ex ante expected loss of getting an $a > a_D$, taking proper account of the fixed innovation costs. See Figure 1, but ignore a_X for the moment.

As usual in the linear demand system of OTT framework, a bigger market supports more firms and tougher competition. Thus the truncation point is lower (i.e. a wider range of inefficient firms are eliminated) in bigger markets. This is a new agglomeration force that Melitz-Ottaviano (MO) have found – one that stems from the combination of heterogeneous firms and the standard pro-competitive effect that arises in OTT (but not in the Dixit-Stiglitz setting). In short, the cluster-productivity correlation in MO due *only* to agglomeration economies, i.e. the big market makes the average firm more productive by expanding production of the most efficient firms and eliminating the worst firms.

To highlight the difference between the MO paper and our own, consider the following thought experiment. Suppose there are two such nations that are fundamentally different in size and they begin to trade but trade is costly. Since trade is costly, the effective marginal cost of selling in the export market is higher, so there will be some firms that find it worth their while to sell locally, but not to export (call these D-type firms), while others – the most efficient firms – find it worth their while to export (call these X-types). Referring to the threshold, maximum marginal cost for X-types as a_X , we see that market size will matter. In particular a_X will be higher in the big market (since they are exporting to the smaller market with weaker competition). See Figure 1.

Figure 1: Melitz-Ottaviano thresholds

Now imagine we allowed firms to relocate in the MO framework. The equilibrium outcome is complex, but it is absolutely obvious that some firms would move (see the appendix for details). To take the easiest example, there is a range of varieties that are ‘lying fallow’ in the big market but which would have positive value in the small market (i.e. these varieties have been developed and could produce if conditions were right).

Of course the relocation will itself change all the thresholds and the mass of active firms, but our basic point is that relocation would change things. MO assumed away relocation because, presumably, they wanted to focus the international trade aspects (the main axis of the heterogeneous-firms trade models).

1.1.2 Nocke (2003)

Most of the paper is spent establishing that the Hopenhayn mechanism and imperfect competition produce the basic within-nation truncation results that are familiar from heterogeneous-firms literature. The big difference between Nocke and Melitz (2003) is that Nocke assumes that the process generates new entrepreneurs and these are without geographical location in the sense that new entrepreneurs find it equally costly to locate in any market after they know their efficiency, whereas the Melitz and MO models firms are tied to the nation in which they are created.

There are four main differences between Nocke (2003) and our paper. First, Nocke explicitly states (p.12) that his sorting result will not go through when consumers have a CES utility function as in the Dixit-Stiglitz model of monopolistic competition. We work with Dixit-Stiglitz monopolistic competition and do show a spatial selection effect and sorting with subsidies. This is not a critique of Nocke but rather *prima facie* evidence that his paper and ours are quite different. Second, Nocke assumes that heterogeneity is associated with ‘entrepreneurs’, i.e. people, not firms. For this reason, his sorting result has been viewed as belonging to the migration-sorting literature. Third, Nocke works in a partial equilibrium setting (e.g., no resource constraints, no wage determination). Fourth, the entrepreneurs in

Nocke never consider the local price index when choosing location, even though comparison of real rewards is standard in the international trade and economic geography literature. Again this is not a critique, but it does show that he is working in a model where things are held constant that cannot be held constant in Economic Geography models. In summary, there is no doubt that Nocke (2003) presents sorting results similar to ours, but his paper is really quite different.

1.1.3 Spatial selection models

A number of papers in the theoretical literature study various forms of heterogeneity in economic geography models. Tabuchi and Thisse (2002) investigate the impact of heterogeneous tastes for living in various regions. They show that this location-taste-heterogeneity acts as a strong dispersion force and removes the unrealistic, bang-bang predictions of the early NEG models.

Amiti and Pissarides (2002) also model heterogeneous labour but the heterogeneity concerns idiosyncratic worker characteristics rather than locational preferences. Since workers are idiosyncratic, workers and firms cannot know how good of a ‘match’ they will make *ex ante*. The quality of the match is assumed to affect worker productivity, so a ‘thick market externality’ arises and acts as an agglomeration force in the spirit of Marshall’s labour-market-pooling. Another paper in this line is Coniglio (2001). This paper allows for high and low skilled workers and assumes that skill premium is an increasing function of the number of high-skilled workers in a region (the implicit story is one of knowledge spillovers). Combes, Duranton and Gobillon (2004) use micro data to show that worker heterogeneity is important and that workers appear to sort themselves geographically. They make the point that failing to control for heterogeneity among workers can bias estimates of agglomeration economies upward.

Heterogeneity among workers and its importance for economic geography is quite clear as the empirical literature and migrant self-selection demonstrates (Borjas, Bronars, Trejo. 1992, and Chiswick 1999). However, at least in Europe, labour mobility is fairly limited both within and especially between nations, so labour heterogeneity is unlikely to be the only form of heterogeneity that is relevant to agglomeration.

Without denying the importance of labour heterogeneity for many forms of agglomeration, we focus on a complementary form of heterogeneity, namely firm-level productivity differences. We believe that this form of heterogeneity may be especially important for aspects of economic geography where labour mobility is not a key factor, but it probably operates even in situations where labour mobility is high. Moreover, as pointed out above, firm-level productivity differences have been well documented and are very large. Our paper is a first step in studying the economic geography implications of such heterogeneity.

Dupont and Martin (2003) study the impact of subsidies in a NEG setting with homogeneous firms. They show that the impact on location of such subsidies is stronger when trade is freer due to home market magnification effect. They also study the income distribution effects finding that although subsidies “constitute an official financial transfer from the rich to the poor region, they actually lead to an income transfer from the poor to the rich region” in certain cases.

The rest of the paper is organised in 4 sections. Section 2 presents the basic model, Section 3 demonstrates the ‘selection effect’ and points out its implications for empirical work, Section 4 demonstrates the ‘sorting effect’ of production subsidies and Section 5 presents our concluding remarks.

2. THE BASIC MODEL

This section introduces a simple new economic geography (NEG) model with heterogeneous firms.

The model can be thought of as the familiar NEG model of Martin and Rogers (1995) – also known as the footloose capital model, or FC model for short – extended to allow for exogenous heterogeneity in firms' marginal costs. The heterogeneity in our model is akin to the heterogeneity in Melitz (2003) but the firm generation and selection process is radically simplified in so as to concentrate attention on the spatial aspects of the model.

2.1. *The footloose capital model with heterogeneous firms*

The FC model assumes a fixed stock of firms in order to spotlight the location effects of freer trade.² It achieves this by assuming that the fixed-cost element for each differentiated variety involves a single unit of capital. Each nation's overall number of varieties/firm is thus pinned down by its capital endowment. We follow this tradition but add exogenous heterogeneity.

2.1.1 Basic set up

As in the standard FC model, we assume two regions, two sectors and two factors. The factors – physical capital K and labour L – are inelastically supplied in each region; capital is assumed to be inter-regionally mobile, while labour is immobile. The regions – referred to as the north and the south – are symmetric in terms of tastes, technology and openness to trade. They are also endowed with the same relative factor supplies, but the north is larger in a pure sense, i.e. its endowment of labour and capital are proportionally larger than the south's. This rules out Heckscher-Ohlin motives for trade and thus allows us to concentrate more precisely on agglomeration forces.

The two sectors are manufacturing and an outside good, referred to as 'agriculture' for convenience. The agricultural sector is kept as simple as possible. It produces a homogeneous good using only labour under constant returns and perfect competition; its output is traded costlessly.

Manufacturing is marked by increasing returns, Dixit-Stiglitz monopolistic competition and iceberg trade costs. The cost function of a typical manufacturing firm in the FC model is non-homothetic; the fixed cost involves *only* capital and the variable cost involves *only* labour. Specifically, each manufacturing firm requires one unit of K and 'a' units of labour per unit of output. This means that the increasing returns sector is intensive in the use of the mobile factor, as in most NEG models.

Capital owners are immobile across regions, so when pressures arise to concentrate manufacturing in one region, physical capital moves but its reward is repatriated to its country of origin.

Because physical capital can be separated from its owners, the region in which capital's income is spent may differ from the region in which it is employed. We must therefore distinguish the share of world capital owned by northern residents (we denote this as $s_K \equiv K/K^W$) from the share of world capital employed in the north. Because we assume that

²See Baldwin, Forslid, Martin, Ottaviano and Robert-Nicoud, 2003, Chapter 3 for a complete analysis of this model and comparison to other NEG models in core-periphery tradition (i.e. those based on Dixit-Stiglitz monopolistic competition) and to NEG models based on the OTT monopolistic competition framework.

each manufacturing variety requires one unit of capital, the share of the world capital stock employed in a region exactly equals the region's share of world manufacturing. Consequently, we can use north's manufacturing share, i.e. $s_n \equiv n/(n+n^*)$, to represent the share of capital employed in the north.

The tastes of the representative consumer in each region are quasi-linear:

$$U = \mu \ln C_M + C_A, \quad C_M \equiv \left(\int_{i \in \Theta} c_i^{1-1/\sigma} di \right)^{1/(1-1/\sigma)}, \quad 0 < \mu < 1 < \sigma$$

where C_M and C_A are, respectively, consumption of the composite of M-sector varieties and consumption of the A-sector good, and σ is the constant elasticity of substitution between any two M-sector varieties; Θ is the set of all varieties produced. This set of varieties is pre-determined by endowments since each variety requires a unit of capital and the world capital stock is fixed.

Since capital moves without its owners, capital moves in search of the highest *nominal* reward rather than the highest real reward since its income is spent in the owner's region regardless of where the capital is employed (here nominal means the reward in terms of the numeraire; real means the reward in terms the ideal price index).

We extend the FC model in two ways; we add heterogeneity in firms' marginal production costs and we assume firms face quadratic adjustment cost when switching regions.

2.1.2 Additional assumptions: firm heterogeneity and delocation costs

Following Melitz (2003), we allow firms to have different unit input coefficients, i.e. different a 's. One of the major contribution of Melitz (2003) is to endogenise the distribution of firm-level productivity and characterise the influence of that openness has on it. For our purposes, however, we are not fundamentally interested in the overall distribution of firm-level productivity; we are interested in how agglomeration and policy affects its geographic distribution.³

To focus on these goals, we take the distribution of firm-level efficiency as part of each region's endowment. Since each firm is associated with a particular unit of capital, it is natural to assign the source of heterogeneity to capital. We assume that each unit of capital in each region is associated with a particular level of productive efficiency as measured by the unit labour requirement, ' a '. The distribution assumed is Pareto:

$$(1) \quad G[a] = \left(\frac{a^\rho}{a_0^\rho} \right), \quad 1 \equiv a_0 \geq a \geq 0, \quad \rho \geq 1$$

where $a_0 < \infty$ is the scale parameter (highest possible marginal cost) and ρ is shape parameter. Since we are free to choose units of M-sector goods, we can normalise a_0 to unity without loss of generality. Note that unlike the Melitz model, all of our firms sell in both markets as long as trade costs are finite since we do not allow for 'beachhead market-entry costs' as in Melitz (2003).⁴

³ We have explored the FC model with a full-blown Melitz model, but found the results were qualitatively identical but the reasoning was much less transparent. The one extra result concerns the fact that some firms with fairly high marginal costs may cease to sell to both regions after they move to the big region.

⁴ Melitz (2003) assumes that firms must incur a fixed cost to establish a 'beachhead' in a market, i.e. to sell in that market, so only sufficiently efficient firms sell in both markets.

Second, we deviate from the FC model in that we assume that relocation is subject to quadratic adjustment costs. In particular, the cost of switching regions is χ units of labour per firm, where χ depends upon the flow of firms relocating. The specific assumption is:

$$(2) \quad \chi = \gamma m$$

where m is the flow of migrating firms. This means that in steady state, when all delocation has ceased, the migration costs are zero on the margin.

2.2. Intermediate results and short run equilibrium

Results for the A sector in this sort of model are simple and well known. Constant returns, perfect competition and zero trade costs equalise nominal wage rates across regions. We choose units of A and the numeraire such that $w=w^*=1$. This means that all differences in M-firms' marginal costs are due to differences in their a 's.

Utility maximisation generates the familiar CES demand functions.⁵ These, together with the standard Dixit-Stiglitz monopolistic competition assumptions on market structure imply 'mill pricing' is optimal and that operating profit earned by a typical firm in a typical market is $1/\sigma$ times firm-level revenue.⁶ Accordingly, operating profit realised by a south-based firm is:

$$(3) \quad \pi^*[a] = \left(\frac{a}{1-1/\sigma}\right)^{1-\sigma} \left(\frac{\phi s_E}{\int_{i \in \Theta} p_i^{1-\sigma} di} + \frac{1-s_E}{\int_{i \in \Theta} p_i^{*1-\sigma} di}\right) \frac{E^w}{\sigma}; \quad s_E \equiv \frac{E}{E^w}, \quad \phi \equiv \tau^{1-\sigma}$$

where E^w is world expenditure on M-goods, s_E is the northern share of this expenditure, ϕ is the free-ness of trade ($\tau \geq 1$ is the iceberg trade cost, so $\phi=0$ with infinite trade costs and $\phi=1$ with costless trade). The consumer prices in north and south are denoted p and p^* , with Θ representing the set of varieties produced (all varieties are sold in both regions).

Using mill pricing and cancelling the $(1-1/\sigma)$ terms, northern and southern operating profit as a function the firm's ' a ' can be written as:

$$(4) \quad \pi[a] = a^{1-\sigma} \left(\frac{s_E}{\Delta} + \frac{\phi(1-s_E)}{\Delta^*}\right) \frac{E^w}{K^w \sigma}, \quad \pi^*[a] = a^{1-\sigma} \left(\frac{\phi s_E}{\Delta} + \frac{(1-s_E)}{\Delta^*}\right) \frac{E^w}{K^w \sigma}$$

where the denominators of the northern and southern demand functions (Δ is a mnemonic for denominator):⁷

$$\Delta \equiv s_K \int_0^1 a^{1-\sigma} dG[a] + (1-s_K) \phi \int_0^1 a^{1-\sigma} dG[a];$$

$$\Delta^* \equiv s_K \phi \int_0^1 a^{1-\sigma} dG[a] + (1-s_K) \int_0^1 a^{1-\sigma} dG[a]; \quad s_K \equiv \frac{K}{K^w}, 1-s_K \equiv \frac{K^*}{K^w}$$

⁵ Individual demand for a typical variety j is $c(j)=p(j)^{-\sigma}\mu/\Delta$, where $\Delta \equiv \int p(i)^{1-\sigma} di$ and the integral is over all available varieties, μ is expenditure on all varieties.

⁶ A typical first order condition is $p(1-1/\sigma)=wa$; rearranging, the operating profit, $(p-wa)c$, equals pc/σ .

⁷ To go from the denominator of the CES demand function defined in terms of an integral over goods to one defined in terms of an integral of marginal costs, we use the density of a 's in each nation, namely K times $G[a]$ for north and K^* times $G[a]$ for south. The support in both cases is the unit interval.

when no firms have relocated to the north. Here K^w is world endowment of K , which is also the mass of varieties produced worldwide; s_K is the share varieties produced in the north since K and K^* are the north's and south's capital endowments, respectively.

Solving the integrals, using (1) and assuming $1-\sigma+\rho>0$ (so the integrals converge) we have:⁸

$$(5) \quad \Delta = \lambda(s_n + \phi(1 - s_n)); \quad \Delta^* = \lambda(\phi s_n + 1 - s_n); \quad \lambda \equiv \frac{\rho}{1 - \sigma + \rho} > 0$$

Intuition. Notice that the deltas can be viewed as a measure of the degree of competition in each market. For example, Δ increases with s_K since an increase in the share of varieties that are locally produced (as opposed to imported) means more varieties are sold without trade costs and this intensifies competition in the local market.⁹ Likewise, $a^{1-\sigma}$ can be viewed as the competitiveness of a firm with marginal cost 'a', since this rises as its marginal cost fall. Combining these points, we see that a firm's market share – which equals $a^{1-\sigma}/\Delta K^w$ in the north – depends upon its relative competitiveness (if all firms had the same marginal cost, the market share of each firm would be $1/K^w$). Alternatively, we can view the firm's market share, $a^{1-\sigma}/\Delta K^w$, as varying with the ratio of its marginal cost to a weighted average of its competitors' marginal costs.

2.3. Delocation tendencies

Starting from the initial situation where no firms have moved, we turn now to considering the delocation tendency of firms. The standard logic of the Home Market Effect (HME) tells us that the big market (north) will have a more than proportional share of industry. In the traditional FC model, the implications of this are completely captured by the share of industry in the big market. But when firms are heterogeneous, an additional question arises. Which firms delocate first?

To work this out, we start from a situation where no delocation has occurred, so $s_n=s_K$ (i.e. the north's share of industry exactly matches its share of capital). Since firms are atomistic, the change in operating profit from a single firm moving from south to north (small region to big) is of function of it the firm's marginal cost, 'a', namely:¹⁰

$$(6) \quad \pi[a] - \pi^*[a] = a^{1-\sigma} \left(\left(\frac{s_E}{\Delta} - \frac{1-s_E}{\Delta^*} \right) (1-\phi) \right) \frac{E^w}{\sigma K^w}$$

Starting from the initial situation where no firms have moved, so $s_n=s_K$ and using the symmetry of region's relative factor endowments i.e. $s_E=s_K>1/2$, (6) simplifies to:

$$a^{1-\sigma} \left(\frac{(1-\phi)E^w}{\lambda \sigma K^w} \right) \frac{2\phi(s - \frac{1}{2})}{((1-\phi)s + \phi)(1-s + \phi s)}$$

where 's' is the north's share of world E and K (i.e., $s \equiv s_E=s_K>1/2$).

There are three key features of this expression. First, the term in large brackets is positive since north is larger, so it is clear that every southern firm would gain from being the first to

⁸ Since firms are atomistic, the first firm to move has no impact on the Δ 's.

⁹ Note that price-cost markets are always fixed with Dixit-Stiglitz competition, so the 'local competition' effect might more precisely be called the 'market crowding' effect since higher local competition results in lower sales per firm with no change in the markup.

¹⁰ Since firms are atomistic, the first firm to move has no impact on the Δ 's.

delocate from south to north. Second, no northern firm would gain from moving south. Third, the size of the gain for south-to-north migration is greatest for the most efficient firms.

2.3.1 Which firms move first?

It is intuitively obvious that the first firms that will find it profitable to pay the quadratic delocation costs will be those that have the most to gain, namely the most efficient southern firms.¹¹ The movement of efficient southern firms to the bigger market in the north changes the degree of competition in the two markets, that is to say, the Δ 's are affected by delocation.

To work out the feedback between migration and the Δ 's, we define the threshold level of marginal costs for migration as a_R where the 'R' stands for 'relocate'. We shall provide the condition characterising this cut-off level, but taking it as given for the moment, we note that the migration of the most efficient southern firms to the north will change the equilibrium Δ and Δ^* . Specifically:

$$\Delta = s \int_0^1 a^{1-\sigma} dG[a] + (1-s) \left\{ \int_0^{a_R} a^{1-\sigma} dG[a] + \phi \int_{a_R}^1 a^{1-\sigma} dG[a] \right\},$$

$$\Delta^* = \phi s \int_0^1 a^{1-\sigma} dG[a] + (1-s) \left\{ \phi \int_0^{a_R} a^{1-\sigma} dG[a] + \int_{a_R}^1 a^{1-\sigma} dG[a] \right\}; \quad K^w \equiv 1$$

The first expression reflects the north's degree of local competition. The first term in the top expression reflects the prices of north-made varieties sold in the north; the 's' in front of the integral reflects the north's share of K, namely s_K , but by symmetry of relative endowments s_K equals the relative size of the north's market, i.e. 's'.¹² The second expression reflects the prices of southern firms' varieties that are produced in the north. To understand this, recall that southern firms with a's in the range $[0, a_R]$ have relocated and thus become north-based firms. The third expression reflects the prices of southern varieties that are made in the south and exported to the northern market. The second expression is the isomorphic formula for the southern market. We have normalised $K^w=1$ to lighten the notation.

Solving the integrals using (1):

$$(7) \quad \Delta = \lambda \left(s + (1-s)a_R^{1-\sigma+\rho} + \phi(1-s)(1-a_R^{1-\sigma+\rho}) \right)$$

$$\Delta^* = \lambda \left(\phi s + \phi(1-s)a_R^{1-\sigma+\rho} + (1-s)(1-a_R^{1-\sigma+\rho}) \right)$$

Given these expressions that link the Δ 's to the range of firms that have migrated, namely a_R , we can write the value of delocation for any atomistic southern firm as a function of its own marginal cost and the range of firms that have already moved. Specifically:

$$(8) \quad v[a, a_R] \equiv \pi[a, a_R] - \pi^*[a, a_R]$$

where

¹¹ More formally, note that all atomistic southern firms would want to move first if the delocation cost were zero. However, if all tried to move at the same time, the flow of migrants would be infinite and so the quadratic delocation costs would also be infinite. This tells us that some, but not all firms will want to be the first to move. Suppose that the flow of firms, call it 'z', is positive and finite. Then the quadratic adjustment cost would also be positive and finite. Consequently, only sufficiently efficient southern firms would want to move. This explains more formally, the claim made in the text.

¹² Recall that we have cancelled price-marginal cost markups, so the a's reflect the prices.

$$\pi[a, a_R] = a^{1-\sigma} \left(\frac{s_E}{\Delta[a_R]} + \phi \frac{1-s_E}{\Delta^*[a_R]} \right) \frac{E^w}{\sigma}, \quad \pi^*[a, a_R] = a^{1-\sigma} \left(\phi \frac{s_E}{\Delta[a_R]} + \frac{1-s_E}{\Delta^*[a_R]} \right) \frac{E^w}{\sigma}$$

Turning from the benefits of relocation to the costs, note that the stock of southern firms in the north is $K^* a_R^\rho$, given (1). Thus the flow of migrating firms is $m = K^* \rho a_R^{\rho-1} \dot{a}_R$, where the ‘dot’ indicates a time derivative as usual. Given this and (2), the cost of moving is:

$$(9) \quad \chi = \gamma K^* \rho a_R^{\rho-1} \dot{a}_R$$

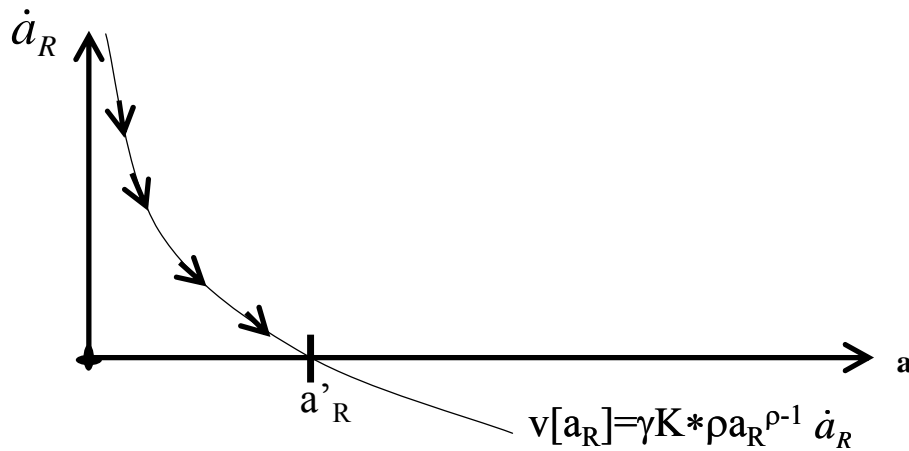
Equilibrium delocation process

Profit maximising firms will delocate if the benefit of doing so is greater than or equal to the costs. Thus, during the transition to the long-run steady state, firms will move in order of rising marginal costs. In particular, the ‘a’ of migrating firms is at any instant is pinned down by the equality of the cost and benefit of migrating, so the value to the marginal firm of migrating will be $v[a_R, a_R]$, which we write as $v[a_R]$ for short. Using (8) and (9), this means:

$$(10) \quad v[a_R] = \gamma K^* \rho a_R^{\rho-1} \dot{a}_R$$

This describes the delocation process fully. Because $v[a_R]$ is declining in a_R , as inspection of (7) confirms, the transitional delocation process is stable and converges to the long-run steady state level a_R' as shown in Figure 2.

Figure 2: Delocation with quadratic adjustment costs.



The intuition for the transitional process is straightforward. Since the most efficient firms move first, progressive delocation reduces competition in the southern market and raises it in the northern market. As long as trade is not too free, the process will stop at some intermediate level of delocation, which we call a_R' . To summary we write:

Result 1: The first firms to delocation from the small region (south) to the large region (north) are the most efficient small-region firms.

2.4. The location condition: long run equilibrium

The key variable to determine in the long run equilibrium is the cut-off level of marginal cost, a_R' . To characterise this, we note that when the firm migration has stopped, marginal

adjustment costs are zero. Thus the long-run equilibrium satisfies the location condition $v[a_R]=0$. The long run a_R can be solved for analytically using (8). It is:

$$(11) \quad a_R^{1-\sigma+\rho} = \frac{2\phi(s-\frac{1}{2})}{(1-\phi)(1-s)}, \quad s_n = s + (1-s)a_R^\rho$$

where s_n is the share of all firms in the big region. Note that a_R rises with ϕ and this means that ever more inefficient firms find it profitable to delocate as trade gets freer. Note also that as in the traditional FC model, the share of firms in the big region rises as trade gets freer and reaches unity before trade is costless. Finally, note that full agglomeration occurs when ϕ equals or exceeds the sustain point, which is:

$$(12) \quad \phi^{CP} = \frac{1-s}{s}$$

Note that the ϕ^{CP} here is exactly the same ϕ^{CP} as in the standard FC model (see Baldwin et al 2003, Chapter 3).

Result 2: The maximum marginal cost of firms that find delocation profitable rises as trade gets freer. Full agglomeration occurs at the same level of trade openness as in the FC model without heterogeneity, namely at $\phi^{CP}=(1-s)/s$.

Interestingly, this means is that heterogeneity by itself does not affect the balance of agglomeration forces and dispersion forces in this model when trade is sufficiently free or restricted. Indeed, when measuring delocation in terms of the value of world production in the big market the heterogeneity does not the degree of agglomeration at any level of openness. It is easy to show that the share of production that has moved to the large market is¹³:

$$(13) \quad s_n = s + 2\phi \frac{s-1/2}{1-\phi}$$

This implication leads immediately to the next result concerning the degree of agglomeration measured in terms of the share of firms in the big market. When trade costs are at an intermediate level of free-ness fewer firms will have moved from the south to the north. The reason is simple and can be seen by considering a small increase in openness from an intermediate level. The extra openness makes the larger northern market more attractive, so some firms must move northwards to restore equality of profitability. Because the first firms to delocate in this are the most efficient, they have an above average impact on the degree of competition in the two regions. As a consequence, fewer firms need to move to restore the balance of profitability in the two regions. Using (11) and the equivalent expression for s_n in the FC model, which is $s+2\phi(s-1/2)/(1-\phi)$, we have:

$$(14) \quad s_n^{FC} - s_n = \frac{2\phi}{1-\phi} \left(s - \frac{1}{2} \right) \left(1 - \left(\frac{2\phi}{1-\phi} \frac{s-1/2}{1-s} \right)^{\frac{\rho}{1-\sigma+\rho}} \right)$$

¹³ Sales to the northern market of northern firms and southern firms located in the north are $\lambda s/\Delta$ and $\lambda(1-s)A/\Delta$ respectively, while their sales to the southern market are $\phi\lambda s/\Delta$ and $\phi\lambda(1-s)A/\Delta$, respectively. Weighting each of these four levels of sales by the relevant market sizes (by s or $1-s$) and adding the terms yields the expression in the text after some simplification.

Note that as long as ϕ is less than ϕ^{CP} , i.e. as long as some firms are in both regions, this quantity is positive. To summarise, we write:

Result 3: Heterogeneity of firm-level productivity does not change the share of world production that moves to the big region compared to the standard FC model with homogeneous firms. However, when measuring agglomeration in terms of the number of firms, heterogeneity can be thought of as a dispersion force in the sense that a smaller share of firms will have delocated from the small to big region for any intermediate level of trade free-ness.

and

Result 4: Heterogeneity dampens the HME (defined in terms of the share of firms) for intermediate levels of trade cost, but has no impact on the HME defined in terms of the share of production.

2.5. *Spatial selection*

The preceding analysis shows how the introduction of heterogeneous firms into an economic geography framework can be used to crystallise thinking about spatial selection effects.

What we showed can be thought of as a refinement on the usual HME. The standard HME states that when delocation is allowed, firms are attracted to the big market to such an extent that the big market ends up with a share of manufacturing firms that more than proportional to its size. Here we showed that the firms that move to the big market are systematically more efficient than the firms that stay behind. In other words, we have added spatial selection to the HME.

Result 5: Allowing for firm heterogeneity adds a spatial selection component to the HME, namely the big market attracts more than its ‘fair share’ of firms overall as usual, but it also attracts all the most efficient firms. Plainly, this suggests a testable hypothesis. Firms in the big market should be larger on average than firms in the small market. Importantly, this is not due to the competitive effects that appear in the OTT framework (markups are fixed with Dixit-Stiglitz competition). It is due entirely to spatial selection whereby the biggest, most efficient firms are the first to move to the bigger market.

The intuition for this spatial selection effect is uncomplicated. The most efficient firms have a stronger preference for location in the big market than do inefficient firms, because efficient firms have higher sales and thus enjoy greater savings on trade costs. Additionally, the extra local competition in the big market is less of a problem for efficient firms.

3. SELECTION BIAS AND THE MIS-MEASUREMENT OF AGGLOMERATION ECONOMIES

We turn now to two important implications of the spatial selection, namely its impact on regional productivity differences and how can this be measured.

3.1. *Testing for agglomeration economies*

The basic approach to testing for agglomeration economies is to see if the average measured productivity of a region is related to the amount of industry in the region. For example, see Ciccone (2002) and Midelfart-Knarvik and Steen (1999). To establish a baseline, suppose we

are in a world where labour is the only measured input and we test for agglomeration economies with the simplest regression:

$$(15) \quad \ln(lprod_r) = c + \alpha \ln(s_{nr}) + \varepsilon$$

where ‘ $lprod_r$ ’ is measured labour productivity in region ‘ r ’, and ‘ s_{nr} ’ is the share of industry in region r . The test for agglomeration economies would be based on α . If α exceeded zero in a statistically significant manner, we would conclude that agglomeration economies were present and would take α as a measure of their strength. A more detailed empirical specification would control for other region-specific factors using region fixed effects or actual data on productivity altering factors such as education and capital stocks; such considerations are tangential to our main point and so are ignored.

3.1.1 Test with the standard footloose capital model

To set the stage, consider how this test would perform if there were agglomeration economies and no heterogeneity. Thus for the moment we suppose that the true model of the world is the standard FC model. North’s manufacturing labour productivity is the region’s real value of manufacturing output divided by the region’s manufacturing labour input. In the FC model with homogenous firms, total northern manufacturing revenue – i.e. the value of output – is $np^{1-\sigma}(E/\Delta+\phi E^*/\Delta^*)$ where n is the mass of firms located in the north (see Baldwin et al 2003, Chapter 3). The total labour input is ‘ a ’ times the units produced $np^{-\sigma}(E/\Delta+\phi E^*/\Delta^*)$. Due to mill pricing, the ratio of the revenue to the labour input will be $1/(1-1/\sigma)$. To convert this to real terms we divide by the north’s manufacturing price index; either the consumer price index or the producer price index, which are, respectively, $(np^{1-\sigma}+\phi n^*p^{1-\sigma})^{1/(1-\sigma)}$ and $(np^{1-\sigma})^{1/(1-\sigma)}$. Thus, measured labour-productivity is (using the producer price index):

$$(16) \quad \ln(lprod) = \ln\left(\frac{s_n^{\frac{1}{\sigma-1}}}{a(1-1/\sigma)}\right)$$

What we can see from this is that agglomeration economies are indeed in operation in the sense that labour productivity increases with the share of firms in the north (recall that $n+n^*=K^w$ in the FC model). A properly specified cross-region regression equation would find that firm-level productivity is increasing in the mass of firms present in a region and this would be interpreted as evidence of agglomeration economies. Specifically, the estimated α would be $1/(\sigma-1)>0$.

In the FC model with heterogeneous firms, total northern manufacturing revenue – i.e. the value of output – is $(E/\Delta+\phi E^*/\Delta^*)\int p(i)^{1-\sigma} di$ where the limits of integration are from zero to n . The labour input would be $(E/\Delta+\phi E^*/\Delta^*)\int a(i)p(i)^{-\sigma} di$ with the same limits of integration. Again due to mill pricing, the ratio of these is $1/(1-1/\sigma)$. Converting to real terms using the producer price index, which in this case will be equal to $(\lambda K+\lambda K^* a_R^{1-\sigma+\rho})^{1/(1-\sigma)}$, measured labour productivity is:

$$(17) \quad lprod|_{het} = \ln\left(\frac{(\lambda K + \lambda K^* a_R^{1-\sigma+\rho})^{\frac{1}{\sigma-1}}}{a(1-1/\sigma)}\right)$$

Again a properly specified regression would detect a positive relationship between the mass of firms in a region and firm-level productivity, however, the resulting estimate of agglomeration economies would be upward biased since firms that had relocated to the big

region would have systematically higher than average productivity (recall that all our a 's are bound between 0 and 1, so $a_R^{1-\sigma+\rho} < 1$). To summarise:

Result 6: Selection Effect Bias – Standard econometric tests for agglomeration economies are likely to overestimate the impact of agglomeration on firm-level efficiency since delocation systematically involves the most efficient firms moving to the large region. In other words, because the most efficient firms are the first to move from the small region to the big, average firm productivity in big regions should be higher even if there are negligible agglomeration economies in operation.

4. SORTING AND SUBSIDIES

This section turns to the implications for regional policy. We continue with the basic model presented in Section 2, but we start from an initial situation of full agglomeration. As we saw above, all firms will be in the large region when trade is freer than ϕ^{CP} . Moreover, we consider a policy that pays firms a subsidy of S (a mnemonic for ‘subsidy’) to move from the large region to the small region. To focus on the impact of the subsidy, we assume the subsidy is financed by lump sum taxation (see Dupont and Martin 2003 for a consideration of tax issues).

4.1. *The new locational equilibrium*

We start the analysis by showing that this relocation subsidy will induce relocation to the small region, if the subsidy is sufficiently large.

Starting with all firms in the north, the change in operating profit (ignoring the subsidy) for an atomistic firm moving from the north to the smaller south would be:

$$(18) \quad a^{1-\sigma} \frac{(1-\phi)E^w}{\lambda\sigma} \left(\frac{1-s}{\phi} - s \right) < 0; \quad \phi > \phi^{CP}$$

for a firm with the marginal cost of ‘ a ’. Since we are considering $\phi > \phi^{CP}$, this difference would be negative as shown (this follows from the definition of ϕ^{CP}). Importantly, the loss from relocation, ignoring the subsidy, is decreasing in the firm’s marginal cost parameter ‘ a ’. The reason is a corollary of Result 1; the most efficient firms find location in the big region most profitable, so they are also the ones that would sacrifice the most by relocating to the small region. However, if we started with a very small subsidy and increased it, the first firms to relocate to the small region would be the most inefficient firms.

Result 7: The first firms to respond to subsidised relocation from the big region to the small one will be the least efficient firms.

To work out the precise relationship between the subsidy S and the cut-off marginal cost, we note that if all firms with marginal costs in excess of a_S move to the south, the Δ 's will be:¹⁴

$$(19) \quad \Delta = \lambda \left(a_S^{1-\sigma+\rho} + \phi(1 - a_S^{1-\sigma+\rho}) \right), \quad \Delta^* = \lambda \left(\phi a_S^{1-\sigma+\rho} + 1 - a_S^{1-\sigma+\rho} \right)$$

¹⁴ For example, Δ involves four integral; the prices of northern and southern firms in the north and northern and southern firms in the south. The first two integrals are $K \int a^{1-\sigma} f[a] da + K^* \int a^{1-\sigma} f[a] da$, where the limits of integration are from 0 to a_S . Solving these equal $(K+K^*)\lambda a_S^{1-\sigma+\rho}$, but $K+K^*=1$. Using similar manipulations for the third and fourth integrals yields to formula in the text.

where a_s is the cut-off level of efficiency above which firms do not move. The implied change in a north-based firm's operating profit when it moves south is (including the subsidy):

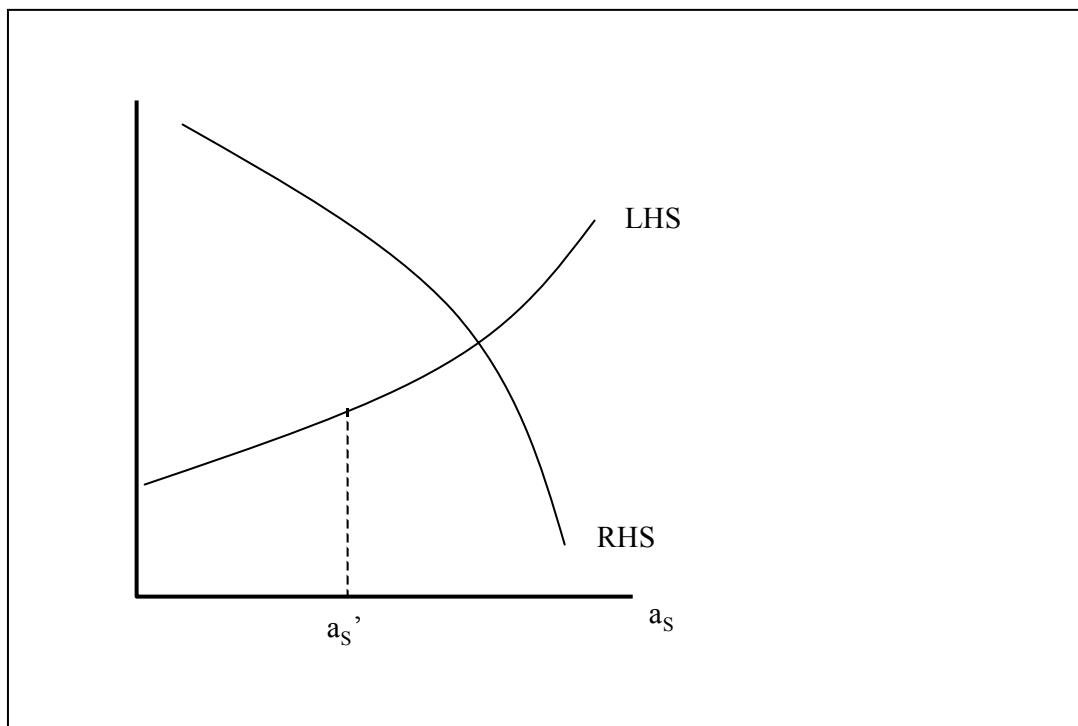
$$a^{1-\sigma} \frac{E^w(1-\phi)}{\sigma} \left(\frac{1-s}{\Delta^*} - \frac{s}{\Delta} \right) + S$$

Since the Δ 's involve a_s raised to the power of $1+\sigma+\rho$, we cannot explicitly solve for a_s , but the condition for it can be implicitly written as:

$$(20) \quad a_s^{\sigma-1} \frac{S\sigma}{E^w} = (1-\phi) \left(\frac{s}{\Delta} - \frac{1-s}{\Delta^*} \right)$$

where $E^w = \mu L^w$, since each individual spends μ on manufactures and there are L^w individuals in the world. Note that the left-hand and right-hand sides of this expression are always positive. The right-hand side is always decreasing in a_s since the degree of competition in the north falls and that in the south rises, as a_s increases. The left-hand side, by contrast, is always increasing in a_s . This tells us that there will be a unique solution for a_s in the economically relevant range (which is the unit interval in this case since we assumed $a_0=1$) as long as S is big enough. The solution is illustrated in a_s' .

Figure 3: Solution for cut-off level a_s



4.1.1 Delocation, subsidy size and openness

Two comparative static exercises are of interest. The first is to see how the cut-off varies with the level of the subsidy. Inspection of (20) reveals that an increase in S will raise the left-hand side without altering the right-hand side, so the result will be a decrease in the cut-off level of inefficiency. This means that a higher subsidy will attract more firms, as expected.

The second exercise is to see how deeper integration affects the effectiveness of a given subsidy. Since higher trade free-ness, i.e. $d\phi > 0$, lowers the right-hand side without altering the left-hand side, we see the subsidy becomes more effective as trade gets freer. This result is

quite intuitive since we know that both the agglomeration and dispersion forces weaken in the FC model as trade gets freer (Baldwin et al 2003, chapter 3), but the subsidy incentive is unaffected by changes in openness. Once trade is sufficiently free, all firms would move to get the subsidy.

To summarise, we write:

Result 8: Starting for a core-periphery situation, a per-firm subsidy aimed at encouraging production in the periphery tends to attract the least efficient firms. The reason is that the most inefficient firms are the ones that have the least to lose from leaving the big region. This may help explain why regional production subsidies are considered so ineffective in improving the competitiveness of remote regions.

Result 9: The subsidy is more effective in promoting relocation the larger the subsidy and the freer is trade.

For completeness, we note that it is possible to find analytic solutions for the minimum effective subsidy and the subsidy that induces all firms to move to the small region. To summarise:

Result 10: The minimum effective subsidy (i.e. the subsidy that just induces some firms to relocate to the periphery) is $2\mu(1-\phi)[(1+\phi)s-1]/\lambda\phi\sigma$. To induce all firms to move to the intrinsically small region, the subsidy would have to be infinite.

4.2. *Sorting equilibria*

Another way of expressing Result 8 is to say that production subsidies will result in what might be called a ‘sorting equilibrium’. Since the most efficient firms have the most to gain from being in the big market and the least efficient firms have the least to lose from leaving, a subsidy tends to sort firms according to their efficiency levels. All the most inefficient firms end up in the periphery and all the most efficient firms end up in the core.

The notion of sorting equilibria has two immediate implications. First, sorting magnifies the econometric difficulties pointed out in the previous section. Since real-world firms do have heterogeneous levels of inefficiency, sorting will lead to an outcome that mimics agglomeration economies. Second, judging the success of regional subsidies such as the EU’s Structural Funds will be tricky. Since such funds will systematically attract the least efficient firms to periphery regions, there will be an important difference between the share of firms in the periphery and their efficiency. Although we do not consider it explicitly, this later suggestion may have growth implications if there is a correlation between a firm’s level of efficiency and its ability to innovate. We leave this for future work.

5. CONCLUSION

Hereto, the new economic geography literature has relied on the assumption of identical industrial firms. While this was viewed as an assumption of convenience, this paper shows that allowing for firm-level heterogeneity has important implications for empirical work and for policy predictions. In particular, we showed that the most efficient firms are the ones that move first to the big region. This non-random ‘selection’ implies that standard empirical methodologies will tend to overestimate agglomeration economies. Moreover, the same selection logic implies that production subsidies aimed at promoting industry in disadvantaged regions can have a ‘sorting effect’. That is, the subsidies will result in a

situation where all the most productive firms, regardless of their region of origin, will choose to locate in the core while all the least productive firms will locate in the periphery.

We believe that the inclusion of firm-level heterogeneity raises many interesting issues in economic geography that should be explored in future work. For example in search for the most appropriate model, we considered adding Melitz-heterogeneity to other NEG models such as the footloose entrepreneur model. Here we found that heterogeneity had complex effects on both demand and cost linkages.

REFERENCES

- Alsleben, C. (2005). "Spatial agglomeration, competition and firm size", mimeo.
- Amiti, M and C. Pissarides (2002), "Trade and Industrial Location with Heterogeneous Labour," CEPR DP 3366, London.
- Baldwin, R. and R. Forslid, P. Martin, G. Ottaviano and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton University Press, Princeton.
- Bernard, A, B. Jensen, and P. Schott (2003). "Falling Trade Costs, Heterogeneous Firms and Industry Dynamics", *CEPR Discussion Papers*, Centre for Economic Performance, LSE.
- Borjas, G.J., S.G. Bronars, and S.J. Trejo. (1992), Self-Selection and Internal Migration in the United States, *Journal of Urban Economics* 32: 159-185.
- Cabral, L. M. B. and J. Mata (2003) "On the Evolution of the Firm Size Distribution: Facts and Theory", *American Economic Review*, pp 1075-1090.
- Chiswick, B. (1999), "Are immigrants favourably self-selected?" *American Economic Review*, Papers and proceedings, vol. 89, pp. 181-185.
- Ciccone, A. (2002) "Agglomeration effects in Europe" *European Economic Review*, pp 213-227.
- Combes, P., G. Duraton and L. Gobillon (2004) "Spatial wage disparities: Sorting matters!" CEPR discussion paper, 4240.
- Coniglio, N. (2001), "Regional Integration and Migration: an Economic Geography Model with Heterogeneous Labour Force," December, University of Glasgow manuscript.
- Dixit, A.K. and J.E. Stiglitz (1977) Monopolistic competition and optimum product diversity, *American Economic Review* 67, 297-308.
- Dupont, V. and Martin, P. 2003. 'Subsidies to Poor Regions and Inequalities: Some Unpleasant Arithmetic'. CEPR Discussion Paper no. 4107. London, Centre for Economic Policy Research.
- Fujita M., Krugman P. and A. Venables (1999) *The Spatial Economy: Cities, Regions and International Trade* (Cambridge (Mass.): MIT Press).
- Fujita, M. and J.-F. Thisse (2002) *Economics of Agglomeration*, (Cambridge: Cambridge University Press).
- Helpman, E., M. Melitz, and S. Yeaple (2004). "Export Versus FDI With Heterogeneous Firms," forthcoming in *American Economic Review*.
- Hopenhayn, Hugo (1992a). "Entry, Exit, and Firm Dynamics in Long Run Equilibrium." *Econometrica* 60:1127-1150.

- Hopenhayn, Hugo (1992b). "Exit, Selection, and the Value of Firms." *Journal of Economic Dynamics and Control* 16:621-653.
- Krugman, Paul (1991) Increasing Returns and Economic Geography, *Journal of Political Economy* 99, 483-99.
- Martin, P and C.A Rogers (1995) 'Industrial location and public infrastructure' *Journal of International Economics* 39: pp335-351.
- Melitz, M and J. Ottaviano (2003), Market size, trade, and productivity, mimeo, Harvard University.
- Melitz, M. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity, *Econometrica*, Vol. 71, November 2003, pp. 1695-1725.
- Midelfart-Knarvik, K.H. and F. Steen (1999), Self-Reinforcing Agglomerations? An Empirical Industry Study, *Scandinavian-Journal-of-Economics*. December 1999; 101(4): 515-32
- Miren Lafourcadey Giordano Mionz (2003), "Concentration, Spatial Clustering and the Size of Plants: Disentangling the sources of co-location externalities," mimeo, December 8, 2003
- Nocke, V (2003) "A Gap for Me: Entrepreneur and Entry", University of Pennsylvania.
- Ottaviano G. , J. Thisse and T. Tabuchi, (2002), "Agglomeration and Trade Revisited", *International Economic Review*, 43, pp.409-436.
- Tabuchi T. and J.-F. Thisse (2002) Taste heterogeneity, labor mobility and economic geography, *Journal of Development Economics* 69, 155-177.
- Venables Anthony (1996). Equilibrium locations of vertically linked industries, *International Economic Review* 37, 341-359.

APPENDIX: SPATIAL SELECTION IN THE MELITZ OTTAVIANO MODEL

This appendix introduces the Melitz and Ottaviano (2003) framework discussed in the introduction and shows that the basic features of spatial selection that arise in our model also arise in the MO framework.

5.1.1 Assumptions and the model's core logic

Melitz and Ottaviano (2003) model can be thought of as the marriage of the Ottaviano, Tabuchi and Thisse (OTT) monopolistic competition framework and the Hopenhayn-Melitz mechanism for the development of firms with heterogeneous marginal costs.

5.1.2 The Ottaviano Tabuchi Thisse monopolistic competition framework

The OTT monopolistic competition set-up works with a linear demand system where income effects have been eliminated via quasi-linear preferences. As usual in the monopolistic competition tradition, there are many, many firms producing differentiated varieties. Since the firms are small, they ignore the impact of their sales on industry-wide variable. Practically, this means that the producer of each differentiated variety acts as a monopolist on a linear demand curve. Indirectly, however, firms face competition in the OTT framework since the demand curve's intercept declines as the number of competing varieties rises.

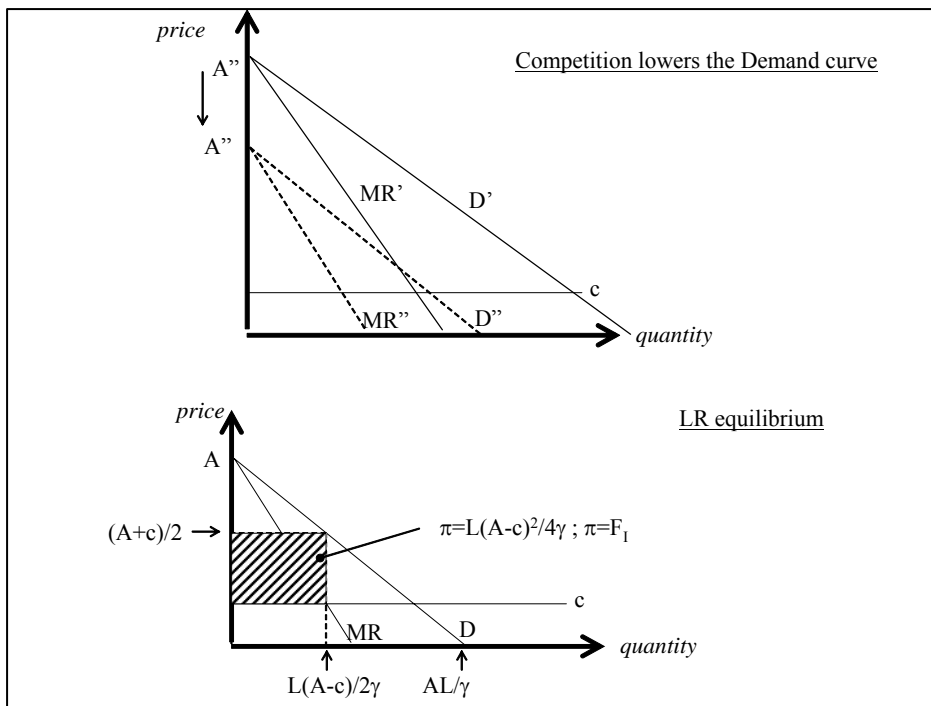
Open-economy general-equilibrium considerations are minimised by assuming that labour is the only primary factor and that the sector producing the 'linear good' in the quasi-linear preferences is marked by perfect competition, constant returns to scale and costless trade. This equalised wages internationally and ensuring trade balances at all times.

Increasing returns is introduced as usual by assuming potential producers must pay a fixed cost to develop a new variety. This cost – which comprises F_1 units of labour – is paid just prior to production.

Free entry and the elimination of pure profits. As in other monopolistic competition models, the OTT framework assumes that the degree of competition rises up to the point where pure profit is eliminated. Given the very convenient preferences assumed, more competition – either in the form of more firms, or in the form of lower trade barriers – lowers the y-axis intercept of the linear demand curve facing each and every firm (see top panel of Figure 4); as usual the marginal revenue curve meets the demand curve at the intercept and has half its slope, so it too falls with extra competition. Although it is not shown in the figure, it is easy to understand that the price charged by a typical firm with marginal cost 'c' is $(A+c)/2$ and this falls as competition rises. Moreover, sales of a typical firm (which is always $b(A-c)/2$ for a linear demand curve, where 'b' is the demand curve's slope), so operating profit (which is always $b(A-c)^2/4$ for a linear demand curve) falls monotonically with the degree of competition.

When the long-run equilibrium is reached, the degree of competition must be such that the operating profit earned by a typical firm, π in the figure, is just sufficient to cover the fixed entry cost F_1 (see bottom panel of Figure 4).

Figure 4: OTT monopolistic competition.



5.1.3 The Hopenhayn-Melitz framework

The Hopenhayn-Melitz approach to heterogeneous firms assigns all heterogeneity to firms’ marginal costs. That is to say, the differentiated varieties are made from labour by firms facing constant marginal production costs; namely, ‘ c_i ’ units of labour per unit of output of variety i . With wages equal to unity, we can refer without ambiguity to ‘ c_i ’ as firm i ’s marginal costs.

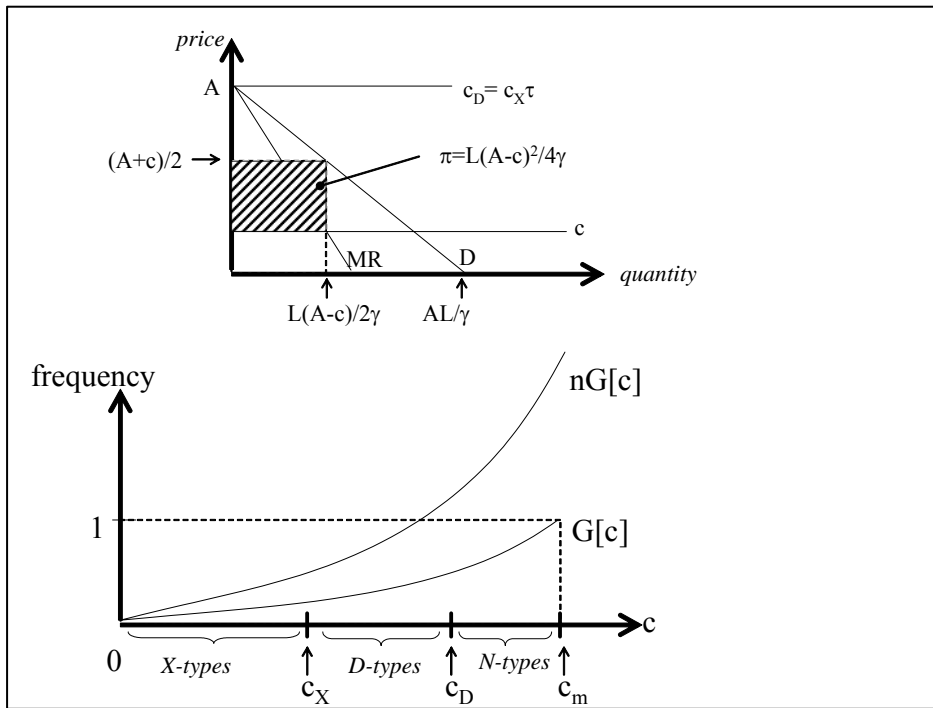
Where does the marginal-cost heterogeneity come from? Following Hopenhayn (1992a, 1992b), it is assumed that once a new variety is developed (by paying F_1), the new variety is randomly assigned a marginal costs, i.e. a ‘ c ’, from a known distribution function, $G[c]$. $G[c]$ has positive probability over the range $0 \leq c \leq c_m$, so maximum conceived marginal cost is c_m and the minimum is zero. Intuition is served by thinking of this as a stochastic variety-innovation process; F_1 buys a ‘blueprint’ for a new variety, but the new variety’s marginal cost is random. The Hopenhayn-Melitz approach focuses exclusively on steady state equilibriums and ignores discounting but keeps present values finite by assuming firms face a constant probability of ‘death’ according to a Poisson process with a hazard rate of δ .

5.1.4 The equilibrium

Entry is continuous since varieties are continually dying and replacements are needed to keep the system at its long-run equilibrium (δn firms die every instant and the same number is needed to maintain equilibrium). The continuous drawing from the underlying distribution $G[c]$ yields a mass of potential firms $nG[c]$, but only the most efficient find it worthwhile to actually produce, as the top panel of Figure 5 illustrates. Specifically, some firms will have marginal costs above the demand curve intersection, A , and so will not sell/produce anything; we call this threshold marginal cost c_D (D is a mnemonic for domestic market). In this way,

competition truncates the distribution at of marginal costs at c_D (recall that A falls as competition rises).

Figure 5: N, D and X types and cut-offs.



In the simplest heterogeneous-firms trade model, there are two nations assumed, for simplicity's sake, to be symmetric in terms of tastes, technology, size and trade costs. The cost of transport goods between nations is $\tau > 1$ (initially assumed identical for trade in both directions). Note that one of the merits of the MO model is that it can handle asymmetries and many nations, but the model's core logic is best presented in a setting uncluttered by exogenous asymmetries.

Given the trade cost τ , only firms with sufficiently low marginal costs can profitably sell in their export market. This second cut off is obviously related to the first since the intercept in the two nations is the same by symmetry; specifically, $c_X \tau = c_D$, where c_X is the maximum marginal cost of an exporting firm (X is a mnemonic for export).

5.1.5 Allowing delocation

In this appendix, we extend the MO model to allow firms to delocate between two regions in search of the highest profits. As in our model, we rule out both firm death and firm entry in order to spotlight the spatial aspects of the problem.

Following to MO, in demand side, the inverse demand for each variety can be written as:

$$(21) \quad q_i = \frac{L}{\gamma}(A - p_i); \quad A \equiv \frac{\alpha\gamma + \eta P}{\eta m_c + \gamma}; \quad P \equiv \int_{\Omega} p_j dj; \quad \eta \equiv 1$$

where L is the number of consumers (and the labour supply), 'A' is the endogenous y-axis intercept, n_c is the mass of varieties consumed (since not all varieties are traded, we need a separate notation for the number of varieties produced in a typical country and the number consumed by a typical consumer), and P is the sum of all prices in the market; Ω is the set of all varieties consumed in the market. By choice of units, we normalise $\eta = 1$.

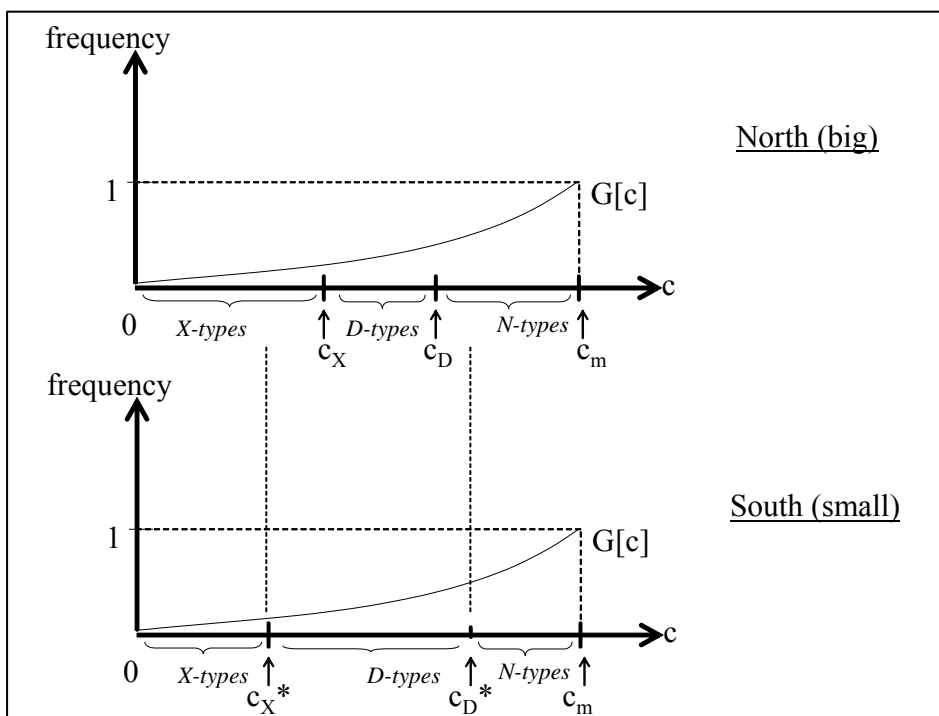
Inspection of (21) reveals two convenient features. First, a ceteris paribus increase in the number of varieties consumed, n_c , lowers the intercept. Second a lowering of the price of competing varieties will lower P and thus lower the intercept.

The linear demand system makes this model extremely simple to work with. Atomistic firms take the mass and distribution of c 's as given, so the A in (21) is viewed as a parameter and they act as monopolists on their linear demand curve. Standard results imply that (see top panel of Figure 5):

$$(22) \quad \begin{aligned} p_i &= \frac{A + c_i}{2}; & p_i^* &= \frac{A + c_i \tau}{2}; & \pi^D(c_i) &\equiv L \frac{(A - c_i)^2}{4\gamma}; & c_X \leq c_i \leq c_D \\ \pi^X(c_i) &\equiv L \frac{(A - c_i)^2}{4\gamma} + L \frac{\tau^2 (A - c_i)^2}{4\gamma}; & 0 \leq c_i \leq c_X; & \pi^N(c_i) &\equiv 0; & c_D \leq c_i \end{aligned}$$

where p is the profit maximising price charged by D- and X-type firm in their local markets, p^* is the price charged by X-types in their export markets, and the π 's are the Ricardian surpluses for D-types, X-types and N-types, respectively.

Figure 6: Cut-offs in the big and small markets.



5.1.6 The cut-off conditions

Firms only sell locally if they can make positive operating profits. Given (22), the zero operating profit condition is where sales are zero and this implies:

$$(23) \quad a_D = A, \quad a_X = A^* / \tau$$

where a_D and a_X are the cut-off for D-types and X-types, respectively. To find the closed form solution for A , we must solve for the price index P and this requires a specific functional form for $G(c)$. However, for the purposes of this appendix, qualitative reasoning is sufficient.

Notice that A is increasing in P and decreasing in n_c . Since competition is tougher in the big market, and more varieties are consumed in the big market, A is lower in the big market, i.e. $A < A^*$. Consequently, we know that $c_D < c_D^*$, where c_D and c_D^* are the cut-offs in the big and small market respectively. Likewise, $c_X^* < c_X$ since firms based in the small nation have to be more competitive to sell into the big competitive market. To summarise:

Result 1: Big country has smaller export-entry cut-off marginal costs, c_X . This implies that the big country has higher share of exporting firms. On the other hand, small country has larger market-entry cut-off marginal costs, c_D .

This results is shown in Figure 6.

5.1.7 Delocation Tendencies with Quadratic Adjustment costs

Now we add geographical delocation to the MO model. Firm can delocate between the countries in search of the highest profit. As in our model, we assume quadratic adjustment costs, as in our model this means that firms move in the order of how much they have to gain from moving and in the long run, the moving costs are zero and profits are equalised between the two regions for any given level of 'c'.

We begin by considering the incentives of various firms to delocate, assuming that no delocation has occurred.

Profit gap function

Inspection of Figure 6 shows that in the initial no-delocation, firms that move from the small south to the big north may change types. For instance, southern firms with c 's less than c_X but more than c_X^* would switch from being D-types in the south to being X-types in the north. We call this DX migration. Southern firms with c 's below c_X^* will remain X types, so we call this XX migration. Southern firms with c 's above c_X and below c_D will remain D types, so we call this DD migration. Some southern firms would never complement moving from the initial situation since they can make positive profits in the small market but not in the big market.

The incentives for delocation are different across migration types. Figure 7 plots the profit gap function for each type in terms of marginal costs. Interestingly, the gap function for XX migration is a hump-shaped curve. This concavity implies that the intermediate productivity of XX migrants have the highest incentive to delocate to the bigger market while the highest productivity firm is indifferent in its location. The gap function for XX migration is tangent to the gap function DX migration at point c_X^* (the cut-off between D and X types in the south). From c_X^* to c_X , migration would be DX. The DX profit gap is tangent to the DD curve at c_X . From c_X to c_D , migration of the DD sort.

It is easy to show and intuitively obvious that as south to north migration occurs, the cut-off points converge since the firm migration equalises the degree of competition in each market. Thus in this model, the first firms to delocate are southern firms who are more efficient than average (they are X-types before and after delocation), but they are not the most productive firms. This confirms that qualitative message in our model that delocation leads to spatial selection such that delocating firms have above average productivity.

The long run equilibrium is complex to illustrate analytically, but we have simulated the model and find that as in our model, the firms that move in equilibrium have above average productivity compared to the firms that remain in the small region.

Figure 7: Migration types and initial incentives to delocate.

