Heterogeneous firms, agglomeration and economic geography: Selection and sorting

Richard E. Baldwin and Toshihiro Okubo
Graduate Institute of International Studies, Geneva

June 2004

ABSTRACT

A Melitz-style model of monopolistic competition with heterogeneous firms is integrated into a simple NEG 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 productivity 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

The great contribution of the new economic geography 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 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; see Fujita and Thisse (2002) for a succinct synthesis of this work, much of it undertaken by the authors themselves.

One of most convenient, but least realistic, assumption 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, Jenson and Schott 2003, Helpman, Melitz and Yeaple 2004).

Our paper argues that this 'assumption of convenience' is neither necessary nor innocuous. 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 productivity firms.

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

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, Duraton and Gobillion (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 low 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 simplified Melitz (2003) model that is adapted to locational analysis. It is easier to manipulate analytically than Melitz (2003), but displays only a restricted range of Melitz model features.

2.1. General set up: FC model

We start from the familiar 'new economic geography' model of Martin and Rogers (1995), also known as the footloose capital model, or FC model for short (see Baldwin, Forslid, Martin, Ottaviano and Robert-Nicoud, 2003, Chapter 3 for a thorough analysis of this model). The model has two regions, two sectors and two factors. The regions are referred to as the north and the south; they are symmetric in terms of tastes, technology, openness to trade, and relative factor endowments. The north, however, is larger in the sense that it is endowed with more of both factors in equal proportions. The factors, physical capital K and labour L are inelastically supplied; only K is inter-regionally mobile.

The two sectors are manufacturing and agriculture. 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, 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. Thus, 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. Worldwide supplies of capital and labour are fixed, with the world's endowment denoted as L^w and K^w.

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 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 and the share of all varieties made in the north.

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

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

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; Ω is the set of all varieties produced. The number of varieties produced is pre-determined by endowments since each variety requires a unit of capital and the capital stock is fixed.

The final assumption concerns factor migration. Physical 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.

2.2. Additional assumption: delocation costs and firm heterogeneity

First, as in 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 firm-level productivity; we are interested in how agglomeration and policy affects its geographic distribution.[1]

[1] 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.

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 a Pareto probability distribution:

$$\textbf{(1)} \qquad f[a] = \rho(\frac{a^{\rho-1}}{a_0{}^{\rho}}), \qquad 1 \equiv a_0 \geq a \geq 0, \quad \rho \geq 1$$

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).[2]

Second, we deviate from the FC model in that we assume that relocation is costly. In particular, a manufacturing firm must pay a fixed cost of χ units of labour to change regions.

2.3. Intermediate results

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.[3] 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/σ times firm-level revenue.[4] Accordingly, operating profit realised by a south-based firm is:

[2] 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.

[3] 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.

[4] A typical first order condition is p(1-1/σ)=wa; rearranging, the operating profit, (p-wa)c, equals pc/σ.

(2)
$$\pi^*[a] = (\frac{a}{1-1/\sigma})^{1-\sigma}(\frac{\phi s_E}{\int_{\in\Omega} p_i^{1-\sigma} di} + \frac{s_E^*}{\int_{\in\Omega} p_i^{*1-\sigma} di})\frac{E^w}{\sigma};$$
$$s_E \equiv \frac{E}{E^w}, \quad s_E^* \equiv 1 - s_E, \quad \phi \equiv \tau^{1-\sigma}$$

where E^w is world expenditure on M-goods, s_E and s_E* are the northern and southern shares 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), p and p* are the consumer prices of varieties in the northern and southern markets with Ω representing the set of varieties produced (all varieties are sold in both regions).

2.4. Short run equilibrium

We start by taking as given the spatial distribution of industry. 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. How does firm heterogeneity change this? In particular, which types of firms delocate?

To work this out, we suppose that the relocation costs, χ, is initially prohibitive so there is no delocation and $s_n=s_K$ (i.e. the north's share of industry exactly matches its share of capital since no delocation has occurred). We then lower χ progressively until the first firm moves from south to north. Since firms are atomistic, the change in operating profit from a single firm moving from south to north (small region to big) is:

(3)
$$\frac{a^{1-\sigma}E^w}{\sigma}\left((\frac{s_E}{\Delta} + \frac{\phi s_E^*}{\Delta^*}) - (\frac{\phi s_E}{\Delta} + \frac{s_E^*}{\Delta^*}) \right)$$

where using mill pricing and cancelling the (1-1/σ) terms, the Δ's can be written as:

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

when no firms have relocated to the north. Here where K^w is world endowment of K with s_K and s_K* being the northern and southern shares of this. Solving the integrals, using (1) and assuming 1-σ+ρ>0 (so the integrals converge) we have:[5]

$$\Delta = K^w \lambda (s_K + \phi(1 - s_K)); \qquad \Delta^* = K^w \lambda (\phi s_K + 1 - s_K); \quad \lambda \equiv \frac{\rho}{1 - \sigma + \rho} > 0$$

Using these Δ's and (3), the net benefit to a southern firm with marginal cost 'a' of delocating to the north is:

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

Since we have assumed regions have symmetric factor endowments, $s_E = s_K \equiv s > ½$ and this reduces to:

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

Noting that the term in large brackets is positive since north is larger, we see that there would be a gain in operating profit for any individual southern firm moving north. Importantly, the gain would be largest for the most efficient southern firm, that is to say, the southern firm with the lowest 'a'. More precisely, if we progressively lowered χ from a very high level, the first firm to relocate would be the most efficient one. Indeed, any firm with a level of inefficiency that was below some cut-off level efficiency would migrate to the north. We summarise this as:

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

The delocation cut-off inefficiency level

We define the cut-off level of inefficiency 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:

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

$$\Delta = s\int_0^1 a^{1-\sigma} f[a]da + (1-s)\{\int_0^{a_R} a^{1-\sigma} f[a]da + \phi\int_{a_R}^1 a^{1-\sigma} f[a]da\},$$
$$\Delta^* = \phi s\int_0^1 a^{1-\sigma} f[a]da + (1-s)\{\phi\int_0^{a_R} a^{1-\sigma} f[a]da + \int_{a_R}^1 a^{1-\sigma} f[a]da\}; \qquad K^w \equiv 1$$

since southern firms with a's in the range [0,a$_R$] become north-based firms (recall that we use the symmetry of region's endowments to set s$_E$=s$_K$≡s). We have normalised the world capital stock to unity for notational convenience. Solving the integrals:

$$(4) \qquad \begin{aligned}\Delta &= \lambda\left(s + (1-s){a_R}^{1-\sigma+\rho} + \phi(1-s)(1-{a_R}^{1-\sigma+\rho})\right),\\ \Delta^* &= \lambda\left(\phi s + \phi(1-s){a_R}^{1-\sigma+\rho} + (1-s)(1-{a_R}^{1-\sigma+\rho})\right)\end{aligned}$$

Given the expressions for the Δ's in (4) we can write the gain to a southern firm with marginal cost 'a' from delocating to the north as:

$$(5) \qquad a^{1-\sigma}(1-\phi)(\frac{E^w}{n^w\sigma})\left(\frac{s}{\Delta} - \frac{1-s}{\Delta^*}\right) - \chi$$

where E^w=2μ given the quasi-linear preferences, and n^w≡K^w=1 by normalisation.

2.5. Long run equilibrium

The key variable to determine in the long run equilibrium is the cut-off level of marginal cost, a$_R$. To improve comparability with the FC model, we first solve for a$_R$ assuming – as in the traditional FC model – that there are no fixed delocation costs, i.e. χ=0.

No delocation costs case

When there are no delocation costs, a$_R$, and the implied share of firms in the big region (north) can be solved for analytically. They are:

$$(6) \qquad \left.{a_R}^{1-\sigma+\rho}\right|_{\chi=0} = \frac{2\phi(s-\frac{1}{2})}{(1-\phi)(1-s)}, \qquad\qquad s_n = s + (1-s){a_R}^{\rho}$$

where s$_n$ is the share of all firms in the big region. Note that a$_R$ rises with ϕ which means ever more inefficiency 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 (1-s)/s; we call this level of openness the core-periphery ϕ, or ϕ^{CP} for short. 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 minimum level of inefficiency at which firms 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 ϕ=(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. However, when trade costs are at an intermediate level of free-ness the heterogeneity acts as a dispersion force. 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 (6) and the equivalent expression for s_n in the FC model, which is s+2ϕ(s-½)/(1-ϕ), we have:

$$(7)\qquad {s_n}^{FC} - s_n = \frac{2\phi}{1-\phi}(s-\frac{1}{2})(1-\left((\frac{2\phi}{1-\phi})\frac{s-1/2}{1-s} \right)^{\frac{\rho}{1-\sigma+\rho}})$$

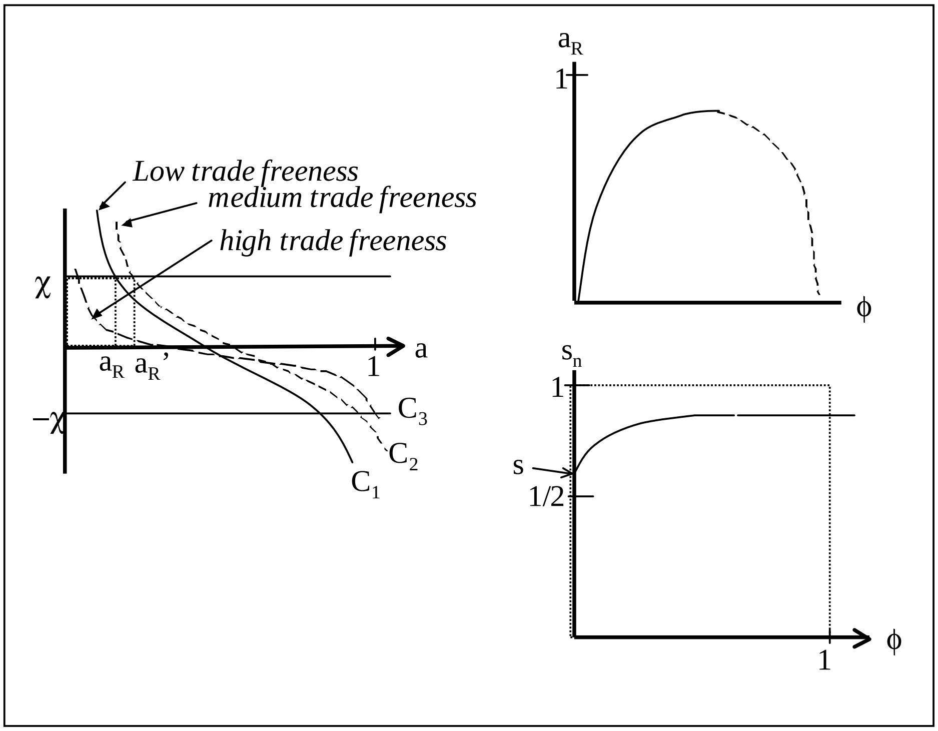
Note that as long as ϕ is less that ϕ^{CP}, i.e. as long as some firms are in both regions, this quantity is positive. In other words, agglomeration is always greater for the homogeneous firm case than it is for the heterogeneous firm case. To summarise:

> **Result 3: Heterogeneity of firm-level productivity is 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.**
>
> **Result 4: Heterogeneity dampens the HME for intermediate levels of trade cost.**

Positive delocation costs

It seems quite plausible that delocation would involve some fixed costs, so we solve the for the cut-off a_R when χ>0. Unfortunately, we cannot analytically solve for the cut-off level of inefficiency since 'a_R' is raised to different, non-integer powers in the denominator and the numerator of (5). It is nonetheless a simple matter to solve for this numerically for specific values of the parameters. The key difference stemming from positive relocation costs concerns the degree of spatial concentration when trade is very free.

Figure 1: Cut-off with relocation costs



The left-panel of Figure 1 plots the change in operating profit from a move from the south to the north as the curve marked C_1 – this corresponds to the first term in (5). The relocation cost χ is shown as a horizontal line. The cut-off marginal cost (i.e. the highest marginal cost for which relocation is just worthwhile) is given by the intersection of the C_1 curve and χ (shown as a_R in the diagram). When we increase the freeness of trade from a very low level, the cut-off marginal cost rises as shown in the top right-panel. However, when trade gets sufficiently free, the change in operating profit from a northward move becomes small, even for firms with very low marginal costs. In the extreme, costless trade sets the change to zero. What this means is that the intersection between the C-curve and χ begins to fall back toward a=0. The C_3 curve shown in the left panel illustrates a case where the intersection is almost at zero. The values of a_R where a is falling as ϕ is rising are shown with a dashed curve in the top right-panel. It is important to note, however, that the solutions of (5) that correspond to the dashed curve in the right panel are not relevant. The point is that although firms that have moved would regret their choice, the operating profit difference is not sufficient to make re-migration worthwhile. In short, some kind of ‘lock in’, or hysteresis effect is in operation.

The evolution of the change in the north’s share of industry as trade gets freer is shown in the bottom right panel. When ϕ=0, there is no relocation for s_n=s, but as trade gets freer, s_n rises up to a point and then remains at that level. We note that this feature stems primarily from the fixed relocation cost rather than the heterogeneity in marginal costs. Indeed, this sort of lock in would occur even in the standard FC model.

3. Selection bias and the mis-measurement of agglomeration economies

We turn now to two of the key questions – how does this affect 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, K.H. and F. Steen (1999). To establish a baseline, suppose we are in a world where labour is the only measured input (as in our model and the FC model) and we test for agglomeration economies with the simplest regression:

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

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 concluded 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.

Test with the standard FC 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):

(9) $$lprod\big|_{\text{hom}} = \ln\left(\frac{s_n^{\frac{1}{\sigma-1}}}{a(1-1/\sigma)}\right)$$

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 mass of 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 producer price index, which in this case will be equal to $(\lambda K+\lambda K^* a_R^{1-\sigma+\rho})^{1/(1-\sigma)}$ given (4), measured labour productivity is:

(10)
$$lprod\big|_{het} = \ln\left(\frac{(\lambda K + \lambda K^{*} a_R{}^{1-\sigma+\rho})^{\frac{1}{\sigma-1}}}{a(1-1/\sigma)} \right)$$

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 5: <u>Selection Effect Bias</u> – 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 in Result 2, all firms will be in the large region when trade is freer than ϕ^{CP}, where this equals (1-s)/s. 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

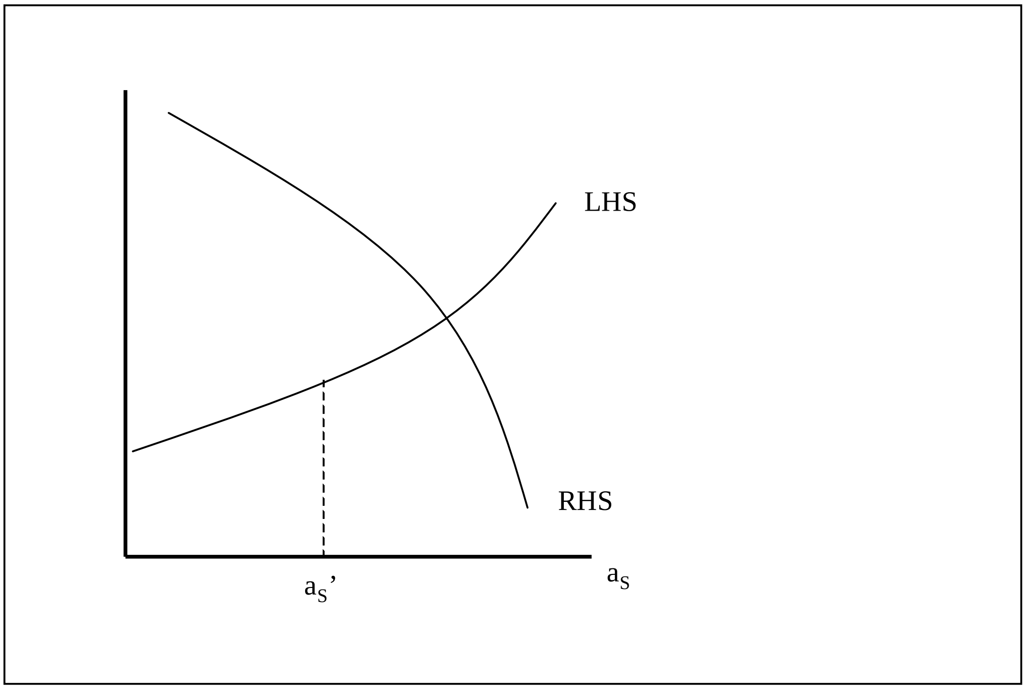
We start the analysis by showing that this relocation subsidy will induce relocation to the small region, if the subsidy is sufficiently large.

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:

(11)
$$a^{1-\sigma}\frac{(1-\phi)E^{w}}{\lambda\sigma}\left(\frac{1-s}{\phi}-s\right)<0; \qquad \phi>\phi^{CP}$$

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 re-locate to the small region would be the most inefficient firms.

Figure 1: Solution for cut-off level a_S



Result 6: 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:[6]

$$(12) \qquad \Delta = \lambda\left(a_S{}^{1-\sigma+\rho} + \phi(1-a_S{}^{1-\sigma+\rho})\right), \qquad \Delta^* = \lambda\left(\phi a_S{}^{1-\sigma+\rho} + 1 - a_S{}^{1-\sigma+\rho}\right)$$

[6] 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:

$$a_S{}^{\sigma-1}\frac{S\sigma}{E^w} \quad = \quad (1-\phi)(\frac{s}{\Delta}-\frac{1-s}{\Delta *}) \qquad (13)$$

where $E^w=2\mu$. 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 Figure 1 as a_S'.

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 (13) 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 7: 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 leave the big region. This may help explain why regional production subsidies are considered so ineffective in improving the competitiveness of remote regions.**
>
> **Result 8: The subsidy becomes more effective in promoting relocation as is increases in size and as trade gets freer.**

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 9: 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 7 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

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.

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. Gobillion (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).

Krugman, Paul (1991) Increasing Returns and Economic Geography, Journal of Political Economy 99, 483-99.

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

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.