The Empirics of Agglomeration and Trade *

Keith Head† Thierry Mayer‡

June 6, 2003

Abstract

This chapter examines empirical strategies that have been or could be used to evaluate the importance of agglomeration and trade models. This theoretical approach, widely known as "New Economic Geography" (NEG), emphasizes the interaction between transport costs and firm-level scale economies as a source of agglomeration. NEG focuses on forward and backward trade linkages as causes of observed spatial concentration of economic activity. We survey the existing literature, organizing the papers we discuss under the rubric of five interesting and testable hypotheses that emerge from NEG theory. We conclude the chapter with an overall assessment of the empirical support for NEG and suggest some directions for future research.

*This paper was prepared as a chapter for the forthcoming *Handbook of Regional and Urban Economics*, vol. 4. A first version was presented at the CEPR "Cities and Geography" conference on December 12, 2002 in Paris. We appreciate the helpful suggestions we received during this event and later from the editors, Jacques Thisse and Vernon Henderson. Pierre-Philippe Combes, Gordon Hanson, Pamina Koenig-Soubeyran, Françoise Maurel, Henry Overman, Stephen Redding, and Federico Trionfetti also provided detailed and useful comments.

†Corresponding author: Faculty of Commerce, University of British Columbia, 2053 Main Mall, Vancouver, BC, V6T1Z2, Canada. Tel: (604)822-8492, Fax: (604)822-8477, Email:keith.head@ubc.ca

‡Université de Paris I Panthéon Sorbonne (TEAM), CEPII, CERAS and CEPR. Email:tmayer@univ-paris1.fr

Contents

List of Tables

List of Figures

1 Introduction

In the 1990s, theorists developed a new approach to understanding why some regions seem to attract a disproportionate share of economic activity. Widely known as "New Economic Geography" (NEG), this approach emphasizes the interaction between trade costs and firm-level scale economies as a source of agglomeration. The dictionary provides two senses for the word agglomeration. The first is that of a process by which things come together. The second is the description of a pattern, namely one in which economic activity is spatially concentrated. NEG starts with the observed pattern of agglomeration and postulates a process through which it might have emerged: Producers and consumers co-locating to exploit plant-level scale economies while minimizing trade costs. NEG therefore specifically focuses on forward and backward trade linkages as causes of observed spatial concentration of economic activity.

"New economic geography has come of age" as Peter Neary (2001) recently wrote in a mildly skeptical review for the *Journal of Economic Literature*.[1] While this statement seems deserved for theory, the empirical literature treating the same questions remains unsettled in both methodology and results. There is no agreed upon regression to estimate, nor even a consensus dependent variable to explain. As a result, empirical papers addressing various aspects of agglomeration and trade are difficult to compare. The Fujita et al. (1999) and Baldwin et al. (2003) books devote a few paragraphs each to empirical work and emphasize that the time has now come to devote greater research efforts to the empirical validation or falsification of the framework. Overman et al. (2001) and Hanson (2001) are early surveys of empirical work on NEG. Brakman et al. (2001) provide the first textbook where many empirical aspects of NEG are covered in detail. Since those surveys were written, the literature has continued to grow in many directions. Here we attempt to weave together the disparate strands of the empirics of agglomeration and trade and outline the important and challenging questions for future research.

The chapter starts with the definition and delimitation of the field in section 2, where we organize the paper around five empirical propositions that we believe capture the essential insights offered by the theory. Section 3 emphasizes the central role of market potential in determining location patterns in those models and provides a method of measurement directly derived from theory. Then, each of the remaining sections covers one of the five empirical propositions identified in section 2. The empirical work on the impact of market potential on factor prices and factor movements is covered in sections 4 and 5 respectively. The benefits that regions can enjoy from a large domestic demand ("home market effects") have been subjected to important empirical tests that we survey in section 6. The impact of trade integration on the level of agglomeration is one of the most sensitive questions of this field. We describe existing results and consider new ways to test this proposition in section 7. The last proposition that has perhaps been most emblematic of NEG models because of its spectacular nature is the possibility of disproportionately strong effects of small, temporary shocks. "Spatial catastrophes," where short-lived shocks can have permanent impacts on location patterns, have been very recently subject to empirical testing, which we cover in our section 8. We conclude the chapter with an overall assessment of the empirical support for NEG and suggest some directions for future research.

[1] There are now at least three monographs—Fujita et al. (1999), Fujita and Thisse (2002), and Baldwin et al. (2003)—authored by combinations of leading theorists in the field that provide thorough analyses of the theoretical aspects of the literature.

2 Defining, Delimiting, and Testing the NEG

The label "new economic geography" is unfortunate in a number of respects. First, it raises hackles by claiming as novel that which some already considered to be well-known, but under-appreciated work. More importantly, the label gives no clear indication of the contents. This means that the same label might be used to describe quite different areas of inquiry. Finally, it is not clear what one should call later work that might supersede the current approach. However, in linguistic choice as with location choice, there is often a gain from following the decisions of predecessors. We therefore adhere to common usage in taking "new economic geography" (or NEG) to refer to theories that follow the approach put forward in Krugman's 1991 book (Krugman, 1991b) and, particularly his *Journal of Political Economy* article (Krugman, 1991a). While we do not wish to denigrate the contributions preceding and following these two pieces, their huge influence is an empirical fact. A *Web of Science* search shows that these two works received a combined total of over 1000 journal citations since they were written.

Ottaviano and Thisse point out in their chapter of this Handbook that many of the ingredients of new economic geography were developed many decades before Krugman's (1991a) paper. Indeed they suggest that the main contribution of NEG was to "combine old ingredients through a new recipe." Krugman and many of the other 1990s contributors to NEG gave little acknowledgement to its antecedents in regional science and location theory. Rather, they approached economic geography with perspectives developed from "new trade" theory. Indeed, the concluding section of Krugman (1979) anticipates many of the model elements and results that would appear over a decade later:

> "...suppose that there are two regions of the kind we have been discussing and that they have the same tastes and technologies. There is room for mutual gains from trade, because the combined market would allow for both greater variety of goods and a greater scale of production. The same gains could be obtained without trade however, if the population of one region were to migrate to the other. In this model, trade and growth in the labor force are essentially equivalent. If there are impediments to trade, there will be an incentive for workers to move to the region which already has the larger labor force. This is clearest if we consider the extreme case where no trade in goods is possible but labor is perfectly mobile. Then the more populous region will offer both a greater real wage and a greater variety of goods, inducing immigration. In equilibrium, all workers will have concentrated in one region or the other. Which region ends up with the population depends on initial conditions; in the presence of increasing returns history matters."
> Krugman (1979), p. 478.

This quote shows that the main elements of the stories formalized in the 1990s economic geography literature had already been anticipated by Krugman in the late 1970s. Krugman certainly did not originate all the ideas currently associated with NEG. However, the approach he popularized drew heavily on his own earlier work on trade patterns.

2.1 Essential ingredients for NEG

Five essential ingredients distinguish NEG models from other approaches to understanding the geography of economic activity. We do not wish to imply that they were novel contributions of NEG or new trade but rather that they are useful indicators for categorization.

1. *Increasing returns to scale (IRS) that are internal to the firm.* NEG models assume a fixed, indivisible amount of overhead required for each plant. NEG models *do not* assume any pure technological externalities that would lead directly to external scale economies.

2. *Imperfect competition.* With internal increasing returns, marginal costs are lower than average costs. Hence, one cannot assume perfect competition because firms would be unable to cover their costs. The vast majority of the literature goes on to assume a particular market structure and accompanying functional forms for demand: Dixit and Stiglitz' (1977) model of monopolistic competition.[2]

3. *Trade costs.* The outputs and inputs used by firms are tradeable over distances but only by incurring costs. These costs are often assumed to be proportional to the value of the goods traded.

4. *Endogenous firm locations.* Firms enter and exit in response to profitability at each possible location. The assumption of increasing returns implies that firms have an incentive to select a single production site and serve most consumers at a distance. If plant-level fixed costs were negligible, the firm would replicate itself everywhere (a la McDonalds).

5. *Endogenous location of demand.* Expenditure in each region depends upon the location of firms. Two mechanisms for the mobility of demand have been proposed.

   (a) Mobile workers who consume where they work (Krugman, 1991a).

   (b) Firms that require the outputs of their sector as intermediate inputs (Krugman and Venables, 1995).

Ingredients 1–4 all appeared in the new trade literature, and in particular gave rise to the *home market effects* identified in Krugman (1980). With these assumptions, agglomeration can arise but only through the magnification of initial region size asymmetries. The key innovation of NEG relative to new trade is assumption 5. Without 5, symmetric initial conditions can be expected to lead to symmetric outcomes. With all five assumptions, initial symmetry can be broken and agglomerations can form through a process of circular causation. This is perhaps the basis for the Davis blurb on the back of Fujita et al. (1999) that, "the work is an even more radical departure from orthodoxy than the new trade theory of the 1980s."

2.2 Alternative explanations of agglomeration

If NEG comprises models with these five ingredients, what are the competing explanations of economic geography? Empirical work testing NEG-based hypotheses benefits from the consideration of a set of *plausible alternatives*. Prominent alternatives to NEG include

[2]Recent work by Ottaviano et al. (2002) shows that a linear model of monopolistic competition retains most of the key predictions obtained from the Dixit-Stiglitz structure. Results by Combes (1997), Head et al. (2002), and Feenstra et al. (2001) suggest that NEG models could also rely on Cournot competition with free entry.

- *Natural advantages* (see Ellison and Glaeser, 1997, 1999)—also known as "First Nature" (Krugman, 1993) and "locational fundamentals" (Davis and Weinstein, 2002)—and the closely related "factor proportions theory" take the geographic distribution of productive resources as exogenous and use it to explain the geographic distribution of production.

- *Human capital externalities* models link the return to skill in a location to the number of skilled workers there. High skill areas tend to attract larger numbers of employers of skilled workers. Marshall (1920) describes this mechanism for agglomeration. Formal models were developed by Krugman (1991b) and Helsley and Strange (1990). Human capital externalities are central in Lucas' (1988) theory of economic development. Empirical applications are covered in the Moretti chapter of this Handbook.

- *Technological externalities/Knowledge spillovers*: Producers benefit from spatial proximity of their counterparts in the same industry via flows of productive knowledge.

The Rosenthal and Strange chapter of this Handbook considers the empirical evidence in favour of each of these microfoundations for agglomeration. Our chapter, in contrast, focuses its attention on work that has a direct bearing on the validity of the NEG approach.

In any type of empirical testing of NEG predictions, we think an important issue is that the researcher should keep in mind the presence of the alternative explanations outlined above. Ideally, the empirical procedures employed should incorporate one or more *discriminating hypotheses* that can help differentiate NEG-type mechanisms from natural advantages or "pure externalities" explanations for the level of agglomeration observed in the data. Davis and Weinstein (1996), which we cover in detail, proposed a first empirical test along this route, trying to discriminate between NEG and the explanatory framework of traditional trade theory. While discrimination often proves difficult in this type of modelling, we believe the literature would progress in an important way by following this path, through the application of discriminating tests to a broader set of issues.

2.3 Testing NEG propositions

For guidance, we think it useful to refer to Leamer and Levinsohn's (1995) influential chapter on the empirical evidence on international trade theory. This chapter is known by many empirical trade economists for its puzzling injunction to "Estimate, don't test." Its more useful contribution is the process of laying out clear and compelling propositions derived from theory that can be subjected to empirical scrutiny (i.e. tested).

Leamer and Levinsohn (1995) council empiricists to steer a middle road between "taking theory too seriously" and "treating theory too casually." A related way to state the problem is in terms of the classical statistical problems of Types I and II error. In doing empirical work on NEG we want to avoid interpreting results as rejecting NEG when it actually offers valuable insights. This might occur if our tests hinge on some highly fragile aspect of the theory rather than its core empirical content. Conversely, we do not want to confirm the validity of NEG based on results that are consistent with NEG but would also be equally consistent with alternative theories.

Two examples illustrate these problems. In terms of "false confirmations" consider the following quote from Baldwin et al. (2003): "Exhibit A is the concentration of economic activity in the face of congestion costs. Two bedroom houses in Palo Alto California routinely change hands for hundreds of thousands of dollars while houses in northern

Wisconsin can be had for a song. Despite the high cost of living and office space, Silicon Valley remains attractive to both firms and workers while economic activity in northern Wisconsin languishes." While high housing prices within agglomerations are consistent with NEG they are also consistent with the three alternative theories of spatial variation in economic activity. Indeed the natural advantages theory seems consistent with the facts above. In particular, the superior climate in the San Francisco Bay Area (temperatures averaging 49F (9.5C) degrees in January versus 14F (-10C) in Green Bay in Northern Wisconsin) could push up housing prices and raise economic activity there.

False rejections can arise from the failure of the actual data to exhibit certain features that models exhibit only as a consequence of simplifying assumptions rather than as a result of the fundamental mechanism the model proposes. For example, Krugman (1991a) predicts that the distribution of manufacturing activity across regions will be either perfect symmetry or complete concentration in one region. Actual data for Europe or North America show that all major regions contain some manufacturing workers but they are far from evenly distributed. Before we reject NEG based on this data, we should recognize that models including all 5 of the identifying features of NEG are consistent with these facts (Tabuchi and Thisse, 2002).

Our review of the empirics of agglomeration and trade is organized around 5 propositions that emerge from the most well-known NEG models. In some cases we include alternative or subsidiary formulations of a given proposition.

1. *Market potential raise local factor prices.* A location whose access to major markets and suppliers is not impeded by large trade costs will tend to reward its factors with higher wages and land rentals.

2. *Market potential induces factor inflows.* Capital will be drawn to areas with good access to major markets for final goods and major suppliers of intermediate inputs (backward linkages). Workers favour locations with good access to suppliers of final goods (forward linkages).

3. *Home market/magnification effect* (HME). Regions with large demand for increasing returns industries account for an even larger share of their production. Put another way, the larger of two regions will be a net exporter to the smaller region in industries characterized by plant-level increasing returns.

4. *Trade induces agglomeration* (TIA). In an industry featuring increasing returns and partially mobile demand, a reduction in trade costs facilitates spatial concentration of producers and consumers.

5. *Shock sensitivity*: A temporary shock to economic activity in a location can permanently alter the pattern of agglomeration.

3 Preliminaries: Defining and Measuring Market Potential

The primary mechanisms at work in NEG are the market size effects first identified in Krugman (1980). Krugman (1980) developed the basic model combining monopolistic competition and trade costs. He then explored two implications, which we will refer to as the "price" and "quantity" aspects of the market size effect.

The price effect emerges in a one sector model. If the resources employed in each country in each sector are fixed by full-employment and trade balance considerations, then the zero profit condition implies that the smaller country must pay lower wages. Otherwise, firms would prefer to locate in the large country and serve the small one through exporting.

Krugman (1980) illustrates the quantity effect in a very stylized setting involving equal-sized countries, two industries, and "mirror-image" preferences. Helpman and Krugman (1985) later provided a more satisfactory development of the quantity market size effect. As with the price version, the country with the larger market is appealing because it allows the producer to economize on trade costs. If wages do not rise to eliminate this advantage, then a disproportionate share of the producers will locate in the large market. This result is usually referred to as the "home market effect" or the "magnification effect."

The rest of this section will proceed as follows. First we will show how trade costs influence trade flows and introduce the critical parameter, ϕ_{ij}, measuring accessability of a given market i to imports from source j. Then we derive what might be thought of as the fundamental equation of NEG: The relationship between the prospective profitability of a location and its "real market potential." Then we consider the effect of market potential on factor prices and location decisions in subsequent sections.

3.1 Measuring access to markets

We employ the Dixit-Stiglitz-Krugman model of monopolistic competition and trade in a multi-region setting. Let $\mu_i Y_i$ denote expenditures by region i on the representative industry. In theoretical models it is standard to make industry level expenditure be exogenous by assuming an upper level utility function that is Cobb-Douglas with expenditure parameter μ_i, thus giving rise to fixed expenditure shares out of income, Y_i. The sub-utility is a constant elasticity of substitution (CES) aggregate of differentiated varieties produced in the considered industry, with σ representing an inverse index of product differentiation.[3] In this model, σ plays several "roles," being in particular an inverse measure of the markup and available economies of scale. This parsimony is useful in theory but dangerous in applications.

The amount spent by consumers from region i for a representative variety produced in region j is given by

$$p_{ij}q_{ij} = \frac{p_{ij}^{1-\sigma}}{\sum_k n_k p_{ik}^{1-\sigma}} \mu_i Y_i, \tag{1}$$

where p_{ij} is the delivered price faced by consumers in i for products from j. It is the product of the mill price p_j and the *ad valorem* trade cost, τ_{ij}, paid by consumers. Trade costs include all transaction costs associated with moving goods across space and national borders. We can see from (1) that trade costs influence demand more when there is a high elasticity of substitution, σ. Indeed many results in Dixit-Stiglitz based models depend on the term $\phi_{ij} \equiv \tau_{ij}^{1-\sigma}$, that Baldwin et al. (2003) punningly refers to as the "phi-ness" of trade.

[3]The *taste for variety* of each consumer represented by the CES functional form is not essential to the model, as the same aggregate demand structure can be obtained with a model of *variety of tastes* when the variance of consumer preferences is described by a logistic distribution (Anderson et al., 1992).

The total value of imports (including trade costs) from all n_j firms based in region j will be denoted m_{ij}.

$$m_{ij} = n_j p_{ij} q_{ij} = n_j p_j^{1-\sigma} \phi_{ij} \mu_i Y_i P_i^{\sigma-1}, \tag{2}$$

where $P_i = (\sum_k n_k p_k^{1-\sigma} \phi_{ik})^{1/(1-\sigma)}$. Fujita et al. (1999) refer to P_i as the "price index" in each location. It is a generalized mean of the delivered costs of all the suppliers to location i that assigns increasing weight to sources that have a large number of suppliers, n_k, or good access to market i, measured by a high ϕ_{ik}. Thus a location that is served by a large number of nearby and low-price sources will have a low P_i and will therefore be a market where it is difficult to obtain a high market share.

Equation (2) can be manipulated to obtain an estimate of ϕ_{ij}. First, divide m_{ij} by m_{ii}, the region i's imports from itself. The $\mu_i Y_i P_i^{\sigma-1}$ cancel since they apply to i's imports from all sources. The remaining expressions involve relative numbers of firms and relative costs in i and j. These ratios can be eliminated by multiplying by the corresponding ratio for region j: m_{ji}/m_{jj}. The result is

$$\frac{m_{ij} m_{ji}}{m_{ii} m_{jj}} = \frac{\phi_{ij} \phi_{ji}}{\phi_{ii} \phi_{jj}}. \tag{3}$$

The standard practice in NEG models is to assume free trade within regions, i.e. $\phi_{ii} = \phi_{jj} = 1$ and symmetric bilateral barriers $\phi_{ji} = \phi_{ij}$. These assumptions lead to a very simple estimator for ϕ_{ij}:

$$\hat{\phi}_{ij} = \sqrt{\frac{m_{ij} m_{ji}}{m_{ii} m_{jj}}} \tag{4}$$

The numerator requires only trade flow data expressed according to industry classifications. The denominator factors are each region's "imports from self" (or, equivalently, "exports to self"). They are calculated as the value of all shipments of the industry minus the sum of shipments to all other regions (exports).

It therefore is fairly easy to give a feeling of the extent of current trade freeness among the biggest industrialized countries for which bilateral trade flows and production figures are readily available. We use here the database recently made available by the World Bank[4] combined with the OECD STAN database (the appendix gives details about this data) in order to calculate values of trade flows and $\hat{\phi}_{ij}$ for distinctive pairs of countries in 1999. We opt for the United States-Canada and France-Germany as our pairs of countries.

Recalling that $0 < \hat{\phi}_{ij} < 1$ with 0 denoting prohibitive trade costs, the overall level of trade costs in Table 1 seems to be very high. We can obtain from $\hat{\phi}$ an estimate of the ad valorem equivalent of all impediments to trade between the United States and Canada. The calculation requires an estimate of the price elasticity σ. Using the lowest Head and Ries (2001) estimate of σ for US-Canada trade in manufactured goods (8), trade costs have an ad valorem equivalent ranging from $\tau - 1 = 0.717^{-1/7} - 1 = 4.9\%$ for Canada-US auto trade to just over 36% for Canada-US trade in clothing and Germany-France trade in autos. With the exception of North American auto trade, the level of trade freeness appears to be quite low, even though we have chosen pairs of countries known for their high levels of formal trade integration.

The starkest predictions of NEG models deal with the possibly dramatic consequences of trade liberalization on agglomeration. It is often assumed that we live in an era of trade

[4] http://www1.worldbank.org/wbiep/trade/data/TradeandProduction.html.

Table 1: The ϕ-ness of trade in 1999 for North America and Europe, selected industries, import values in millions of US$.

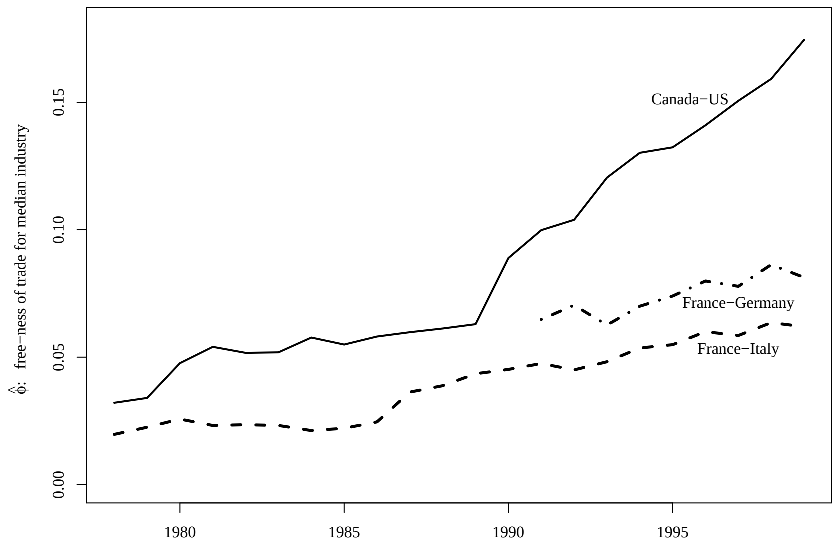
Textiles, apparel & leather				
country	foreign source (m_{ij})	domestic source (m_{ii})	odds (m_{ij}/m_{ii})	$\hat{\phi}$
Canada	3232.48	6275.55	0.515	0.111
USA	3437.03	144731.00	0.024	
Germany	1955.14	9742.13	0.201	0.130
France	1604.72	19095.45	0.084	
Motor vehicles & parts				
country	foreign source (m_{ij})	domestic source (m_{ii})	odds (m_{ij}/m_{ii})	$\hat{\phi}$
Canada	41069.02	13257.20	3.098	0.717
USA	58776.02	354653.00	0.166	
Germany	7468.84	101719.28	0.073	0.114
France	9842.01	55179.41	0.178	

integration and that would here translate into a trend of rising $\hat{\phi}$ over time. Do we actually observe this trend in the $\hat{\phi}$ data?

We consider, in Figure 1, the evolution of trade freeness for three distinctive country pairs. We can indeed see that international trade is getting easier over the recent period. The rate of progress is not the same for all country pairs, with North America being the fastest integrating region since the end of the eighties. The pace of trade integration also seems to be more important since the late eighties in the European Union, as can be seen from the France-Italy combination for which a longer time period is available.[5] It is noteworthy that the change in the pace of integration for the median industry seems to correspond in both regions to the starting date of implementation of a major trade liberalization agreement (the U.S.-Canada Free Trade Agreement in January 1989 and the Single European Act in January 1987). This observed rise in $\hat{\phi}$ is a sort of pre-requisite for any test of the main predictions of NEG models: Although remaining at surprisingly low levels, the integration of the world economy is rising, which corresponds to the typical thought experiment of NEG theoretical predictions.

[5]Note that the fact that trade is consistently freer in the Franco-German than in the Franco-Italian combination is consistent with the smaller bilateral distance in the former as compared to the latter (411 against 550 miles in Head and Mayer, 2000).

Figure 1: Evolution of median-industry ϕ-ness of trade



3.2 Profits as a function of market potential

Returning to the firm's location decision, total production cost in each region is assumed to take the form $c_j q_j + F_j$. Increasing returns come from a plant-specific fixed costs F_j, q_j is the total output of the representative firm in j and c_j is the constant marginal cost of production. Each firm maximizes the following gross profit function for each market: $\pi_{ij} = (p_j - c_j)\tau_{ij}q_{ij}$. The resulting mill prices are simple mark-ups over marginal costs:

$$p_j = \frac{c_j \sigma}{\sigma - 1}.$$

The gross profit earned in each market i for a variety produced in region j is given by $\pi_{ij} = (p_{ij}q_{ij})/\sigma$. Substituting in equation (2) and then summing the profits earned in each market and subtracting the plant-specific fixed cost, F_j, we obtain the net profit to be earned in each potential location j:

$$\Pi_j = \frac{1}{\sigma} c_j^{1-\sigma} \mathrm{RMP}_j - F_j, \tag{5}$$

where $\mathrm{RMP}_j = \sum_i \phi_{ij}\mu_i Y_i P_i^{\sigma-1}$. RMP is an abbreviation of Real Market Potential. Redding and Venables (2000) derive the same term (except they do not use ϕ_{ij} notation) and call it market access. To maintain continuity with prior work (from Harris, 1954, to Fujita et al. 1999), we instead employ the term *market potential*. The "real" is added in order to contrast it with an alternative formulation that we refer to as Nominal Market Potential or $\mathrm{NMP}_j = \sum_i \phi_{ij}\mu_i Y_i$. The "nominal" refers to the absence of an adjustment for variation in the price index term P_i.

Davis and Weinstein (2003a) use a variant of NMP in which they set $\phi_{ij} = d_{ij}^{\delta}$ where d_{ij} is the distance between locations i and j and δ is the coefficient on $\ln d_{ij}$ in a gravity equation estimation using industry level bilateral trade. Since usual estimates of δ do not differ greatly from minus one[6], $1/d_{ij}$ is a reasonable approximation for ϕ_{ij}. Further assuming the share of income devoted to each industry does not vary across countries, one obtains $\mathrm{NMP}_j \propto \sum_i Y_i/d_{ij}$, where Y_i is an aggregate measure of demand such as GDP or retail sales. Thus, NMP is proportional to the original formulation of market potential used by Harris (1954) and in subsequent work of geographers.

Nominal Market Potential is intuitively appealing and not very difficult to implement empirically. However, the omission of the price index adjustment $P_i^{1-\sigma}$ effectively severs the link with the underlying profit maximization problem. The reason is simple. Large demand translates into large profits if profit margins and market shares are high. The more competitors there are in a given location, and the more competitors that have low-cost access to that location, and the lower the marginal costs of those local and nearby competitors—that is to say the lower is P_i—the lower will be any particular firm's share of market i. In other words, a large market that is extremely well-served by existing firms might offer considerably less potential for profits than a smaller market with fewer competitors in the vicinity. NMP might still be useful for some purposes. For instance since it does not depend on locations of firms or on industry level costs, both of which are endogenous in economic geography models, NMP might be a good instrument for RMP. However, a regression that includes just NMP is, at best, a reduced form whose coefficients must be interpreted with great caution.

4 Market Potential Raises Factor Prices

The impact of market potential on factor prices can be seen by solving for the variable costs in region j that would set the profit equation, (5), equal to zero.

$$c_j = \left(\frac{\mathrm{RMP}_j}{\sigma F_j}\right)^{1/(\sigma-1)}.$$

Suppose, following Redding and Venables (2000), that c_j is function of wages (w), prices of other primary factors (v), and intermediate input prices. If all firms use the same basket of intermediates, then P_j is also the appropriate intermediate price index. Assuming a Cobb-Douglas form, we obtain $c_j = P_j^{\alpha} w_j^{\beta} v_j^{\gamma}$. Suppose further that fixed costs are proportional to variable costs, i.e. $F_j = fc_j = fP_j^{\alpha} w_j^{\beta} v_j^{\gamma}$ where f is a constant determining the strength of increasing returns. After making substitutions and rearranging we have

$$\beta \ln w_j + \gamma \ln v_j = -\frac{1}{\sigma}\ln(\sigma f) + \frac{1}{\sigma}\ln \mathrm{RMP}_j - \alpha \ln P_j. \tag{6}$$

Redding and Venables (2000) and Hanson (1998) proceed to the empirical implementation of various versions of this equation linking factor prices to market potential.[7] There are three terms to be estimated in the complete version of this equation. The two most

[6] Disdier and Head (2003) find a mean of -0.87 in meta-analysis of 896 coefficients supplied by 55 different papers.

[7] The first published derivation of the wage-potential equation seems to be the 1991 working paper version of Krugman (1993).

important concern the real market potential on one hand and the price index on the other hand. Note that (6) is closely linked to (5): High RMP for a region predicts a relatively high profit for firms located there. In the long run, when free entry drives profits everywhere towards zero, the input prices have to rise to absorb those extra profits in high RMP regions.

Note also that the price index, P_j, appears twice in the factor price equation. It first appears in the RMP term where it acts as a weight on NMP, accounting for the number of potential suppliers to each market, discounted if they have poor access or charge high prices. The price index also enters as the aggregate prices of intermediate inputs. Based on the assumption that firms consume all varieties of competitors as inputs, costs are lower when those input-output linkages are relatively free from trade costs, i.e. when the price index is low, signifying that (input-supplying) competitors are relatively close from your place of production.

Equation (6) bears a close resemblance to the equation estimated by Dekle and Eaton (1999). They relate a share-weighted index of wages and land rents in Japanese prefectures to a term that sums across incomes discounted by distance. There are important differences in theoretical motivation that also result in subtle, but important, differences in specification. Dekle and Eaton (1999) assume agglomeration economies taking a technological form in which the production function has a neutral shift term that depends on nearby economic activity. Their term is a variant of NMP that assumes an exponential distance decay function. Thus, it differs from RMP because of the absence of the price index term. In principle this distinction might be used to break the observational equivalence between NEG approaches to factor price determination and approaches that invoke spatial technological externalities.

4.1 Market Potential and international income inequality

The left hand side of (6) is a cost-share weighted sum of logged primary factor prices. A natural proxy for this is the log of GDP per capita or ln GDPC. Adding an error term we have

$$\ln \mathrm{GDPC} = \zeta + \frac{1}{\sigma} \ln \widehat{\mathrm{RMP}}_j + \frac{\alpha}{\sigma - 1} \ln \widehat{\mathrm{SP}}_j + \varepsilon_j, \tag{7}$$

where $\mathrm{SP}_j \equiv P_j^{1-\sigma}$ stands for supply potential (referred to as "supplier access" by Redding and Venables, 2000). This is the key equation estimated in Redding and Venables (2000). The authors obtain the RMP and SP terms needed for this regression using estimates from the bilateral trade equation of the model.[8] Indeed, using the fact that bilateral trade volumes $m_{ij} = n_j p_{ij} q_{ij}$, where $p_{ij} q_{ij}$ is given by (2), we obtain

$$\ln m_{ij} = \mathrm{FX}_j + \ln \phi_{ij} + \mathrm{FM}_i, \tag{8}$$

where the variables FX_j and FM_i are exporter and importer fixed effects respectively, with theoretical correspondence $\mathrm{FX}_j = \ln\left(n_j c_j^{1-\sigma}\right)$ and $\mathrm{FM}_i = \ln\left(P_i^{\sigma-1} \mu_i Y_i\right)$. Therefore, a bilateral trade flow regression in a first step provides estimates of ϕ_{ij} and fixed effects that can be used to construct $\widehat{\mathrm{SP}}_j = \sum_i \exp(\mathrm{FX}_i)\phi_{ij}$ and $\widehat{\mathrm{RMP}}_j = \sum_i \exp(\mathrm{FM}_i)\phi_{ij}$.

[8]This two step procedure where the first step makes use of the gravity-like prediction of bilateral trade patterns in empirical implementations of the NEG model originates in the 1998 NBER working paper version of Davis and Weinstein (2003a) and is also adopted by Head and Mayer (2002).

Equation (8) is fitted by Redding and Venables (2000) on a sample of 101 countries for the year 1994 with bilateral distance and contiguity being used to estimate ϕ_{ij} in this gravity-like regression. The two variables of interest SP_j and RMP_j are then constructed for the same set of countries in 1996 with a distinction between the *domestic and foreign components* of those potential variables. Three different sets of variables are constructed which use different proxies for ϕ_{ii}, the internal trade costs, supposed to be linked in alternative ways to the internal distance of a country approximated by $d_{ii} = 2/3\sqrt{\text{area}/\pi}$.[9] Supplier and market potential are regressed separately because of strong correlation in the series, most of the analysis uses RMP regressions. Note that the method of calculating ϕ_{ii} relative to ϕ_{ij} has important implications for the results. It can be seen from the definition of RMP that any overestimate of freeness of internal trade relative to international trade will give higher relative weight to local GDP in the RMP calculation. In the limit, if ϕ_{ii} approaches 1 and ϕ_{ij} approaches 0 for all international trade flows, only local GDP will be considered in market potential and the estimation will be dangerously approaching a regression of GDP per capita against GDP. This issue is extremely clear when comparing results from the two last RMP specifications in their Figures 3 and 4 graphing the log of GDP per capita against the log of RMP(2) and RMP(3) respectively. The definitions of those variables are such that RMP(3) divides the coefficient applied to internal distance in ϕ_{ii} calculation by two compared to RMP(2). This systematically increases the weight of local GDP in market potential calculation and not surprisingly increases the fit of the regression as remote but high-income countries like Australia see their remoteness reduced through the larger weight put on local GDP.

A natural way to correct for this problem is to run regressions with only the foreign component of market potential. The authors show that this component alone can explain an impressive 35% of GDP per capita variation across the sample. The full specification with RMP(3) makes the figure rise to near 75%. Robustness checks are conducted with first inclusion of recently successful variables of the cross-country growth literature (endowments in natural resources, physical geography, quality of institutions). The NEG variables measuring access retain their influence. The problem with focusing on the foreign component is that the theory clearly calls for local wages to be increasing in the size of the local market. *Based on foreign market potential only, one would expect Canada to pay substantially higher wages than the United States.*

The ideal solution would be to construct RMP using domestic *and* foreign market potential, but to instrument for it in the regression analysis to solve the endogeneity problem (income enters on both sides of the equation). Redding and Venables (2000) take this approach, using distance to New York City, Brussels, and Tokyo as the instruments. This approach removes contemporaneous shocks to local income per capita. Since the location of these centers of high income are not themselves exogenous in the long run (i.e. why not Rio de Janeiro, Lagos, and Delhi?), we see much scope for future development of the instrumental variable approach.

4.2 Market potential and interregional real wage differences

In a seminal paper on estimation of wage equations, Hanson (1998) adopts a similar model structure for his study of market potential inside the United States.[10] One difference is that

[9] This measure makes the assumption that each country is a disk where all producers are located in the center and consumers are located uniformly over the area.

[10] We refer here to the 2001 revised version of a paper first issued as a 1998 NBER working paper.

he omits intermediate inputs and primary factors other than labor from the production function. Imposing $\alpha = \gamma = 0$ and $\beta = 1$ in (6), the iso-profit condition reduces to

$$\ln w_j = -\frac{1}{\sigma}\ln(\sigma f) + \frac{1}{\sigma}\ln \mathrm{RMP}_j = -\frac{1}{\sigma}\ln(\sigma f) + \frac{1}{\sigma}\ln\left(\sum_i \mu_i Y_i \phi_{ij} P_i^{\sigma-1}\right). \tag{9}$$

Hanson (1998) then imposes two additional equilibrium conditions. First, he assumes that free migration equalizes real wages across locations. The model follows Helpman (1998) in replacing the agricultural good in the upper level utility function of individuals with housing. Denoting H_i as the housing stock in i and P_i^H the price of housing, real wage equalization implies $w_i P_i^{-\mu_i}(P_i^H)^{-(1-\mu_i)} = C, \quad \forall i$, where C is a constant.[11] The second equilibrium condition is that housing payments equal housing expenditure: $P_i^H H_i = (1-\mu_i)Y_i$. It is then possible to replace the two price terms in (9), and obtain:[12]

$$\ln w_j = B + \frac{1}{\sigma}\ln\left[\sum_i Y_i^{\frac{\sigma(\mu-1)+1}{\mu}} w_i^{\frac{\sigma-1}{\mu}} H_i^{\frac{(\sigma-1)(1-\mu)}{\mu}} \phi_{ij}\right], \tag{10}$$

with B a function of C, μ, σ and f, constant over all locations i. The first difference of this last equation as well as a simplified version omitting the price index in the market potential (and therefore closer to nominal market potential) are estimated using a nonlinear least squares estimation procedure. Concentrating on the specification most linked to theory, the principal result is that the estimated coefficients imply parameters consistent with the underlying theoretical framework, with a reasonably good overall fit (0.347 for the 1980-1990 period). The analysis produced for all 3075 US counties shows that the higher are personal incomes, wages and housing stocks in proximate locations, the greater will be the local wage. An appealing feature of the approach is that the estimation of the wage equation provides estimates of key parameters of the model. This is useful *per se* but also can be used as a device to check the consistency of the results with the underlying theoretical framework. The estimate of σ ranges between 4.9 and 7.6, which corresponds to recent estimates in the literature by Head and Ries (2001) or Lai and Trefler (2002) for instance. Those values of σ are interesting first because they confirm recent results through a very different estimation strategy, but also because those estimates are consistent with reasonable values for other equilibrium relationships in the model: With those high σ, the equilibrium markup of prices over marginal costs in the model, $\sigma/(\sigma-1)$, is between 1.15 and 1.25. The expenditure share of the IRS traded good, μ is estimated to be between 0.91 and 0.97, which lies within the 0-1 range but is much higher than the actual share of expenditures on manufactured goods.[13] Another interpretation on parameters values

[11] This assumption, present in Helpman (1998), is somewhat restrictive. Indeed, imposing equality in real wages in the original Krugman (1991a) framework forces the model to be at the symmetric equilibrium, for if the equilibrium was agglomerated, it would yield a core-periphery outcome and the region hosting the manufacturing industry would pay a higher real wage as in Figure 5.2 in Fujita et al. (1999). However, due to the presence of the housing sector, the Helpman (1998) model can exhibit interior agglomerated equilibria that satisfy the real wage equalization assumption. This assumption, while not innocuous, seems defensible as long as the sample under consideration does not exhibit a core-periphery outcome (which is obviously very rare).

[12] Note that Hanson (1998) makes the additional assumption that $\mu_i = \mu, \quad \forall i$. Also the ln function is missing for the market potential term in this version.

[13] The set of unreported robustness checks include a specification replacing the housing sector by a Krugman (1991a) freely traded agricultural product. Estimates of σ and τ are similar but μ is estimated to be between 1.5 and 2.

is that, in the Helpman (1998) model, the equivalent of the "no black hole condition" of Fujita et al. (1999) is that $\sigma(1-\mu) < 1$. If this condition is not satisfied, the equilibrium is always dispersed (remember that the prediction of the model, in terms of relationship between agglomeration and trade costs, in Helpman (1998) is the reverse of Krugman's (1991a) one), independently of trade costs. All specifications yield values of parameters satisfying this condition. The econometric analysis therefore reveal that the wage equation seems to fit well the spatial variation of wages within the United States. It is noteworthy that the full implementation of the theory-based wage equation obtains a better fit than the simpler wage equation based on Harris (1954) market potential.

To give a clearer view of the extent to which geography matters in the determination of wages in the United States, Hanson (1998) then proceeds to simulations of the model with the parameters estimated. The exercise simulates a negative 10% shock on the income in Illinois. The impact of this shock on wages drops very rapidly with distance: 74 kilometers away from the initial shock, wages fall by only 0.43%, in St Louis (345 kilometers away), the fall is down to 0.32%, and wages are unchanged at a distance of 885 kilometers. This extremely strong impact of distance on wage response to localized shocks is the translation of the trade cost parameter estimated which implies that travelling 2 kilometers multiplies the price of a good by $\exp(2 \times 1.97) = 51.4$ (using Hanson ,1998, specification of trade costs and his estimate for the 1970-1980 regression). This disturbingly large estimate may be a consequence of the function form of the distance decay function. Hanson (1998) assumes $\tau_{ij} = \exp(td_{ij})$, the formulation used by Krugman (1993) in his original theoretical derivation of the wage equation and by Dekle and Eaton (1999) in their empirical work. The vast empirical literature estimating gravity equations suggests that τ_{ij} should be a *power* function of distance of the form $\tau_{ij} = d_{ij}^{\delta}$, as the log of trade flows is unanimously found to decrease linearly with the *log* of distance (usually with slope near -1).

The above analysis by Gordon Hanson is a structural estimation of the wage equation linking positively nominal wage to market access within the United States. The chapter by Combes and Overman discusses some recent work that applies the Hanson (1998) methodology to wages in European countries. While the Hanson (1998) and Redding and Venables (2000) papers both draw on the iso-profit equation's implications for spatial wage variation, they make different assumptions about worker mobility and use very different econometric strategies. Future work should evaluate these differences and their implications.

Non-structural methods can also be useful for assessing the relationship between access and factor prices. With an important external trade liberalization, the internal geography of production is likely to change rapidly as foreign markets rise in importance to domestic producers. The quality of access to foreign consumers may gain weight in the location decision relative to former domestic centers of consumption.

Hanson (1997) takes the example of trade liberalization in Mexico which provides a natural experiment of this process, as the country experienced a 40-year period of protectionism ending abruptly in 1985, when the country liberalized foreign trade dramatically. The fact that centrally-located Mexico City concentrated such a large proportion of industrial activity before liberalization combined with the proximity, location and size of the US economy, makes the evolution of wage gradients inside the Mexican economy the basis of a reduced form estimation of market access forces at work.

Hanson (1997) uses as a dependent variable the wage in each Mexican region relative to Mexico City wage in the same industry. The explanatory variables are distances to the capital and to the nearest major United States border crossing, together with the same variable interacted with a post-1985 dummy variable. Industry and year fixed effects are included. The results show indeed that distance to industry centers has a negative influence on relative wages. A 10% increase in distance to Mexico City reduces wages by 1.92% whereas the same increase in distance to the US border reduces wages by 1.28%. Access to markets indeed matters for local wages. The other main prediction however receives less support. The change in trade policy occurred in 1985, but there seems to be no strong evidence of a strong change in wage gradients (diminished impact of distance to Mexico City and increased impact of distance to US) after this date.[14] There is some evidence of wage compression over time from 1965 to 1988 in the country, but this movement is not much more pronounced in the end of the sample and was in fact quite stable over the period. The interpretation Hanson (1997) favors is that the older maquiladora programme (launched in 1965 and providing massive liberalization on input imports for exporting plants) had already contributed to the theoretically expected wage compression.

Even if all methodological issues have not yet been resolved, the results surveyed in this section point to an apparent empirical success of the wage equation (and more generally of the price aspect of market size effects), which constitutes an important mechanism of NEG models.

5 Market Potential Attracts Factor Inflows

With micro data one may explore how firms' and workers' location decisions depend on market and supply potential. Two types of location choices can be studied, location choice of production units by firms and location choice of individuals through migrations. Such analysis can be interpreted as empirical test of the existence of: 1) *Backward linkages* (are firms attracted to locations with large demand for their products?) 2) *Forward linkages* (are consumer-workers attracted to locations with high industrial production?).

5.1 Firm locations and downstream demand

Consider first the location decisions of firms. Firms wish to choose the location that leads to highest expected profits. Thus a firm will choose location j if it expects Π_j to exceed Π_k for all $k \neq j$. Firms that appear identical to the econometrician (same investment timing and industry, etc.) often choose different locations. As a result, it is conventional to assume that there are unobserved components to the profit function. When those unobserved components have a distribution given by a multivariate extreme value, parameters of the profit equation yielding location choices can be estimated by the conditional logit model initiated by Daniel McFadden. Carlton (1983) was the first to apply this model to choice of production sites by firms. Most recent work following this methodology studies the determinants of foreign affiliates' location choices (foreign direct investment samples have the advantage of presenting relatively footloose location choices that are often concentrated over a relatively short period of time) and incorporate a variable or a set of

[14] The evidence of a diminishing importance of distance to Mexico City is slightly stronger in Hanson (1996). This paper (the first of this stream of work on changes in regional manufacturing activity in Mexico) focuses on regional wages in the apparel industry, for which the strength of linkages with upstream US firms might be sufficiently high to yield rapid and measurable changes in location patterns.

variables accounting for the "quality" of access of each alternative location to downstream demand. Table 2 gives some examples of such work.

Table 2: Downstream demand variables used in location choice studies

Reference	Origin of investors	Location choices	Demand measurement
Coughlin et al. (1991)	All foreign investors	American States	State per capita income
Devereux and Griffith (1998)	United States	U.K., France, Germany	Share of total apparent consumption in the industry
Friedman et al. (1992)	All foreign investors	American States	Gravity measure of states' per capita income
Head et al. (1999)	Japan	American States	State personal income + sum of the contiguous states' personal income
Henderson and Kuncoro (1996)	Indonesia	Indonesian districts	District population + distance to the nearest large town

In general, this kind of literature only considers rather simple demand specifications that either consists solely of local income or of *ad hoc* constructions that try to incorporate both local size of demand and more distant sources of consumption. The precise construction used varies from adding the income from contiguous locations (Head et al., 1999) to using a gravity-type measure of incomes bearing some resemblance with NMP (Friedman et al., 1992). The vast majority of the results yield positive coefficients, confirming the intuition that firms value proximity to consumers. The structural interpretation of those coefficients is however problematic as even NMP is only a rough approximation of what a fully specified market potential variable (RMP) should be.

For many suppliers of intermediate inputs, the relevant "consumers" are downstream firms. Smith and Florida (1994) examine the location decisions of about 200 Japan-based auto parts suppliers that established factories in the United States during the 1980s. They found a strong attractive influence of the location of Japan-owned auto assembly factories. Head et al. (1995) exploit the Japanese institution of vertical *keiretsu* to examine the co-location of vertically related factories of 751 Japanese plants established in the U.S. between 1980 and 1990. They found that members of the same *keiretsu* tended to choose the same states. This tendency was large and statistically significant even after controlling for agglomeration effects at the industry level. Moreover, states that were adjacent to locations that had attracted *keiretsu* investment were more likely to be chosen than states with no nearby *keiretsu* investment. This suggests that a spatial nature to the linkage rather than mere emulation of location choices.

Head and Mayer (2002) explore the firms' side of location decisions based on a structural model of the market access motive. This paper studies a sample of 452 affiliates that Japanese firms established in 57 regions belonging to 9 European countries (Belgium, France, Germany, Ireland, Italy, the Netherlands, Spain, Portugal and the United Kingdom) during the period 1984–1995. When an affiliate chooses its location, the only relevant information is the ordering of profits over alternative locations. Monotonic transformations can therefore be made to the profit function (5) in order to obtain an additive expression for the profitability of each location. Specifically, we add F_j, multiply by σ, and take logs, yielding

$$V_j \equiv \ln[\sigma(\Pi_j + F_j)] = -(\sigma - 1)\ln c_j + \ln \mathrm{RMP}_j. \tag{11}$$

As in Hanson (1998) and Redding and Venables (2000), the central issue is the construction of the RMP_j variable. Head and Mayer (2002) make use of the trade equation of this model seen in section 3.1. The methodology is, in this respect, close to Redding and Venables (2000). Estimation of a transformed version of the bilateral trade equation m_{ij} enables to obtain the parameters needed for the calculation (whereas Redding and Venables, 2000, can be interpreted as a direct estimation of the entire RMP_j term).

Using $v_j = n_j p_j q^*$ as a notation for the value of production in the considered industry in region j (q^* standing for the individual output of firms, constant in this model if firms share the same technology), CL_{ij} as a dummy variable set to one for countries sharing a common language, and the assumption that trade costs are positively influenced by distance (with elasticity δ) and negatively related to common language (with elasticity λ), the estimated trade equation derived from (1) is

$$\ln\left(\frac{m_{ij}}{m_{ii}}\right) - \ln\left(\frac{v_j}{v_i}\right) = -b - (\sigma - 1)\ln\left(\frac{p_j}{p_i}\right) - \delta \ln\left(\frac{d_{ij}}{d_{ii}}\right) + \lambda \mathrm{CL}_{ij} + \epsilon_{ij}, \tag{12}$$

where $\exp(-b)$ gives, everything else equal, the ratio of intra-national to international trade (the, large, negative impact of the political borders on trade flows first uncovered by McCallum, 1995). The estimated parameters $(\hat{\sigma}, \hat{b}, \hat{\delta}, \hat{\lambda})$ are then used (together with industry-level apparent consumption, wages and number of competitors for each region j needed in RMP_j) to construct the market potential variable included in the location choice analysis of Japanese firms in Europe. The formulas used for the construction of trade costs are

$$\phi_{ij} = \tau_{ij}^{1-\sigma} = e^{-\hat{b}+\hat{\lambda}\mathrm{CL}_{ij}} d_{ij}^{-\hat{\delta}} \text{ for } i \neq j \text{ and } \phi_{ii} = \tau_{ii}^{1-\sigma} = d_{ii}^{-\hat{\delta}}.$$

The c_j variable can be given several specifications in empirical work. In their most complete setting, Head and Mayer (2002) consider typical labour market and fiscal determinants of production costs (wages, unemployment rate, social charges, corporate tax rate and regional subsidy eligibility). The paper also incorporates two proxies for other intra-industry externalities into the cost function intended to capture the possibility that clustering leads to direct economic benefits such as access to workers with specialized skills or knowledge sharing between competitors.

There are three main specifications of the market potential estimated, the first one corresponds to the theoretical equation (11), the second reduces market potential to the Harris (1954) formula which simplifies the assumed trade costs and neglects the impact of competitors on the location choice. The third specification follows Redding and Venables (2000) and separates RMP_j into local and nonlocal components. In nested logit estimates (a discrete choice model that allows for the correlation of error terms among location alternatives inside a same country) Head and Mayer (2002) obtain a point estimate of 1.26 on the RMP_j term. This implies that a 10% rise in the market potential of a European region yields to a 10.5% increase in the probability of this region being chosen by a Japanese investor. The near unitary effect of market potential corresponds to the theoretical prediction and the goodness of fit of the different regressions as well as the coefficients suggest a small preference for the RMP_j specification of market potential over the Harris (1954) version. However the specification yielding the highest fit is the one separating local and nonlocal components of the market potential term. The striking result of this specification is that the local component of demand has a clearly dominant influence on location choices.

A last result is that the variables embodying other intra-industry externalities retain a strong positive effect on location choices regardless of the estimation technique and market potential formulation. The previous findings of agglomeration effects using the same type of variables are very common in the literature (Head et al. 1995, Devereux and Griffith, 1998, Guimarães et al., 2000, Crozet et al., forthcoming, for instance). Those previous results could have been caused by a mis-specification of the demand term, described as (various forms of) local income of the locations, and therefore be proven invalid when considering market potential properly. Head and Mayer (2002) show evidence of the contrary: Even when final demand linkages are appropriately controlled for through the market potential term, direct agglomeration effects appear to retain a powerful role in location choices. This suggests that the *backward linkage* NEG mechanism might not be the only or even the main driver of clustering behavior by firms (at least by foreign investors).

5.2 Worker locations and forward linkages

What about the empirical validity of the *forward linkage*? Workers choose locations to maximize expected real wages after taking into account mobility costs. Let us denote real wages in i with ω_i, given by nominal wages divided by the aggregate price index. The latter depends on the modern sector's price index, P_i, with expenditure share μ and the traditional (often considered to be agriculture) sector price p_i^Z with share $1-\mu$. Under standard assumptions p_i^Z is normalized to one. Hence the log real wage is given by

$$\ln \omega_j = \ln w_j - \mu \ln P_j. \tag{13}$$

In the Krugman (1991a) version, worker movement is governed by a very simple equation: $\dot{\lambda}_i = \kappa(\omega_i - \bar{\omega})\lambda_i$. This formulation does not consider mobility costs and the high likelihood of heterogeneity in the cost of re-locating. Discrete choice models offer a much more realistic treatment of mobility without sacrificing tractability.[15] Suppose that utility in region j is given by $\ln \omega_j + \epsilon_j$, where ϵ_j describes heterogeneous nonpecuniary benefits living in region j. The probability of moving from i to j is given by the probability that $(\ln w_j - \ln w_i) + \mu(\ln P_i - \ln P_j) - \rho + \epsilon_j - \epsilon_i > 0$, where ρ represents the mean relocation costs. Making distributional assumptions for $\epsilon_i - \epsilon_j$, we can estimate the responsiveness of location choice to proximity to producers. Although an estimation of this sort lies at the core of the propositions of NEG, we are not aware of any papers to have done so with micro data. Crozet (2000) has estimated a similar proposition with *aggregate* flow data in a structural estimation of a NEG model.

His framework incorporates a third industry, a nontraded service sector, which (as in Hanson, 1998) enables the predictions of the theoretical model to be less dramatic, the periphery always maintaining some production in the manufacturing industry. The real wage equation is then transformed to be $\ln \omega_j = \ln w_j - \mu \ln P_j^X - \psi \ln P_j^Y$, X and Y being

[15] Tabuchi and Thisse (2002) show that allowing for probabilistic migration due to taste heterogeneity has important effects in the NEG model they consider. Murata (2003) confirms this finding in a Dixit-Stiglitz-Krugman framework.

the manufacturing and service industries respectively. Crozet (2000) envisions a mobility cost specification similar to the one in discrete choice models with a deterministic and a random component. Keeping the same notation as above, the number of migrants from j to i is shown to be equal to

$$\mathrm{mig}_{ji} = \frac{\omega_i}{\omega_j}\frac{\epsilon_i}{\rho_{ji}},$$

where ϵ_i is the probability of finding a job in location i (assumed proportional to the employment rate in i) and $\rho_{ji} \equiv (d_{ij} + cB_{ij})^{\nu}$, the bilateral mobility cost, assumed positively related to bilateral distance and non-contiguity (B_{ij} is a dummy variable set to one for contiguous regions). The central equation of the paper is obtained by using mig_{ji} together with traditional pricing rules, the proportional relationship between the number of varieties and the labour force of this model, and the definition of price indexes:

$$\mathrm{mig}_{ji} = \left(\frac{w_i}{w_j}\right)^{1-\psi}\left(\frac{L_i^Y}{L_j^Y}\right)^{\frac{\psi}{\sigma^Y-1}}\left(\frac{\sum_k L_k^X \phi_{ki} w_k^{1-\sigma^X}}{\sum_k L_k^X \phi_{kj} w_k^{1-\sigma^X}}\right)^{\frac{\mu}{\sigma^X-1}}\frac{\epsilon_i}{\rho_{ji}} \tag{14}$$

Several things can be learned from this equation. First, leaving aside relative nominal wage, this expression has some resemblance with the gravity equation. Bilateral flows of workers are positively related to the relative size of the hosting region (because a large host region produces a large share of available varieties and has therefore a low overall price index). The distance term is here related to the mobility costs. Second, the large central term in (14) is related to the nominal market potential (NMP) defined above. It indeed consists of trade cost-weighted sums of market sizes (number of workers here instead of incomes in NMP). Access (which Crozet, 2000, refers to as "centrality") is an attractive characteristic for regions here, like in the location choice of firms. However, it should be noted that the reason is intrinsically different: Good access is attractive *for firms and for workers* because of their high market potential, which translates either in high expected profits (equation 11) or higher nominal wages (equation 7). This dimension appears in the first nominal wage term of (14), but there is an additional aspect that makes access attractive for workers yielded by the large availability of nearby producers and corresponding low price index. Note that the corresponding effect for firms is yielded by low price of inputs in central places, which is given by the supply potential term ($\mathrm{SP}_j \equiv P_j^{1-\sigma}$) in equation (7).

Crozet (2000) proceeds to a transformation of (14), yielding a structurally estimable (through nonlinear least squares, very much in the spirit of Hanson, 1998 and Lai and Trefler, 2002) specification. The sample consists of internal bilateral migration flows between regions of Germany (period 1983–1992), Spain (period 1983–1993), Italy (period 1983–1993), The Netherlands (period 1988–1994), and the United Kingdom (period 1980–1985).

The main results of interest here are the impact of market access on migration inflows. The estimated parameters correspond to signs and magnitudes predicted by the theoretical framework, with a good overall predictive power of the regressions. The most interesting parameters are σ and δ, the CES and the elasticity of trade costs to distance respectively. All estimates of σ lie significantly above 1, ranging from 1.3 for the United Kingdom to 4.3 for the Netherlands. Estimates of δ are also systematically significantly positive across countries, with a very high average value of 1.8 but considerable variation across countries (over 3 in Germany to 0.5 in Spain).

Like Hanson's (1998) simulations of the geographical dissipation factor of a negative shock to Illinois income, Crozet (2000) proceeds to use parameter estimates to evaluate theoretical predictions numerically. The prediction he examines is the break point of trade costs below which the symmetrical equilibrium of his model is not stable anymore and the country should exhibit an increasing core-periphery structure. This is done for each country in terms of relative distance below which the core-periphery should be the only equilibrium. The idea can be summarized as a calculation, for each major region in the country, as a radius defining a surrounding area where the activity would tend to be "attracted" to the central region. It appears that those relative distances are very small, which means that significant core-periphery patterns can only happen on very small distances. An example from the paper is that the German region of Bayern with an internal distance of about 100 kilometers is predicted to attract all IRS activities located within a radius of 120 kilometers from its center. It does not seem to threaten any other important region. This calculation does not incorporate migration costs, and a last exercise conducted by Crozet (2000) uses all estimated parameters to calculate the equilibrium predicted number of migrants for equalized nominal wages arising from (14). The number of predicted migrants (actually roughly consistent with real ones) is strikingly low, even for very large differences in size and very small relative distances. Put together, *these results point to the empirical relevance of agglomeration forces operating through forward linkages, but those forces are likely to stay very localized, unable to generate core-periphery patterns in Europe at a large geographical level*, at least as long as labour remains so sensitive to migration costs.

6 Home Market/Magnification Effects

There are three closely related predictions regarding the effects of market size asymmetries on the geographic distribution of industry activity that have come to be known as "home market effects." Krugman (1980) initiates the literature by demonstrating that the country with the larger number of consumers of an industry's goods will run a trade surplus in that industry. Further development of the model in Helpman and Krugman (1985) shows that the larger country's share of firms in the increasing returns industry exceeds its share of consumers. They also show that increases in a country demand lead to more than one-for-one increases in production.

6.1 The magnification of production

The Krugman (1980) formulation relates ratios of numbers of firms to ratios of numbers of consumers. In particular, Krugman imagines two equal size countries with different preferences. The relative size of country i's home market in terms of our notation would be μ_i/μ_j. Its relative number of firms (and relative production) would be n_i/n_j. Expressed in our notation, Equation (25) of Krugman (1980) shows that

$$\frac{n_i}{n_j} = \frac{\mu_i/\mu_j - \phi}{1 - \phi\mu_i/\mu_j}.$$

Davis and Weinstein (1996, 1999, 2003a) use the derivative to motivate their estimation:

$$\frac{d(n_i/n_j)}{d(\mu_i/\mu_j)} = \frac{1-\phi^2}{(1-\phi[\mu_i/\mu_j])^2} > 1.$$

They term this result the "magnification effect". Note that starting from a point of symmetric preferences, that is where $\mu_i = \mu_j$,

$$\frac{d(n_i/n_j)}{d(\mu_i/\mu_j)} = \frac{1+\phi}{1-\phi}.$$

Inspection of this expression reveals that increasing "free-ness" of trade leads to a magnification of the magnification effect.

Empirical work based on Krugman (1980) must confront three important issues related to the difference between model and data dimensionality. The model (as most of the subsequent theoretical work on the topic) assumes 1 factor, 2 industries, and 2 countries. These assumptions raise the following questions for empirical analysis.

1. How can one allow for factor proportions to influence the trade pattern as in the traditional model of trade? Tests of new trade need to be compared with an alternative of H-O trade.

2. How do we model the relationships between multiple industries? Krugman (1980) considered one industry models where balanced trade required higher wages in the large country. He also considered a two-industry model where balanced trade was achieved by making one country having "mirror image" differences in preferences. This allowed for idiosyncratic demand to determine the location of production at the industry level while retaining equal wages in equilibrium. Helpman and Krugman (1985) gives a role for absolute differences in country size by assuming that there is a zero trade cost constant returns sector that equates wages and absorbs any trade imbalances caused by home market effects operating on the IRS industry. In actual data the mirror image assumption certainly fails and the CRS sector probably does not have zero trade costs or the ability to absorb all trade imbalances.

3. How do we construct demand measures in the presence of more than two countries? Indeed how does one even formulate the home market effect hypothesis? The ratios and shares of the theoretical formulations neglect third country effects.

In addition to these three conceptual problems, there is also a practical problem. What data should be used to measure demand differences? The obvious approach is to use something called "apparent consumption" or "domestic absorption." This starts with domestic production adds imports and subtracts exports. While the measure is sensible, it requires trade data measured in a way so as to make them comparable with production data. The underlying data are collected by different agencies (survey and census for production, customs offices for trade) using different classification systems. To make trade and production data comparable across multiple countries is even more difficult.[16]

Given these four challenges one can understand why the first empirical tests of home market effects did not appear until 16 years after the publication of the Krugman (1980)

[16]There are several data sets that attempt to provide compatible trade and production data for a broad set of countries, industries and years. The World Bank provides a recently assembled data set, with wide overall coverage (referenced above). The database covers the 1976–1999 period with compatible bilateral trade and production data at the ISIC 3-digit level (28 industries) for 67 countries and ISIC 4-digit level (81 industries) for 24 countries.

paper that coined the term.[17] The pioneering paper, Davis and Weinstein's 1996 NBER working paper "Does Economic Geography Matter for International Specialization," was never published. Nevertheless, its methodology and discussions of its chief results appear in Davis and Weinstein (1999, 2003a). Prior to considering the results of these three papers we shall discuss the common method and how it confronted each of the three "dimensionality" issues posed above.

Davis and Weinstein (1996, 1999, 2003a) estimate equations that they describe as being "inspired" by Krugman (1980) but adapted to allow for an important role for factor endowments. Their specification is estimated in levels, rather than the ratios analyzed by Krugman (1980). In particular they propose that production of "goods" (the most disaggregated classification of industries available) is linearly related to variables called SHARE and IDIODEM.

$$X_{gr} = \beta_1 \text{SHARE}_{gr} + \beta_2 \text{IDIODEM}_{gr} + \epsilon_{gr}. \tag{15}$$

SHARE_{gr} is not actually a share. Rather it gives a prediction for region r's production of a good if its output at the more aggregated level (X_r) were allocated across goods in the same proportion as the rest of the "world". Thus suppose we denote X_{gr} as the value of r's production of good g. Then production of the aggregate sums the goods for a given region: $X_r = \sum_g X_{gr}$. Production of rest of world for the good and the aggregate are given by $X_{gR} = \sum_{s \neq r} X_{gs}$ and $X_R = \sum_g X_{gR}$. Expressed in our notation (which suppresses the n subscript they use for industry aggregates), we have

$$\text{SHARE}_{gr} = \frac{X_{gR}}{X_R} X_r.$$

The key variable in the analysis is IDIODEM. It is defined as a deviation from rest-of-world demand patterns. Recall that we defined $E_r = \mu_r Y_r$ where μ_r is an expenditure share parameter and Y_r is total income. Applying the same notation as with X, we have

$$\text{IDIODEM}_{gr} = \left(\frac{E_{gr}}{E_r} - \frac{E_{gR}}{E_R} \right) X_r.$$

The term in parentheses measures demand differences. Note that it does not depend on absolute differences in country size (Y_r and Y_R cancel out in the E ratios). In the absence of idiosyncratic demand differences, that is for $\text{IDIODEM}_{gr} = 0$, Davis and Weinstein (1996, 1999, 2003a) expect $X_{gr} = \text{SHARE}_{gr}$ and therefore expect β_1 to be approximately one.

The specification is augmented by a vector of endowments of land, capital, and labor by education category called **Factors**.

$$X_{gr} = \beta_1 \text{SHARE}_{gr} + \beta_2 \text{IDIODEM}_{gr} + \Omega_g \mathbf{Factors} + \epsilon_{gr}. \tag{16}$$

[17] Justman (1994) is an early empirical contribution emphasizing the importance of local demand in the production hosted by a particular location. His analysis is primarily centered on correlation coefficients between industry level output and local demand for the good considered among 318 US MSAs. Results point to some explanatory power of local demand for the good in industry supply location pattern but the link to theory is rather thin and no reference is actually made to HME or NEG theories. In fact, as emphasized by Davis and Weinstein (1999), simple correlations are not the appropriate device for hypothesis testing of agglomeration theories. All sensible theories yield a *positive relationship* between the location of supply and the location of demand, thus implying positive correlation coefficients. But correlation coefficients cannot distinguish between more than proportional from less than proportional relationships between supply and demand. A very high correlation can be associated with either a steep or a flat slope of the link between output and demand.

They consider this specification to nest comparative advantage (via **Factors**) and increasing returns (via IDIODEM) in the same specification. Note that the maintained assumption throughout these studies is that **Factors** determine production at the level of the industry aggregates X_r.

The estimate of β_2 is the focus of the analysis. A coefficient on IDIODEM above one provides evidence of home market effects. The Davis and Weinstein (1996, 1999, 2003a) specification might be thought of as a kind of linear approximation of the true model. Around the point of symmetry then, it may be the case that β_2 provides a rough estimate of what Helpman and Krugman (1985) show to be the (magnified) response of the share of production in r with respect to an increase of the share of demand located in r: $\mathcal{M} = (1+\phi)/(1-\phi)$. Davis and Weinstein (1996, 1999, 2003a) argue that a coefficient between zero and one implies a comparative advantage world with trade costs. Subsequent theoretical results by Feenstra et al. (2001), Trionfetti (2001), and Head et al. (2002) all cast doubt on this implication. With a fixed number of firms, asymmetric home bias in preferences, or national product differentiation, it is possible to observe production respond less than one-for-one to demand even in models of imperfect competition without comparative advantage. The implication *does* seem to run strongly in the opposite direction. Models of comparative advantage with constant returns are inconsistent with the magnifying effect of market size, i.e. $\hat{\beta}_2 > 1$ supports increasing returns models.

Coefficients of 0 and 1 on IDIODEM correspond, respectively, to a frictionless and autarkic CRS world. Davis and Weinstein (1996, 1999, 2003a) do not discuss the interpretation of negative coefficients. We are not aware of any model that generates a negative relationship between demand and the location of production. Consequently, a finding of $\hat{\beta}_2 < 0$ suggests either sampling error or mis-specification.

The Davis and Weinstein (1999, 2003a) results are considered by many to provide strong support for home market effects and NEG more generally. In an endnote, Fujita et al. (1999) write of the paper on Japanese regions published in 1999 that Davis and Weinstein "measure the importance of the home market effect and find surprisingly strong impacts." In their review of empirical evidence Baldwin et al. (2003) state that these papers find "econometric evidence that one agglomeration force—the so-called home market effect—is in operation."

Table 3: Davis and Weinstein's "pooled" HME estimates

	OECD, DW '96		Japan, DW '99		OECD, DW '03	
SHARE (β_1)	1.103	0.259	1.033	-1.744	0.96	—
	(0.002)	(0.198)	(0.007)	(0.211)	(0.01)	
IDIODEM (β_2)	1.229	0.712	1.416	0.888	1.67	1.57
	(0.005)	(0.033)	(0.025)	(0.070)	(0.05)	(0.10)
Factors	No	Yes	No	Yes	No	Yes

Table 3 presents the pooled results from the three Davis and Weinstein (1996, 1999 and 2003a) papers on home market effects (with standard errors in parentheses). The result from the 1996 and 1999 studies both indicate that controlling for factors drives the coefficient on IDIODEM below one. Thus, if one accepts the nested specification proposed by Davis and Weinstein, the pooled estimates do not support home market effects. The 2003 paper retains $\hat{\beta}_2 > 1$ after controlling for **Factors** but it should be noted that this specification omits SHARE. The problem with the pooled results, as noted by the authors, is that it gives a single answer as to the presence or absence of home market effects. Since industries differ, it would seem more attractive to let the data indicate which industries have home market effects and which ones have production patterns determined mainly by factors.

Table 4: Summary Statistics on Davis and Weinstein's Disaggregated IDIODEM estimates

Paper/Table	Mean $\hat{\beta}_2$	Median $\hat{\beta}_2$	N	% > 1	% Sig > 1	% < 0	% Sig < 0
DW (1999), Japan							
Table 6	1.63	0.45	20	45%	40%	40%	5%
DW (2003), OECD							
Table 2 (4-digit runs)	1.47	0.95	50	50%	22%	38%	4%
Table 3 (4-digit pooled)	1.20	1.02	13	54%	31%	0	0
Table 4 (3-digit runs)	4.23	0.71	24	37.5%	8.3%	37.5%	12.5%

Table 4 presents some summary statistics on the disaggregated results. The average values of $\hat{\beta}_2$, the coefficient on IDIODEM, are greater than one in all four sets of results, suggesting that manufacturing industries on average have home market effects. Means can be strongly influenced by outliers. The median coefficients on IDIODEM are less than one in three sets of regressions and marginally over one in just Table 3 of the OECD study. In summary, more than half of the industry level coefficients are less than one and a disturbingly large share are negative. One interpretation of the results is that a sizeable number of industries (11 out of 50) appear to exhibit home market effects. Another take is that the industry level estimates are just too noisy to provide solid support for the HME.

Head and Ries (2001) consider home market effects in the context of trade liberalization between Canada and the United States, phased in over ten years starting in 1989. They base their specification on Helpman and Krugman's (1985) share equation. Helpman and Krugman (1985) developed a piece-wise linear formulation in terms of shares of consumers and producers. We reproduce here the formulation expressed in the terms of our modeling from section 3.1 in a two region (noted *i* and *j*) framework. Let λ denote the share of producers in country *i* and θ to denote its share of demand. Thus, we would have $\lambda = n_i/N$ and $\theta = (\mu_i Y_i)/E$, where $E = \sum_k \mu_k Y_k$ and $N = \sum_k n_k$. For interior solutions, i.e. where n_i and n_j are both positive, the *spatial equilibrium* arises at $\lambda^* \in (0,1)$ such that $\Pi_i(\lambda^*) - \Pi_j(\lambda^*) = 0$. Using (5), the difference in profits function writes:

$$\Pi_i(\lambda^*) - \Pi_j(\lambda^*) = \frac{1}{\sigma}(c_i^{1-\sigma}\mathrm{RMP}_i - c_j^{1-\sigma}\mathrm{RMP}_j) - (F_i - F_j).$$

To solve for the spatial equilibrium, the literature typically relies upon a particular specification of the *other* sector, referred to as agriculture or "traditional." The other sector has constant returns to scale, perfect competition, zero transport costs and a unitary labor requirement technology. It is also assumed to account for a large share of total consumer expenditures. All these conditions are used to ensure that, with this sector staying active in both economies, the price of this good is equalized, therefore equal wages prevail in both economies for the manufacturing sector as well. With identical technologies this equates marginal and fixed costs across countries ($c_i = c_j = 1$ and $F_i = F_j = F$)

The difference in profits between locations i and j is then given by

$$\Pi_i(\lambda^*) - \Pi_j(\lambda^*) = \frac{E}{\sigma N}\left[\frac{\lambda(\phi-1)-\phi+\theta(\phi+1)}{\lambda(1-\phi)(1-\lambda)+\frac{\phi}{1-\phi}}\right], \tag{17}$$

In the equilibrium of the Helpman and Krugman (1985) model, producer and demand shares are therefore related by

$$\lambda^* = 1/2 + \mathcal{M}(\theta - 1/2), \tag{18}$$

where $\mathcal{M} = (1+\phi)/(1-\phi)$. This equation illustrates a number of key ideas. First we now have a magnification effect that does not vary with the share of demand; $\mathcal{M}$ depends solely on the "phi-ness" of trade:

$$d\lambda^*/d\theta = \mathcal{M} = (1+\phi)/(1-\phi) > 1.$$

The share magnification effect, $\mathcal{M}$, is strictly increasing in ϕ and therefore decreasing in transport costs. Ottaviano and Thisse refer to this as the HME magnification result. To avoid confusion with the primary magnification effect, $d\lambda^*/d\theta > 1$, we suggest calling the $d\mathcal{M}/d\phi > 0$ result "secondary magnification." Industry can agglomerate entirely in one country if the other country is small enough. In particular, the home country would be pushed out of the "modern" good and specialize in the "traditional" sector if $\theta < (1/2)(1 - 1/\mathcal{M})$.

In some respects this result is remarkably robust. The linear demand monopolistic competition model developed by Ottaviano et al. (2002) and the Brander (1981) model of segmented markets Cournot competition both deliver the same linear share equation but with different $\mathcal{M}$.[18] Head and Ries (2001) show that the Helpman and Krugman (1985) model can be contrasted with an alternative of perfect competition with national product differentiation (also known as the Armington assumption). In that model, $d\lambda^*/d\theta = (1-\phi)/(1+\phi) < 1$.

Equation (18) has a natural empirical counterpart, that is a linear share equation for a panel of industries (denoted i) and years (t), formulated as the following regression equation:

$$\lambda^*_{it} = \beta_1 + \beta_2\theta_{it} + \epsilon_{it}.$$

In the shares equation, β_2 corresponds fairly closely to the way it is used in the Davis and Weinstein (1999, 2003a) levels equation. Here, however, it can be related directly to the underlying parameters of the model, trade costs and the elasticity of substitution between varieties. Head and Ries (2001) first estimate ϕ_{it} for three-digit manufacturing industries in North America.[19] They use the median industry to obtain an idea of what $\hat{\beta}_2$ one should expect in light of the observed freeness of trade. The median ϕ in their data is 0.07 and the Helpman-Krugman case predicts a $\hat{\beta}_2 = 1.15$. Under perfect competition and national product differentiation, $\hat{\beta}_2 = 0.87$. Panel data vary along "between" (cross-industry in this case) and "within" (over time) dimensions. Head and Ries (2001) investigate each dimension separately estimating a between regression corresponding to

$$\bar{\lambda}^*_i = \beta_1 + \beta_2\bar{\theta}_i + \bar{\epsilon}_i,$$

[18] For a derivation and comparison see Head et al. (2002).

[19] Actually Head and Ries (2001) calculate $1/\phi_{ij} \geq 1$ along the lines the trade freeness computations in section 3.1 of this paper.

and a within specification given by

$$(\lambda^*_{it} - \bar{\lambda}^*_i) = \beta_1 + \beta_2(\theta_{it} - \bar{\theta}_i) + (\epsilon_{it} - \bar{\epsilon}_i).$$

The variables under bars are the six-year averages for the corresponding industry. Between and within results for the share equation are strikingly different. Using the share of shipments as the proxy for λ^*, the between estimate of β_2 is 1.13. With a standard error of 0.07, this result provides some support for the hypothesis that North American manufacturing exhibits "on-average" home market effects. The impact of demand is slightly smaller than the already small value predicted by the calculated $\hat{\phi}_{it}$. The within dimension of the data, which allows each industry to have its own fixed effect, reveals a $\hat{\beta}_2$ of 0.84.

One way to read these results is supportive of the home market effect. This reading emphasizes the between results and dismisses the within results with the argument that six years is not long enough for the magnification effect to manifest itself. Alternatively, a skeptic would critique the between specification, pointing out that fairly small correlations between omitted determinants of comparative advantage and the demand shares could deliver a spuriously high coefficient on demand. The within specification's industry-specific fixed effects might be interpreted as controls for comparative advantage. That specification has a standard error (RMSE) that is less than a fifth of the standard error of the between specification.

To resolve this impasse, Head and Ries (2001) offer a third, "tie-breaking" specification in which they relate changes in production shares to changes in trade barriers interacted with the *initial* share of demand. In both short and long-run versions of the Krugman (1980) model, higher tariffs are more helpful the lower is the home country's level of demand. In the data, however, tariff protection offers greater benefits to relatively large demand industries. That is, when Canada-US border costs declined, the low demand industries in Canada fared better than their high demand counterparts. This result is consistent with a constant returns model of the manufacturing sector in which varieties are differentiated according to the nation of production.

6.2 The impact of "home biased" demand

Trionfetti (2001) also employs a specification based on Helpman and Krugman (1985). He introduces a novel means of discriminating between the increasing returns and comparative advantage hypotheses: The impact of "home biased" demand. Trionfetti (2001) shows that, in the increasing returns, monopolistic competition framework, for a given share of demand from all sources (θ), a country with a higher share of customers that "buy domestic" will tend to have a higher share of the firms in the industry. Trionfetti (2001)'s specification can be expressed (in terms of the notation we have already been using) as

$$\lambda^*_i = \beta_1 + \beta_2\theta_i + \beta_3\mathrm{HB}_i + \epsilon_i.$$

The new variable HB_i measures the share of the "home-biased" demand in industry i residing in the home-country. The coefficient on HB_i should be positive if and only if the increasing returns monopolistic competition model applies. The magnification effect, β_2, from prior specifications need not be greater than one in the presence of increasing returns *and* home bias. The key issue is not the mere existence of home-biased demand. Any symmetric avoidance of foreign varieties is observationally equivalent to a common non-tariff barrier. In the Helpman and Krugman (1985) model such symmetric home bias lowers ϕ

and therefore $\mathcal{M}$ as well. The Trionfetti (2001) specification relies upon one country having more home biased customers than the other. If that is the case, firms would not simply follow demand to be close to the larger market. This might put them into a position of having to reach home-biased customers by exporting to them which is not profitable. Rather, firms will care about locating near customers in general but particularly near those who refuse to buy non-local goods. Asymmetric home bias is like an asymmetric tariff. The larger is the tariff on imports holding the tariff on exports constant, the more the incentive to locate in the protected market. Thus, Trionfetti (2001) is relying on the "import protection as export promotion" feature of increasing returns models.

Like Davis and Weinstein (2003a), Trionfetti (2001) estimates using cross-country variation in demand and production to identify coefficients at the industry level. His sample comprises eight European countries and he identifies home biased demand using input-output tables for those countries isolating for each industry the sources of demand for which the import share is below average or twice below average. Trionfetti (2001)'s results offer mixed support for the home market effect. The magnification effect, β_2, is never significantly greater than one and often significantly less than one. The home-bias effect β_3 is positive and significant for 7 out of 18 industries.

Brülhart and Trionfetti (2002) propose a similar test based on a different estimate of home biased demand. They proceed in two steps, first estimating a gravity trade equation where, in the spirit of Wei (1996), flows internal to countries are added and identified by a dummy. The exponential of the coefficient on this dummy gives the extent to which countries trade "excessively with themselves," which is identified with the home bias. With such an estimate for each industry of each importing country, they can construct an IDIOBIAS variable on the same model as the IDIODEM variable (capturing deviations from the median home bias in the sample) from Davis and Weinstein (1999, 2003a) papers and run the same type of regressions as Davis and Weinstein with this new variable added.[20] The hypothesis tested is very similar to the above: Models of trade characterized by increasing returns and home bias should exhibit a positive coefficient on the IDIOBIAS variable as opposed to models of comparative advantages. Their sample comes from the OECD COMTAP database and the HME regressions concerns 6 countries (Belgium, France, Germany, Italy, Netherlands, UK), 18 manufacturing industries and 4 years (1970, 1975, 1980 and 1985). Five out of the 18 industries exhibit a response to home biased demand and therefore validate the discriminating test in favor of increasing returns. Those industries (Office machinery, Motor Vehicles, Meat products, Dairy products, Paper and Printing) taken together represent around a quarter of manufacturing output of the zone.

6.3 The magnification of exports

As stated in the very beginning of this section, the original formulation of the home market effect by Krugman (1980) focused on the impact of market size on net exports of a country in IRS industries. In a two region framework, this "trade version" of the home market effect states that the region with a share of demand for the IRS good superior to one half will be a net exporter of this good. This prediction, as with the one on production shares seen in the above subsections, extends to other (though not all) popular imperfect competition models with trade costs (Head et al., 2002). Lundbäck and Torstensson (1998)

[20]Note that the 2001 paper by Trionfetti follows a share regression specification which has a closer link to theory than the Davis and Weinstein (1999, 2003a) type of regression adopted in the later paper.

implement this prediction empirically for 17 OECD countries over 49 industries (using the STAN database). Their version of the theoretical setup includes possibly different home biased preferences across countries and, as in Trionfetti (2001), this yields an additional HME prediction. In this setup, a country will produce disproportionately and be a net exporter of the goods for which the home bias of its customers is most pronounced compared to its trade partners. The empirical specification links the net trade in industry/country combinations to a measure of "demand bias" (intended to capture how demand in a country deviates from the sample's average demand for the considered industry), and a measure of home bias. This last variable is given by the residuals from a first stage regression of domestic producers market share of domestic demand on their market share of world demand for each industry. The regressions also include three more variables, two for factor endowments and one for scale economies. Results are again mixed for the HME: The demand bias variable is positive and significant in 6 out of 17 countries, significantly negative in 3 countries, and insignificant for the 8 remaining countries. The variable intended as a proxy for home bias asymmetries in preferences offers much greater support, being very significantly positive in all countries.

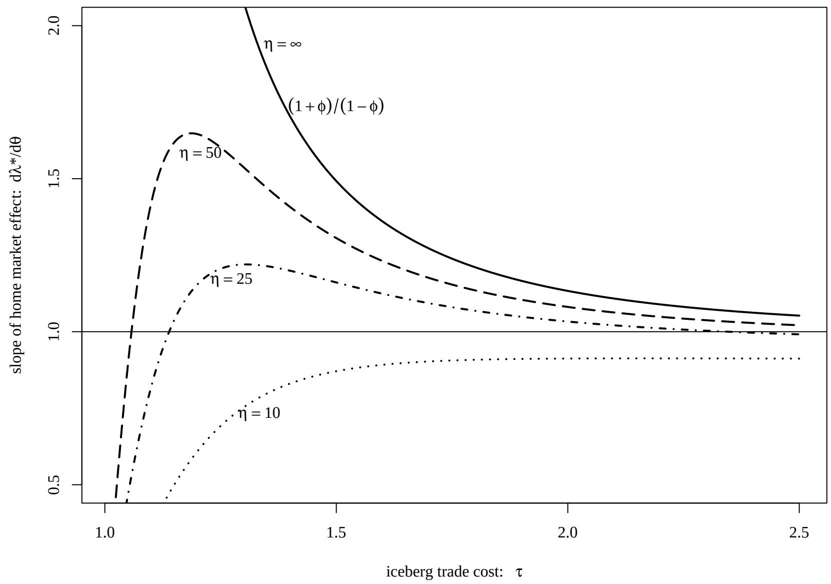
Three recent papers—Feenstra, et al. (2001), Weder (forthcoming), and Hanson and Xiang (2002)—propose tests for the HME using *bilateral* export patterns. Feenstra et al. (2001) estimate gravity equations and interpret a larger coefficient on exporter GNP than importer GNP as evidence of the home market effect. They find this coefficient pattern in differentiated products but not for homogeneous products. Weder (forthcoming) finds that the ratio of UK to US exports to third markets are increasing in the relative size of the UK market. It is not clear, however, whether this result violates a model of national product differentiation and constant returns if the latter allows larger countries to produce (and export) a larger number of varieties. Hanson and Xiang (2002) adopt a different definition of home market effects from what has been standard in the theory and empirical literature. This makes their results difficult to compare with those of prior studies. One important finding of this paper is that demand measures based only on national demand give quite different results from summations of proximate demand subject to a distance discount. The latter approach corresponds to the concept of nominal market potential defined earlier in this chapter and also utilized in Davis and Weinstein (2003a).

6.4 The robustness of the relationship

We have summarized the methods and results of ten papers that test for the home market effects (HMEs) implied by increasing returns models using the relationship between production, exports, and home demand. *The evidence on HMEs accumulated by these papers is highly mixed.* One can see some support for HMEs in some industries in some specifications. However reverse HMEs (coefficients on demand of less than one or on home biased demand of less than zero) are more frequent. These overall unsupportive results should be contrasted with the more robust results arising from wage equations seen in section 4. The empirical success of wage equations and the less successful attempts to validate home market effects in production regressions are entirely consistent with each other. They can be interpreted in a positive way as a sign that market access mechanisms of NEG are empirically important, but generally take the form of higher factor incomes in large demand areas rather than magnified production shares of IRS industries.

Despite its robustness to alternate market structures and demand formulations, the home market effect turns out to be quite fragile in one key respect. The theoretical litera-

Figure 2: Home market effects with imperfectly elastic labor supply.



ture following Helpman and Krugman (1985) makes assumptions that lead to a perfectly elastic supply of labor to the increasing returns sector. This is necessary to obtain the linear share equation. In contrast, the literature on income-access effects may be seen as holding quantity constant and letting wages adjust. The more general case where market access influences the number of firms in a location, and thereby net exports, as well as the prevailing wage is much more difficult to carry out. However, Fujita et al. (1999) provide an illuminating investigation, that when pushed a little bit further, yields a result that can help to make sense of the results of the two empirical literatures.

Start from a symmetric equilibrium. Then totally differentiate, linearize and make substitutions. The result, as shown in Fujita et al. (1999), is

$$\left[\sigma\frac{1+\phi}{1-\phi} + (1-\sigma)\frac{1-\phi}{1+\phi}\right]\frac{dw}{w} + \left[\frac{1+\phi}{1-\phi}\right]\frac{dL}{L} = \frac{dY}{Y}. \tag{19}$$

The supply of labour to the modern sector, L, has a wage elasticity of

$$\eta \equiv \frac{dL/L}{dw/w}.$$

Since, around the symmetry point, firms have all the same employment, we can write $dn/n = dL/L$. Finally around the point of symmetry, $d\lambda^*/d\theta = \frac{dn/n}{dY/Y}$. Thus we can combine all these substitutions and re-express equation (19) as

$$d\lambda^*/d\theta = \mathcal{M}/(1 + (1 + (\mathcal{M}^2 - 1)\sigma)/\eta). \tag{20}$$

As $\eta \to \infty$ we obtain $d\lambda^*/d\theta = \mathcal{M} = (1+\phi)/(1-\phi) > 1$. However, smaller elasticities of labour supply lead to bigger wages in the large market and this dampens or even destroys the home market effect. This can be seen by letting the trade costs get very large. When $\phi \to 0$, we obtain $d\lambda^*/d\theta = 1/(1+1/\eta) < 1$; as long as η is of finite (positive) value, there always exists a level of trade costs above which we obtain reverse HMEs. We illustrate in Figure 2, where $d\lambda^*/d\theta$ is graphed against τ for different values of the labour supply elasticity. It can be seen in this Figure that large trade impediments associated with low labour supply elasticities will yield reverse home market effects (slopes inferior to one). Furthermore the monotonically decreasing relationship between the HME and trade barriers (that we refer to as "secondary magnification") is only valid in the limit when $\eta \to \infty$.

We conclude that market access is an important determinant of both the locations of producers and their factor returns. However, the prediction of a more than one-for-one response of production to demand only arises under extreme versions of more general models. Since less than unitary responses are consistent with constant returns models, the HME test is not ideal for discriminating between increasing returns and traditional models. While consistently larger than one estimates of the HME would have militated in favor of an increasing returns model, the highly mixed pattern of estimated coefficients neither supports nor falsifies the new trade foundations of NEG.

7 Trade-induced Agglomeration

The work reviewed in sections 4, 5, and 6 all consider the impact of the geographic distribution of demand as an *explanatory* variable. While this empirical approach is useful and justifiable in certain contexts, it is also problematic. The key idea of NEG is that *the location of demand is jointly determined with the location of production.* In particular, the opportunity to export at low cost to immobile sources of demand allows all the mobile consumers and producers to congregate in the so-called manufacturing core. The predicted relationship between the free-ness of trade and agglomeration motivated the title of this chapter. Indeed, a large part of European academic interest in agglomeration stems from the question of whether a more united European market will lead to more spatially concentrated industry.

We begin this section with a review of work that has examined the relationship between agglomeration and trade costs, as well as the related issues of plant-level increasing returns and demand mobility. Existing work of this type is loosely related to the underlying theory. Later in this section, we consider steps that might be taken to treat the theory more seriously and review two papers that move in this direction.

7.1 Concentration regressions

The papers we present in this section may be thought of as reduced-form approaches to the hypotheses expressed verbally above. They construct concentration indexes to measure the strength of agglomeration forces over different industries and time periods, and then check whether those patterns are broadly consistent with predictions of NEG models or with other plausible stories. All papers reviewed here can be grouped as doing the following type of regression:

$$\mathrm{CONC}_s = a + b\mathrm{TRCOSTS}_s + c\mathrm{IRS}_s + d\mathrm{LINKAGES}_s + \ldots + e_s.$$

The dependent variable, CONC_s is the particular geographic concentration index of industry s. $\mathrm{TRCOSTS}_s$ and IRS_s are proxies for trade costs (τ in the model) and the degree of increasing returns ($1/\sigma$ in the model) respectively. LINKAGES$_s$ measures the industry's reliance on intermediate inputs sometimes distinguishing between those that are mobile versus those that are tied to immobile natural resources. A variety of other variables (represented above as ...) can be added to this type of regressions, some intended alternative explanations for agglomeration such as endowments or technological spillovers.

7.1.1 Concentration indexes of agglomeration

Measuring spatial concentration of activity is a far less trivial exercise that might seem at first sight. Duranton and Overman (2002) list five properties we would expect from a meaningful concentration index. Combes and Overman add four additional *desiderata* in their chapter in this Handbook. Most indexes are constructed by dividing up geographic space into regions and comparing the share of activity (measured by number of firms, production, or employment) in each region with a benchmark. Two problems deserve special attention. The first is that an industry with a small number of establishments may appear to be concentrated purely by chance. This so-called lumpiness problem makes it problematic to compare industries with commonly used measures such as the locational Gini index. Ellison and Glaeser's (1997) solution to the lumpiness problem has led to wide adoption of their index, hereafter referred to as the EG index.

A second important issue that still awaits a satisfactory solution is the dependence of concentration indexes on the level and method of geographical disaggregation. When geographic units lack economic relevance, actual clusters of industries that take place across borders of those units are artificially separated. Furthermore, standard concentration indexes fail to account for the spatial proximity of those units. A concrete example illustrates these problems. In 1995, 76 establishments produced watches in France, employing 5406 people. The first département for this industry hosted 45 of those firms, accounting for 64% of national employment of the industry (against approximately 1% of France's GDP and area). This extreme concentration pattern would be partly captured by Gini or EG indexes, and it is indeed, as this industry appears to have among the highest EG index in Table 1 of Maurel and Sédillot (1999), who use very comparable data. One thing those indexes miss is that the considered départment is Doubs, which is contiguous to Switzerland. It is therefore quite likely that the real agglomeration in the watch industry spills over the political border, a feature this type of index cannot account for. In addition, the two following départements in terms of the number of firms for the watch industry are Haute-Savoie and Jura (7.1% and 3.3% of industry's employment respectively), which are also contiguous to Switzerland and very close to Doubs. The EG index cannot control for this additional dimension of agglomeration, as its computation would be exactly the same if Haute-Savoie and Jura were located hundreds of kilometers apart and away from Doubs.

Duranton and Overman (2002) construct a "continuous-space concentration index" that alleviates the problems associated with standard indexes. Their index uses the actual location of firms at the most detailed level available and compares bilateral distances between all pairs of firms to a counterfactual distribution emerging from a random relocation of all firms. There are two practical problems that will limit adoption of this method. First, only a few data sets provide the precise address of each producing establishment. Second, the use of simulations to construct the benchmark raises issues of replicability.

Is the picture of relative spatial agglomeration of industries actually different when using different indexes? Duranton and Overman (2002) calculate (in addition to their own index) EG indexes using 120 postcodes in the UK as geographical units. The most interesting result of the comparison for our purpose is that the two measures of agglomeration are almost uncorrelated when considering rankings of industries. The correlation between the two rankings improves significantly when only large plants are considered, but still the Spearman rank correlation between the two indexes is only equal to 0.4. This means that the results given by a discrete space index and a continuous space index might be very different. This should draw our attention to the fact that the spatial scale is very important in results using the EG index. In particular, taking a level of location units that is "too fine" can lead to an underestimation of agglomeration levels because it artificially separates clusters that sprawl over the border between units. Even worse, the ranking of industries can be radically changed by the choice of units, which endangers any attempt to *explain different concentration levels across industries*. This important problem is also apparent in Rosenthal and Strange (2001) who calculate EG indexes at the state, county and zip code level for 4-Digit industries in the United States. The mean EG index goes from 0.0485 at the state level to 0.0101 at the Zip code level. The correlation between the two being only 0.58.[21] Rosenthal and Strange (2001) interpret this as a possible change in determinants across geographical levels, but the inadequacy of the EG index to deal properly with spatial aggregation problems is another plausible explanation.

7.1.2 Results of concentration regressions

All users of concentration indexes acknowledge that multiple phenomena (endowments, spillovers, and NEG-type linkages for instance)—usually considered separately in theory—probably act simultaneously in a great number of industries. High values of indexes *per se* are therefore not very informative on the prevalence of NEG mechanisms in the economy. What needs to be done is to disentangle the share of each possible explanation in the observed concentration index. We now consider papers that have related concentration indexes to proxies of trade costs, increasing returns, and vertical linkages while controlling for other possible sources of agglomeration.

Since trade costs have tended to decline over time due to improvements in transport technology, and—since the end of WWII—due to reductions in trade barriers, a crude strategy is to measure how spatial concentration has changed over time. Kim (1995) examines the period from 1860 to 1987. Concentration, measured by a locational Gini index, falls until 1900, then rises to a peak around 1927 and then declines steadily until 1987, reaching a level approximately a third lower than in 1860. This non-monotonic evolution of concentration presents a puzzle for the basic Krugman (1991a) model. Examining European data from 1972 to 1996, Brülhart (2001) finds that the average employment Gini index grows by about 18%. Interestingly, there is no evidence that the growth rate accelerates in the sub-period following the signing the Single Market Programme. In fact, contrary to the fears of increased agglomeration with trade liberalization, the average growth rate is about one third *lower* after 1986. These results are interesting but their interpretation as evidence for or against NEG relies upon the untenable assumption that trade costs are the only variable changing over time.

[21] Maurel and Sédillot (1999) also found that the average EG index rises with the level of spatial aggregation (from 0.06 for the 95 French départements to 0.09 for the 22 French régions. Ellison and Glaeser (1997) found their index to have a median value of 0.005 at the US county level and 0.023 at the state level.

A more direct approach is to relate industry-level spatial concentration to industry-level proxies for trade costs. Brülhart (2001) relates variation in the Gini indexes across industries and over time the Buigues et al. (1990) classification of industries as characterized by high, medium or low NTBs. Seemingly contradictory with the basic predictions of NEG models, concentration is positively related to NTB level. Haaland et al. (1999) find no effect for the same NTB measure. Given the crude nature (one year, low informativeness about the level of protection) of the Buigues et al. (1990) measure, we should not infer much from these inconsistent results.

Ades and Glaeser (1995) provide more persuasive results using a substantially different methodology. Their data comprises a cross-section of 85 countries. Instead of concentration indexes, their dependent variable is the log of the size of the country's largest city. As they control for the population of the rest of the country, this is like measuring the share of the population in the main city. Three variables capture transport costs within each country. The first is area. Holding population constant, larger areas amount to greater average distances between buyers and sellers, and therefore larger transport costs (lower ϕ). The second and third variables measure transportation infrastructure, using, respectively expenditures on transport and communication and road density. All three variables point to a *positive* relation between trade costs and agglomeration (in the main city). This contradicts the prediction of Krugman (1991a) but is consistent with the Helpman (1998) model that reverses the relationship between agglomeration and trade costs.

Kim (1995) is one of the first papers to investigate empirically the relative explanatory power of alternative theoretical frameworks in a panel data setting. Kim (1995) regresses Gini indexes calculated in 1880, 1914, 1947, 1967, and 1987 for twenty 2-digit industries on a proxy for internal scale economies (production workers per plant), a resource intensity variable (cost of raw materials divided by value added), and two sets of industry and year fixed effects. The significant positive influence of scale economies offers some support for NEG.

Using national data from Europe, several papers have attempted to corroborate Kim's (1995) finding of a positive relationship between spatial concentration and measures of scale economies. Amiti (1999) follows Kim (1995) in using firm size as the proxy for increasing returns and in controlling for industry fixed effects. Her work suggests that European industries also exhibit a positive correlation between changes in increasing returns and changes in spatial concentration. Brülhart and Torstensson (1996) find a 0.69 rank correlation between locational Gini indexes and returns to scale estimates of Pratten (1988) in a cross-section of 18 industries. They also find a 0.63 correlation between the degree of increasing returns and a "centre-periphery bias" variable that the authors constructed by relating each industry's geographic distribution of employment to the corresponding distribution of market potential. Thus it appears that increasing returns industries are both spatially concentrated and centrally located. Moving beyond simple correlations and adding a 25 year temporal dimension to the concentration data, Brülhart (2001) finds however no significant effect for the Pratten measure of increasing returns. Haaland et al. (1999) find that their scale economies proxy has a consistently negative impact on concentration.

Trade costs and increasing returns are the two key parameters determining agglomeration in the Krugman (1991a) version of NEG. The Venables (1996) version focuses on input-output linkages between industries. In the Puga (1999) model, this corresponds to a parameter we call α in equation (21). Ellison and Glaeser (1997) establish a relationship between a variant of their index capturing co-agglomeration and the input-output linkages between the considered industry pairs. They construct two lists of 100 industries pairs, one consisting of the 100 downstream industries that receive the largest value of inputs per dollar value of output from a single upstream industry. The second list consists of the 100 upstream industries selling the largest portion of their output to a single industry. Out of the first (downstream) list of industry pairs, 77 industry pairs show a tendency to co-agglomerate, whereas the figure is 68 for the second list. Rosenthal and Strange (2001) use manufactured inputs per dollar of shipments as a proxy for the strength of input-output linkages in the industry (what they call "input sharing"). They find weak empirical evidence of such linkages, with statistical significance only at the state level (the significance is slightly improved when considering "young firms" that have less than 5 years of existence). Using a similar measure, Amiti (1999) finds significant positive effects of linkages on spatial concentration in Europe. In Haaland et al.'s (1999) paper, input-output linkages always have a small and barely significant coefficient.

What is the take-away from the concentration regressions relating spatial concentration to proxies for the key NEG parameters? First, there is little persuasive evidence that the degree of increasing returns raises spatial concentration. Whether the absence of a statistical relationship reflects poor proxies for increasing returns or inadequate concentration indexes or the absence of an economic relationship is uncertain. Second, vertical linkages do seem to have a fairly robust relationship with concentration. We would hope that future work would follow the approach of Ellison and Glaeser (1997) in exploiting the precise nature of input-output linkages, rather than just summing over all intermediate input purchases. Last, trade costs have a highly mixed impact on geographic concentration. As will be discussed in the following section, this is not inconsistent with some versions of NEG theory. Greater concern over functional form is warranted here, as well as better measures of trade costs. Somewhat surprisingly to us, the most convincing evidence—provided by Ades and Glaeser (1995)—militates in favor of the Helpman (1998) model.

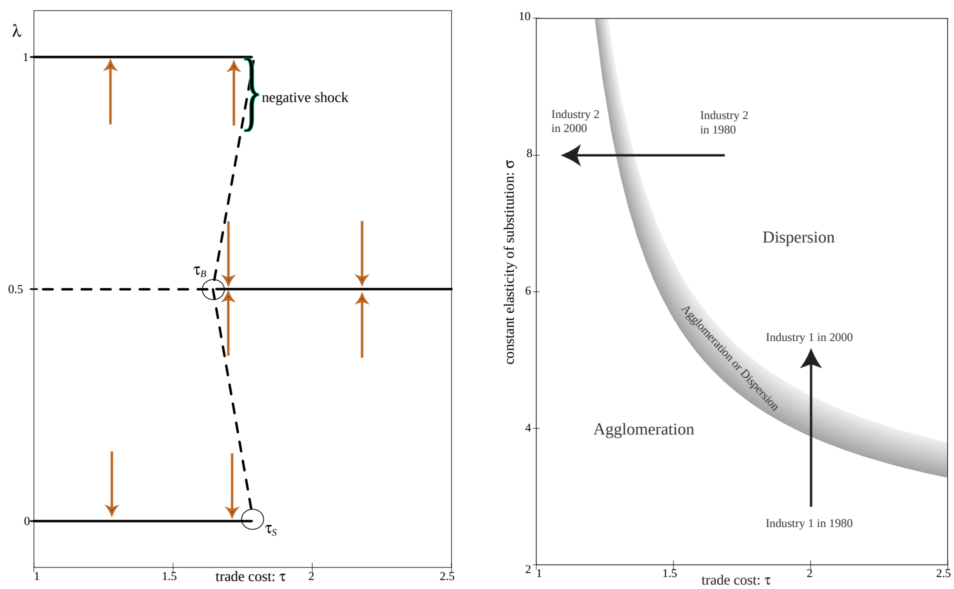
7.2 Taking NEG theory seriously

The models described in the previous subsection do not take NEG theory "too seriously." Brülhart (2001), for example, explains his goal is to "..look for stylized facts that might or might not be consistent with theoretical predictions rather than for rigorous tests of competing models." We think this approach is quite understandable for first-generation empirical assessments of NEG theories. Nevertheless, it seems worthwhile to take a closer look at the predictions of NEG.

7.2.1 Concentration predictions of NEG models

Consider first the simplest NEG model, namely the Krugman (1991a) model and its conditional concentration prediction often illustrated with the "tomahawk" subcritical bifurcation diagram. We present an example of this diagram in panel (a) of Figure 3 which replicates the version presented as Figure 5.4 in Fujita et al. (1999) (drawn for $\mu = 0.4$ and $\sigma = 5$). Three equilibrium configurations for the share of firms λ are associated with a gradual fall in trade costs from an initially high level: Stable dispersion only, followed by a multiple equilibria range where both dispersion and agglomeration are possible outcomes, and last stable agglomeration only, for high levels of trade integration. Location adjustment dynamics towards stable equilibria are indicated by the arrows.

Figure 3: Equilibrium configurations in the Krugman (1991a) model.



(a) Bifurcation for $\mu = 0.4$ and $\sigma = 5$ (b) Bifurcation for $\mu = 0.4$ and varying σ

Inspection of this diagram reveals that there are indeed predictions on concentration variation, but those are at the same time too simple to be verified and not simple enough to be easily implemented empirically. The basic prediction is that for levels of trade costs above the sustain point (τ_S), only dispersion can be an equilibrium, while under the break point (τ_B), only full agglomeration can be sustained as a stable equilibrium. While this simplest prediction of an abrupt and immediate change from complete symmetry to extreme agglomeration is clearly too stark to be verified, the pattern suggests a positive relationship between trade integration and concentration, which is the rationale behind much of the empirical work reviewed in section 7.1.2. Note that this prediction can in principle be subjected to empirical test using time series (focused on the evolution of concentration indexes within each industry) or cross section data (focused on assessing which industries are correctly predicted to be dispersed of agglomerated).

There are however important issues in the implementation of such tests. Indeed, a key concern is that this model predicts *nothing like* a simple linear relationship between concentration and trade integration. In fact, it is immediately apparent from the diagram that for the vast majority of admissible parameter values, “nothing will happen” in terms of concentration after a small fall in trade costs τ. It is only somewhere between τ_S and τ_B that a considered industry will discontinuously jump from the symmetric to the agglomerated equilibrium. This prediction, sometimes referred to as catastrophic agglomeration, is summarized in the following quote:

> Catastrophe is the most celebrated hallmark of the CP model—probably because it is so unexpected. Specifically, starting from a symmetric outcome and very high trade costs, marginal increases in the level of trade free-ness ϕ has no impact on the location of industry until a critical level of ϕ is reached. Even a tiny increase in ϕ beyond this point causes a catastrophic agglomeration of

industry in the sense that the only stable outcome is that of full agglomeration. [Baldwin et al. (2003)]

A linear regression is therefore severely *mis-specified* as even if there existed an industry where the simplest NEG model applied perfectly, the linear estimation would presumably yield a coefficient not statistically different from zero, which would likely be misinterpreted as a rejection of NEG. The expectations that should be derived from this theory are truly more complicated than a simple linear relationship and heavily dependent on the values of the parameters. The vast empirical literature (covered above and in much more depth in Combes and Overman chapter of this Handbook) trying to find evidence of NEG through linear relationships with concentration indexes as the dependent variable is therefore weakly grounded in theory (while being often both insightful and instructive). Our belief is that time has come for this type of research to now re-consider their methodological strategy and in particular think about ways to improve the specifications with a closer concern about what the models actually predict.

There are however several possible empirical implementations of the simple NEG model more consistent with theory. The bifurcation diagram can again be used here, in a version accounting for the likely variance across industries in *both* τ and σ dimensions. Panel (b) of Figure 3 uses the same sustain and break point equations to divide the σ-τ parameter space into ranges where (i) full *agglomeration* in one location or the other are the only stable equilibria, (ii) symmetric *dispersion* is the only stable equilibrium, and (iii) the shaded area in which agglomerated *and* dispersed equilibria are stable. This representation can be seen as a graphical version of Table 5.1 in Fujita et al. (1999).

Several empirical implementations seem possible when considering the (b) panel of Figure 3. Industry-level estimates of τ and σ can be used to give coordinates for each industry to be placed in the Figure for a given set of trade partners. Measuring these parameters is not a trivial task, of course. Ideally, τ should capture a variety of sources of trade costs including transport costs, tariffs, non-tariff barriers, communication costs. Those costs are bilateral in nature (depending for instance on bilateral distance between the trading partners) and industry-specific (transport costs of concrete and semiconductor chips differ drastically).[22]

Measuring σ raises difficult issues because this parameter fulfills multiple roles in the Dixit-Stiglitz model. It is not only a differentiation parameter, but also the price elasticity of demand, an inverse index of scale economies, and an inverse measure of equilibrium markups. Using a gravity equation, coefficients on the origin country price term or on bilateral tariffs or freight can then be used to infer σ.[23] Alternatively, one can exploit the fact that the Lerner index in the Dixit-Stiglitz model is given by $(p - c)/p = 1/\sigma$, with p denoting price and c marginal cost. Consequently, multiplying by the output of symmetric firms in the industry, one can calculate σ at the industry level as shipments/(shipments-variable costs).[24]

With estimates of τ and σ in hand, the next step is to assess whether a cross section of industry-level concentration indexes match the predictions of the model (for instance,

[22]Hummels (1999) and Limao and Venables (2001) are two papers that grapple with the issue of measuring international transport costs correctly.

[23]For details on several variants of this method, see Hummels (1999), Head and Mayer (2000), Head and Ries (2001), Lai and Trefler (2002), and Erkel-Rousse and Mirza (2002).

[24]In other market structures, such as Ottaviano et al. (2002), this simple relationship between markups and the substitution parameter does not exist.

that industry 1 was more agglomerated than industry 2 in 1980 as predicted by the model illustrated in Figure 3). One may also look *within* industries to verify whether changes in parameters over time delivered the predicted change in agglomeration patterns. In the examples considered in Figure 3, concentration indexes can be used to assess whether industry 1 became more dispersed over the period and industry 2 more agglomerated.

7.2.2 The diagonal Puga model

The Krugman (1991a) model is however probably too restrictive to be used directly in empirical work following the lines just mentioned. Indeed, a particularly important feature of this model is that it predicts that high trade costs will generate dispersion and low trade costs agglomeration. The problem with this is that the Krugman (1991a) model continues to predict full agglomeration even as transport costs become tiny. This is because the "centrifugal" forces that would promote dispersion decline with trade costs at an even more rapid rate than the "centripetal" forces that promote agglomeration. With *any other congestion force unrelated to trade costs*, the equilibrium pattern of location will return to dispersion for some (low) trade costs threshold where all trade-related forces become so weak that they must be dominated by the congestion force.[25] These additional congestion forces cause *dispersion* to have a U-shaped relationship with trade costs. Reciprocally, spatial *concentration* has what Ottaviano and Thisse in this Handbook describe as a *bell-shaped* relationship with trade costs.[26]

Linear regressions of concentration indexes on trade costs remain inappropriate in the Puga (1999) model. The good point of the bell shape prediction in terms of empirical testing and specification is that there is at least a continuous relationship between trade costs and concentration over some range of the parameters. Unfortunately, this relationship is *not linear and worse, not monotonic*.

The Puga (1999) version of the NEG model removes the exotic dynamics of the Krugman (1991a) model while remaining analytically tractable. It is sufficiently detailed and complete to nest the Krugman (1991a) and Krugman and Venables (1995) approaches. To extend the Puga (1999) model to accommodate multiple increasing returns industries, we do have to impose a strong assumption about the input-output structure: Firms in an industry source all their intermediate inputs from their own industry. This implies a diagonal input-output (I-O) structure. We also must assume that industry expenditure shares are fixed by preferences (i.e. the upper level utility is Cobb-Douglas). Those assumptions are restrictive, being more acceptable as approximations only for highly aggregated industries. This suggests the need for more detailed modelling of actual input-output linkages and demand substitution patterns between industries (as detailed below, this is an important contribution of Forslid et al., 2002 to provide predictions of a "full" model with I-O linkages between 14 industries calibrated on real data).

[25]Examples of congestion forces giving rise to the bell shape include Helpman (1998), where the housing sector makes agglomeration unsustainable for very low trade costs, or comparative advantage as in Forslid and Wooton (forthcoming). The Ottaviano and Thisse chapter also analyzes mechanisms yielding the bell. The bell-shaped prediction can be obtained through the inclusion in the NEG model of different realistic features such as impediments in inter-regional workers' mobility (Krugman and Venables, 1995, enriched considerably in Puga, 1999) or heterogeneity in the tastes of workers which translates into their migration patterns (Tabuchi and Thisse, 2002).

[26]Describing the bell as an "inverted U" is both awkward and potentially confusing and should therefore be avoided.

The parameters of interest include τ (trade costs), σ (the elasticity of substitution between varieties), μ (the share of consumer expenditure on manufactured goods), α (the share of costs constituted by purchases of intermediate goods from one's own industry, which is zero in the Krugman (1991a) model), and η, (the elasticity of a region's labor supply to the manufacturing sector with respect to local agricultural wages, which Krugman (1991a) and Krugman and Venables (1995) assume to be infinite).

We therefore implement the analysis presented in Puga (1999), where he identifies the threshold transaction costs between which dispersion is unstable and we should therefore expect to observe agglomeration. This analysis is intended to illustrate what this unexplored path of empirical implementation of NEG theory could be. Let us follow his notation and define $\underline{\phi}_s$ and $\overline{\phi}_s$ as the lower and upper break points for sector s. Puga (1999) shows that these break points are solutions to the following quadratic equation in ϕ:

$$\begin{aligned} &[\sigma(1+\alpha)-1][(1+\alpha)(1+\eta)+(1-\alpha)\mu)]\phi^2 \\ &-2\{[\sigma(1+\alpha^2)-1](1+\eta)-\sigma(1-\alpha)[2(\sigma-1)-\mu\alpha]\}\phi \\ &+(1-\alpha)[\sigma(1-\alpha)-1](\eta+1-\mu) = 0. \end{aligned} \tag{21}$$

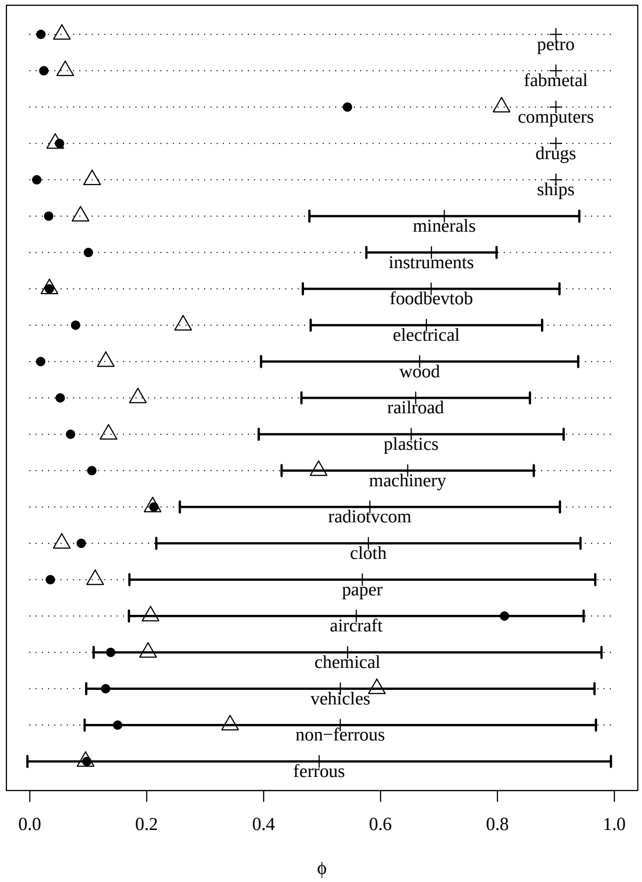
The roots of this equation give the degrees of trade freeness $\underline{\phi}_s$ above which complete symmetry is unstable and activity starts to agglomerate, and $\overline{\phi}_s$ for which trade is so easy that the process of re-dispersion is completed and the equilibrium reverts to perfect symmetry. Although the analytical expressions of solutions to equation (21) are not easy to manipulate, they can be calculated very easily for each sector s when one plugs in values of parameters of main interest, σ_s, μ_s and α_s. This gives for each industry the range defined by $[\underline{\phi}_s, \overline{\phi}_s]$ over which agglomeration is expected and that we can compare with $\hat{\phi}_s$ calculated from observed trade flows of country pairs representative of ongoing regional integration (namely USA-Canada and Germany-France) following equation (4) (see the appendix for a complete description of sources of parameters and data). The results for all industries are represented in Figure 4.

Horizontal solid lines (sorted by midpoint) show the range, for each industry, over which symmetric equilibria are unstable in Puga (1999), and therefore agglomeration is expected. Industries without solid lines had undefined break points (no real roots existed for their values of parameters).[27] Dots (•) for France-Germany and triangles (△) for Canada-US show estimated $\hat{\phi}$ using 1995 trade and production data gathered from World Bank and OECD sources.

Therefore, we can first identify, with the position of the $\hat{\phi}$ symbols, the industries that are predicted to be in a symmetric equilibrium and the ones that are predicted to be in an agglomerated equilibrium for the two integrating regions. Furthermore, among those industries predicted to be dispersed, we can in theory draw a clear distinction between the industries for which the trade integration level is so low that they did not even enter the agglomeration zone yet, and those for which the integration process is so advanced that they are already out of the agglomeration zone. Note that this first very rough empirical implementation of the Puga (1999) model predicts most of the industries to be near the lower end of the agglomeration range, where more trade integration will yield more agglomeration. Those break points calculations can be quite sensitive to chosen parameters values, which pleads for cautious interpretations of the results. More experimentation

[27] Inspection reveals that, for those five industries, equation (21) is positive for all values of $0 < \phi < 1$. This corresponds to local stability of the *symmetric equilibrium* for all admissible values of ϕ.

Figure 4: "Where in the bell are we?"



with different sets and sources of key parameters is in this respect clearly needed to check the robustness of those predictions.

Can these results be related to observed agglomeration of the considered industries in order to check if theoretical predictions arising from Figure 4 match with real data? Returning back to the bell shape curve of Puga's (1999) Figure 6, we can first relate a measure of concentration of the industry to its position on the bell curve. This is however maybe taking the theory "too seriously." In the actual data, it is for instance highly unlikely that we would observe some industries to be totally dispersed and some totally agglomerated. A perhaps more sensible test of those predictions would be to try to fit a bell-shaped function to the data. Thus we might relate a geographic concentration index of industry s, CONC_s, to a bell-shaped function, $f(\cdot)$, of the gap between actual free-ness of trade and the midpoint of the two breakpoints:

$$\text{CONC}_s = f(\hat{\phi}_s - [\underline{\phi}_s + \overline{\phi}_s]/2) + \epsilon_s,$$

where $f(\cdot)$ peaks at $f(0)$. This equation could also be estimated using time-series data instead of a cross-section of industries.

7.2.3 Simulations of higher dimension models

NEG theory mostly deals with the case of two locations, two industries, and two factors. This simplifies models in which it is already difficult to obtain analytical results. It is therefore quite difficult to envision what the theoretical predictions would be in a framework of a higher order dimension. Nevertheless, we have to confront higher dimensional data in almost any sensible empirical verification of the theory. This is especially important as it is well known from traditional trade theory and new trade theory that $2 \times 2 \times 2$ model predictions often do not have simple counterparts when expanding the dimensions of the model. Forslid et al. (2002) present a simulation exercise where a large scale computable general equilibrium (CGE) model is calibrated on EU data using various 1992 external sources for parameters. The aim is to obtain "numeric intuition" of higher order properties of those models.

Also important for empirical implementation is to depart from the assumption that countries "are all alike." One of the important goals of the NEG literature was to show that agglomeration could arise endogenously, starting from a situation of perfectly symmetric countries or regions. This mirrored the effort of new trade theories a decade before to design models able to generate (intra-industry) trade in a world of seemingly identical countries in terms of endowments and technology. In empirical work, natural advantages have to be brought back in the analysis, because in the real world countries differ in their initial conditions in ways that can be expected to alter the final outcome.

A quite important point is that traditional comparative advantage constant returns with perfect competition models also give rise to predictions of increasing agglomeration accompanying trade liberalization. The increased specialization of countries in the production of the goods for which they have a comparative advantage will indeed translate into increased agglomeration of industries across space. However this relationship is predicted to be monotonic as opposed to the NEG models of the increasing returns with imperfect competition type outlined above, where the bell shape emerges. Forslid et al. (2002) provide a framework encompassing both input-output linkages in a Venables (1996) type model and comparative advantage patterns in order to assess which industries are predicted to exhibit the bell shape and which industries are predicted to agglomerate monotonically with trade integration. There are 14 industries linked with region-specific input-output tables (the regions are groupings of 17 West European countries into 4 European regions called Central, North, South and West). Of those 14 industries, ten are assumed to have the Dixit-Stiglitz-Krugman usual market structure, two are traded perfectly competitive sectors without trade costs and with decreasing returns to scale, and two are nontradable monopolistic competition services sectors. Capital, unskilled labour and skilled labour—the three primary factors of production—are assumed to be internationally immobile. Data for calculating parameters mostly comes from Eurostat, GTAP and NBER world trade flows databases. The parameters of primary interest, trade costs and elasticities of substitution, respectively come from GTAP and from scale elasticities calculation based on Pratten (1988).

The main result of interest for our purpose lies in Forslid et al. (2002)'s Figure 3, which depicts the path of the agglomeration of each industry (as measured by the standard deviation of the distribution of the share of production of the industry in each region) with respect to trade costs. Metals, chemicals, transport equipment and machinery all exhibit a distinct bell shape in the agglomeration index with decreasing trade costs, while the other increasing returns industries in their model (and specially so textile, leather, and food products) show a monotonic increase in agglomeration. The bell-shaped industries show, as expected, the highest degree of increasing returns to scale and relatively high share of own output in their intermediate goods consumption. We can also note from this paper that those industries are predicted to be at the start of the agglomeration process; that is, in the beginning of the range of trade costs for which agglomeration increases with trade liberalization. Note however that *the amount of predicted changes in the concentration patterns is much lower in the bell-shaped industries than in the others that seem to follow more closely the predictions from comparative advantage theories*. Thus, while the theoretical interest is primarily focused on those industries, it might be that the major part of the action concerning spatial distribution of activities in Europe will take place in more traditional industries exhibiting considerably larger concentration trends. This pattern is also observed in some of the papers investigating concentration patterns in a more descriptive way (like the ones covered in section 7.1.2).

Combes and Lafourcade (2001) also propose simulations based on a model featuring input-output linkages between imperfectly competitive industries operating in a multiple location space. Their modeling strategy however differs notably from the usual approach as they use a Cournot, segmented markets, homogenous goods model as their theoretical framework. The paper proceeds in two steps: A structural estimation of the model is followed by a simulation of transport costs reduction effects. The estimated equation relates employment per firm in each of the 341 French regions considered to two terms capturing final demand and input-output linkages. The econometric analysis involves estimation for each industry of the sole unobserved element in the model: Industry-specific transport costs (a parameter for each industry multiplying an observed average transport costs). If this parameter is insignificant, the industry is estimated to be unaffected by transport costs and the linkages at the heart of agglomeration in this model are irrelevant. Significant and positive parameters are interpreted as empirical validation of the model. The results exhibit 47 significantly positive coefficients out of 64 industries in the full version of the model.

The second step use the transport cost sensitivity estimates to simulate the effects of a uniform transport cost decrease in France (up to 30%). For computational reasons, simulations have to be run for the short-run version of the model (keeping the number of plants in each location-industry at its actual level in 1993). The change in production patterns and extent of agglomeration therefore entirely arises from changes in prices and individual production by firms (both of which would be unchanged in a Dixit-Stiglitz-Krugman framework).

The simulation results show a fall in production concentration for all industries. However, spatial scale matters. At the national level, the authors offer the stark prediction of a gradual switch from a monocentric structure to a duocentric one, the area around Lyon emerging as a second important center more comparable to the area around Paris. Meanwhile, at a finer geographic scale, increased polarization of activity around the main cities of France arises from the simulations. The overall picture is therefore one of an increased number of large centers of more even size, with surrounding areas loosing their industrial base to the benefit of the local center.

Those last papers seem to correspond to the kind of "computable spatial equilibrium" work that Fujita et al. (1999) called for in the conclusion of their book. They use ambitious NEG theoretical modelling, extended to account for important characteristics such as precise input-output linkages between a great number of industries, in order to give insights of what those models predict when a particular experiment, such as a drop in trade costs, occurs. In that sense they bear a large inheritance from modern computable general equilibrium modelling of trade liberalization accounting for market structure imperfections. This kind of work can be viewed, as Forslid et al. (2002) nicely put it, as "theory with numbers, rather than empirical results." These papers show how to generate empirical predictions that are tightly linked to rich versions of underlying theory. The next step would be to find cases where reality has conducted the same type of experiment as the simulation. Then one can confront predictions from calibrated models with actual data on concentration indexes to assess the empirical validity of predictions that are tightly linked to theory.

8 Instability, Persistence, and Agglomeration

The existence of multiple equilibria, only some of which are stable, is a very general feature of the NEG framework. Evidence of multiple equilibria in economic geography would not directly support the NEG approach since human capital and technological externalities also generate the self-reinforcing processes that create multiple equilibria. However, evidence refuting multiple equilibria would support the "natural advantages" approach in which agglomerations occur where they do because of exogenous and unchanging features of the natural setting.

Davis and Weinstein (2002) recently examined Japanese history and devised several tests designed to detect multiple equilibria. Their results, summarized in the quote below, clearly indicates that the authors do not find a lot of support for the existence of multiple equilibria.

> "An important practical question, then, is whether such spatial catastrophes are theoretical curiosa or a central tendency in the data. Our results provide an unambiguous answer: Even nuclear bombs have little effect on relative city sizes over the course of a couple of decades. The theoretical possibility of spatial catastrophes due to temporary shocks is *not* a central tendency borne out in the data." Davis and Weinstein (2002) (p. 1284 emphasis is in the original)

The basic question is whether the geographic pattern of agglomeration is stable over time periods featuring large shocks. Natural advantages models should exhibit such stability since there is a single equilibrium, which is globally stable and should change slowly given that nature changes slowly. In contrast, NEG models—and others of similar ilk—*might* exhibit instability. Referring back to the panel (a) of Figure 3, suppose the economy has parameter values that situate it in the region of three stable equilibria. Then a moderate negative shock to an agglomeration (that is a decline in λ from a starting point of $\lambda = 1$, as illustrated with a "}") could move the economy past the dashed line to a region of the parameter space where the dynamics (shown by the arrows) now push towards the symmetric dispersed equilibrium. Thus, while a small shock would rapidly be reversed (agglomeration is locally stable), a moderate shock could cause the agglomeration to unravel. An extremely large shock could even reposition the agglomeration from one location to another.

There are two related statistical methods for examining the issues of persistence and responsiveness to shocks. First, one can simply look at the correlations between the size of current agglomerations and their size in the past. Second we can estimate the extent to which locations recover from measured shocks.

8.1 Stability of historical location rankings

The long-run correlation method calculates the raw or rank correlation between city i's current share of the relevant population, λ_{it}, and its share b years before, $\lambda_{i,t-b}$. While a high correlation is expected for small b, it seems likely that over a longer period, featuring general population increase, important economic transitions and shocks, the correlation would decline dramatically. In calculating λ_{it} one may use cities as the geographic unit as long as they are consistently defined over time. Lacking such data, Davis and Weinstein (2002) use 39 regions for which they divide regional population by regional area and obtain population density as the agglomeration measure. The most striking result from the intertemporal correlations is that 1998 population density has a 0.76 raw correlation with population density in 1600 (i.e. $b = 398$); at 0.83, the rank correlation is even higher. Thus, over a four century period in which the total population of Japan increased tenfold, the economy shifted from agriculture to manufacturing and services, the ranking of regions remained remarkably stable.

Brakman et al. (2002) investigate stability of city sizes in Germany. Unlike Japan, where mountainous terrain substantially constrains where its 126 million residents might live, Germany's physical geography exerts a less dominant influence. For 60 cities, the authors find a a 0.841 rank correlation between their 1939 and 1999 populations. Since Davis and Weinstein (2002) find 0.93 rank correlation between 1920 and 1998, this suggests that Germany's agglomeration pattern is somewhat less persistent than Japan's.

One of the main messages of Davis and Weinstein (2002) is that physical geography matters a great deal for economic geography. They quote from recent theoretical monographs to establish that NEG theorists have given inadequate attention to the importance of physical geography in explaining agglomeration. Another set of researchers, most notably Jeffrey Sachs (2001) and Jared Diamond (1997), have been pushing a "geography as destiny" viewpoint. Acemoglu et al. (2002) illustrate one case where early geographic advantages translated into subsequent *disadvantages*. Their experiment is the European colonization of much of the Americas, Africa, and Oceania following 1500. One might expect, under some models of NEG, that Europeans would colonize areas that already provided good markets and supplies of inputs. In that case, we might expect colonizers to choose areas that already had relatively dense and urbanized populations. In a natural advantages setting, one would expect Europeans to choose the areas with strong fundamentals, which again would probably be the areas of relatively dense inhabitation.

Acemoglu et al. (2002) also argue that the urbanized areas were very likely to be the more prosperous areas based on both theory and current cross-sectional correlations between urbanization and income per capita. They raise the question of whether prosperity in 1500 would be a good forecaster of prosperity in 1995. The answer they find is a resounding no. Incomes in 1995 are negatively related to both urbanization and population density in 1500. The currently prosperous countries tend to be ones that attracted European colonists who brought with them European institutions. The Europeans tended to treat existing population centers as locations to extract resources from and this resulted, according to Acemoglu et al. (2002) in investment-depressing institutions. While this study has only a tangential connection to NEG empirics, we think it worth mentioning to elaborate on the type of historically focussed empirical work that might help to disentangle the different roles of natural advantages, self-reinforcing processes, and shocks in determining the pattern of agglomeration.

Dumais et al. (2002), following up on Ellison and Glaeser (1997), study the evolution of the EG concentration index over the period 1972–1992. One of the objectives of Dumais et al. (2002) is to investigate the pattern of industry mobility to assess "how important historical accidents are in practice and whether Krugman's charming examples are representative." An important preliminary finding is that the measured level of agglomeration of industries is very stable over time: They find a correlation coefficient of 0.92 between 1972 and 1992 EG indexes across industries (Kim, 1995, finds a corresponding striking correlation of 0.64 with a different localisation index between 1860 and 1987 values).

As emphasized by the authors, this dynamic stability is compatible with different, and informative, patterns of underlying "firms demographics." One possible pattern is that, in each industry, new firms replace old or dead ones in the same locations. Another possibility is that the underlying economic forces in each industry persist over time and therefore yield this great stability in the *levels* of agglomeration, despite important changes in the precise *location* of the industry. NEG models are often characterized by historical accidents, in which a region taking an accidental lead in the production share of the IRS industry might end up attracting all firms of this industry. The linkages creating the agglomeration forces thereafter make it very difficult to "break the core" into a more dispersed pattern or relocate this core in another region. Concentrated industries because of NEG linkages should therefore be expected to be very immobile over time.

Contrary to those NEG-type expectations, Dumais et al. (2002) find that the most geographically concentrated industries do *not* exhibit any less mobility than a typical unconcentrated industry. This result therefore sheds some doubt on the hypothesis that spatial concentration would be mainly explained by mechanisms locking-in industries in the locations historically chosen by pioneering firms.

8.2 The long-term impact of temporary shocks

The long-run correlations are interesting especially when we have strong reason to believe that there were important city-specific shocks that might have impacted agglomeration patterns. It is more compelling to examine these city-specific shocks directly using the shock persistence regressions. Assuming multiplicative shocks, taking natural logs, and calculating before and after differences, one obtains

$$(\ln \lambda_{i,t+a} - \ln \lambda_{it}) = \alpha + \beta(\ln \lambda_{it} - \ln \lambda_{i,t-b}) + e_{it}, \tag{22}$$

where a is the time elapsed *after* the split point (t) and b is the time elapsed *before* the split. Thus, b is the duration of the period in which the shock occurs. The estimated value of β tells us about the dynamics. An estimate of $\hat{\beta} \approx 0$ suggests a random walk in city sizes. That is all shocks have permanent effects. On the other hand, $\hat{\beta} \approx -1$ suggests shocks undo themselves over the time frame of a years.

For Davis and Weinstein (2002), the shock period is 1940 to 1947 (i.e. $b = 7$) when Japan experienced intense bombing from Allied forces that devastated many cities. The shock recovery period is 1947 to 1960 (i.e. $a = 13$). While their motivating algebra is in terms of the log shares, they replace the difference in log shares with the growth rates in their regressions. While these will be approximately the same for small changes, we think it advisable to retain the difference-in-logs specification for contexts such as their study where there were large changes.

Note, that it can be shown that if the three values of $\ln \lambda_i$ ($t+a$, t, and $t-b$) were completely independent of each other (say just random noise) then the expected coefficient on β would be -0.5. This is because λ_{it} enters negatively in the dependent variable and positively in the explanatory variable. To deal with simultaneity, Davis and Weinstein (2002) instrument for $\lambda_{it} - \lambda_{i,t-b}$ using city-specific death and destruction measures.

Davis and Weinstein (2002) estimates $\hat{\beta} \approx -1$. Thus, cities experiencing the largest population declines due to bombing tend to have the fastest postwar growth rates. By 1960, on average the population shocks have been fully reversed. Even Nagasaki and Hiroshima, victims of atomic bombs that reduced populations by 8.5% and 20% respectively, saw their populations come back in line with their 1925–1940 growth trends as early as 1960 for Nagasaki and 1975 for Hiroshima.

Those fascinating, albeit macabre, results exhibit no evidence for the catastrophe phenomena that are possible in NEG models. However, the distance from theory of this work commands some caution in interpretation: How large should the shock be for the model to predict a change in equilibrium? One should probably employ a simulated version of the model to examine this question. An additional difficulty is that the size of the shock needed depends on the level of integration of the zone. As shown in the left panel of Figure 3, the region of high sensitivity to shocks is only for a narrow range of τ ($1.627 < \tau < 1.807$ for the case considered in Figure 3). Outside that range, two situations are possible: When τ is very high, symmetry cannot be broken, *whatever the size of the shock*. When the actual τ is lower than the bottom of the range, much larger shocks would be required to reverse the pattern of agglomeration. Indeed, the Davis and Weinstein (2002) paper is unclear as to whether they assume Japan in that period to be in the zone where both symmetric and agglomerated equilibria coexist or in the zone where there is no stable dispersed equilibrium. The two cases have different implications: In the former, the equilibrium can jump from agglomerated to dispersed (or the reverse) with a relatively small shock compared with the shock needed in the latter situation to make the equilibrium switch from agglomeration in one region to agglomeration in another. Therefore, it is at least possible that Japan was at the time in a parameters zone where only a *reversal* of agglomerated equilibrium was possible, a switch that could only result from shocks even larger than nuclear bombing. Further empirical investigation of NEG-type persistence of temporary shocks needs to take into account that the predictions of those models are conditional on values of the parameters. This recommendation parallels the one made above about concentration index regressions.

Another caveat regarding inferences to be drawn from Davis and Weinstein's (2002) study is that this is a case where physical geography matters an exceptional amount. Japan's mountainous topography, with a small share of overall land actually suitable for large scale city locations, makes it possible that activity reverts to its original location because there is no other suitable location left to occupy. While this point might certainly have some validity for activity and population growth as a whole, it should have less importance at the industry level. This is investigated in a follow-up paper by Davis and Weinstein (2003b). In the aftermath of allied bombing on Japanese cities, they show a tendency for *specific industries* to locate back where they initially were (despite massive destruction that drastically changed the distribution of industries across cities). This finding further undermines the case for multiple equilibria in location patterns.

Brakman et al. (2002) study the impact of wartime bombing in Germany. With respect to the persistence of related shocks, they find an estimate of β equal to -0.42 for West

German cities when they assume $a = 4$. This goes to -0.52 when the authors broaden the "after" window to 17 years. They use house destruction as their instrumental variable for the population shock. Oddly, in East Germany, there was no tendency towards shock reversal and urban populations appear to follow random walks. We see a value to more studies of shock persistence. From these two studies, it is tempting to conclude that the greater the constraint imposed by physical geography, the greater will be the tendency for shocks to undo themselves over time.

Combining the key results of this section, it seems that this set of recent papers shows no evidence of either catastrophes (city sizes persist despite large shocks) or historical accidents (same level of mobility between concentrated and dispersed industries). It suggests that those two celebrated characteristics of NEG models should perhaps be considered more as fascinating theoretical "exotica" rather than as robust elements of economic geography.

9 Conclusion

Theoretical work on economic geography has a long and productive history. The last decade has seen a torrent of new papers, many of which expand upon the framework developed by Krugman (1991a). This literature, often referred to by the not very descriptive title of "new economic geography", is exciting because it generates results that contrast markedly with the traditional analyses involving exogenous factor supplies and constant returns to scale. NEG theories are characterized by magnification, bifurcation, multiple equilibria, and the possibility of catastrophe.

At its conclusion, the authors of *The Spatial Economy* argued that a vital part of "the way forward" from their work would involve empirical examination of the "intriguing possibilities" raised by the new theory. They did not specify the form these examinations should take, nor has any consensus emerged on the empirical methods to be applied to NEG.

Although the theory is still being digested, a large new serving of empirical work has arrived over the last five years. This survey has attempted to organize the new empirics of agglomeration and trade into categories and then assess the collective support it provides for NEG. The diversity in approaches that characterizes this literature probably stems in large part from the difficulties inherent in testing theories involving circular causation. In terms of the results, our sense is that the dust has not settled yet. One can see a number of supportive findings but there are just as many findings that appear to undermine the new theory. The positive relation between wages and market potential looks like a sturdy result but the response of production to demand, while certainly positive, is not consistently greater than one for one. Economic activity concentrates spatially but this agglomeration cannot yet be seen as confirmation of the theories that were constructed to explain the phenomenon. There are a number of other explanations that are consistent with the data and not much yet that strongly points to the explanation offered by NEG.

The lesson to be learned from past work (and from Leamer and Levinsohn, 1995) is that methods need to be designed to connect closely to the theory but should not be reliant upon features of models that were included for tractability or clarity of exposition instead of realism. Rather we need to focus on testing the essential distinguishing features of the models that allow one to falsify them or their alternatives.

What elements of the existing empirical literature will and should continue to figure prominently in future empirics? First, trade costs are a critical parameter and further work will continue to try to estimate how they vary across industries and over time. In models based on CES demand, it is critical to identify the free-ness of trade which is a compound parameter, $\tau^{1-\sigma}$, depending on trade costs as moderated by the elasticity of substitution between varieties. Second, the concept of real market potential (demands that are summed up while discounting for distance, borders, and supply alternatives) should continue to figure in studies of the location decisions of firms and workers, as well as the determination of factor prices. More work will be required to decide how to estimate each location's real market potential. In addition we need tests to discriminate between market potential as a motive for agglomeration in contrast to other mechanisms that might generate similar empirical relationships. Indeed, while structural estimation of NEG models is a valuable approach, we believe the biggest advances will come from approaches like David and Weinstein's (1996, 2003) where estimates of a single parameter can allow us to choose between plausible alternative mechanisms of economic geography.

References

Acemoglu, D., S. Johnson, and J. Robinson (2002), "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution", *Quarterly Journal of Economics* 117(4):1231–1294.

Ades, A. and E. Glaeser (1995), "Trade and Circuses: Explaining Urban Giants", *Quarterly Journal of Economics* 110(1):195–227.

Amiti, M. (1999), "Specialisation Patterns in Europe", *Weltwirschaftliches Archiv* 134(4):573–593.

Anderson, S., A. De Palma and J-F. Thisse (1992), *Discrete Choice Theory of Product Differentiation* (MIT Press, Cambridge).

Baldwin, R., R. Forslid, P. Martin, G. Ottaviano and F. Robert-Nicoud (2003), "The Core-Periphery Model: Key Features and Effects", draft chapter 1 of *Public Policies and Economic Geography* (Princeton University Press, Princeton).

Brakman, S., H. Garretsen and C. van Marrewijk (2001), *An Introduction to Geographical Economics* (Cambridge University Press, Cambridge).

Brakman, S., H. Garretsen and M. Schramm (2002), "The Strategic Bombing of German Cities during WWII and its Impact on City Growth", mimeo (http://www.few.eur.nl/few/people/vanmarrewijk/geography/further%20reading/bombingsept2002def.pdf).

Brülhart, M. and J. Torstensson (1996), "Regional Integration, Scale Economies and Industrial Economies", Center for Economic Policy Research Discussion Paper # n1435.

Brülhart, M. (2001), "Evolving Geographical Specialisation of European Manufacturing Industries", *Weltwirschaftliches Archiv* 137(2):215–243.

Brülhart, M. and F. Trionfetti (2002), "A Test of Trade Theories When Expenditure is Home Biased", mimeo, University of Lausanne.

Carlton, D. (1983), "The Location and Employment Choices of New Firms : an Econometric Model with Discrete and Continuous Endogenous Variables", *Review of Economics and Statistics* 65(3):440–449.

Combes, P-P. (1997), "Industrial Agglomeration under Cournot Competition", *Annales d'Economie et de Statistique* 45:161–182.

Combes, P-P. and M. Lafourcade (2001), "Transport Cost Decline and Regional Inequalities: Evidence from France", Center for Economic Policy Research Discussion Paper # 2894 (revised version: http://www.enpc.fr/ceras/combes/pdf/artinf.pdf).

Coughlin, C.C., J.V. Terza and V. Arromdee (1991), "State Characteristics and the Location of Foreign Direct Investment in the United States", *Review of Economics and Statistics* 73(4):675–683.

Crozet, M. (2000), "Do migrants believe in market potential?" Cahier de la MSE–Série Blanche # 2000-30 (revised version: http://team.univ-paris1.fr/trombi/crozet/Do_migrants.pdf).

Crozet, M., T. Mayer and J.L. Mucchielli (forthcoming), "How do firms Agglomerate? A Study of FDI in France", forthcoming in *Regional Science and Urban Economics*.

Davis, D. and D. Weinstein (1996), " Does Economic Geography Matter for International Specialization?", National Bureau of Economic Research Working Paper # 5706.

Davis, D. and D. Weinstein (1999), "Economic Geography and Regional Production Structure: An Empirical Investigation", *European Economic Review* 43(2):379–407.

Davis, D. and D. Weinstein (2002), "Bones, Bombs, and Break Points: The Geography of Economic Activity", *American Economic Review* 92(5):1269–1289.

Davis, D. and D. Weinstein (2003a), "Market Access, Economic Geography and Comparative Advantage: An Empirical Assessment", *Journal of International Economics*, 59(1):1–23.

Davis, D. and D. Weinstein (2003b), "A Search for Multiple Equilibria in Urban Industrial Structure", mimeo presented at the CEPR workshop 'The Economic Geography of Europe: Measurement, Testing and Policy Simulations' held in Villars 17/18 January 2003.

Dekle, R. and J. Eaton (1999), "Agglomeration and Land Rents: Evidence from the Prefectures", *Journal of Urban Economics* 46(2):200–214.

Devereux, M. and R. Griffith (1998), "Taxes and the Location of Production : Evidence from a Panel of US Multinationals", *Journal of Public Economics* 68(3):335–367.

Diamond, J. (1997), *Guns, Germs and Steel: The Fate of Human Societies* (Norton, New York).

Disdier, A.C. and K. Head (2003), "Exaggerated Reports on the Death of Distance: Lessons from a Meta-Analysis", University of British Columbia mimeo.

Dixit, A. and J. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity", *American Economic Review* 67(3):297–308.

Dumais, G., G. Ellison and E. Glaeser (2002), "Geographic Concentration as a Dynamic Process", *Review of Economics and Statistics*, 84(2):193–204.

Duranton, G. and H. Overman (2002), "Testing for Localization Using Micro-Geographic Data", Center for Economic Policy Research Discussion Paper # 3379.

Ellison, G. and E. Glaeser (1997), "Geographic Concentration in U.S. Manufacturing Industries: a Dartboard Approach", *Journal of Political Economy* 105(5):889–927.

Ellison, G. and E. Glaeser (1999), "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?", *American Economic Review Papers and Proceedings* 89(2):311–316.

Erkel-Rousse, H. and D. Mirza (2002), "Import price-elasticities: reconsidering the evidence", *Canadian Journal of Economics* 35(2):282–306.

Feenstra, R., J. Markusen and A. Rose (2001), "Using the gravity equation to differentiate among alternative theories of trade", *Canadian Journal of Economics* 34(2):430–447.

Forslid, R., J. Haaland and K-H Midelfart-Knarvik (2002), "A U-shaped Europe? A simulation study of industrial location", *Journal of International Economics* 57(2):273–297.

Forslid, R., and I. Wooton (forthcoming), "Comparative Advantage and the Location of Production", *Review of International Economics*.

Friedman, J., D. Gerlowski and J. Silberman (1992), "What Attracts Foreign Multinational Corporations", *Journal of Regional Science* 32(4):403-418.

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

Fujita , M., and J-F Thisse (2002), *Economics of Agglomeration; Cities, Industrial Location and Regional Growth* (Cambridge University Press, Cambridge).

Guimarães, P., O. Figueiredo and D. Woodward (2000), "Agglomeration and the Location of Foreign Direct Investment in Portugal", *Journal of Urban Economics* 47(1):115–135.

Haaland J., H-J. Kind, K-H Midelfart Knarvik (1999), "What Determines the Economic Geography of Europe?", Center for Economic Policy Research Discussion Paper # 207.2

Hanson, G. (1996), "Localization Economies, Vertical Organization, and Trade", *American Economic Review* 86(5):1266–1278.

Hanson, G. (1997), "Increasing Returns, Trade and the Regional Structure of Wages", *Economic Journal* 107(440):113–133.

Hanson, G. (1998), "Market Potential, Increasing Returns, and Geographic Concentration", National Bureau of Economic Research Working Paper # 6429, (revised Dec. 2001 version: http://www-irps.ucsd.edu/faculty/gohanson/Uscnty.pdf.

Hanson, G. (2001), "Scale Economies and the Geographic Concentration of Industry", *Journal of Economic Geography* 1(3):255–276.

Hanson, G. and C. Xiang (2002), "The Home Market Effect And Bilateral Trade Patterns", NBER Working Paper # 9076.

Harris, C. (1954), "The Market as a Factor in the Localization of Industry in the United States", *Annals of the Association of American Geographers* 64:315–348.

Head, K. and T. Mayer (2000), "Non-Europe : The Magnitude and Causes of Market Fragmentation in Europe", *Weltwirschaftliches Archiv* 136(2):285–314.

Head, K. and T. Mayer (2002), "Market Potential and the Location of Japanese Investment in the European Union", Center for Economic Policy Research Discussion Paper # 3455.

Head, K., T. Mayer and J. Ries (2002), "On the Pervasiveness of Home Market Effects", *Economica* 69(275):371–390.

Head, K. and J. Ries (2001), "Increasing Returns Versus National Product Differentiation as an Explanation for the Pattern of US-Canada Trade", *American Economic Review* 91(4):858–876.

Head, K., J. Ries and D. Swenson (1995), "Agglomeration Benefits and Location Choice : Evidence from Japanese Manufacturing Investment in the United States", *Journal of International Economics* 38(3-4):223–247.

Head, K., J. Ries and D. Swenson (1999), "Attracting Foreign Manufacturing: Investment Promotion and Agglomeration", *Regional Science and Urban Economics* 29(2):197–218.

Helpman, E. (1998), "The Size of Regions", in: D. Pines, E. Sadka and Y. Zilcha, eds., *Topics in Public Economics, Theoretical and Applied Analysis* (Cambridge University Press, Cambridge) 33–54.

Helpman, E. and P. Krugman (1985), *Market Structure and Foreign Trade* (MIT Press, Cambridge).

Helsley, R., W. Strange (1990), "Matching and Agglomeration Economies in a System of Cities", *Regional Science and Urban Economics* 20(2):189–212.

Henderson, V. and A. Kuncoro (1996), "Industrial Centralisation in Indonesia", *World Bank Economic Review* 10(3):513–540.

Hummels, D. (1999), "Toward a Geography of Trade Costs", mimeo, University of Chicago.

Justman, M. (1994), "The Effect of Local Demand on Industry Location", *Review of Economics and Statistics* 76(4):742–753.

Kim, S. (1995), "Expansion of Markets and the Geographic Distribution of Economic Activities: The Tends in U.S. Regional Manufacturing Structure, 1860-1987", *Quarterly Journal of Economics* 110(4):881–908.

Krugman, P. (1979), "Increasing Returns, Monopolistic Competition, and International Trade", *Journal of International Economics*, 9(4) : 469–479.

Krugman, P. (1980), "Scale Economies, Product Differentiation, and the Pattern of Trade", *American Economic Review* 70:950–959.

Krugman, P. (1991a), "Increasing Returns and Economic Geography", *Journal of Political Economy* 99:483–499.

Krugman, P. (1991b), *Geography and Trade* (MIT Press, Cambridge).

Krugman, P. (1993), "First Nature, Second Nature, and Metropolitan Location", *Journal of Regional Science* 33(2):129–144.

Krugman, P. and A. Venables (1995), "Globalization and the Inequality of Nations", *Quarterly Journal of Economics* 110(4):857–880.

Lai, N. and D. Trefler (2002), "The Gains from Trade with Monopolistic Competition: Specification, Estimation, and Mis-Specification", NBER Working Paper # 9169.

Leamer, E. and J. Levinsohn (1995), "International Trade Theory: The Evidence" in: G. Grossman and K. Rogoff, eds., *Handbook of International Economics* vol. 3 (North-Holland, Amsterdam) 1339–1394.

Limao, N. and A. Venables (2001), "Infrastructure, Geographical Disadvantage, Transport Costs, and Trade", *World Bank Economic Review* 15(3):451–479.

Lucas, R. (1988), "On the Mechanics of Economic Development", *Journal of Monetary Economics* 22:3–42.

Lundbäck, E. and J. Torstensson (1998), "Demand, Comparative Advantage and Economic Geography in International Trade: Evidence from the OECD," *Weltwirtschaftliches Archiv* 134(2):230–249.

Marshall, A. (1920), *Principles of Economics* (Macmillan, London).

Murata Y. (2003), "Product Diversity, Taste Heterogeneity, and Geographic Distribution of Economic Activities: Market vs. Non-market Interactions", *Journal of Urban Economics* 53(1):126–144

Neary, P. (2001), "Of Hype and Hyperbolas: Introducing the New Economic Geography", *Journal of Economic Literature* 39(2):536–561.

Ottaviano, G., T. Tabuchi and J-F. Thisse (2002), "Agglomeration and Trade Revisited", *International Economic Review* 43(2):409–436.

Overman, H., S. Redding and A. Venables (2001), "The Economic Geography of Trade, Production and Income: A survey of Empirics", Center for Economic Policy Research Discussion Paper # 2978, forthcoming in: J. Harrigan and K. Choi, eds., *Handbook of International Trade* (Basil Blackwell, London).

Puga, D. (1999), "The rise and fall of regional inequalities", *European Economic Review* 43(2):303–334.

Redding, S. and A. Venables (2000), "Economic Geography and International Inequality", Center for Economic Policy Research Discussion Paper # 2568 (revised version, April 2002, http://econ.lse.ac.uk/staff/ajv/wincrev6.pdf).

Rosenthal, S. and W. Strange (2001), "The Determinants of Agglomeration", *Journal of Urban Economics* 50:191–229.

Sachs, J. (2001), "Tropical Underdevelopment", National Bureau of Economic Research Working Paper # 8119.

Smith, D. and R. Florida (1994), "Agglomeration and industry location: An econometric analysis of Japanese-affiliated manufacturing establishments in automotive-related industries", *Journal of Urban Economics* 36(1):23–41.

Tabuchi, T., and J-F. Thisse (2002), "Taste Heterogeneity, Labour Mobility and Economic Geography", *Journal of Development Economics* 69(1):155-177.

Trionfetti, F. (2001), "Using Home-Biased Demand to Test Trade Theories", *Weltwirtschaftliches Archiv* 137(3):404–426.

Venables, A. (1996), "Equilibrium Locations of Vertically Linked Industries", *International Economic Review* 37(2):341–359.

Weder, R. (forthcoming), "Comparative Home-Market Advantage: An Empirical Analysis of British and American Exports", forthcoming in *Weltwirtschaftliches Archiv*.

Wei, S-J (1996), "Intra-National Versus International Trade: How Stubborn Are Nations in Global Integration?", National Bureau of Economic Research Working Paper # 5531.

A Data appendix of Figure 4

Figure 4 uses several data sets that make feasible an industry-level collection of parameters values combined with trade freeness calculation. The main issue is to find a sufficiently flexible industry classification that allows both for a reasonable level of detail in the study and good data availability. The classification used by the OECD for its IO tables is quite attractive in those respects as it has a very easy correspondence with UN industry classifications ISIC rev2 and rev3, which are widely used and are quite easy to match with trade data.

The parameters needed are μ (the share of the industry's good in overall consumption), α (the share of own industry inputs expenditures in overall costs) and σ (which has the many interpretations emphasized above). For this graph, we use the OECD IO table for Japan in 1990 which is the latest table available. μ_s is calculated as the share of domestic demand for industry s goods in total domestic demand (domestic demand being defined as private final consumption + government purchases plus purchases for investment of goods from industry s). α represents the share of inputs purchased from own industry in overall costs (proxied by total purchases on intermediates plus compensation of employees). σ_s is taken from Table 4 of Hummels (1999) which gives estimates for 2 digit SITC rev3 industries in 1992, easy to match with the IO classification (the average of σ values are taken when multiple SITC goods categories map into a single IO industry). The last parameter needed to compute the range defined by $[\underline{\phi}_s, \overline{\phi}_s]$ is η, fixed to 200. More work is needed on getting estimates of η from the literature and ensuring that real roots exist for lower, more realistic, values.

The calculation of $\hat{\phi}_s$ involves comparable bilateral trade and production data in a common classification for our country pairs. The trade data from the World Bank trade and production database and production figures extracted from STAN OECD database both map into IO industries and give data for a quite long time period. For Figure 4, we use

Table 5: Values of parameters used in Figure 4

IO code	Description	μ	α	σ	$\phi_{\text{fr-de}}$	$\phi_{\text{us-ca}}$
3	foodbevtob	6.78%	18.52%	4.53	0.033	0.034
4	cloth	2.34%	34.66%	6.62	0.088	0.055
5	wood	0.36%	20.38%	3.64	0.019	0.130
6	paper	0.53%	36.61%	4.34	0.035	0.112
7	chemical	0.49%	42.93%	3.89	0.138	0.202
8	drugs	0.43%	7.56%	9.53	0.051	0.044
9	petro	0.72%	6.77%	5.01	0.019	0.055
10	plastics	0.26%	22.55%	5.36	0.070	0.135
11	minerals	0.10%	15.10%	2.65	0.032	0.087
12	ferrous	0.00%	58.59%	2.32	0.098	0.095
13	non-ferrous	0.04%	49.19%	6.66	0.150	0.343
14	fabmetal	0.49%	7.49%	4.85	0.024	0.061
15	machinery	3.83%	22.61%	7.87	0.106	0.494
16	computers	1.50%	19.38%	11.02	0.543	0.807
17	electrical	1.71%	19.30%	5.88	0.078	0.262
18	radiotvcom	1.89%	32.94%	9.44	0.212	0.210
19	ships	0.16%	0.12%	7.40	0.012	0.107
20	railroad	0.31%	21.01%	7.40	0.052	0.185
21	vehicles	2.67%	49.08%	7.11	0.130	0.594
22	aircraft	0.22%	39.63%	7.40	0.812	0.207
23	instruments	0.59%	17.80%	7.43	0.100	.

1995 data to evaluate freeness of trade (except for aircraft, for which we use 1996 for the France-Germany ϕ), a recent year that is not too remote from the years for which parameters μ, α and σ are available.