Alma Mater Studiorum - Università di Bologna

Dipartimento di Scienze Economiche

Dottorato di Ricerca in Economia - XIX Ciclo

Esame finale anno 2007

Explaining regional productivity differentials: four essays

Dottorando: Dino Pinelli

Relatore: Prof. Gianmarco I.P. Ottaviano

Coordinatore: Chiar.mo Prof. Luca Lambertini

Settore scientifico disciplinare: SECS-P/01 Economia Politica

TABLE OF CONTENT

CHAPTER 1

A 'NEW ECONOMIC GEOGRAPHY' PERSPECTIVE TO GLOBALIZATION

1. INTRODUCTION 4
2. THEORY 6
3. COUNTRIES 11
4. REGIONS 14
5. GLOBALIZATION 17
6. WELFARE 20
7. EVIDENCE 23
8. CONCLUSION 31

REFERENCES 33

CHAPTER 2

MARKET POTENTIAL AND PRODUCTIVITY: EVIDENCE FROM FINNISH REGIONS

1. INTRODUCTION 39
2. THE MODEL 40
3. FINLAND 45
4. DATA 47
5. REGRESSION ANALYSIS 50
6. CONCLUSION 58

REFERENCES 59

ANNEX: THE DATA 61

CHAPTER 3

MEASURING DIVERSITY: A CROSS-DISCIPLINARY COMPARISON OF EXISTING INDICES

1. INTRODUCTION 65

2. A FRAMEWORK FOR INVESTIGATION 66

3. MEASURING DIVERSITY AS RICHNESS AND EVENNESS 67

4. DEALING WITH TYPES DIFFERENCES: INTRODUCING ‘DISTANCE’ 81

5. AN APPLICATION 87

6. CONCLUSIONS 91

REFERENCES 92

ANNEX: THE DATA 96

CHAPTER 4

DIVERSITY AND PRODUCTIVITY: EVIDENCE FROM EUROPEAN CITIES

1. INTRODUCTION 101

2. THE LITERATURE ON DIVERSITY 102

3. THE DATASET 104

4. MEASURING DIVERSITY 105

5. DIVERSITY IN EUROPEAN REGIONS 107

6. THEORETICAL MODEL 112

7. EMPIRICAL RESULTS 117

8. CONCLUSIONS 124

REFERENCES 126

EXPLAINING REGIONAL PRODUCTIVITY DIFFERENTIALS: FOUR ESSAYS

Introduction

After WWII the barriers to the international mobility of goods, factors and ideas have steadily fallen. Interactions among customers and suppliers around the world have become more and more tight, due also to technological progress in ICT. Contemporaneously, migration flows have fostered the global interaction of a growing and increasingly diversified number of people. To many observers, these phenomena underpin the creation of a unique global market (‘globalisation’). At European level, globalization is reinforced by the twin processes of deeper integration and enlargement.

The central questions addressed in this thesis are whether and how these phenomena can be expected to change the geographical distribution of economic activities.

Chapter 1 and 2 tackle the foregoing questions from the specific point of view of ‘new economic geography’ (henceforth, simply NEG), an approach to economic geography firmly grounded on recent developments in mainstream industrial organization and international trade theory.

Chapter 1 starts from a comprehensive review of NEG to assess its theoretical predictions in the light of available empirical evidence. The paper reviews recent development in NEG and assesses its theoretical predictions in the wake of empirical relevance. The paper considers that globalisation can be expected to have a non-linear effect on the degree of geographical agglomeration of economic activities. Initially, lower transport costs, lower institutional barriers, and lower communication costs foster agglomeration. When all these costs become negligible, agglomeration unfolds. The paper finds that the agglomeration of economic activities in core areas damages immobile people in peripheral ones. However, opposing agglomeration is not always socially desirable. Indeed, when these activities benefit from localized ‘knowledge spillovers’, the most efficient way to take care of the periphery is to allow for agglomeration in the core while redistributing to the periphery some of the gains of the core. Finally, the paper discusses the empirical support for NEG insights. It is argued that no conclusive evidence is available yet and this is mainly due to a communicational gap between theoretical and applied investigations. Indeed, until very recently theorists have shown little interest in translating their insights in clear-cut testable predictions. At the same time, empiricists have made little efforts to understand what theory exactly implies.

Chapter 2 takes this research agenda forward by testing NEG's theoretical predictions on Finnish NUTS 4 regions. NEG models consider various types of linkages as agglomeration forces. However, which agglomeration force dominates in reality is left nonetheless unspoken as the empirical implications of different models have not been entirely spelled out. The paper uses a typical NEG model to design an empirical methodology aimed to assess whether linkages are relevant at all and, if so, whether they are more important for firms or workers in terms of, respectively, productivity or amenity. The proposed methodology is applied to Finnish NUTS 4 regions from 1977 to 2002. Results show that linkages are very relevant and that firm-related demand and cost linkages are more important than worker-related cost-of-living linkages. Results also bring support to two main predictions of NEG models. First, by fostering the agglomeration of workers and firms, labour mobility and specialization in new footloose sectors hamper the process of regional convergence in productivity and amenity. Second, with or without labour mobility, agglomeration happens in places that enjoy better market and supplier access.

Chapters 3 and 4 explore the economic implications of the growing diversity of people that live and work in our cities. The issue is at the core of current public debates. On the one hand, official rhetoric looks at diversity as a main asset for development and human welfare. On the other hand, the general public perceives issues such as migration and enlargement as very problematic. Similarly, economic literature shows that diversity entails potential costs as well as potential benefits. Which of those prevail is still subject to empirical investigation.

Chapter 3 addresses the issue of measuring diversity. Quantitative measures of diversity are necessary to many fields of scientific investigation. This has led to the development of a variety of indices. In this paper we review the different indices and approaches proposed. We do not discuss why diversity is important in the different fields. The focus is on how diversity is measured. We propose a systematisation of the main statistical indicators of diversity beyond the different names that different disciplines often give to very similar measures. In doing so, we have clarified that the crucial distinctions between indices arises from the specific components of diversity that they aim at capturing: richness, evenness or distance, or combinations of the three. We show that, when targeted at the same component(s) of diversity, different indices yields very similar results. Most naturally, differences emerge only when the components of diversity addressed are in fact different. In particular, the indices measuring only evenness might differ substantially from those measuring only richness or richness and evenness together. By showing how many indices are closely related, our results provide a framework to compare the methodology and the results of existing studies on diversity across disciplines. To future studies, our results also offer a toolbox that simplifies greatly the choice of the correct diversity measures.

Chapter 4 uses a newly developed database to study the relationships between diversity and productivity across European cities. A number of cross-country studies points to a negative relationship between diversity and economic performance. However, cross-country regressions are likely to be affected by institutional differences (see Chapter 4 for discussion). Recent evidence on US data at city level show that richer diversity is associated with higher wages and productivity of natives with causation from the former to the latter. Using a new regional database for Europe, we take this research agenda forward and look for the first time at the relationship between diversity and productivity across European regions. The dataset includes demographic and economic data for over 300 NUTS 3 European regions. Demographic data are collected from national censuses of 1991 and 2001. Economic data are mostly from the Eurostat REGIO database. Prices on non-tradeable are proxied by prices in restaurants and hotels from Michelin Guides of 1991 and 2001. We find results that are broadly consistent with those on US cities: richer diversity is associated with higher productivity also in EU regions. We provide evidence that causation again runs from the former to the latter.

CHAPTER 1

A 'NEW ECONOMIC GEOGRAPHY' PERSPECTIVE TO GLOBALIZATION

1. Introduction

After WWII the barriers to the international mobility of goods, factors and ideas have been steadily falling. Interactions among customers and suppliers around the world have become increasingly tight, due also to technological progress in ICT. To many observers, all this is fostering the creation a unique global market for goods, factors and ideas ('globalization'). The central questions addressed in this paper are whether and how these phenomena can be expected to change the geographical distribution of economic activities both within and between countries.

The foregoing questions are tackled from the specific point of view of 'new economic geography' (henceforth, simply NEG), an approach to economic geography firmly grounded on recent developments in mainstream industrial organization and international trade theory. Section 2 explains that the hallmark of NEG is the focus on the interactions among firms and workers in markets where producers face increasing returns to scale and enjoy market power. Intense scale returns and strong market power may generate self-sustaining processes of agglomeration that make firms cluster in space.

Section 2 also introduces the key concept of market potential as a measure of the location appeal of a region in terms of customer and supplier proximity. Specifically, the market potential of a region measures the sales and the profits an average firm can make if located in that region. Hence, interregional differences in market potentials predict the future of the economic landscape as firms are attracted towards higher market potential regions.

The logical implication is that, according to NEG, understanding the geographical evolution of the economy requires understanding the causes of the changes in market potential. Section 3 and 4 apply this insight to the inter- and the intra-national context respectively. The fundamental difference between the two contexts is that labour is

The paper was published as: Ottaviano G.I.P and D. Pinelli (2005), A 'new economic geography' perspective to globalization, *Scienze Regionali,* 4 (1), pp. 71-106. The paper has been prepared within the framework of study *Finland in the Global Economy*, Prime Minister's Office, Helsinki. We have benefited from comments by Tarja Cronberg, Pertti Haaparanta, Heikki Loikkanen, Paavo Okko, Hannu Tervo, Vesa Vihriälä, Raija Volk, Pekka Ylä-Anttila, John Zysman and other seminar participants in Budapest, Helsinki, Mariehamn, Regensburg. We also are grateful to the editor and two referees of the publishing journal for their comments.

much less mobile between than within countries. As labour mobility is shown to promote agglomeration, NEG forces are more likely to foster interregional rather than international imbalances. In both contexts, however, trade liberalization initially fosters agglomeration, whereas further reductions in trade impediments trigger a reverse process of dispersion. Under this respect, transport infrastructures play an important role not only because they make goods and factors mobility easier, but also because the presence of ‘transport hubs’ and ‘gate regions’ makes clustering more likely.

Section 5 analyzes in detail the various channels through which ‘globalization’ can be expected to affect the spatial distribution of economic activities. It identifies five main channels associated with different costs of doing business: search and matching costs, direct shipping costs, control and management costs, the costs of personal interactions, and the costs of relocation. In the wake of the previous sections, it argues that globalization can be expected to have a non-linear effect on the degree of geographical agglomeration of economic activities. Initially, lower transport costs, lower institutional barriers, and lower communication costs foster agglomeration. As all those costs become negligible, agglomeration unfolds. Moreover, agglomeration is more pronounced and more persistent in sectors characterized by intense scale economies, strong market power, tight input-output relations, higher relative intensity of mobile than immobile factors (such as capital and skilled labour versus land and unskilled labour), rapidly changing products and tasks (as in hi-tech industries), high value added.

Section 6 discusses the welfare implications of NEG. In particular, it addresses two crucial policy questions: Is agglomeration desirable from a social point of view? Should policy makers foster or control it? It argues that the agglomeration of economic activities in core areas damages immobile people in peripheral ones. However, opposing agglomeration is not always socially desirable. Indeed, when those activities benefit from localized ‘knowledge spillovers’, the most efficient way to take care of the periphery is to allow for agglomeration in the core while redistributing to the periphery some of the gains of the core.

The empirical support for NEG insights is discussed in Section 6. It is argued that no conclusive evidence is available yet and this is mainly due to a communicational gap between theoretical and applied investigations. Indeed, until very recently theorists have shown little interest in translating their insights in clear-cut testable predictions. At the same time, empiricists have made little effort in understanding what theory exactly implies. As a result, the empirical evidence on NEG is still quite patchy. The aim of the section is to compose the available pieces of information within a coherent framework. The overall picture that emerges is rather promising for NEG. First, regions with higher market potential exhibit higher productivity and attract both firms and workers. Second, the resulting agglomeration is more pronounced for intermediate levels of trade impediments. Third, it is more pronounced in industries characterized by stronger scale economies, tighter input-output linkages, higher technological intensity, and higher skill intensity. Fourth, high densities of economic activities are associated with productivity-enhancing knowledge spillovers, so market potential is not the only driver of agglomeration. However, such spillovers fade away quite rapidly with distance. Finally, agglomeration make regions group in ‘convergence clubs’ depending on their long-run growth rates. Reciprocal distances play a role in determining club affiliation as closer regions tend to belong to the same clubs. [1]

2. Theory

Places differ in terms of their 'first nature', that is, in terms of their relative abundance of natural resources, their proximity to natural means of communication, and their climatic conditions. However, 'first nature' seems to be an inadequate explanation of the dramatic differences in economic development that one observes even between areas that are not very different in terms of those exogenous attributes. That is why scholars have come to the conclusion that the observed regional unbalances must be driven by some other forces ('second nature') that are inherent to the functioning of economic interactions and that, in principle, are able to generate uneven development even across ex-ante identical places.[2]

Various 'second nature' forces have been studied by economists, geographers and regional scientists.[3] However, in the last decade a specific approach, so called 'new economic geography' (henceforth, NEG) has played centre stage in mainstream economics. What distinguishes NEG from alternative approaches is the focus on market rather than non-market interactions within a 'general equilibrium' set-up, i.e. a framework of analysis that stresses the endogenous determination of good and factor prices and the importance of economy-wide budget constraints.[4] The aim of this section is to clarify the theoretical underpinnings of NEG and to highlight its differences with respect to alternative approaches.

2.1. Firm location

The location decision of a firm gives rise to a non-trivial economic problem when two things are true. First, the shipment of goods and factors across space is costly. Second, the fragmentation of production is also costly, which is the case whenever there are increasing returns to scale at the plant level. Costly transportation gives physical substance to the concept of geography: with no transport costs space would be immaterial. Together with plant-level scale economies, costly shipments generate an economic trade-off between 'proximity' and 'concentration'. When customers and

[1] There exist many surveys of NEG and alternative approaches to spatial issues. Theoretical surveys are more focused on NEG: Ottaviano and Puga (1998), Fujita et al (1999), Neary (2001), Ottaviano and Thisse (2001, 2004), Fujita and Thisse (2002), Baldwin et al (2003), Ottaviano (2003). Empirical surveys are generally less focused on NEG per se: De la Fuente (2000) Audresch and Feldman (2004), Head and Mayer (2004), Magrini (2004), Moretti (2004), Overman et al (2003). As it will be discussed, the different focuses of theoretical and empirical surveys reflect the different stages of development of the corresponding literatures.

[2] The distinction between 'first nature' and 'second nature' is due to Cronon (1991).

[3] See Fujita and Thisse (2002) for a thorough assessment of the relative merits of the different approaches.

[4] The advantage of the general equilibrium approach is nicely summarized by Fujita and Krugman (2004): "you want a general-equilibrium story, in which it is clear where the money comes from and where it goes".

intermediate suppliers are geographically dispersed, proximity allows one to minimize transportation by patronizing them through many small local plants. This strategy, however, foregoes the economies of scale that could be achieved by concentrating production in few large plants. Notice that both transport costs and scale economies are necessary for a location problem to arise.[5] On the one hand, costless transportation would allow production to be concentrated at a single plant with no penalty in terms of shipping costs. On the other hand, without scale economies, a firm could serve each market by a small local plant with no penalty in terms of high production costs (‘backyard capitalism’). More generally, a firm will tend to fragment production across many plants when transport costs are high and returns to scale are weak. Analogously, it will prefer to concentrate production in few plants when transport costs are low and returns to scale are strong.

Building Block 1 - Plant-level scale economies and shipping costs generate a trade-off between ‘proximity’ and ‘concentration’.

While transport costs and scale economies are necessary for the existence of a non-trivial location problem, such problem is complicated by the presence of competing firms. The reason is that proximity and concentration generate the basic trade-off for a firm even abstracting from any interaction with other firms. However, once competitors enter the picture, the location choice of the firms has to take into account also their potential threat. In particular, the geographical positioning of a firm with respect to its competitors affects the market power that necessarily stems from plant-level economies of scale.

Generally speaking, firms have market power when they do not take market prices as given as perfectly competitive firms would. Under such a price-making behaviour, called ‘imperfect competition’, firms trade quantity against price in making their profit-maximizing decisions. Specifically, in their product markets, firms trade higher prices against higher quantities sold. Analogously, in their input market, they trade lower prices against higher quantities demanded. NEG mainly concentrates on product markets where market power derives from product differentiation (so-called ‘monopolistic competition’), from few competitors (‘oligopoly’), or both (‘differentiated oligopoly’). In these cases, location is a crucial dimension of the profit-maximizing decision as it allows a firm to increase its market power by careful positioning.

Building Block 2 – A firm can increase its market power with respect to its competitors by careful geographical positioning.

[5] Scotchmer and Thisse (1992) call this the ‘folk theorem of spatial economics’.

From a social welfare point of view, a key implication of market power is that the prices, on which consumers and firms base their consumption, production and location decisions, do not fully reflect the corresponding social values. As a result, under imperfect competition, market interactions generate 'side effects' for which no quid-pro-quo is paid. Such side effects that are associated with market transactions are called 'pecuniary externalities'. For example, the relocation of a firm decreases competition in the place it leaves while increasing competition in the place it joins. This raises the profits of its competitors in the former place while reducing them in the latter place. There exists a pecuniary externality in so far as, under imperfect competition, the relocating firm disregards those effects. Specifically, the firm imposes a positive externality on its competitors in its place of origin and a negative externality on its competitors in its place of destination.

Rather than stressing the role of imperfect competition and pecuniary externalities, approaches other than NEG prefer to focus on 'technological externalities'.[6] These are independent from any market interaction as they materialize through sheer physical proximity. Being the outcome of non-market interactions, also for them, by definition, no quid-pro-quo is paid. This is the case whenever the productivity of a firm is influenced by the presence of other firms nearby even though no market relation is established with them. Also technological externalities can be either negative or positive. For example, nearby firms may reduce a firm's productivity through the pollution they generate or through the congestion they cause in the use of local public goods and infrastructures. On the other side, nearby firms may increase a firm's productivity through informal knowledge transmission ('spillover') generated as a by-product of their contacts with the surrounding environment.[7]

To sum up, no matter whether through market or non-market interactions, the geographical distribution of other firms determines the relative attractiveness to a firm of alternative locations. This creates a feedback mechanism among firms' location decisions through which firms' interactions ('second nature') may alter the economic landscape implied by natural resources, natural means of communication, and climatic conditions ('first nature').[8]

Building Block 3 – Firms' location decisions jointly generate localized externalities that determine regional attractiveness.

Moreover, since 'second nature' is driven by externalities, the free-market economic landscape is generally inefficient and appropriate public intervention is generally needed. Once more, this is true no matter whether the externalities are pecuniary or

[6] This point has been raised quite forcefully by Marshall (1890). See Henderson (1978) as well as Ciccone and Hall (1996) for recent reassessments.

[7] The distinction between pecuniary and technological externalities is due to Scitovsky (1954).

[8] Accordingly, traditional trade theories in the wake of Ricardo, Heckscher and Ohlin can be interpreted as stressing 'first nature' aspects.

technological. Nonetheless, pecuniary externalities do have a logical advantage with respect to technological externalities. Their advantage lies in the possibility of relating their emergence to a set of well-defined microeconomic parameters as the next section will show. So far that has proven to be quite difficult in models based on technological externalities as these still remain mostly 'black boxes'.[9]

2.2. Micro-founded agglomeration

As already discussed, the main tenet of NEG is that the evolution of the economic landscape is mainly driven by pecuniary externalities. These are generated by market interactions among imperfectly competitive firms that make their profit-maximizing choices with three objectives in mind: proximity to customers and suppliers, concentration of production in few plants, and distance from competitors. The key insight of NEG is that such choice is not trivial not only because the three objectives are often in conflict but also because their relative impact on profits depends on a set of underlying industry characteristics.

To understand this point, consider a production chain in which there are three vertically linked activities: intermediate production, final production, and consumption.[10] For simplicity, assume that final production uses only intermediate inputs, intermediate production employs only labour, workers are the only source of final demand and they are geographically immobile. If, for any reason, a new firm starts producing intermediates, it will increase labour demand and intermediate supply. Due to excess demand and supply respectively, wages will go up while intermediate prices will fall. This is bad news for the other intermediate producers ('market crowding effect' due to competitor proximity). However, it is good news for final suppliers, who experience falling production costs and higher demand by richer workers. As new final producers are lured to enter the market, the expansion of final production will feed back into stronger intermediate demand so that also intermediate suppliers will benefit ('market expansion effect' due to customer proximity). Clearly, when the latter effect dominates the former, both final and intermediate firms will end up being agglomerated in the same place. Accordingly, circular causation among firms' location decisions can generate persistent differences even among initially identical places ('second nature').

The crucial contribution of NEG is that such simple arguments are translated into general equilibrium models with solid microeconomic foundations. This allows the evolution of the spatial landscape to be related to observable microeconomic parameters: agglomeration is more likely to take place in sectors where increasing returns are intense and market power is strong. The reason is that more intense returns to scale and stronger market power weaken the market crowding effect.

[9] See, e.g., Ottaviano and Thisse (2001) as well as Duranton and Puga (2004) for a recent assessment.

[10] The example is borrowed from Ottaviano (2003) in the wake of Venables (1996).

Result 1 – Positive externalities are stronger in sectors with pronounced scale economies and strong market power. These sectors tend to be more clustered.

In other words, increasing returns and market power give strength to ‘second nature’ against ‘first nature’. This detaches the emerging economic landscape from the physical attributes of its underlying geography. Thus, there is a priori great flexibility on where particular activities locate. However, once the agglomeration process has started, spatial differences take shape and become quite rigid.[11]

2.3. Accessibility, attraction, and competition

In addition to the role of scale economies and market power, the most celebrated insight of NEG is probably the impact of transportation improvements and trade liberalization on geographical unbalances. The reason is that, with respect to alternative approaches, NEG adds a more detailed understanding of how the economic landscape evolves as trade impediments are gradually eliminated.[12] In particular, it argues that the level of trade impediments affects the balance between market expansion and market crowding effects in a non-linear way, thus changing non-linearly the relative importance of first and second natures in determining the spatial distribution of economic activities. As it will be discussed in the next section, the relative strength of second nature is maximized for intermediate levels of trade costs.

The basic concept underlying the analysis is the so-called ‘market potential’. This has both nominal and real definitions.[13] Whereas the ‘nominal market potential’ (henceforth, NMP) is a measure of customer proximity, the ‘real market potential’ (henceforth, RMP) is a combined measure of customer and competitor proximity.[14] Formally, consider a group of locations. The nominal market potential of a certain location A is the weighted average expenditures across all locations that plants can tap if located in A. Differently, the real market potential of A is the weighted average real expenditures (‘purchasing power’) across all locations that plants can tap if located in A. In both cases, the weight of each location is a decreasing function of its distance from A. The underlying idea is that NMP is a good proxy of the value of sales that plants can expect to make on average if located in A. Differently, RMP is a good proxy of the profits than an average firm can make if located in A. In the long run, since firms can freely pick plant locations, profits should reach the same normal level everywhere.

[11] This is what Fujita and Thisse (1996) call ‘putty clay’ geography.

[12] Fujita and Thisse (2002).

[13] The notion of ‘market potential’ is due Harris (1954) and has been recently refined by Head and Mayer (2004).

[14] The concepts of NMP and RMP are closely related to spatial interaction theory (Smith, 1975). The NMP of a certain area captures both the size of its local market (‘attraction’) and its connection to other markets (‘accessibility’). In addition, the RMP captures the intensity of competition faced by firms located in that area (‘repulsion’). Attraction, accessibility and, to a lesser extent, repulsion are also the main ingredients of gravitational models of international trade (e.g., Head and Mayer, 2004).

Therefore, in the long run RMP differences should eventually vanish as NMP differentials are capitalized in local price differences. Accordingly, short-run RMP differences should predict the future evolution of the economic landscape as firms are attracted towards areas temporarily boasting higher RMP.

Result 2 – The sales and the profits an average firm can make if located in a certain area are measured by the area's nominal and real market potentials (NMP and RMP) respectively. Differences in RMP across areas predict the future evolution of the economic landscape as firms are attracted towards higher RMP areas.

To sum up, according to NEG, understanding the geographical evolution of the economy requires understanding the causes of the changes in market potentials. It is in the wake of those changes that firms phase in and out their plants in different areas.

3. Countries

The traditional approach to international trade considers two countries only. Let us call them 'home country' and 'foreign country'. To focus on the role of international trade barriers, the analysis also abstracts from the internal geography of countries by assuming away any internal transportation cost. Finally, it rules out any labour mobility between countries, as very limited migration seems to be a common feature of the actual world.[15]

3.1. Home market effect

With two countries the determination of the nominal market potential is quite straightforward. To see this, consider two initially identical countries exhibiting the same levels of expenditures and the same numbers of plants. Now let expenditures grow exogenously in the home country.[16] Since firms have costless access to local customers but face trade barriers to reach external customers, the NMP (i.e. the distance-weighted average expenditures) and, in the short-run, also the RMP (i.e. the average profit) in the home country will grow. As profit rises, supply will expand until the resulting increase in competition brings profits back to their normal level. The opposite will happen in the foreign country. During the period of adjustment, the home country will grow faster as higher profit increase the return to investment in both physical and human capital accumulation as well as the return to innovation.[17]

[15] See Baldwin and Martin (1999) for an historical perspective.

[16] This could be caused, for example, by an increase in productivity due to technological progress.

[17] This implication is highlighted by Baldwin (1999) as well as Baldwin et al (2001) in the wake of Grossman and Helpman (1991). Monfort and Ottaviano (2004) show that faster human capital accumulation is also fuelled by higher participation to the labour force. This is matched by higher unemployment and vacancy rates as the expected return to skilled jobs rises.

An important implication of imperfect competition is that the resulting supply gap between the two countries will have to be larger than their expenditure gap in order to level the real market potential differential. Such an amplified impact of demand on supply is called ‘home market effect’: small national demand shocks can generate large international supply unbalances.[18] Starting from prohibitive trade barriers, that effect is initially strengthened by lower trade barriers as trade liberalization make the supply response to demand shocks more pronounced.[19]

As supply grows, the two countries undergo a complex process of industrial restructuring: new plants open, whereas old plants expand, contract, or even shut down depending on their relative productivities.[20] In the end, the larger country will eventually host more plants and, on average, these will be larger, more productive, more profitable even though their unit profit margin will be lower.[21] Also the composition across types of firms will change as in the home country larger expenditures foster inward foreign direct investment and multinational activity.[22]

Result 3 – Markets with higher NMP host more firms. These are larger, more productive and more profitable than firms in lower NMP areas.

3.2. Cumulative causation

The fact that, after the initial exogenous demand shock, the home country increases its stocks of both physical and human capitals creates the possibility of cumulative causation. The reason is that the additional income generated by newly accumulated capital feeds into additional expenditures. These trigger a second supply response via a second round of capital accumulation. If the market expansion effect due to new income is more pronounced than the market crowding effect due to new and enlarged firms, a self-sustaining cycle of income and expenditures growth may eventually arise in the home country. Analogously, a symmetric cycle of income and expenditures contraction may arise in the foreign country. As a result, small transitory country-specific shocks can give rise to large permanent international unbalances.

[18] Krugman (1980).

[19] Baldwin at al (2003) call ‘home market magnification’ the enhancing effect of lower trade barriers on the home market effect.

[20] See Helpman and Krugman (1985, 1989) for a survey of the effects of trade liberalization under imperfect competition.

[21] The selection effects of trade liberalization are modelled by Melitz (2003) with identical countries and fixed mark-ups. They are analysed in an multi-country setting with variable mark-ups by Melitz and Ottaviano (2003).

[22] Barba Navaretti and Venables (2004).

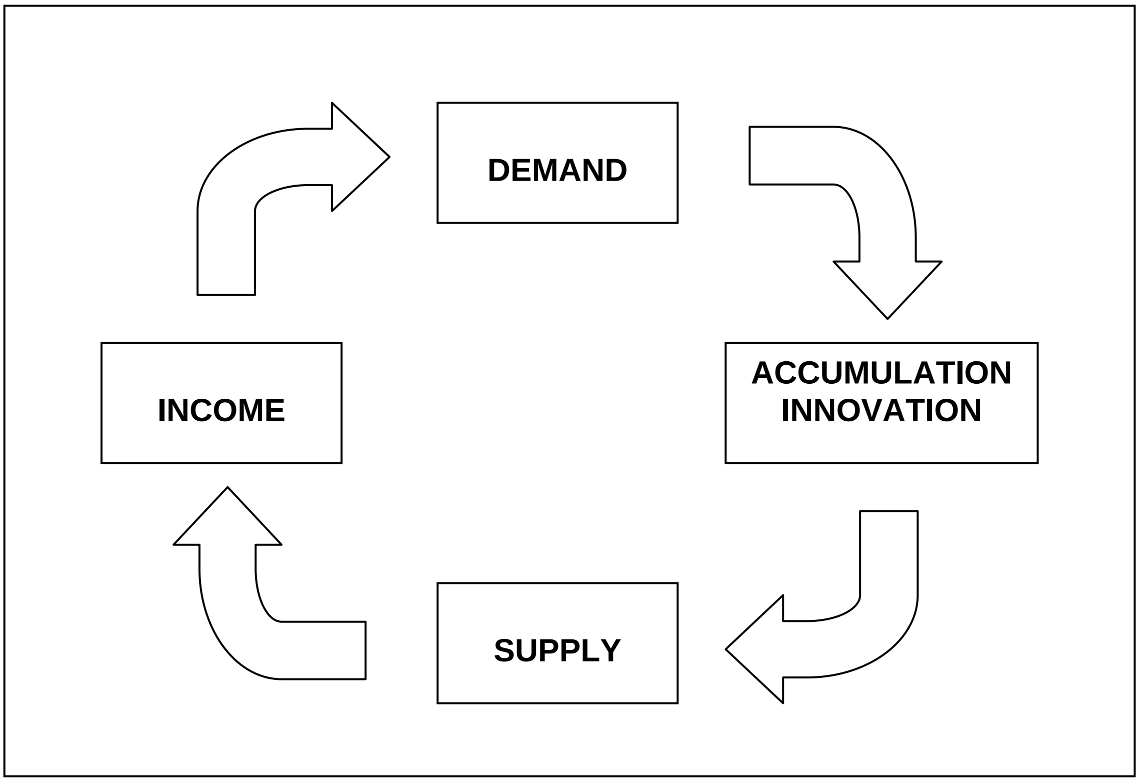


Figure 1 – Cumulative causation

Whether the market expansion effect is stronger than the market crowding effect or viceversa crucially depends on the level of trade barriers. In particular, the likelihood of circular causation is maximized for intermediate trade barriers.[23]

To understand this point, it is useful to go back to the example described in Section 2.2, which features a production chain with three vertically linked activities: intermediate production, final production, and consumption. Final production uses only intermediate inputs, intermediate production employs only labor, workers are the only source of final demand. Both intermediate and final firms are geographically mobile, whereas workers are not.

What makes the effect of declining trade barriers on agglomeration change sign below some threshold is precisely the presence of immobile workers. Their role is twofold. On the one hand, they generate localized labour supply. On the other hand, their expenditures also generate localized final demand. Therefore, as long as they are geographically dispersed, immobile workers create dispersed patterns of labour supply

[23] Puga (1999).

and product demand that hamper agglomeration by luring firms away from congested areas.[24]

Against this background, lower trade barriers make it easier for firms to reach dispersed demand without local production, thus weakening the anti-agglomeration impact of dispersed final demand. Differently, since immobile labour is non-tradable by definition, the level of trade barriers has no influence on the anti-agglomeration impact of dispersed labour supply. Thus, when trade barriers are high, the clustering of supply in the home country is hampered by the incentive that some firms still find in locating close to customers in the foreign country. When trade barriers are low, labour market pressures in the home country makes foreign location attractive. That is why agglomeration is sustainable only for intermediate trade barriers when there is scope for using location to boost firm market power.[25]

Result 4 – Initially international trade liberalization fosters cross-country agglomeration. However, further reductions in trade impediments trigger a reverse process of dispersion.

This is due the fact that, during the initial phases of international trade liberalization, positive externalities gain strength and firms tend to cluster.[26] As trade impediments are further reduced, externalities get weaker and clusters unfold. In other words, ‘first nature’ is dominant when countries are either isolated or highly integrated.

4. Regions

In order to study the evolution of the economic landscape within countries, NEG had to face two main difficulties. First, while labour immobility is a good approximation to reality at the international level, it is much less so at the regional level. Second, the processes of interregional and international integration may have very different impacts on the spatial distribution of economic activities. Of course, such possibility is hard to investigate through arguments that consider two locations only.

[24] More generally, the anti-agglomeration effect of labour immobility is stronger the larger the share of immobile workers in the labour force (Krugman, 1991).

[25] Krugman and Venables (1995) as well as Venables (1996) point out that cumulative causation is more likely in the presence of vertically-linked industries. The reason is that demand shocks and supply responses propagate and get amplified along the vertical production chain.

[26] Divergence may come as an abrupt change once trade barriers fall below a certain threshold value (‘break point’). Nonetheless, such a catastrophic behaviour should be probably considered as a rare event. The reason being that it is based on simultaneous identical decisions by firms and workers that require an extreme degree of homogeneity in tastes and technologies (Murata, 2003; Tabuchi and Thisse, 2003).

4.1. Labour mobility

The previous section has shown how the two-country set-up can be used as a fruitful analytical tool when the question is how international trade liberalization affects the location of industry between countries disregarding what happens within them. Most naturally, the same tool can be used when the question is how interregional trade liberalization affects the location of industry within countries disregarding what happens between them. In this case the above results carry through virtually unchanged with a single major caveat: labour cannot be considered immobile at the regional level.[27]

To understand the implications of labour mobility, recall one of the key insights of NEG: capital accumulation and innovation can give rise to self-enforcing agglomeration even between initially similar countries or regions. The reason is cumulative causation through which higher expenditures stimulate capital accumulation and this feeds back into even higher expenditures through the associated increase in income. For intermediate trade costs, such a market expansion effect dominates the market crowding effect due to the increased number and sizes of firms that a larger capital stock allows to operate.

When workers are mobile, capital accumulation is not necessary for cumulative causation to take place. The reason is that income differences can be driven by migration.[28] Consider again an initial situation with two identical regions that is altered by a positive demand shock to one of them. Assume that production is labour intensive. Specifically, for the sake of argument, assume that production employs only labour and no capital. As before, the demand shock will create an incentive for supply to expand in the shocked region. However, output expansion will require additional employees. This will push wages up, thus attracting workers from the other region. As workers immigrate, local income rises and this feeds back into higher expenditures. The larger the immigration flow for a given wage differential, the more local expenditures expand, which makes cumulative causation more likely.

Result 5 – Labour mobility fosters regional agglomeration.

This is true whatever the intensity of scale economies, the strength of market power, and the level of interregional trade impediments. Moreover, as skilled workers are typically more mobile than unskilled ones, skill-intensive sectors tend to be more clustered.[29]

[27] The distinction between countries and regions in terms of labour mobility dates back at least to Ohlin (1933).

[28] Indeed, in NEG’s seminal paper by Krugman (1991) cumulative causation is sustained by labour migration rather than capital accumulation.

[29] Forslid and Ottaviano (2003), Ottaviano et al (2002).

4.2. Transport infrastructures

In addition to labour mobility, studying the effects of regional integration raises the issue of considering a more realistic description of the geographical space. This is achieved by investigating the behaviour of multi-location economies, which simultaneously gives substance to the two ideas. On the one hand, countries and regions do have an internal spatial dimension. On the other hand, real world phenomena do involve many regions and many countries at the same time.

Having more than two locations does not affect the clustering effects of more intense scale economies and stronger market power. It does not affect either the non-linear impact of trade liberalization as clustering is still likelier for intermediate trade impediments. Nonetheless, when locations belong to different countries, one has to distinguish between international and interregional trade liberalization while also keeping in mind that labour mobility is negligible at the international level but much less so at the regional level. In this more complex scenario, agglomeration within countries is mainly shaped by interregional trade impediments. Vice versa, agglomeration between countries is mainly shaped by international trade barriers.[30]

Result 6 – Initially the implementation of interregional transport infrastructures fosters cross-region agglomeration. However, further improvements in transportation trigger a reverse process of dispersion.

Dispersion may, however, come at the cost of slower innovation and slower capital accumulation. Indeed, that would happen whenever skill-intensive sectors benefit from positive technological externalities, such as localized ‘knowledge spillovers’, whose work is undermined by the geographical dispersion of plants, labs and skilled workers.[31] This adverse effect could be offset if better transport infrastructures improved the international attractiveness of the national market.[32]

4.3. Hubs and gates

Another insight that scholars have gained from the study of multi-location models is that the home market effect does not generally survive scrutiny.[33] This is due to the fact that, even in the presence of a third location only, an increase in one location’s

[30] Krugman and Livas (1996), Monfort and Nicolini (2000), Paluzie (2001), Crozet and Koenig-Soubeyran (2002), Behrens et al (2003). Similar results hold true in the absence of interregional migration whenever firms are linked by strong input-output ties (Puga and Venables, 1997; Monfort and Van Ypersele, 2003).

[31] Martin (1999), Braunerhjelm et al. (2000), Manzocchi and Ottaviano (2001).

[32] Martin and Rogers (1995).

[33] See, e.g., Behrens et al. (2004) for the multi-country extension of the two-country model by Krugman (1980).

expenditure share may well map into a less than proportionate increase in its output share as the third location drains away some firms. In more extreme cases, an increase in the expenditure share of a location may even lead to a decrease in its output share. This is the case, for instance, when one of the locations is a 'transport hub', that is, a location with better accessibility to all other locations. As a result of international economic integration, the main transport hubs tend to coincide more and more with the so-called 'gates', that is, locations through which goods mostly flow in and out of a country. The presence of a hub or a gate implies that a positive demand shock to any other location may result in supply expanding in the hub or the gate and contracting elsewhere. Therefore, agglomeration is more likely to take place in the presence of hubs and gates.[34]

Result 7 – The presence of 'transport hubs' and 'gate regions' makes cross-region agglomeration more likely.

Once more, firms cluster only for intermediate levels of interregional trade impediments. When this is the case, clustering takes place in the gate region.[35] The clustering of firms in hubs and gate regions happens through intense industry restructuring. As less productive firms are competed out, surviving firms end up being more productive, bigger, and more profitable. This selection process improves local aggregate performance as average productivity grows and average prices fall.[36]

While the home market effect does not survive scrutiny, other related predictions remain valid. The first is the so-called 'dominant market effect'. This implies that a location with a sufficiently large expenditure share attracts all firms in sectors characterized by scale economies and imperfect competition. The second prediction is the 'magnification effect', according to which, starting with prohibitive trade barriers, freer trade initially leads to a more uneven spatial distribution of those sectors.

5. Globalization

The previous sections have presented the detailed logic and insights of NEG. To prepare the discussion of its empirical relevance (Section 7), it is now useful to provide a synthesis. An effective way to proceed is by referring to the traditional taxonomy of regional studies, according to which the alternative explanations of the spatial distribution of economic activities can be classified in terms of the relative weights they give to a set of 'centripetal' and 'centrifugal' forces. The taxonomy will then be used to assess the impact of globalization on the economic landscape.

[34] Krugman (1993).
[35] Behrens at al (2003).
[36] Melitz and Ottaviano (2003).

5.1. Centripetal and centrifugal forces

Centripetal forces are all the effects that foster the geographical agglomeration of firms and workers. According to the traditional taxonomy there are three main centripetal forces:[37]

- *Market-size effects*. As firms and workers cluster in a certain area, the local market expands. This makes the cluster attractive in terms of both customer proximity (‘demand linkages’) and supplier proximity (‘cost linkages’). In both cases, by clustering, firms generate positive pecuniary externalities.
- *Matching effects*. The density of firms and workers in a cluster makes search and matching of complementary needs easier. This reduces the expected search costs. Thus, by clustering firms generate positive technological externalities.
- *Spillover effects*. The density of economic activities in a cluster creates informational spillovers benefiting all local firms and workers. It also makes it easier for firms and workers to benchmark each other performances. In both cases, by clustering firms generate positive technological externalities.

If only centripetal forces were at work, the final result would be a single huge cluster. Of course, in reality that does not happen because the expansion of a cluster is limited by centrifugal forces:

- *Factor market-crowding* effects. The clustering of firms in a certain area increases the local prices of immobile factors, such as land, natural resources, and to some extent also labour especially if unskilled. Higher factor prices increase the local production costs, thus limiting the process of agglomeration.
- *Product market-crowding* effects. The presence of many firms in a cluster makes local competition fierce. By cutting into firms’ revenues and profits, competitive pressures limit the dimension of the cluster, at least insofar as some customers and suppliers are tied to geographically disperse immobile factors.
- *Congestion effects*. The clustering of firms and workers in a certain area generates traffic, congestion, pollution, and crime. The associated additional costs of living and producing rise with the size and the density of the cluster, thus limiting its expansion.

5.2. Globalization and its impact

The basic message of NEG can be simply rephrased in terms of the above taxonomy: the level of ‘trade costs’ affects the balance between centripetal and centrifugal forces.

The expression ‘trade costs’ should be interpreted in a comprehensive way as all costs associated with the exchange of goods and factors among agents located in different

[37] The three main centripetal forces are described by Marshall (1890), hence they are also known as the ‘Marshallian triad’. In the wake of Maignan et al (2003), the presentation of those forces as well as the subsequent discussion of centrifugal forces is adapted from Krugman (1998) and Venables (2001).

places. Some of these costs are due to the sheer existence of distance (e.g., the costs of transportation and communication), others arise from institutional barriers (e.g., the costs due to tariffs or different quality and safety standards) or even from linguistic and cultural differences (e.g., the costs of communication again or those due to different business practices). Since globalization is dramatically reducing all these costs, it has the potential impact of altering the current equilibrium of centrifugal and centripetal forces, and therefore of re-designing the existing economic landscape. This may happen through various channels:[38]

- *Search and matching costs*: identifying a potential trading partner.

  Lower communication costs significantly reduce the search and matching costs. This weakens the positive matching effect of agglomeration and, thus, the associated centripetal force. Such impact is particularly relevant for dynamic skill-intensive sectors in which the complementary needs of firms, workers and customers change rapidly.

- *Direct shipping costs*: moving inputs and outputs.

  Lower transport costs and lower institutional barriers decrease the delivery costs of goods and services. On the one hand, that weakens the positive market-size effect of agglomeration as the relevance of customer and supplier proximity falls. On other hand, that also weakens the negative market-crowding effect of agglomeration. In product markets, lower delivery costs make the intensity of competition increasingly independent of actual location. In factor markets, they make factor prices increasingly independent of the actual production site. However, as already discussed, while shipping costs fall, the market-crowding effect weakens faster than the market-size effect.

- *Control and management costs*: monitoring and management.

  Lower transport costs, lower institutional barriers, and lower communication costs affect the internal organization of firms by making it easier to split production and administration into spatially different units. Such a geographical fragmentation allows a firm to choose the locations of the different stages of the production process independently according to their specific needs. This makes the different forces operate at the level of the single stage of production rather than at the level of the firm as a whole. Accordingly, clusters become increasingly specialized with the same firms placing different production stages in different clusters.

- Costs of personal interactions: knowledge spillovers

  Lower communication costs foster personal interaction and knowledge transmission beyond geographic proximity. This weakens the positive spillover effects of agglomeration, thus weakening the corresponding centripetal force.

- *Costs of* relocation: changing location

[38] The discussion is structured along the lines drawn by Venables (2001) in a different context.

> Lower transport costs, lower institutional barriers, and lower communication costs make firms increasingly footloose. To some extent, this is true for workers too, especially for skilled ones. This does not affect directly any of the centripetal and centrifugal forces per se. However, it makes firms and workers more reactive to any change in those forces, which puts additional pressure on public policies to deliver.

To sum up, according to NEG:

Globalization - Globalization can be expected to have a non-linear effect on the degree of geographical agglomeration of economic activities. Initially, lower transport costs, lower institutional barriers, and lower communication costs foster agglomeration. As all those costs become negligible, agglomeration unfolds.

Agglomeration will be more pronounced and more persistent in sectors characterized by: intense scale economies, strong market power, tight input-output relations, higher relative intensity of mobile than immobile factors (such as capital and skilled labour versus land and unskilled labour), rapidly changing products and tasks (as in hi-tech industries), high value added (that is, small congestion cost per euro produced).

Admittedly, to learn that globalization is expected to have a ‘non-linear effect’ on the degree of geographical agglomeration is too vague to be helpful. Moreover, most NEG models abstract from some features that many observers consider as essential characteristics of globalization (e.g., fragmented production processes by multi-plant firms, far-flung multi-modal supply chains that cross borders many times, intensive use of ICT and logistical services, etc.). Accordingly, some authors have felt the urgency of tightening the implications of theoretical speculation. A first step in this direction is the calibration of ‘agglomeration ranges’, that is, intervals of trade barrier values that, according to simple NEG models, should support the agglomeration of different sectors. Such ranges have been compared with estimated values of current trade barriers. When run on bilateral data for Canada and the US or France and Germany, such experiments show that most industries are closer to the lower end of the agglomeration range, where more trade integration would lead to more agglomeration. While the calibration of ‘agglomeration ranges’ is still in its infancy, it represents an promising attempt to extract tighter predictions from NEG models.[39]

6. Welfare

The previous section has discussed the preditions of NEG on the effects of globalization on the geographical distribution of economic activities. However, the crucial policy questions have remained so far unanswered: Is agglomeration desirable from a social welfare point of view? Should policy makers foster or control it?[40]

[39] Head and Mayer (2004).

[40] Baldwin et al (2003).

The answers are not straightforward as they involve both efficiency and equity considerations. Indeed, while the distinction between equity and efficiency is fundamental, it is often misunderstood in the policy debate. If one pictures the welfare of the economy as a pie, equity is about the relative sizes of the slices that go to different people, irrespective of the overall size of the pie. On the contrary, efficiency is about the overall size of the pie, irrespective of the sizes of the slices of different people. Thus, under an *equity* perspective, one identifies the winners and the losers from agglomeration. Under an *efficiency* perspective, one evaluates whether the winners gain enough to be able to compensate the losers.

6.1. Equity

In terms of equity, the crucial distinction is between mobile and immobile people.[41] Mobile people, who are typically young and skilled, are the winners from agglomeration. They can take care of themselves by moving to the areas that provide them with the best working conditions and the highest quality of life. When clustering in core areas, they enjoy the associated benefits: richer variety and quality of (both private and public) goods and services, lower prices for tradables, more productive jobs, better matching in the labour market. All these benefits are capitalized in higher prices of non-tradables (such as land).

Immobile people, who are typically old and unskilled, are the losers from agglomeration. When mobile people cluster in core areas, those who cannot follow are left behind in peripheral areas facing poorer variety and quality of (both private and public) goods and services, higher prices for tradables, less productive jobs, and worse matching in the labour market. All these disadvantages are capitalized in lower prices of non-tradables.

6.2. Efficiency

There are two main ways for policy makers to take care of those who are left behind. One way is to hamper agglomeration. This can be achieved through the direct control of migration flows as in China, or through subsidies to peripheral location as in the EU. The alternative way is to allow for agglomeration and then redistribute some of the associated gains from winners to losers.

Which way to go depends on the specific economic activities involved in the agglomeration process. As discussed in Section 4.2, skill-intensive sectors benefit from positive technological externalities, such as localized 'knowledge spillovers', whose work is undermined by the geographical dispersion of plants, labs and skilled workers. In those sectors dispersion is obtained at the cost of slower innovation and slower capital accumulation. When that is the case, allowing for agglomeration and

[41] Ottaviano and Thisse (2002).

redistributing its gains is the socially desirable (i.e., most efficient) way to deal with regional disparities: agglomeration achieves efficiency, redistribution supports equity.[42]

Things are more complicated in the presence of pecuniary externalities. First, when efficient agglomeration is driven by pecuniary rather than technological externalities, the foregoing redistributive strategy may not be viable. The reason is that, if market interactions are the driving forces of agglomeration, any relevant redistribution of income from centre to periphery is bound to lead to a more even spatial distribution of economic activities, hence reducing the efficiency gains from localized pecuniary externalities.[43]

Second, with pecuniary externalities, whether agglomeration is efficient or not depends on the level of trade costs. This stems from the non-linear relation between the level of trade costs and the strength of pecuniary externalities underlying Results 4 and 6. Specifically, the free market outcome is socially desirable when trade costs are either high or low. In the former case activities are dispersed, in the latter they are agglomerated. For intermediate trade costs, however, the market delivers agglomeration whereas dispersion is efficient. In this case, the equity-efficiency trade-off disappears: efficiency is achieved through equity and viceversa.[44]

This suggests that whether efficient regional intervention should hamper agglomeration or simply redistribute some of the associated gains depends on spatial scale. Indeed, low trade costs may be viewed as corresponding to shipping costs between locations belonging to different small-sized areas. Large costs would instead be the counterpart of shipping costs between locations belonging to different large-sized areas. Intermediate values would, therefore, correspond to shipping costs between locations belonging to different medium-sized areas. This interpretation implies that efficient regional policy should aim at reducing agglomeration between medium-sized areas only, otherwise confining itself to simple redistribution.

To summarize, according to NEG:

Welfare – The agglomeration of economic activities in core areas damages immobile people in peripheral ones. The most efficient way to take care of the periphery depends on whether agglomeration is driven by localized market or non-market interactions and on the level of trade costs. When non-market interactions (e.g. ‘knowledge spillovers’) dominate and, in any case, when trade costs are either high or low, policy makers should achieve efficiency by allowing for agglomeration while pursuing equity through

[42] In the limit, when the positive impact of agglomeration on innovation is strong enough, no redistribution is actually needed as very fast growth in the core improves the welfare of the periphery through a strong (Ricardian) terms-of-trade effect. In such case, as both the core and the periphery gain, agglomeration dominates dispersion in the sense of Pareto. See Martin (1999), Braunerhjelm et al. (2000), Manzocchi and Ottaviano (2001).

[43] In the absence of technological externalities, only strong vertical linkages among firms can rule out this win-lose situation (Carlot et al, 2004; Ottaviano and Robert-Nicoud, 2004).

[44] Ottaviano and Thisse (2002).

interregional redistribution. By contrast, when market interactions dominate and trade costs are intermediate, agglomeration should be hindered on both equity and efficiency grounds.

7. Evidence

The empirical assessment of NEG is still at an infant stage and no conclusive evidence is available yet. This has been mainly due to the gap between theoretical and applied investigations: theorists have shown little interest in translating their insights in clear-cut testable predictions; empiricists have made little effort in understanding what theory exactly implies.[45]

As a result, empirical results on NEG are quite patchy. Pieces of evidence are scattered across many studies, often hidden as by-products of analyzes with completely different focuses. The aim of the present section is to compose these pieces within a coherent framework, while knowing that in the end the puzzle will still be incomplete.

7.1. Market potential

As discussed previously (Results 2, 3, and 7), the crucial concepts underlying NEG are the ‘nominal market potential’ (NMP), which captures customer/supplier proximity, and the ‘real market potential’ (RMP), which captures both customer/supplier proximity and competitor proximity. The former predicts the sales that firms can make if located in certain area. The latter predicts the profits than firms can make if located in that area. In the long run, since firms can freely pick plant locations, RMP differences should eventually vanish as NMP differentials are capitalized in local price differences.

A similar argument can be applied to labor after realizing that higher sales and profits are typically associated with higher nominal and real wages. Accordingly, NMP predicts the nominal wages that workers can earn if employed in certain area, whereas RMP predicts the real wages than workers can make if located in that area. In the long run, if workers can freely relocate, real wage differences should eventually disappear as nominal wage differentials are capitalized in local price differences.

The foregoing predictions identify two natural tests of the empirical validity of NEG arguments.[46] On the price side, higher NMP should be associated with higher revenues and higher nominal wages both in the short and the long runs. It should also be associated with higher local prices in the long run, especially in the presence of labor mobility. On the quantity side, positive shocks to NMP should attract both firms and workers.

[45] See the discussion in Head and Mayer (2004).

[46] See Head and Mayer (2004) for a detailed survey.

7.1.1. Price effects

The price predictions have been tested at both international and interregional levels. In cross-country studies labour mobility is negligible and capital mobility limited, which means that RMP differentials do not vanish even in the long run. Accordingly, across countries higher RMP (as well as higher NMP) should be related to higher profits and wages. When brought to the data, these predictions are quite successful: RMP variations explain around 35 per cent of the cross-country income variation. This result is independent of institutions, natural resources, and physical geography. In other words, 'second nature' considerations matter irrespective of 'first nature' attributes. Interestingly, a country's access to the coast raises the local nominal wage by over 20 per cent, which reveals the dominant role of gate regions. [47]

In cross-region investigations, labour mobility plays an important role. This implies that real wages should equalize across regions in the long run. In other words, in the long run NMP-driven nominal wage differences should be capitalized in local price differentials. These differentials are essentially determined by the interregional variations in the prices of non-traded goods and services with a dominant role played by land values. Therefore, higher NMP should be associated with both higher wages and higher land rents. This prediction finds indeed empirical support.[48] Cross-region studies also highlight the dominant role of transport hubs and gates: a 10 per cent increase of the distance from them reduces the nominal wage by 1-2 per cent.[49]

The fact that, with labour mobility, wages and rents are both positively correlated with NMPs can be interpreted as evidence that pecuniary externalities generate higher productivity in areas that offer better customer and supplier proximity. The argument is the following. In principle, mobile workers could command higher wages when employed in a certain area for two different reasons. First, they may dislike the area ('disamenity'). Second, they may be more productive when employed by firms located in that area. However, nobody would ever pay a higher rent to live in a place she dislikes. Thus, higher wages and higher rents must signal higher productivity.

Evidence 1 – In countries with higher market potentials, wages are higher. In regions with higher market potential also rents are higher. When labour is mobile, higher wages and higher rents are associated with higher productivity.

Higher average productivity is due to the availability of cheaper and more varied intermediate inputs. As discussed in Section 3.1, it also stems from the selection caused by competitors' density, which makes more productive firms thrive.[50]

[47] See, e.g., Redding and Venables (2000) for an investigation of 101 developed and developing countries in 1996.

[48] See, e.g., Hanson (1998) for a study of US counties from 1980 to 1990.

[49] See, e.g., Hanson (1997) for a study of Mexico from 1965 to 1988.

[50] Syverson (2002), Campbell and Hopenhayn (2002).

7.1.2. Quantity effects

The quantity predictions stem from the idea that local shocks to final demand or intermediate supply generate short-run RMP variations. The associated variations in profits and real wages cause the relocation of firms and workers, which move towards higher NMP and temporarily higher RMP areas. In the long run NMP differences persist, while RMP differences disappear as firms and workers crowd higher NMP areas.

As to firms, most studies target what is considered the relatively footloose part of their activities: foreign direct investment (FDI).[51] The focus on FDI is crucial in that, whenever their impact on local market conditions is negligible, the spatial allocation of foreign plants can not be expected to lead to RMP equalization even in the long run. In general, FDI analyzes show that foreign firms indeed favour locations with higher RMP. In so doing, they take into account both customer and supplier proximity. According to the estimated impact, a 10-per-cent rise in RMP yields a 10.5-per-cent increase in the probability of a region being chosen by foreign investors.

As to workers, the number of studies addressing the impact of customer and supplier proximity is very small. Preliminary results suggest that migrants respond to RMP differentials in the predicted way. However, their response is limited by distance, which signals the dampening effects of distance-related mobility costs and migration barriers.[52]

Evidence 2 – Firms and workers are attracted to higher market potential areas.

To sum up, the empirical literature that closely matches the theoretical predictions based on market potentials and specific statistical tests is still quite thin. Nonetheless, the existing results support the insights of NEG.

7.2. Trade barriers

As discussed previously (Results 4 and 6), NEG arguments imply a non-linear effect of trade liberalization on the geographical agglomeration of economic activities. Initially, lower trade costs foster agglomeration. As those costs become negligible, agglomeration unfolds.

Since trade costs have declined over time due to both improvements in the transport technology and, after the end of WWII, reductions in trade barriers, most naturally some scholars have tried to investigate their impact on agglomeration by simply observing the evolution of industrial location over time. In the US the spatial concentration of

[51] Coughlin et al (1991) study the location decision of all foreign investors across US states. Head et al (1999) concentrate on Japanese firms only. Head and Mayer (2002) analyze the behaviour of Japanese firms across European regions.

[52] See, e.g., Crozet (2000) for a study of European regions, which shows that a region with 100 Km radius attracts workers within a radius of no more than 120 Km.

manufacturing across states fell until 1900, then rose to a peak around 1927, and finally declined again until 1987 to reach its level in 1860.[53] In the EU the geographical concentration of manufacturing across countries rose sharply between 1972 and 1996 with a slowdown after the start of the Single Market Programme in 1986.[54] While these results are broadly in line with NEG predictions, they are hard to interpret as evidence of any clear-cut impact of trade costs on agglomeration. Indeed, any interpretation in that direction would rely on the implicit assumption that no other variable has affected industry location over time.

So-called ‘concentration regressions’ take a more direct approach by regressing alternative indices of geographical concentration on different measures of ‘trade costs’ (such as administrative barriers, geographical size - larger areas imply greater average distances -, expenditures on transport and communication as well as road/railway/communication density). In so doing, they control for the potential impact of additional variables (such as development stages, industrial compositions, and institutions). The analysis is typically cross-country. Some studies focus on the effects of external trade barriers on cross-country agglomeration. They find results on transactions costs that are inconclusive and somewhat contradictory.[55] Other studies focus instead on the effects of internal and external trade barriers on within-country agglomeration. Their general result is that agglomeration is more pronounced when both external and internal interactions are harder. This would be consistent with NEG in so far as the average integration of the sampled countries is low enough.[56]

Evidence 3 – Agglomeration is more pronounced for intermediate than for high/low trade costs.

To sum up, there is some evidence of a non-linear relation between economic integration and agglomeration. However, the evidence is far from conclusive. The most important caveat concerns spatial aggregation problems. Specifically, the above results obtained at the national level may hide even opposite results at the regional level as concentration indices are sensitive to the spatial scale of analysis.[57]

7.3. Sector characteristics

A possible reason why the foregoing evidence is not conclusive is the high level of aggregation of the analysis. Accordingly, attention has been increasingly devoted to more disaggregated data. In section 6.2 (building on Results 1 e 5), the sector implications of NEG have been summarized as stating that agglomeration should be more pronounced and more persistent in sectors characterized by: more intense scale

[53] Kim (1995).
[54] Brülhart (2001).
[55] Combes and Overman (2004).
[56] Ades and Glaeser (1995), Rosenthal and Strange (2001).
[57] Ellison and Glaeser (1997).

economies, stronger market power, tighter input-output relations, higher relative intensity of capital and skilled labor, faster innovation, and higher value added.

Most of these features have been investigated. First, the majority of studies find a positive correlation between increasing returns and agglomeration.[58] Second, input-output linkages have a fairly robust positive correlation with agglomeration.[59] Third, there is little evidence that labor, capital or resource intensive activities are more agglomerated. Fourth, technology intensive and science-based industries are more agglomerated than average.[60] Finally, trade costs have still a mixed impact on agglomeration.[61] Nevertheless, trade liberalization enhances average productivity through firm selection due to tougher foreign competition.[62]

Evidence 4 –Agglomeration is more pronounced in sectors that exhibit stronger scale economies as well as tighter input-output linkages, and that are technology intensive as well as science-based.

7.4. Spillovers

The fact that technology intensive and science-based industries are more agglomerated than average is consistent with NEG predictions. At the same time, as argued in Section 2.1 (Building Block 3), that fact is also consistent with the presence of localized spillovers only. If that were the case, however, the positive correlation between market potentials and agglomeration should vanish once the impact of spillovers were also taken into account. While this does not happen, localized spillovers do play a role on their own right.[63]

Two main research strategies have been devised in order to assess the relevance of spillovers. A first approach exploits the information that can be indirectly extracted from wage and price variations as in the case of market potentials. A second approach captures the presence of spillovers directly in terms of knowledge creation.

7.4.1. Wage and rent gradients

Localized spillovers make firms and workers more productive when geographically clustered. Accordingly, local shocks to the density of economic activities generate short-run geographical variations of profits and real wages with more productive areas offering higher profits and higher real wages. In the long-run, as firms and workers move to those areas, their local prices rise until the geographical variations of profits and real wages disappear. In the end, productivity differences are entirely capitalized in

[58] Kim (1995), Combes and Overman (2004).
[59] Amiti (1999), Ellison and Glaeser (1997).
[60] Brülhart (1998), Haaland et al (1999), Combes and Overman (2004).
[61] Haaland et al (1999), Brülhart (2001).
[62] Tybout (2002).
[63] Head and Mayer (2002).

local price differences. Therefore, the key question becomes: Are wages and prices higher in areas with a high density of firms and workers? A positive answer would reveal a productivity-enhancing spillover. Moreover, should wages and prices be positively associated with the density of human capital, there would be specific evidence of a productivity-enhancing knowledge spillover.

In general, both skilled and unskilled wages do tend to be higher in locations where the labour force is more educated. The quantitative effect is not negligible. A one-year increase in local average education increases the average wage by 3 to 5 per cent. A one-per-cent point increase in the local share of college educated workers raises the average wage by 0.6 to 1.2 per cent. At the same time, the presence of more educated workers is associated with higher local prices. As argued above, that signals the presence of productivity-enhancing knowledge spillovers.[64]

Some insight on the channels through which non-market knowledge transmission takes place can be gauged from the relative behaviours of young and old workers. The former are paid less than the latter in denser areas such as cities. Yet, they are over-represented in those areas. The fact that young people accept lower wages in denser areas indicates that they value the learning opportunities density offers. As people get older, the expected return to learning falls. Accordingly, they give more weight to the congestion costs associated with density and leave to less dense areas.[65] More generally, in decreasing order of importance, learning spillovers, better matching between firms and workers, and selection effects are all responsible of the wage premia observed in denser areas.[66]

Evidence 5 – In regions with higher densities of firms and workers, wages and rents are higher. When labour is mobile, this is associated with productivity-enhancing spillovers.

Some scholars have also been able to measure the distance decay of spillovers. Non-market knowledge transmission between two individuals vanishes starting from 90-minute-trip distances.[67]

7.4.2. Knowledge creation

The second approach to spillover measurement targets the process of knowledge creation itself. Such process is modelled through knowledge production functions.[68]

A knowledge production functions explains the output of innovation (e.g., patents) in terms of knowledge inputs (e.g., R&D spending and human capital). In the real world such explanation works at the level of areas and industries, but it does not work at the level of firms. This can only happen if firms in an area benefit from research carried out

[64] Rauch (1993), Moretti (2004).

[65] Peri (2002).

[66] Glaeser and Maré (2001), Combes et al (2004).

[67] Conley, Flyer and Tsiang (2004).

[68] See Audretsch and Feldman (2004) for a detailed survey.

by other institutions (universities or firms) located in the same area, therefore pointing out the existence of localized knowledge spillovers. This phenomenon is particularly evident in the case of small firms. These are able to generate innovative output with negligible amounts of R&D by exploiting the knowledge created in universities and large corporations.[69]

As in the case of wage and rent gradients, the positive impact of spillovers appears to fade away quite rapidly with distance. This is revealed by analyzing the location pattern of patent families (i.e., patents that reference or cite each other). Indeed, the probability of cross-citation is much higher when inventors come from the same area, which suggests that cross-fertilization is highly localized. Thus, proximity clearly matters in exploiting knowledge spillovers.[70]

Evidence 6 – The productivity-enhancing impact of spillovers fades away quite rapidly with distance.

Some measure of the overall impact of knowledge spillovers on plant productivity is available. Each year, the contribution of spillovers to aggregate output growth is 0.1 per cent. The estimated effect comes essentially from high-tech plants, as it is virtually zero in low-tech plants.[71]

7.5. Growth

The impact of human capital on output growth has attracted a lot of attention.[72] Indeed, cross-country, cross-region, and cross-city studies generally detect a robust positive correlation between per-capita income growth and the initial level of human capital.

The standard tool of analysis is the ‘growth regression’, which explains per-capita output growth in terms of human capital and a long list of other variables. These can be partitioned in two broad groups, ‘proximate sources of growth’ and ‘wider influences’.[73] In addition to human capital, proximate sources of growth are physical capital and R&D. The evidence confirms that all proximate sources are important: higher investments in human capital, physical capital and R&D all lead to faster long-run growth. This appears to be true across countries, regions and cities.[74] Besides the proximate sources, a variety of ‘wider influences’ affect growth indirectly by improving knowledge and technology transfer as well as the efficiency of input allocation. Government spending (overall size and composition), infrastructures, and socio-political factors are examples of such wider influences.

[69] Acs, Audretsch, Feldman (1994).

[70] Jaffe et al (1993), Jaffe and Trajtenberg (2002)

[71] Moretti (2002).

[72] Barro and Sala-i-Martin (1995), Sala-i-Martin (1996), Temple (1998), Durlauf and Quah (1999).

[73] Temple (1998).

[74] See Temple (1998) for a critical review of what has been assessed on the role of human capital.

Growth regressions provide specific results that complement the empirical evidence presented in the previous sections. First, they provide indirect support to the positive role of knowledge spillovers. Urban growth is faster in areas with a more diverse industrial base, the reason being that local diversity allows knowledge to spill over across industries.[75]

Second, growth regressions explicitly study the impact of labour mobility on the evolution of regional unbalances.[76] In particular, they show that the rate of convergence in income per capita across US states, Japanese prefectures, and European regions does not depend on the rate of migration.[77] This does not rule out the presence of localized externalities. The reason is that, as discussed in Sections 7.1.1 and 7.4.1, when firms and workers are mobile, real income differences vanish in the long run as the effect of localized externalities is capitalized in local price differentials. Hence, when workers migrate, the appropriate measure of local economic success is not income growth but rather population growth. Since there is no evidence of convergence in population growth, converging incomes are indeed consistent with diverging local productivities and localized externalities.[78] The distinction between income and population is more relevant for US states than EU regions as labour is much more mobile across the former than across the latter. This implies that per-capita income and unemployment rate differentials are much larger and more persistent in the EU than in the US.[79]

Finally, growth regressions highlight the presence of localized interactions also under an additional respect. Specifically, they find strong evidence that a region’s per-capita income level and growth depend not only on the region’s own characteristics, but also on the characteristics of other neighbouring regions. This creates spatial clusters of regions that are homogenous in terms of income levels and growth rates.[80]

Evidence 7 – Regions can be grouped in ‘convergence clubs’ depending on their long-run growth rates. Reciprocal distances play a role in determining club affiliation as closer regions tend to belong to the same clubs.

[75] See, e.g., Glaeser et al (1992) for growth in US cities.
[76] Magrini (2004).
[77] Barro and Sala-i-Martin (1995).
[78] Glaeser at al (1995). The absence of convergence is city sizes is a well-known phenomenon called ‘rank size rule’ or ‘Zipf’s Law’ (Gabaix and Ioannides, 2004).
[79] See Blanchard and Katz (1992) for the US as well as Bentivogli and Pagano (1999) for the EU.
[80] Quah (1997), Rey and Montuori (1999), Magrini (2004).

8. Conclusion

Among other effects, globalization is bound to change the world economic geography as we know it. From a NEG perspective three main drivers are particularly relevant:

- falling international and interregional trade costs;
- incomplete international and, to some extent, also interregional mobility of (unskilled) labour;
- increasing importance of knowledge in the production processes.

In terms of falling trade costs, theory predicts a non-linear relationship between trade costs and agglomeration. At country level, Result 4 shows that initially international trade liberalization fosters cross-country agglomeration. However further reduction in trade impediments triggers a reverse process of dispersion. Therefore, the effect of falling international barriers to trade will depend on whether agglomeration has reached or not its peak. The empirical evidence is not yet conclusive. At the regional level, Result 6 describes a similar bell-shaped relationship between trade costs and agglomeration. The empirical analysis based on ‘concentration regressions’ (see Section 7.2) suggests that, as external trade barriers keep on falling, further reductions in internal trade costs (e.g., due to improved infrastructure) will reduce the agglomeration of economic activities.

NEG argues that trade costs matter because they affect the appeal (‘market potential’) of regions in terms of proximity to customer, suppliers, and competitors (Results 2 and 3). For example, the creation of transport networks increases the appeal of hub and gate regions (Results 7). On the price side, NEG predicts that higher market potential should be associated with higher profits and higher nominal wages. Vice versa, on the quantity side, higher market potential should attract both firms and workers. These effects should be more pronounced in sectors characterized by more intense scale economies and stronger firm market power (Result 1). The empirical evidence supports these predictions.

As to labor mobility and agglomeration, NEG argues that labor mobility fosters agglomeration (Result 5). Moreover, as skilled workers are typically more mobile than unskilled ones, skill-intensive sectors should be more clustered. These predictions are strongly supported by the empirical evidence.

Finally, theoretical arguments and empirical evidence suggest that market interactions are not the only determinants of agglomeration processes. Local non-market interactions in the form of informal exchange of knowledge (‘knowledge spillovers’) are important contributors to innovation and growth. The growing knowledge intensity of production processes will increase the agglomerative impact of such spillovers. However, as the distance decay is much steeper for non-market than for market interactions, the impact will be felt more across regions than across countries.

All this has important welfare implications. The agglomeration of economic activities in core areas damages immobile people in peripheral ones. The most efficient way to take care of the periphery depends on whether agglomeration is driven by localized market or non-market interactions and on the level of trade costs. When non-market interactions dominate and, in any case, when trade costs are either high or low, policy makers should achieve efficiency by allowing for agglomeration and pursue equity through interregional redistribution. Differently, when market interactions dominate and trade costs are intermediate, agglomeration should be hindered on both equity and efficiency grounds.

References

Acs Z., D. Audretsch and M. Feldman (2004), “R&D Spillovers and recipient firm size”, *Review of Economics and Statistics*, n. 100, pp. 336-367.

Ades A. and E. Glaeser (1995), “Trade and circuses: explaining urban giants”, *Quarterly Journal of Economics*, n. 110, pp. 195–227.

Amiti M. (1999), “Specialisation patterns in Europe”, *Weltwirschaftliches Archiv*, n. 134, pp. 573–593.

Audretsch D. and M. Feldman (2004), “Knowledge spillovers and the geography of innovation”, in Henderson V. and J-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Baldwin R. (1999), “Agglomeration and endogenous capital”, *European Economic Review*, n. 43, pp. 253-280.

Baldwin R. and P. Martin (1999), “Two waves of globalization: superficial similarities, fundamental differences”, NBER Working Paper No. 6904.

Baldwin R., P. Martin and G. Ottaviano (2001), “Global income divergence, trade and industrialization: The geography of growth take-off”, *Journal of Economic Growth*, n. 6, pp. 5-37.

Baldwin R., R. Forslid, P. Martin, G. Ottaviano and F. Robert-Nicoud (2003), *Economic Geography and Public Policy,* Princeton University Press, Princeton.

Barba Navaretti G. and A. Venables (2004), *Multinational Firms in the World Economy*, Princeton University Press, Princeton.

Barro R. and X. Sala-i-Martin (1995) *Economic growth*, McGraw-Hill, New York.

Behrens K., C. Gaigné, G. Ottaviano and J.-F. Thisse (2003), “Interregional and international trade: Seventy years after Ohlin”, CEPR Discussion Paper No. 4065.

Behrens K., A. Lamorgese, G. Ottaviano and T. Tabuchi (2004), “Testing the home market effect in a multi-country world: The theory”, CEPR Discussion Paper No. 4468.

Bentivogli C. and P. Pagano (1999), “Regional disparities and labor mobility, the EURO-11 versus the USA”, *Labor*, n. 13, pp. 737-760.

Blanchard O. and L. Katz (1992), “Regional evolutions”, *Brooking Papers on Economic Activity*, n. 1, pp. 1-75.

Braunerhjelm P., R. Faini, V. Norman, F. Ruane and P. Seabright (2000), *Integration and the Regions of Europe: How the Right Policies Can Prevent Polarization*, CEPR, London.

Brülhart M. (2001), “Evolving geographical specialisation of European manufacturing industries”, *Weltwirschaftliches Archiv*, n. 137, pp. 215-243.

Campbell J. and H. Hopenhayn (2002), “Market size matters”, NBER Working Paper No. 9113.

Charlot S., C. Gaigné, F. Robert-Nicoud and J.-F. Thisse (2004), “Agglomeration and welfare: The core-periphery model in the light of Bentham, Kaldor, and Rawls”, *Journal of Public Economic*, forthcoming.

Ciccone A. and R. Hall (1996), “Productivity and the density of economic activity”, *American Economic Review*, n. 87, pp. 54-70.

Combes P-P. and H. Overman (2004), The spatial distribution of economic activities in the European Union, in Henderson V. and J.-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Combes P., G. Duranton and L. Gobillon (2004), “Spatial wage disparities: Sorting matters!”, CEPR Discussion Paper No. 4240.

Conley T., F. Flyer and G. Tsiang (2004), “Local market human capital and the spatial distribution of productivity in Malaysia”, *Contribution to the Economics and Growth of Developing Areas*, forthcoming.

Coughlin C., J. Terza and V. Arromdee (1991), “State characteristics and the location of foreign direct investment in the United States”, *Review of Economics and Statistics*, n. 73, pp. 675-683.

Cronon W. (1991)*, Nature’s Megalopolis: Chicago and the Great West*, Norton, New York.

Crozet M. (2000), “Do migrants follow market potential? An estimation of a new economic geography model”, Cahier de la MSE-Serie Blanche 2000-30.

Crozet M. and P. Koenig-Soubeyran (2002), “Trade liberalization and the internal geography of countries”, CREST Discussion Paper N. 2002-37.

Duranton G. and D. Puga (2004), “Micro-foundations of urban agglomeration economies”, in Henderson V. and J.-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Durlauf S. and D. Quah (1999), “The new empirics of economic growth”, in Taylor J. and M. Woodford, *Handbook of Macroeconomics*, North Holland, Amsterdam.

Ellison G. and E. Glaeser (1997), “Geographic concentration in US manufacturing industries: A dartboard approach”, *Journal of Political Economy*, n. 105, pp. 889–927.

Forslid R. and G. Ottaviano (2003), “An analytically solvable core-periphery model”, *Journal of Economic Geography*, n. 3, pp. 229-240.

Fujita M. and P. Krugman (2004), “The new economic geography: Past, present and future”*, Papers in Regional Science*, n. 83, pp.139-164.

Fujita M. and J-F. Thisse (1996), “Economics of agglomeration”, *Journal of the Japanese and International Economies*, n. 10, pp. 339-378.

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

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

Gabaix X. and Y. Ioannides (2004), “The evolution of city size distributions”, in Henderson V. and J.-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Glaeser E. and D. Maré (2001), “Cities and skills”, *Journal of Labor Economics*, n. 19, pp. 316-342.

Glaeser E., H. Kallal, J. Scheinkmann and A. Shleifer (1992), “Growth in cities”, *Journal of Political Economy*, n. 100, pp. 1127-1152.

Glaeser E., J. Scheinkman and A. Shleifer (1995), “Economic growth in a cross-section of cities”, *Journal of Monetary Economics*, n. 36, pp. 117-143.

Grossman G. and E. Helpman (1991), *Innovation and Growth in the World Economy*, MIT Press, Cambridge, Massachusetts.

Haaland J., H-J. Kind and K.-H. Midelfart Knarvik (1999), “What determines the economic geography of Europe?”, CEP Discussion Paper 207.2

Hanson G. (1997), “Increasing returns, trade and the regional structure of wages”, *Economic Journal*, n. 107, pp. 113-133.

Hanson G. (1998), “Market potential, increasing returns, and geographic concentration”, NBER Working Paper No. 6429.

Harris C. (1954), “The market as a factor in the localization of industry in the United States”, *Annals of the Association of American Geographers*, n. 64, pp. 315-348.

Head K., J. Ries and D. Swenson (1999), “Attracting foreign manufacturing: investment promotion and agglomeration”, *Regional Science and Urban Economics*, n. 29, pp. 197–218.

Head K. and T. Mayer (2002), “Market potential and the location of Japanese investment in the European Union”, CEPR Discussion Paper No. 3455.

Head K. and T. Mayer (2004), “The empirics of agglomeration and trade”, in Henderson V. and J.-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

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

Helpman E. and P. Krugman (1989), *Trade Policy and Market Structure*, MIT Press, Cambridge, Massachusetts.

Henderson V. (1978), *Economic Theory and the Cities*, Academic Press, London.

Jaffe A. and M. Trajtenberg (2002) *Patents, Citations and Innovation: a Window on the Knowledge Economy*, MIT Press, Cambridge, Massachusetts.

Jaffe A., M. Trajtenberg and R. Henderson (1993) “Geographic localization of knowledge spillovers as evidenced by patent citations”, *Quarterly Journal of Economics,* n. 63, pp. 577-598.

Kim S. (1995), “Expansion of markets and the geographic distribution of economic activities: The trends in US regional manufacturing structure, 1860-1987”, *Quarterly Journal of Economics,* n. 110, pp. 881-908.

Krugman P. (1980), “Scale economies, product differentiation and the pattern of trade”, *American Economic Review*, n. 70, pp. 950-959.

Krugman P. (1991) “Increasing returns and economic geography”, *Journal of Political Economy*, n. 99, pp. 483-499.

Krugman P. (1993), “The hub effect: or, threeness in international trade”, in Ethier W., E. Helpman, and P. Neary, *Theory, Policy and Dynamics in International Trade*, Cambridge University Press, Cambridge.

Krugman P. (1998), “Space: The final frontier”, *Journal of Economic Perspectives,* n. 12, pp. 161-174.

Krugman P. and R. Livas (1996), “Trade policy and the third world metropolis”, *Journal of Development Economics,* n. 49, pp. 137-150.

Krugman P. and A. Venables (1995), “Globalization and the inequality of nations”, *Quarterly Journal of Economics,* n. 110, pp. 857-880.

Magrini S. (2004), “Regional (Di)Convergence”, in Henderson V. and J.-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Maignan C., G. Ottaviano, and D. Pinelli (2003), “ICT, clusters and regional cohesion: A summary of theoretical and empirical research”, FEEM Working Paper No. 58.2003.

Manzocchi S. and G. Ottaviano (2001), “Outsiders in economic integration: The case of a transition economy”, *Economics of Transition*, n. 9, pp. 229-249.

Marshall A. (1890), *Principles of Economics,* London, Macmillan.

Martin P. (1999), “Are European regional policies delivering?”, *European Investment Bank Papers,* n. 4, pp. 10-23.

Martin P. and C. Rogers (1995), “Industrial location and public infrastructure”, *Journal of International Economics*, n. 39, pp. 335-351.

Melitz M. (2003), “The impact of trade on intra-industry reallocations and aggregate industry productivity”, *Econometrica*, n. 71, pp. 1695-1725.

Melitz M. and G. Ottaviano (2003) “Market size, trade, and productivity”, Harvard University, mimeo.

Monfort P. and R. Nicolini (2000), “Regional convergence and international integration”, *Journal of Urban* Economics, n. 48, pp. 286-306.

Monfort P. and G. Ottaviano (2004), “Spatial mismatch and skill accumulation”, revised version of CEPR Discussion Paper No. 3324, 2002.

Monfort P. and T. Van Ypersele (2003), “Integration, regional agglomeration and international trade”, CEPR Discussion Paper No. 3752.

Moretti E. (2002), “Workers, education, spillovers and productivity: evidence from plant-level production function”, UCLA, mimeo.

Moretti, E. (2004), “Human capital externalities in cities”, in Henderson V. and J.-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Murata Y. (2003), “Product diversity, taste heterogeneity, and geographic distribution of economic activities: Market vs. non-market interactions”, *Journal of Urban Economics*, n. 53, pp. 126-144.

Neary P. (2001), “Of hype and hyperbolas: Introducing the new economic geography”, *Journal of Economic Literature*, n. 39, pp. 536-561.

Ohlin B. (1933), *Interregional and International Trade*, Harvard University Press, Cambridge, Massachusetts.

Ottaviano G. (2003), “Regional policy in the global economy: Insights from New Economic Geography”, *Regional Studies*, n. 37, pp. 665-674.

Ottaviano G. and D. Puga (1998), “Agglomeration in the global economy: A survey of the ‘new economic geography’”, *World Economy*, n. 21, pp. 707-731.

Ottaviano G. and F. Robert-Nicoud (2004), “The ‘genome’ of NEG models with vertical linkages: A positive and normative synthesis”, *Journal of Economic Geography*, forthcoming.

Ottaviano G. and J.-F. Thisse (2001), “On economic geography in economic theory: increasing returns and pecuniary externalities”, *Journal of Economic Geography*, n. 1, pp. 153-179.

Ottaviano G. and J.-F. Thisse (2002), “Integration, agglomeration and the political economics of factor mobility”, *Journal of Public Economics*, n. 83, pp. 429-456.

Ottaviano G. and J.-F. Thisse (2004), “Agglomeration and economic geography”, in Henderson V. and J.-F. Thisse, *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Ottaviano G., T. Tabuchi and J.-F. Thisse (2002), “Agglomeration and trade revisited”, *International Economic Review*, n. 43, pp. 409-435.

Overman H., S. Redding and A. Venables (2003), “The economic geography of trade, production and income”, in Harrigan J. and K. Choi, *Handbook of International Trade*, Blackwell, London.

Paluzie E. (2001), “Trade policies and regional inequalities”, *Papers in Regional Science*, n. 80, pp. 67-85.

Puga D. (1999), “The rise and fall of regional inequalities”, *European Economic Review*, n. 43, pp. 303-34.

Puga D. and A. Venables (1997), “Preferential trading arrangements and industrial location”, *Journal of International Economics*, n. 43, pp. 347-368.

Peri G. (2002), “Young workers, learning, and agglomeration”, *Journal of Urban Economics*, n. 52, pp. 582-607.

Quah D. (1997), “Empirics for growth and distribution: stratification, polarization and convergence clubs”, *Journal of Economic Growth*, n. 2, pp. 27-59.

Rauch J. (1993), “Productivity gains from geographic concentration of human capital: evidence from the cities”, *Journal of Urban Economics*, n. 34, pp. 380-400.

Redding S. and A. Venables (2000), “Economic geography and international inequality”, CEPR Discussion Paper No. 2568.

Rey S. and B. Montuori (1999), “US regional income convergence: A spatial econometrics perspective”, *Regional Studies*, n. 33, pp. 143-156.

Rosenthal S. and W. Strange (2001), “The determinants of agglomeration”, *Journal of Urban Economics*, n. 50, pp. 191-229.

Sala-i-Martin X. (1996), “The classical approach to convergence analysis”, *Economic Journal*, n. 106, pp. 1019-1036.

Scitovsky T. (1954), “Two concepts of external economies”, *Journal of Political Economy*, n. 62, pp. 143-151.

Scotchmer S. and J.-F. Thisse (1992), “Space and competition: a puzzle”, *Annals of Regional Science*, n. 26, pp. 269-286.

Smith T. (1975), “A choice theory of spatial interaction”, *Regional Science and Urban Economics*, n. 5, pp. 137-176.

Syverson C. (2002), “Price dispersion: The role of product substitutability and productivity”, University of Chicago, mimeo.

Tabuchi T. and J.-F. Thisse (2002), “Taste heterogeneity, labor mobility and economic geography”, *Journal of Development Economics*, n. 69, pp. 155-177.

Temple J. (1998), “The new growth evidence”, *Journal of Economic Literature*, n. 37, pp. 112-156.

Tybout J. (2002), “Plant and firm-level evidence on new trade theories”, in Harrigan J., *Handbook of International Economics*, Blackwell, London.

Venables A. (1996), “Equilibrium locations of vertically linked industries”, *International Economic Review*, n. 37, pp. 341-359.

Venables A. (2001), “Geography and international inequalities: The impact of new technologies”, paper prepared for the *World Bank Annual Conference on Development Economics*, Washington , May 2001.

CHAPTER 2

MARKET POTENTIAL AND PRODUCTIVITY: EVIDENCE FROM FINNISH REGIONS

1. Introduction

‘New economic geography’ (henceforth, simply NEG) is an approach to economic geography firmly grounded on recent developments in mainstream industrial organization and international trade theory. After more than a decade since the seminal work by Krugman (1991), NEG has grown into a mature body of literature as testified by a rich list of surveys and textbooks. Nevertheless, its empirical assessment is still at an infant stage and no conclusive evidence is available yet. This has been mainly due to the gap between theoretical and applied investigations: theorists have shown little interest in translating their insights in clear-cut testable predictions; empiricists have made little effort in understanding what theory exactly implies.

A typical example of the state-of-the-art is the empirical investigation of agglomeration forces. The central idea of NEG is that, in the presence of trade costs and increasing returns to scale, market interactions draw firms towards places characterized by higher ‘market potential’, that is, better access to customers (‘demand or backward linkages’) and suppliers (‘cost or forward linkages’). Also workers are attracted to places with higher market potential as these offer better access to final products (‘cost-of-living or amenity linkages’). This generates an incentive for firms and workers to co-locate, thus supporting the agglomeration of economic activities. Different NEG models stress different linkages as the main agglomeration forces. For example, in the presence of labour mobility, Krugman (1991) focuses on demand and cost-of-living linkages; without labour mobility, Krugman and Venables (1995) as well as Venables (1996) highlight demand and cost linkages. Which agglomeration force dominates in reality is

This paper was published as: Ottaviano G.I.P. and Pinelli D. (2006), Market Potential and Productivity: Evidence from Finnish Regions, *Regional Science and Urban Economics*, 36, pp. 636-657. The paper has been prepared within the framework of study *Finland in the Global Economy*, Prime Minister’s Office, Helsinki. We have benefited from comments by Paavo Okko, Heikki Loikkanen, Hannu Tervo, Vesa Vihriälä, Raija Volk, Pekka Ylä-Anttila, and other seminar participants in Helsinki, 21 June 2004. Additional comments by the referees and the editors of this journal have helped us to further improve the paper.

left nonetheless unspoken as the empirical implications of different models have not been entirely spelled out.[1]

The aim of this paper is to fill this gap by proposing a methodology to assess not only whether linkages are relevant but also whether they are more important for firms or workers. This is achieved by showing that a NEG model featuring all three types of linkages can be used to design an empirical identification strategy à la Roback (1982): if market potential boosted firm productivity only, higher values would be associated with higher wages and higher land rents; if market potential boosted amenity only, higher values would be associated with lower wages and higher land rents.[2]

We test that theoretical prediction by estimating income, population, and real estate value growth regressions on Finnish NUTS 4 regions from 1977 to 2002. Finland is an interesting case because it allows us to study two different scenarios while relying on rich comparable data and holding fundamental institutional variables costant. The reason is the role of the 'recession' of the early Nineties commonly perceived as a watershed in recent Finnish economic history. Specifically, Finland entered the recession as an economy characterized by traditional industries, low skills, and limited labour mobility. It emerged as an economy increasingly characterized by high-tech sectors, high skills and mobile workers. We face, thus, an 'old economy' before the recession and a 'new economy' thereafter. Despite such differences, however, we find that the impact of the market potential on regional performance is positive and significant in both periods. What changes is the set of relevant controls. Moreover, according to our identication strategy, the impact of market potential can be interpreted in terms of a dominant positive effect on productivity. Therefore, demand and cost linkages rather than cost-of-living linkages seem to sustain agglomeration in both 'old' and 'new' Finland. Finally, growth regressions also allow us to conclude that, after the recession, increased labour mobility and the rise of new 'footloose' industries (i.e. industries less dependent on natural resources) have hampered the process of regional convergence, as NEG would also predict.

The paper is organized in five sections after the introduction. Section 2 presents the theoretical model and its empirical implications. Section 3 surveys the salient features of Finnish recent economic history. Section 4 describes the data set. Section 5 reports the results of the growth regressions. Section 6 concludes.

2. The model

What distinguishes NEG from alternative approaches to regional issues is the focus on market ('pecuniary') rather than non-market ('technological') interactions within a 'general equilibrium' set-up, i.e. a framework of analysis that stresses the endogenous

[1] There exist many relevant surveys. Theoretical surveys are more focused on NEG. See, e.g., Ottaviano and Thisse (2004). Empirical surveys are generally less focused on NEG per se. Indeed, the different focuses of theoretical and empirical surveys reflect the different stages of development of the corresponding literatures. See the discussion in Head and Mayer (2004) for a survey of current achievements and a to-do list in empirical NEG.

[2] See Moretti (2004) for a survey of studies à la Roback (1982).

determination of good and factor prices and the importance of economy-wide budget constraints. In particular, NEG is based on increasing returns to scale, trade costs and imperfect competition. Plant-level scale economies and shipping costs generate a trade-off between, on the one hand, the ‘concentration’ of production in few plants, and, on the other hand, the ‘proximity’ of plants to customers and suppliers. Given imperfect competition, firms can increase their market power (and thus their profits) with respect to their competitors by careful geographical positioning. In so doing, they generate localized externalities that determine the attractiveness of regions to firms and workers and can give rise to cumulative processes of agglomeration. Such externalities are stronger the higher the returns to scale and the more differentiated the products (as in both cases market power is enhanced). Besides, they more readily cause cumulative agglomeration the higher the share of footloose industries and mobile workers.

2.1. A simple NEG model

The foregoing insights can be brought to data by considering a simple NEG model. This is obtained by extending the set-up of Redding and Venables (2004) by introducing labour mobility and land à la Hanson (1998) and Helpman (1998).

The economy consists of i = 1,…, R regions. On the demand side, in region j the representative worker consumes a set of horizontally differentiated varieties and land services (‘housing’). Her utility function is:

$$U_j = \left(X_j\right)^{\mu}\left(L_j\right)^{1-\mu},\ 0<\mu<1$$

where L_j is land consumption and

$$X_j = \sum_{i=1}^{R}\left\{\int_o^{n_i}\left[x_{ij}(z)\right]^{\frac{\sigma-1}{\sigma}}dz\right\}^{\frac{\sigma}{\sigma-1}} = \sum_{i=1}^{R}\left(n_i x_{ij}{}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}$$

is a CES quantity index of the $\sum_{i=1}^{R} n_i$ varieties available in region j with x_{ij} labelling the consumption in region *j* of a typical variety produced in region *i*. The associated exact CES price index is:

$$P_j = \sum_{i=1}^{R}\left\{\int_o^{n_i}\left[p_{ij}(z)\right]^{1-\sigma}dz\right\}^{\frac{1}{1-\sigma}} = \sum_{i=1}^{R}\left(n_i p_{ij}{}^{1-\sigma}\right)^{\frac{1}{1-\sigma}}$$

where p_{ij} is the delivered price in region *j* of a typical variety produced in region *i*. In the above expressions the second equality exploits the fact that in equilibrium quantities and prices are the same for all varieties produced in country *i* and consumed by country *j*.

Utility maximization then gives the demand in *j* for a typical variety produced in *i*:

(1) $x_{ij} = p_{ij}^{-\sigma} E_j P_j^{\sigma-1}$

where E_j is expenditures on X_j, which is a fraction μ of income I_j, while σ>1 is both the own and the cross price elasticity of demand.

On the supply side, each variety is produced by one and only one firm under increasing returns to scale and monopolistic competition. In so doing, the firm employs labour, land and, as intermediate input, the same bundle of differentiated varieties that workers demand for consumption. Specifically, in region *i* the total production cost of a typical variety is:

$$TC_i = P_i^{\alpha} r_i^{\beta} w_i^{\gamma} c_i (F + x_i),\ \alpha, \beta, \gamma > 0,\ \alpha + \beta + \gamma = 1$$

where x_i is total output, r_i and w_i are land rent and wage, while c_i and c_iF are marginal and fixed input requirements respectively.[3] Trade faces iceberg frictions: for one unit of any variety to reach destination when shipped from region *i* to region *j*, τ_{ij}>1 units have to be shipped. Hence, $x_i = \sum_{j=1}^{R} x_{ij}\tau_{ij}$.

Firm profit maximization yields the standard CES mark-up pricing rule:

(2) $p_i = \dfrac{\sigma}{\sigma - 1} P_i^{\alpha} r_i^{\beta} w_i^{\gamma} c_i,\ p_{ij} = \tau_{ij} p_i$

Free entry then implies that in equilibrium firms are just able to break even, which happens when they operate at scale $\overline{x} = (\sigma - 1)F$. Together with (1) and (2), that allows us to write the free entry condition in region *i* as:

(FE) $\overline{x}\left(\dfrac{\sigma}{\sigma - 1} r_i^{\beta} w_i^{\gamma} c_i\right)^{\sigma} = MA_i\ SA_i^{\frac{\alpha\sigma}{\sigma-1}}$

where $MA_i = \sum_{j=1}^{R} \tau_{ij}^{1-\sigma} E_j P_j^{\sigma-1}$ is the 'market access' of region *i*. This is a measure of customer competitor proximity ('demand linkages') that predicts the quantity a firm sells given its production costs. The term $SA_i = P_i^{1-\sigma} = \sum_{j=1}^{R} n_j p_j^{1-\sigma} \tau_{ji}^{1-\sigma}$ is, instead, the 'supplier access' of region *i*, a measure of supplier proximity. This inversely predicts the prices a firm pays for its intermediate inputs ('cost linkages') and a worker pays for her consumption bundle ('cost-of-living linkages') when located in a certain region

Workers work and consume in the region where they reside and can pick their residence freely. This implies that in equilibrium they are indifferent about location as they would achieve the same level of indirect utility *V* wherever located. Given the chosen utility, if

[3] In the cross-country study by Redding and Venables (2004), the parameter c_i is allowed to vary to capture Ricardian productivity advantages across countries. This interpretation is hard to defend within the same country, so its variation across Finnish regions will be interpreted as the outcome of localized technological externalities. These will be introduced as controls in the empirical analysis.

we further assume that the land of a region is owned by locally resident landlords, free mobility then gives:[4]

$$\textbf{(FM)}\quad \frac{w_i}{SA_i^{\frac{\mu}{1-\sigma}} r_i^{1-\mu}} = V$$

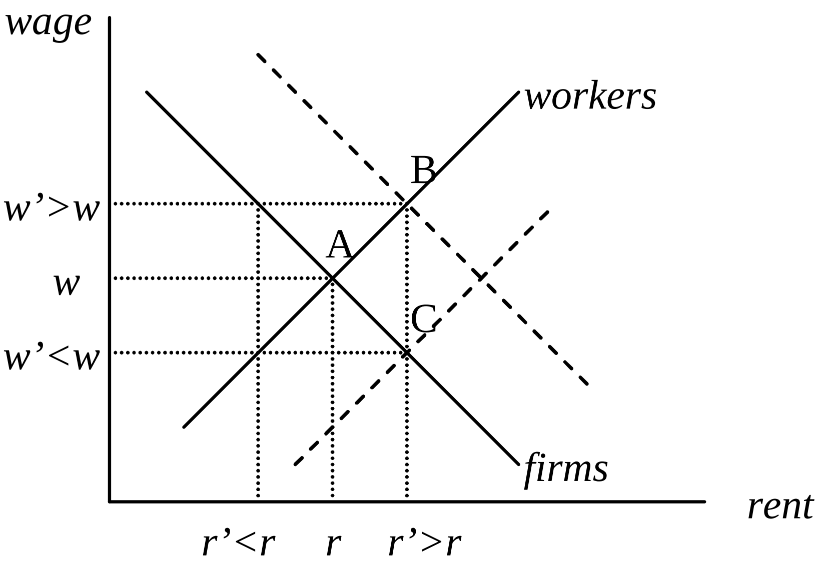
After log-linearization, conditions (FE) and (FM) are depicted in Figure 1, which measures the logarithm of regional nominal wages (w) along the vertical axis and the logarithm of regional land rents (r) along the horizontal one. Downward sloping lines are derived from (FE) and depict the combinations of wages and rents that make firms indifferent about regions. Their downward slope reflects the fact that firms can break even in different regions provided that higher wages correspond to lower rents and vice versa. Upward sloping lines are derived from (FM) and depict the combinations of wages and rents that make workers indifferent about regions. Their upward slope reflects the fact that workers can achieve the same utility (‘real wage’) in different regions provided that higher rents correspond to higher wages and vice versa.

The exact positions of the two lines depend on regional market access and supplier access. Better market access (larger MA) shifts FE up, increasing both wages and land rents. Better supplier access (larger SA) shifts both FE and FM up, also increasing rents. The effect on wages is, instead, ambiguous: they increase (decrease) if the shift in FE dominates (is dominated by) the shift in FM. This theoretical ambiguity makes it pointless to try to disentangle the effects of MA and SA on equilibrium wages and rents. What we can do, instead, is to check whether their combined effect is indeed positive on rents as predicted by the model. In addition, we can use information about migration flows. Since land values capitalize the attractiveness of a place, land rents rise also because immigration increases the demand for land.

More interestingly, we can also check whether the combined effect of MA and SA is positive or negative on wages, which would point at a dominant impact on firms (point B) or on workers (point C) respectively. Demand and cost linkage would dominate in the former case; cost-of-living linkages in the latter.

[4] This assumption is made only for analytical convenience. What is crucial for what follows is that the rental income of workers, if any, is independent of locations and, thus, it does not affect the migration choice. The alternative assumptions of absentee landlords or balanced ownership of land across all cities would also serve that purpose

Figure 1. The geographical equilibrium



2.2. Growth regressions

The discussion in the previous section suggest to identify the combined effects of MA and SA on productivity and amenity through their impacts on the levels of wages, rents and migration flows using panel techniques. Under the assumption that regions have been fluctuating around a balanced growth path (BGP) during the observed period, the panel estimation of those impacts can be interpreted as their long-run effects along the BGP. This interpretation allows us to use growth regressions instead of panel regressions with a double advantage. First, endogeneity would potentially affect the panel estimates since higher productivity and amenity could be the causes rather than the effects of better market and supplier access. For example, if booming regions attracted firms and workers, then the positive correlation between access and immigration could arise due to reverse causation from the latter to the former. Second, the focus on levels would obscure the dynamic evolution of productivity patterns across regions, which is an interesting issue in itself as NEG stresses the possibility of cumulative agglomeration. In this respect, growth allows us to use a variety of existing works on Finland as benchmarks for our results.

Both issues can be dealt with by estimating standard growth regressions over a set of explanatory variables including some measure of market and supplier access. For instance, as to wages, we will estimate the following equation:

$$(3)\ \ln(w_t) - \ln(w_{t-1}) = \alpha + \beta \ln(w_{t-1}) + \gamma \ln(access_{t-1}) + \delta \ln(controls_{t-1}) + \varepsilon_t$$

where the growth rate of regional wages on the left hand side is regressed on its initial value and other 'initial conditions' including some measure of market and supplier access (details are provided in Section 4).

The idea is that along a BGP productivity grows at a constant rate across regions so that these may differ only in terms of wage levels. Then, under the assumption that the economy fluctuates around its BGP, the growth equation captures transitional growth: if a certain city exhibits a higher growth rate than the other, then the former has a higher level of wage in BGP than the latter and it is converging to that level, given its initial conditions. As anticipated, while modelling the dynamics of the economy, the above equation also allows us to partially tackle the endogeneity problem. The reason is that, whereas market and supplier access is measured at the beginning of period (at time t-1), the growth of wage is measured during the period of observation (from times t-1 to t). In other words, the independent variables are predetermined relative to the dependent one.

As argued in the previous section, to disentangle productivity from amenity effects, the above equation has to be matched by similar regressions for land values and migration flows.

3. Finland

Finnish regions provide an attractive scenario for testing the above predictions for the following reasons. First, as the units of analysis belong to the same country, differences in regional development are unlikely to be driven by institutional differences or Ricardian comparative advantage, which have both been shown to play an important role in cross-country studies.[5] Second, during our period of observation, Finland was hit by a dramatic exogenous shock, the 'recession', which is considered a 'watershed' under several respects (more on this below). Such shock is exogenous to any region-specific development. Third, Finland entered the recession as an economy characterized by traditional industries, low skills, and limited labour mobility. It emerged as an economy increasingly characterized by skill-intensive sectors, high skills and mobile workers. This allows us to test the role of market and supplier access in two rather different economies within quite a homogenous data set.

Given its role, it is worth spending a few words on the recession. First of all, the shock was huge. Between 1990 and 1993, Finnish GDP plunged by 9.5 per cent and unemployment surged from 3.2 to 16.6 per cent. This was the worst recession since the 1930s.

The recession was the effect of both 'bad luck' and 'bad policies'. The collapse of the USSR brought to an abrupt end the long-standing bilateral trading system between the two countries. The system was based on five-year agreements with quotas balancing

[5] See Alcalà and Ciccone (2004) for a recent assessment of the relation between trade and productivity at the international level.

imports (mainly oil) and exports (a variety of primary and manufactured products). The system was quite lucrative for Finland: there is evidence that the price of exports to former USSR was slightly above market prices (reaching a premium of nearly 16 per cent for pulp and paper). As a result of the shock, the value of manufacturing exports to the USSR fell by 65 per cent in 1991, accounting for a fall of 8 per cent in the value of total manufacturing exports. Traditional industries, such as textile and forestry (and related engineering), were the industries that suffered most (OECD Economic Surveys, 1992).

The collapse of the USSR was not the only negative shock to the economy in the period. The generalised slowdown of industrialised economies and the rise in German interest rates that followed the reunification also contributed to the recession. However, ‘bad luck’ cannot explain the whole story. ‘Bad policies’ also played a key role, acting pro-cyclically both before and after the recession (Honkapohia and Koskela, 1999). Before the recession, relaxed fiscal policies and bad financial deregulation (contributing to increasing bank lending) overheated the economy as testified by the growing indebtedness of households and firms, the bubble of real estate prices, and large capital inflows. These eventually led to the revaluation of the markka in March 1989. After the recession started, the strong markka and a tightened fiscal policy exacerbated the crisis. In particular, interest rates were kept artificially high to defend the pegged exchange rate, further weakening the financial position of households and firms and leading to the collapse of aggregate demand (consumption and investment) and real estate prices.

The recession treated all regions quite equally. Despite differences in timing (the recession first affected export industries and the industrial regions of the south, then spread to the the rest of the country), output and the number of people in work fell by, respectively, 5-10 per cent and about 20 per cent everywhere (Economic Council, 2001).

The recession was followed by a boom. Between 1994 and 2000 the average annual growth in GDP was nearly 5 per cent. The boom was driven by fast growth in high-tech industries, with manufacturing of electric and electronic products (especially telecommunication equipment) being the fastest sector (Rouvinen and Ylä-Anttila, 2003; Kangasharjiu and Pekkala, 2004). Nokia alone is estimated to account for around 1.5 percentage point of GDP annual growth rate. This transformed the Finnish economy (traditionally based on primary products) into a innovation-driven economy, with high-tech products accounting for 20.4 per cent of exports in 1999 (only 12.4 per cent in 1994). High private and public investment in R&D and a strong commitment to education were at the base of the transformation. In particular, following investment in education in previous decades, young Finns entering the labour market in the post-recession period were among the most educated in the world (Rouvinen and Ylä-Antttila, 2003). Higher educational attainment and industrial restructuring promoted intermunicipal mobility. Between 1995 and 2000, about 1.5 million people changed municipality whereas only 1.2 million did the same over 1985-1990 (Nivalainen, 2003).

The boom had a strong regional dimension. The concentration of fast growing high-tech industries (and related business services) favoured areas such as Salo, Oulu and

Helsinki, while rural and traditional areas suffered from the poor output and (to a much larger extent) employment performance of primary and traditional manufacturing industries. The regional dimension of the boom was reinforced by several changes affecting the policy environment (Rouvinen and Ylä-Anttila, 2003; Tervo, 2004). Firstly, efforts to balance the public economy, privatize operations and produce public services more efficiently led to a decrease of over 100,000 jobs over 1990-1995 (mostly concentrated in administrative centres and service centres in northern Finland). Secondly, while general government policy is still balancing out regional disparities (richer regions still contribute more than proportionally to and receive less than proportionally from government accounts), the scope and structure of direct regional government intervention was re-shaped with the accession to the European Union, with Structural Funds largely replacing national instruments as the adoption of the euro in 1995 imposed stricter constraints of national budgets (Economic Council, 2001). Thirdly, accession to the Common Agricultural Policy further limited the scope for direct intervention to maintain agricultural production in rural areas. All this was associated with an abrupt stop of the process of regional convergence observed before the recession (Kangasharju et al., 2001; Taipale, 2002).

4. Data

We now investigate the forces that have driven the regional performance of Finland from 1977 to 2002.[6] The time spanned by the analysis is partitioned in two periods, 1977-1990 (pre-recession period) and 1994-2002 (post-recession period). Following the consensus approach for Finnish studies, the three years from 1991 to 1993 are removed as all regions were in recession (Suomen Kuntaliitto, 1999). The analysis is carried out at the level of NUTS 4 of the European Union. This classification corresponds to subregional units whose borders follow closely those of commuting districts.[7]

4.1. Performance measures

To implement our identification strategy, we jointly use the following three measures of regional economic performance:[8]

- Income per capita growth. Since wages are not available at the level of NUTS 4, two alternative measures are used to proxy them in terms of income per capita. First, we use taxable income, which refers to gross income accruing from personal, corporate, and property sources less deductions. We use this measure instead of the more commonly used gross regional product (GRP). The reason is that the time series available for taxable income is longer. The key difference between the two

[6] Data are kindly supplied by the Pellervo Economic Research Institute (PTT). See the appendix for details.

[7] Because of their peculiarities, the three islands of the Ahvenanmaa region (Mariehamns stad (211), Ålands landsbygd (212), Ålands skärgård (213)) are excluded from the sample.

[8] For each measure Y annual growth rates are calculated by fitting a linear regression $\ln(Y)=a+b\cdot t$ where t is time. The growth rate is then defined as $g=100\ [\exp(b)-1]$. This way the growth rate does not depend only on the initial and final values of Y over the period of observation (see Temple, 1998). The results are virtually unchanged by using the simple log growth rate.

measures is that gross regional product refers to production, whereas taxed income refers to earnings accruing from production. The main shortcoming in using taxable income is that it includes income from stock options. The regional distribution of this type of income is very random and might influence substantially overall income in small regions (at least for what concerns the period after the recession). We use, therefore, primary income as an alternative measure to control for this effect. Primary income is available only since 1995 and the corresponding regressions are only estimated for the second period. Both measures of income are deflated by the national price index, which does not affect the nominal cross-region variation predicted by the theoretical model.

- Population growth. We use two measures of population growth. The first measure is simply the annual average growth rate of the number of inhabitants in a region. This measure is determined by both birth/mortality rates and net migration flows. However, only the latter are likely to respond to economic factors in short periods of time. We therefore calculate also an adjusted measure of population growth based on net migration flows (i.e., net of newborns and deaths).
- House price growth. Rents are generally available only for a small subset of urban areas and very limited time periods. We proxy them by average house prices for which data availability is slightly better. Nonetheless, house prices are not collected for NUTS 4 regions but only at NUTS 3 level, and for the main NUTS 4 subunits in each NUTS 3 unit. Therefore, each NUTS 4 subunit within the same NUTS 3 region is assigned the same value, calculated so that the population weighted average of house prices in the NUTS 4 gives the reported NUTS 3 value. Moreover, house prices are only available from 1987.

4.2. Explanatory variables

The macroeconomic literature (see, e.g., Temple, 1999) explains differences in economic growth across geographical areas in terms of two main sets of variables: proximate sources of growth and wider influences. We enlarge the list of the latter to take into account a richer array of geographical variables. In particular, we introduce ‘first nature’ and ‘second nature’ explanatory variables. The former variables capture the exogenous attractiveness of a region due to its abundance of natural resources, its proximity to natural means of communication, and its climatic conditions. The latter capture the endogenous attractiveness of a region determined by economic interactions.

4.2.1. Proximate sources of growth

Proximate sources are production factors that directly affect regional performance:

- Human capital. We measure the stock of human capital in two ways: by the share of population with at least a secondary education degree; and by the share of population with at least a tertiary education degree. Following recent literature (see, e.g., Temple, 2001), we introduce (alternatively) the level of human capital (to capture the so-called ‘technology adoption effect’) and its change over the period (to capture the so-called ‘neo-classical accumulation effect’).

- Knowledge capital. We measure the stock of knowledge capital by R&D expenditure per capita and by the number of patents per capita.[9]
- Physical capital. The initial level of income is introduced to control for decreasing returns to capital accumulation.

4.2.2. Wider influences

Wider influences affect regional performance indirectly by improving knowledge and technology transfer as well as the efficiency of input allocation.

Policies

We capture the impact of local policy along the following dimensions:

- Labour market. The unemployment rate is used to proxy the efficiency of the local labour market.
- Regional policy. The level of central government expenditure and the level of central government grants to municipalities (both in per capita terms) are used as proxies of interregional redistribution.[10]
- International openness. Distance from the Russian border (specifically, from the closest point with passport control) is used to control for proximity to Western Europe and collapsing trade with the former USSR.
- Infrastructures. The availability of physical infrastructures is captured by the distance from airports and train stations for the fastest trains. In particular, short distance from airports signals a ‘gate’ function of the region.

First nature

Geographers stress the role of natural means of communication and climate in determining the economic performances of different areas:

- Natural communications. The proximity to natural means of communication is captured by the distance from ports.
- Climate. We measure the climatic conditions by the share of land covered by lakes and by the average temperature.

Second nature

Geographical economics stresses two types of localized externalities, ‘pecuniary’ and ‘technological’, that endogenously determine the economic attractiveness of a region. We capture the two types of externalities by:

[9] Data on patents are available from 1990. Data on R&D expenditure are available from 1995.

[10] Grants to municipalities include grants for health care and social services and education and the so called general grants. Central government expenditure includes central government grants and all kind of subsidies (to agriculture, R&D activities, infrastructure, basic unemployment, etc). Data on government expenditures are available from 1994, and only at NUTS 3 level. The same figure is applied to all NUTS 4 subunits within the same NUTS 3 unit.

- Market potential. At the international level, where labour mobility is not an issue, Redding and Venables (2004) construct measures of both MA and SA using bilateral trade flows data. Such data are not available for Finnish regions. More fundamentally, we have seen in Section .2.1 that, with labour mobility, it is pointless to try to disentangle the separate effects of MA and SA. On both counts, we use a joint measure of market and supplier access, the so-called ‘nominal market potential’.[11] For region *i* this is defined as

$$MP_i = \sum_{j=1}^{j=n} Size_j / d_{ij}$$

  where d_{ij} is the distance between region *i* and region *j*. Distances between NUTS 4 regions are calculated as follows. First, distances along main roads are measured between centres of NUTS 5 regions. Second, distances between NUTS 4 regions are computed as population-weighted average distances between NUTS 5 centres within NUTS 4 regions. Third, own distances d_{ii} are weighted average distances between NUTS 5 centres within each NUTS 4 region. Finally, *Size* is measured by aggregate income.

- Population density. Non-market interactions are more frequent in densely populated areas. Therefore, population density is used to capture the role of technological externalities. Local density may seem too a restrictive measure as ICT promote informal contacts even between remote locations. However, existing empirical evidence suggests that the impact of those contacts appears to fade away quite rapidly with distance (Jaffe et al., 1993; Jaffe and Trajtenberg, 2002; Conley, Flyer and Tsiang, 2003).

5. Regression analysis

The results of the estimation of the growth regressions are reported in Tables 1 and 2 for 1977-1990 and 1994-2002 respectively. We present results from the OLS estimation only. As heteroskedasticity often characterises cross-regional analyses, both tables report t-statistics based on robust standard errors. For each dependent variable we present a benchmark regression selected on the basis of explanatory power and robustness. The results of alternative specifications are discussed when relevant to the assessment and the interpretation of results.

There are two potential problems with OLS. Firstly, our theoretical model shows that equilibrium wages and rents are simultaneously determined. This suggests that there may be correlation between the unobservable idiosyncratic shocks to wages and rents. This potential source of inefficiency in OLS estimations has been tackled with SUR

[11] Head and Mayer (2005) compare alternative measures of market potential. Complex measures lead to results that are essentially the same as the ones associated with the simple measure we adopt in the wake of Harris (1954).

estimation. Results are virtually unchanged from OLS and, therefore, not reported. Secondly, the residuals may exhibit spatial correlation due to interactions among regions that are not captured by the market potential measure. Nevertheless, the analysis based on error/lag models à la Anselin (1988) substantially confirms the OLS results, so we do not report it either.

5.1. Before the recession

Table 1 shows the results of the growth regressions for the first period.[12] Since data on house prices are only available from 1987, we also show the results for population and income growth regressions estimated over the sub-period 1987-1990.

5.1.1. Population

In Table 1 Columns 1, 2 and 3 show the results of the population regressions. As to second nature, NEG-related effects seem to explain most of population growth differentials in the first period. In particular, the coefficient on market potential is positive and significant, which indicates that workers tend to move towards higher market potential locations, as suggested by the NEG literature. Moreover, the negative coefficient on distance from airports confirms that agglomeration takes place at or close to transport hubs.

There is, instead, no evidence of positive technological externalities as the coefficient on the density of population is actually negative (and significant in the 1987-1990 regression – Column 3). This result, however, holds only when the market potential is included in the regression. When it is excluded, the density term bears a positive coefficient, as consistent with the common view that migrants tend to move to higher density areas.

First nature effects are also important. The percentage of land covered by lakes appears to be relevant and positively influences population growth. On the other hand, the positive coefficient on distance from ports seems counterintuitive. However, it can be explained in the light of the bad economic situation of ports during the last decades due to industrial restructuring. This interpretation is supported by the fact that higher rates of unemployment and a higher share of manufacturing industries appear to depress population growth.

As to proximate sources, there is no evidence of a positive relationship between the level of education at the beginning of the period and population growth in the subsequent period. However, when we introduce the change in educational levels, this shows a significantly positive correlation with population growth in the period.

Turning to policy variables, the attractiveness of good infrastructures is revealed by the

[12] With respect to the list of explanatory variables discussed in the main text, we have tried to capture the potential relevance of the effects of knowledge accumulation and policy intervention by including, in all regressions, the level of central government grants per capita and a dummy variable identifying the regions with at least one university (the limited availability of data for this period imposing strong constraints). However, those variables are never significant, so the outcomes of the corresponding regressions are not reported.

negative impact on distance from airports. As to international openness, the positive coefficient on the distance from the Russian border signals both the disadvantage of being far from Western Europe and the backslash of collapsing trade with the USSR for border regions. More comments on policies at the end of Section 5.3.

Finally, the two measures of population growth lead to broadly similar patterns of results. The only difference concerns median age, which becomes insignificant in the regression for adjusted population growth (Column 2). It suggests that an older composition depresses population growth by reducing the demographic balance, rather than by altering the economic attractiveness of a region.

5.1.2. Income

Columns 4 and 5 report the results of per-capita income regressions. The negative and highly significant coefficient on the initial value of per-capita income reveals that the pre-recession period is characterised by a process of regional convergence in income per capita. Indeed, when included alone in the regression, initial income explains over 70 per cent of the variation in regional income per capita growth rates, thus signalling unconditional convergence.[13] Nonetheless, decreasing returns to capital accumulation are not the only force at work. NEG-related effects are also important. The coefficient on market potential is positive and strongly significant as in the population regressions.

As in the population regressions, the coefficient of population density is significantly negative. However, the coefficient becomes not significant when market potential is dropped from the regression. Also the unemployment rate and the share of manufacturing have significant impacts as in the population regression. However, their signs are no longer both negative as income growth appears to be positively related to unemployment. Finally, distance-related variables other than market potential have no longer significant impacts.

5.1.3. House prices

The results of the house price regression in Column 5 complete the picture. The coefficient on the starting level of house prices is strongly negative. The fact that house prices grew faster where they were lower matches the population finding on people moving to less densely populated regions. The result might also reflect the fact that the overshooting of house prices in the growth centres observed in the early 1970s started to smooth down as the flow supply of housing increased in these areas and migration flows declined.

The role of market interactions stressed by NEG receives additional support, whereas there is still no evidence of the relevance of technological externalities. In particular, the coefficient on market potential is again positive, whereas the density of population has once more a negative coefficient. As in the income regressions, the latter becomes not significant when market potential is dropped from the regression.

[13] These results are consistent with Kangasharju (1998) who finds evidence of convergence over the period 1973-1993 (and in the subperiod 1983-1993, although at a slower rate).

Table 1. Before the recession: 1977-1990

Variables: Explanatory (⇓) \ Explained (⇒)	Population growth (1977-1990)	Adjusted$^+$ population growth (1977-1990)	Population growth (1987-1990)	income per capita growth (1977-1990)	Income per capita growth (1987-1990)	House prices growth$^{\$}$ (1987-1990)
Income per capita	-0.226	-0.009	-0.418	-2.368 ***	-1.627 **	
	(-0.83)	*(-0.11)*	*(-0.96)*	*(-6.12)*	*(-2.72)*	
Density of population	-0.067	-0.038	-0.174 **	-0.354 ***	-0.494 ***	-1.411 ***
	(-0.79)	*(-1.31)*	*(-2.07)*	*(-3.71)*	*(-4.13)*	*(-4.05)*
House price						-13.02 ***
						(-4.76)
Median age	-4.191 ***	-0.375	-6.144 ***		-3.182 **	
	(-5.47)	*(-1.35)*	*(-4.58)*		*(-2.66)*	
Level of education						108.7 ***
						(4.48)
Market potential	1.288 ***	0.422 ***	1.691 ***	0.896 ***	1.577 ***	4.33 ***
	(5.43)	*(4.90)*	*(5.18)*	*(4.77)*	*(4.05)*	*(5.11)*
Share of employment in ICT						
Distance from main airports	-4.589 ***	-0.939 ***	-2.316 *			
	(-8.13)	*(-4.52)*	*(-1.91)*			
Distance from Russian crossing borders	2.95 ***	0.673 ***	1.975 **			
	(7.47)	*(4.82)*	*(2.99)*			
Distance from ports	2.983 ***	0.647 ***	2.325 **			
	(6.34)	*(3.70)*	*(2.53)*			
Unemployment rate	-0.095 ***	-0.028 ***	-0.098 ***	0.055 ***	0.068 **	
	(-5.37)	*(-5.06)*	*(-4.8)*	*(3.79)*	*(2.45)*	
Share of manufacturing and construction	-0.903 **	-3.480 **	-0.743	-1.014 **	-0.941	5.697 **
	(-2.18)	*(-2.48)*	*(-1.14)*	*(-2.49)*	*(-1.24)*	*(2.42)*
Lake covered land	1.302 ***	0.505 ***	2.621 ***			
	(3.23)	*(3.76)*	*(4.84)*			
Cons	-9.75 **	-6.510 ***	-11.72	-3.095	-4.639	27.59
	(-2.07)	*(-3.74)*	*(-1,22)*	*(-1.11)*	*(-1.19)*	*(1.90)*
Number of observations	79	79	79	79	79	76
R^2	72%	69%	72%	83%	41%	51%
Note(s):	All explanatory variables are in log terms (apart from shares) t-statistics are in parentheses (based on robust standard errors) *** = significant at 1% level ** = significant at 5% level * = significant at 10% level + = Population growth due only to net migration flows (net of natural balance) $ = Excluding outliers. Regions 56, 61, 79					

5.2. After the recession

The results for the post-recession period are presented in Table 2.[14]

5.2.1. Population

In Table 2 Columns 1 and 2 report the outcomes of population regressions. There are four major changes with respect to the findings of the pre-recession period.

First, while the market potential maintains its positive significant coefficient, other proximity variables such as the distances from ports, airports, and the Russian border, are no longer significant, which points out a weakening of distance-related effects.

Second, the initial share of employment in ICT has a positive influence on growth.[15] This result is very strong and very robust to changes in the specification of the regressions. Since ICT employment shares are not available before the recession, they were not included in the pre-recession regressions. These include instead manufacturing shares, which, as we have seen, have a negative impact on population growth. Together with the positive impact of ICT after the recession, that reveals the relevance of industrial restructuring.

Third, the level of education at the beginning of the period has now a strong positive effect on population growth.[16] We find a positive impact also when we introduce the change in educational levels.

Fourth, population density does not have a significantly negative coefficient anymore. Moreover, the coefficient becomes significantly positive as soon as the market potential is dropped from the regression (more in Section 5.3).

The foregoing results hold for both measures of population growth. As in the first period, the only difference concerns the median age effect. However, differently from before, now the median age has a positive and significant coefficient in the regressions for adjusted population growth (Column 2).

[14] With respect to the list of explanatory variables discussed in the main text, in the second period we were able to include additional variables measuring knowledge accumulation and regional policy. In particular, we could control for the umber of patents, R&D expenditures, as well as central government expenditures and grants. In addition to these controls, as in the pre-recession period, we also included dummy variables for regions having at least a university or a polytechnic. All those variables turn out insignificant, so the corresponding results are not reported. The only exception concerns the number of patents, which appears to be (weakly) significant when introduced together with secondary education in the population regressions. This suggests that the number of patents and tertiary education capture the same effect.

[15] ICT consist of: Manufacture of office machinery and computers; Manufacture of radio, television and communication equipment and apparatus; Manufacture of medical, precision and optical instruments, watches and clocks; Telecommunication services; Data processing services.

[16] The reported impact refers to tertiary education. A similar but weaker impact is obtained when using secondary education instead.

5.2.2. Income

The results of the income regressions are reported in Columns 3 (taxable income) and 4 (primary income) of Table 2. Regressions estimated using all observations performed very poorly because of influential outliers. Columns 3 and 4 report the results of regressions estimated excluding the outliers.[17] The market potential, the initial specialization in ICT, and the distance from the Russian border have positive impacts on income growth. Median age and unemployment rate have negative impacts. Population density has now no significant effect. Again, this coefficient becomes significantly positive when the market potential term is dropped from the regression. This is consistent with previous findings in the literature such as those in Ciccone and Hall (1996) and Ciccone (2002).

The negative coefficient on the starting level of income per capita reaffirms the convergence effect observed in the first period. However, it is interesting to note that, before the recession, the coefficient of initial income is negative and significant even if initial income were included as the only explanatory variable ('unconditional convergence'). Differently, after the recession, such coefficient is negative and significant only after controlling for other region-specific variables. This implies that in the post-recession period income differentials across regions have become persistent being determined by the differences in local characteristics ('conditional convergence').

5.2.3. House prices

In Table 2 Column 5 reports the results for the house price regressions. Three main points are worth noticing. First, the coefficient of initial house prices is positive and strongly significant. This implies that, after the recession, house prices have been growing faster where already initially higher. This result is robust to changes in specification and exactly opposite to what we obtained before the recession. It matches the lost significance of the population density coefficient in the second period. As it was the case in the population and income regression, the population density coefficient is not significant. However, once more, it becomes significantly positive when market potential is dropped from the regression.

Second, the positive impact of market potential is confirmed, while the distance from the Russian border has now a significant positive impact. In this respect, it is interesting to recall that the distance from Russia has also significant positive effects on population growth before the recession and on income growth thereafter but no effects otherwise. This reveals the role of migration in leading the transition from traditional activities (mainly linked to forestry) closer to the Russian border to new knowledge-based activities closer to the coast.

[17] The excluded outliers are Regions 1 (Helsinki), 55 (Härmänmaa), 56 (Järviseutu) and 59 (Sydösterbottens kustregion).

Table 2. After the recession: 1994-2002

Variables: Explanatory (⇓) \ Explained (⇒)	Population growth (1994-2002)	Adjusted Population growth^{+} (1994-2002)	(Taxable) Income per capita growth$^{\$}$ (1994-2002)	(Primary) Income per capita growth$^{\$}$ (1995-2002)	House prices growth$^{\$\$}$ (1994-2002)
Income per capita	-1.421 **	-0.994 ***	-2.566 ***	-3.75 ***	
	(-2.25)	*(-3.48)*	*(-5.01)*	*(-6.18)*	
Density of population					
House price					5.388 ***
					(3.57)
Median age	-3.836 ***	0.722 ***	-1.310 **	-3.136 ***	
	(-6.54)	*(2.75)*	*(-2.14)*	*(-2.92)*	
Level of education	20.54 ***	9.931 ***			
	(5.17)	*(6.52)*			
Market potential	1.029 ***	0.378 ***	0.330 ***	0.937 ***	2.22 ***
	(9.34)	*(8.00)*	*(3.48)*	*(5.05)*	*(7.93)*
Share of employment in ICT	0.056 ***	0.021 ***	0.085 ***	0.192 ***	
	(4.71)	*(3.66)*	*(9.00)*	*(11.93)*	
Distance from main airports					
Distance from Russian crossing borders			0.856 ***	1.294 ***	1.323 **
			(4.34)	*(3.74)*	*(2.02)*
Distance from ports					
Unemployment rate			-0.056 ***	-0.069 ***	
			(-4.47)	*(-3.58)*	
Share of manufacturing and construction					
Lake covered land					
Cons	0.592	-6.606 ***	4.750	3.467	-79.77 ***
	(0.32)	*(-7.72)*	*(1.45)*	*(0.73)*	*(-5.53)*
Number of observations	79	79	75	75	73
R^2	89%	86%	66%	70%	64%
Note(s):	All explanatory variables are in log terms (apart from shares) t-statistics are in parentheses (based on robust standard errors) *** = significant at 1% level ** = significant at 5% level * = significant at 10% level + = Population growth due only to net migration flows (net of natural balance) $ = excluding outliers. Regions 1, 55, 56, 59 $$ = excluding outliers. Regions 3, 79				

5.3. Interpretation

In what follows we discuss our results under a twofold perspective. Firstly, with respect to the role of recession, the results suggest that it is indeed a watershed. Before the recession, our analysis uncovers a distinct pattern of convergence for income, house prices, and population. After the recession, income convergence goes from unconditional to conditional, implying that regional differences in levels become permanent. Moreover, there is no evidence of convergence in population anymore and house prices even diverge. This is consistent with a process of agglomeration that raises productivity and amenity in places crowded by firms and workers.

Secondly, with respect to the main drivers of regional asymmetries, we are able, as discussed in Section 2, to determine the nature of their influence on regional performance by comparing the signs of the coefficients of the explanatory variables in the income and house price (or population) regressions. If a variable has positive (negative) coefficients in both regressions, then it has a positive (negative) impact on firm productivity. If a variable has a positive (negative) coefficient in the income regression and a negative (positive) coefficient in the house price regression, then it has a negative (positive) impact on worker utility.

Our key variable is the market potential, which turns out to have a positive influence on income, house prices and population growth in both periods. This is clear evidence of a dominant positive impact of that variable on productivity: in the long run regions that enjoy better market and supplier access tend towards higher levels of productivity. Thus, demand and cost linkages rather than cost-of-living linkages seem to sustain agglomeration in both ‘old’ and ‘new’ Finland. We do not find, instead, evidence of an independent role of technological externalities as proxied by population density. However, at least in the second period, when the market potential term is dropped from the regressions, the density of population influences positively population, income and house prices growth. This implies a positive impact on productivity, which is consistent with previous finding in literature (Ciccone and Hall, 1996; Ciccone, 2002). The results for the first period are similar but less clear cut, as the coefficient of population density is significantly positive only in the population regressions.

Turning to the other variables, there are clear indications of the effects of education and industrial structure. The level of education positively affects house prices in the first period and population growth in the second. Similarly, the change in the educational level also positively influences population growth in the second period and population and house prices in the first period. In both cases, the absence of any effect on wages signals a positive impact on both productivity and amenity (see Figure 1). The results therefore support the existence of both technology adoption and neo-classical accumulation effects of human capital. The negative impact of manufacturing and the proximity to ports in the first period as well as the positive impact of ICT in the second period reveal that the specialization in sunset industries is detrimental to regional productivity growth while sunrise industries have the opposite effect. As already mentioned, the fact that sunrise activities are disproportionately represented in urban areas close to the coast explains the evolution of the coefficients on the distance from Russia as migration flows promote the geographical reallocation of resources.

Some other variables have mixed effects. Unemployment has a negative impact on population growth and a positive impact on income growth in the first period. It has a negative impact on income in the second period. All this signals a dominant negative effect on amenity in the first period. This effect turns positive in the second period but is accompanied by a negative effect on productivity. Median age has a negative impact on income and (unadjusted) population growth in both periods. This points at a negative influence on productivity. However, the negative impact on (unadjusted) population growth could also simply reflect a negative impact on the natural demographic balance, rather than on the economic attractiveness of the region. Indeed, the positive impact of age on (adjusted) population growth in the second period suggests a positive association of age with utility (with older people living preferably in higher amenity areas).

Finally, there are variables that lose their explanatory power in the second period. Distance from main airports has a negative effect on population growth in the first period but no effect whatsoever in the second period. This points at a negative influence on both productivity and amenity in the first period only. Lake covered land has a positive effect on population growth in first period but no effect in second one. This signals a positive influence on both productivity and amenity in the first period only. On the contrary, we do not find evidence of climate on productivity and amenity, neither in the first nor in the second period.[18]

6. Conclusion

We have focused on two predictions of NEG models. First, by fostering the agglomeration of workers and firms, labour mobility and specialization in new footloose sectors hamper the process of regional convergence in productivity and amenity. Second, with or without labour mobility, agglomeration happens in places enjoying better market and supplier access.

We have tested these predictions on Finnish regional data from 1977 to 2002. We have argued that Finland represents an interesting case due to its rapid transformation at the beginning of the Nineties. In a very short period of time, Finland changed from an economy characterized by traditional industries, low skills, and limited labour mobility to an economy increasingly characterized by high-tech sectors, high skills and mobile workers. Overall, we have found that both predictions are supported by Finnish data. Using a new identification strategy, we have also been able to argue that demand and cost linkages rather than cost-of-living linkages seem to sustain agglomeration in both ‘old’ and ‘new’ Finland.

[18] The result is consistent with Knaap (2004) who finds that climate (in terms of the frequency of exceptionally hot or cold days) does not influence wage differentials across US states, once controlling for market access. On the contrary, Roback (1982) finds that the number of clear days and total snowfall have respectively negative and positive effects on wage differentials across US cities but no influence on house prices.

References

Anselin, L. (1988), *Spatial econometrics: methods and models*, Kluwer Academic Publishers, Amsterdam.

Alcalà, F. and A. Ciccone (2004), Trade and productivity, *Quarterly Journal of Economics* 119, 613-646.

Ciccone, A.(2002), Agglomeration effects in Europe, *European Economic Review* 46, 213-227.

Ciccone, A. and R. Hall (1996), Productivity and the density of economic activity, *American Economic Review* 86, 54-70.

Conley, T., Flyer F. and G. Tsiang (2003), Spillovers from local market human capital and the spatial distribution of productivity in Malaysia, *Advances in Economic Analysis & Policy* 3, Article 5.

Economic Council (2001), Regional development and regional policy in Finland, Prime Minister's Office Publications 2001/2.

Hanson, G. (1998), Market potential, increasing returns, and geographic concentration, NBER Working Paper No. 6429.

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 (2004), The empirics of agglomeration and trade, in: J.V. Henderson and J.-F. Thisse, eds., *Handbook of regional and urban economics*, Vol. 4, Elsevier, Amsterdam.

Head, K and T. Mayer (2005), Regional wage and employment responses to market potential in the EU, CEPR Discussion Paper No. 4908.

Helpman, E. (1998), The size of regions, in: D. Pines, E. Sadka and I. Zilcha, eds., *Topics in Public Economics. Theoretical and Applied Analysis*, Cambridge University Press, Cambridge.

Honkapohja, S. and E. Koskela (1999), The economic crisis of the 1990 in Finland, *Economic Policy* 14, 399-436.

Jaffe, A. and M. Trajtenberg (2002), *Patents, citations and innovation: a window on the knowledge economy*, MIT Press, Cambridge.

Jaffe, A., M. Trajtenberg and R. Henderson (1993), Geographic localization of knowledge spillovers as evidenced by patent citations, *Quarterly Journal of Economics* 63, 577-598.

Kangasharju, A. (1998), Regional economic differences in Finland: variations in income growth and firm formation, Pellervon Economic Research Instituute Publication n. 17.

Kangasharju, A. and S. Pekkala (2004), Increasing regional disparities in the 1990s: The Finnish experience, *Regional Studies* 38, 255-267.

Kangasharju, A, S. Laakso, H. Loikkanen, M. Riihelä and R. Sullström (2001) Economic crisis of the 1990s: what happened to regional convergence and inequality, and housing market phenomena in boom and bust?, in: J. Kalela, J. Kiander, U. Kiuvikuru, H. Loikkanen and J. Simpura, eds., *Down from the heavens, up from the ashes: the Finnish economic crisis of the 1990s in the light of economic and social research* (VATT-Publications, Government Institute for Economic Research: Helsinki).

Knaap, T. (2005), Trade, location, and wages in the United States, Utrecht School of Economics Working Paper No.05-06.

Krugman, P.(1991), Increasing returns and economic geography, *Journal of Political Economy* 99, 483-499.

Krugman, P. and A. Venables (1995), Globalization and the inequality of nations, *Quarterly Journal of Economics* 110, 857-880.

Moretti, E. (2004), Human capital externalities in cities, in: J.V. Henderson and J-F. Thisse, eds., *Handbook of regional and urban economics*, Vol. 4, Elsevier: Amsterdam.

Nivalainen, S., 2003, Where do migrants go? An analysis of rural and urban destined/originated migration in Finland in 1996-1999, PTT Working Paper n. 66.

OECD (1992), *Economic Surveys 1991-1992: Finland,* OECD Publications: Paris.

Ottaviano, G. and J.-F. Thisse (2004), Agglomeration and economic geography, in: J.V. Henderson and J.-F. Thisse, eds., *Handbook of Regional and Urban Economics* Vol. 4, Elsevier: Amsterdam.

Redding, S. and A. Venables (2004), Economic geography and international inequality, *Journal of International Economics* 62, 53-82.

Roback, J. (1982), Wages, rents and the quality of life, *Journal of Political Economy* 90, 1257-1278.

Rouvinen, P. and P. Ylä-Anttila (2003), Case-study: little Finland’s transformation to a wireless giant, in: S. Dutta, B. Lanvin and F. Paua, eds., *The global information technology report* 2003-2004, Oxford University Press, New York.

Suomen Kuntaliitto (1999), Menestys kasaantuu-alueet erilaistuvat. Aluekehityksen suunta 1990-luvulla, Kuntaliiton painatuskeskus: Helsinki.

Taipale M., (2002), Convergence of production and incomes between Finnish subregions, PTT Working Paper 58.

Temple, J. (1999), The new growth evidence, *Journal of Economic Literature* 37, 112-156.

Temple, J. (2001), Generalizations that aren't: evidence on education and growth, *European Economic Review* 45, 905-918.

Tervo, H. (2004), Regional policy lessons from Finland, in: D. Felsenstein, B.A. Portnov, eds., *Regional Disparities in Small Countries* (Springer Verlag: Berlin).

Venables, A. (1996), Equilibrium locations of vertically linked industries, *International Economic Review* 37, 341-359.

Annex: The data

Data are kindly supplied by Pellervon Taloudellinen Tutkimuslaitos (PTT, Pellervo Economic Research Institute, Helsinki).

List of variables

pop77_90	Population OLS growth trend	1977-1990	% pa
pop87_90		1987-1990	% pa
pop94_02		1994-2002	% pa
pope77_90	Population expon. growth rate	1977-1990	% pa
pope87_90		1987-1990	% pa
pope94_02		1994-2002	% pa
apop77_90	Adjusted population OLS growth trend	1977-1990	% pa
apop87_90		1987-1990	% pa
apop94_02		1994-2002	% pa
tinc77_90	Taxable income per capita OLS growth trend	1977-1990	% pa
tinc87_90		1987-1990	% pa
tinc94_02		1994-2002	% pa
tince77_90	Taxable income per capita expon. growth rate	1977-1990	% pa
tince87_90		1987-1990	% pa
tince94_02		1994-2002	% pa
pinc77_90	Primary income per capita OLS growth trend	1977-1990	% pa
pinc87_90		1987-1990	% pa
pinc94_02		1994-2002	% pa
pemp77_90	Primary income per employed OLS growth trend	1977-1990	% pa
pemp87_90		1987-1990	% pa
pemp94_02		1994-2002	% pa
rent87_90	Rental growth	1987-1990	% pa
rent94_02	Rental growth	1994-2002	% pa

PriEmp95	Log (Primary Income per employed)	1995	'000 Euro
PriInc95	Log (Primary Income per capita)	1995	'000 Euro
TaxInc77	Log (taxable Income per capita)	1977	'000 Euro
TaxInc87	Log (taxable Income per capita)	1987	'000 Euro
TaxInc94	Log (taxable Income per capita)	1994	'000 Euro
MPTinc77	Log (Mk Potential), based on Taxable Income	1977	
MPTinc87	Log (Mk Potential), based on Taxable Income	1987	
MPTinc94	Log (Mk Potential), based on Taxable Income	1994	
MPIncPC77	Log (Mk Potential), based on Taxable Income per capita	1977	
MPIncPC87	Log (Mk Potential), based on Taxable Income per capita	1987	
MPIncPC94	Log (Mk Potential), based on Taxable Income per capita	1994	
MPPop77	Log (Mk Potential), based on population	1977	
MPPop87	Log (Mk Potential), based on population	1987	
MPPop94	Log (Mk Potential), based on population	1994	
MPDens77	Log (Mk Potential), based on density	1977	

MPDens87	Log (Mk Potential), based on density	1987	
MPDens94	Log (Mk Potential), based on density	1994	
EduSec77	share pop with at least upper sec degree (at least 10-11 years of education)	1977	ratio
EduSec87	share pop with at least upper sec degree (at least 10-11 years of education)	1987	ratio
EduSec94	share pop with at least upper sec degree (at least 10-11 years of education)	1994	ratio
EduTer77	share pop with at least tertiary degree (at least 13 years of education)	1977	ratio
EduTer87	share pop with at least tertiary degree (at least 13 years of education)	1987	ratio
EduTer94	share pop with at least tertiary degree (at least 13 years of education)	1994	ratio
Age77	Median age	1977	years
Age87	Median age	1987	years
Age94	Median age	1994	years
Pat90-93	Number of patents per capita	Average, 1990-93	patents/inhab
Pat91-94	Number of patents per capita	Average, 1991-94	patents/inhab
RSD95	Research & Development expenditure per capita	1995	mil Euro/inhab
RSD9598	Research & Development expenditure per capita	Average, 1995-1998	mil Euro/inhab
RSD9502	Research & Development expenditure per capita	Average, 1995-2002	mil Euro/inhab
Den77	Density of population	1977	inhab/Km2
Den87	Density of population	1987	inhab/Km2
Den94	Density of population	1994	inhab/Km2
Lake	% land covered by lakes		Ratio
Temp	Average temperature	Average, 1971-2000	℃
Une77	Unemployment rate	1977	ratio
Une87	Unemployment rate	1987	ratio
Une94	Unemployment rate	1994	ratio
Gov94	Government expenditure	1994	'000 Euro/capita
Gov94_02	Government expenditure	Average, 1994-2002	'000 Euro/capita
Agr77	Share of *Agriculture, hunting, forestry and fishing* in total employment	1977	%
Agr87	Share of *Agriculture, hunting, forestry and fishing* in total employment	1987	%
Agr94	Share of *Agriculture, hunting, forestry and fishing* in total employment	1994	%
Man77	Share of *Manufacturing and construction* in total employment	1977	%
Man87	Share of *Manufacturing and construction* in total employment	1987	%
Man94	Share of *Manufacturing and construction* in total employment	1994	%
Air	Average distance from main airports		Km
Bord	Average distance from Russian crossing borders		Km
Port	Average distance from ports		Km
Hst	Average distance from high-speed railways train		Km
ICT	Share of employment in ICT	Average 1987-1995	%
Pop77	Number of inhabitants	1977	inhab
Pop87	Number of inhabitants	1987	inhab
Pop94	Number of inhabitants	1994	inhab
rent87	Rental level	1987	'000 Euro/m2
rent94	Rental level	1987	'000 Euro/m2
gra77	Central government grants per capita	1977	Mil Euro/capita
gra87	Central government grants per capita	1987	Mil Euro/capita
gra94	Central government grants per capita	1994	Mil Euro/capita
gra77_90	Central government grants per capita	Average 1977-	% pa

		1990	
gra87_90	Central government grants per capita	Average 1987-1990	% pa
gra94_02	Central government grants per capita	Average 1994-2002	% pa
emp77	Employment rate (employment/working age population)	1977	%
emp87	Employment rate (employment/working age population)	1987	%
emp94	Employment rate (employment/working age population)	1994	%

Note(s): All values are expressed in constant 2000 prices.

List of Regions

n.	**Code**	**Name**
1	011	Helsinki
2	012	Lohja
3	013	Tammisaari
4	021	Åboland-Turunmaa
5	022	Salo
6	023	Turku
7	024	Vakka-Suomi
8	025	Loimaa
9	041	Rauma
10	042	Kaakkois-Satakunta
11	043	Pori
12	044	Pohjois-Satakunta
13	051	Hämeenlinna
14	052	Riihimäki
15	053	Forssa
16	061	Luoteis-Pirkanmaa
17	062	Kaakkois-Pirkanmaa
18	063	Etelä-Pirkanmaa
19	064	Tampere
20	068	Lounais-Pirkanmaa
21	069	Ylä-Pirkanmaa
22	071	Lahti
23	072	Heinola
24	081	Kouvola
25	082	Kotka-Hamina
26	091	Lappeenranta
27	092	Länsi-Saimaa
28	093	Imatra
29	094	Kärkikunnat
30	101	Mikkeli
31	102	Juva
32	103	Savonlinna
33	105	Pieksämäki
34	111	Ylä-Savo
35	112	Kuopio
36	113	Koillis-Savo
37	114	Varkaus
38	115	Sisä-Savo
39	121	Outokumpu
40	122	Joensuu
41	123	Ilomantsi
42	124	Keski-Karjala

43	125	Pielisen Karjala
44	131	Jyväskylä
45	132	Kaakkoinen Keski-Suomi
46	133	Keuruu
47	134	Jämsä
48	135	Äänekoski
49	136	Saarijärvi
50	137	Viitasaari
51	141	Suupohja
52	142	Pohjoiset seinänaapurit
53	143	Eteläiset seinänaapurit
54	144	Kuusiokunnat
55	145	Härmänmaa
56	146	Järviseutu
57	151	Kyrönmaa
58	152	Vaasa
59	153	Sydösterbottens kustregion
60	154	Jakobstadsregionen
61	161	Kaustinen
62	162	Kokkola
63	171	Oulu
64	173	Ii
65	174	Raahe
66	175	Siikalatva
67	176	Nivala-Haapajärvi
68	177	Ylivieska
69	178	Koillismaa
70	181	Kehys-Kainuu
71	182	Kajaani
72	191	Rovaniemi
73	192	Kemi-Tornio
74	193	Torniolaakso
75	194	Itä-Lappi
76	196	Tunturi-Lappi
77	197	Pohjois-Lappi
78	201	Porvoo
79	202	Loviisa

CHAPTER 3

MEASURING DIVERSITY: A CROSS-DISCIPLINARY COMPARISON OF EXISTING INDICES

1. Introduction

Quantitative measures of diversity are necessary to many fields of scientific investigation. This has led to the development of a variety of indices of diversity. In this paper we review the different indices and approaches proposed. The objective is to provide a common framework for their selection, use and interpretation in empirical analyses.

The literature is large and spans several discipline. We will therefore set the following boundaries to our investigation. First, we will not discuss why diversity is important in the different fields. The focus is on how diversity is measured. Second, we will not discuss whether and how similarities and differences can be identified. In biology, this is done by classifying individual into types (‘species’). The underlying criteria are clear: individuals belong to the same species (i.e., are similar) if and only if they are able to reproduce. In other fields, the feasibility and implications of classifying individuals is more controversial. In particular, psychologists and anthropologists have shown that one’s identity is defined dynamically and in relation with other people, which contradicts the use of fixed categories and typologies. We will not enter this debate. Our focus is on how a synthetic index of diversity can be constructed once the key individuals’ characteristics and the correspondent types are identified.

We show that the crucial distinctions between the plethoras of indices available across various disciplines arise from the specific components of diversity they aim at capturing: richness, evenness or distance, or combinations of the three. Indeed, when targeted at the same component(s) of diversity, different indices yield very similar results. Most naturally, differences emerge only when the components of diversity addressed are in fact different. In particular, the indices measuring only evenness differ substantially from those measuring only and from those that consider distance between species as well.

This paper is co-authored with Gianmarco I.P. Ottaviano, University of Bologna, CEPR and FEEM. Elena Bellini provided excellent research assistance. We thank Philippe van Parijs for his advice. We thank also Guus Extra, Eliana La Ferrara and all participants to the first EURODIV Conference ‘Understanding diversity: Mapping and measuring’, Milano, January 2006. Patrizia Zanoni, Maddy Janssens and Arie de Ruijter and all members of the ENGIME network contributed greatly to our understanding of diversity.

The rest of the paper is organised as follows. Section 2 sets out the framework of investigation. Section 3 reviews the indices of diversity that take into account only the number of types (richness) and their relative abundances (evenness). Section 4 deals with indices that take also into account the extent to which types are different (distance). Section 5 reports an application of the indices presented to a dataset of cultural diversity in US metropolitan areas. Section 6 concludes.

2. A framework for investigation

An early conceptualisation of diversity is set out in Whittaker (1972) who distinguishes between: 1) Inventory diversity, concerning diversity *within* defined geographical (or temporal) units (which can be defined at different resolution). Depending on the geographical unit, Whittaker (1972) further distinguishes between α–diversity (diversity within a habitat), γ–diversity (landscape) and ε–diversity (bio-geographic province); and 2) Differentiation diversity, concerning the variation of diversity *across* geographical (or temporal) units. Whittaker (1972) Depending on the geographical unit, Whittaker (1972) further distinguishes between β-diversity (across habitats within a landscape), and δ-diversity (across landscapes within a bio-geographic region).

In this paper, we will only deal with Inventory (or α–diversity). A simple example will help in identifying its key components.

Consider a population A of 30 individuals and assume that 10 individuals speak English, 10 speak Italian and 10 speak French. Consider now a population B that includes also Spanish speakers. Since the number of types represented in population B is larger than in population A, it is rational to consider population B *more diverse* than population A. On the contrary, consider a population C constituted by 28 English speakers, 1 Italian speaker and 1 French speaker. The number of types represented in population C is the same than in population A. However, two types have a very small number of individuals. Therefore, it is rational to consider population C *less diverse* than population A. Finally, consider a population D where 10 individuals speak English, 10 Italian, and 10 Japanese. The number of types is the same than in population A. As in population A, the population is evenly distributed across types. However, Japanese is (in any language taxonomy we can think of) more different than French from English and Italian. Population D should therefore be considered *more diverse* than population A. In more general terms, the diversity of a population will depend on:

- the number of types represented (which we will refer to as the *richness* dimension of diversity). Diversity increases with the number of types in the population. The determination of richness requires the identification of types on the basis of a set of criteria;
- the relative abundance of types (which we will refer to as the *evenness* dimension of diversity). Diversity increases with the evenness of the distribution of individuals across types. Given richness, diversity reaches its maximum when all types are

equally represented. The determination of evenness requires the distribution of the population across types;

- the differences that characterise one type from the others (which we will refer to as the *evenness* dimension of diversity). The more types are different from each other, the more diverse is the population. The determination of distance requires some form of metric of differences between types.

In what follows, we consider a population Ω of N individuals belonging each to one and only one of S types. Types are identified on the basis of a given criterion (or set of criteria). [1] Let:

N = the number of individuals in the community;

S = the number of types;

n_i = the number of individuals in the s_i type (abundance);

$(p_1, \dots p_s)$ = the vector of ordered (from the least to the most) relative abundances, where $p_i=n_i/N;$

d_{ij} = the distance between type i and type j.

An index can therefore be defined a function that maps population Ω in the domain of real numbers (positive). Section 3 will discuss those indices that consider only the *richness* and *evenness* dimension of diversity. The underlying assumption is that. d_{ij} are constant across all i and all j. This assumption will be relaxed in Section 4 that will consider those indices that take into account *distance* as well.

3. Measuring diversity as richness and evenness

Pielou (1975) identifies the following two properties of a diversity index:

- P1: if the relative abundances are equal then the index is an increasing function of S (ie, diversity increases with ‘richness’);
- P2: for fixed S, the index increases as the relative abundances become more equal (ie, diversity increases with ‘evenness’).

Richness is a well defined concept and Property P1 is straightforward. On the contrary, the concept of evenness underlying property P2 may require further investigation. The large literature on income inequality is of help. Such literature identifies the *transfer* principle (or Pigou-Dalton principle) as the essence of inequality. The principle states that inequality should increase (equality should decrease) for any rank-preserving transfer of income from poorer to richer individuals (and vice-versa). Similarly, we can state that ‘evenness’ should increase following any rank-preserving transfer from less to more abundant types and should decrease for any rank-preserving transfer from more to less abundant types. This property can be expressed in mathematical terms by the strict

[1] As discussed in the Introduction, we will not deal with the issue of whether and how it is possible to identify the types. We will simply assume that types are identified.

Schur-convexity of the index functional form. We will discuss other desired properties of the indices in Section 3.5.

In what follows, we start discussing a class of indices satisfying both Pielou properties. Such indices measure both richness and evenness and, as such, are the most used in the biological literature. Second, we discuss the indices tackling the evenness dimension of diversity only. They are mostly derived from the socio-economic literature on inequality. Third, we discuss a set of indices addressing two additional dimensions of diversity (dominance and polarisation) that, although related to abundance distributions, cannot be categorised under the 'evenness' dimension as they imply a violation of the Pigou-Dalton principle. Finally, we discuss a set of desirable properties that a diversity index should satisfy. The choice of the appropriate index should then be made with respect to this set of desirable properties.

We consider three types of transformations of the indices. First, some of the indices in the original form might assume values falling outside the interval [0,1], or are dependent on the unit of measurement and the scale of the phenomenon. When this is the case, we calculate the *relative* form of the index as:

(1) $$I^r = \frac{I-\alpha}{\beta-\alpha},$$

where α is the minimum value of the index (when all individuals are of one type) and β is its maximum (when all types have the same relative abundance). The values of I^r fall within the [0,1] interval and are independent on the unit of measurement and the scale of the phenomenon. Superscript *r* indicates such *relative* forms of the indices. Second, in their original form some of the indices measure un-evenness rather than evenness (i.e., the index increases when diversity decreases). In these cases, we adopt the complement *(1-I),* the reciprocal *(1/I)* or the opposite (-*I*) of the original index (depending on literature and intuitive appeal of the transformation). Finally, additional transformations are provided in literature to isolate the evenness dimension in some of the indices measuring both richness and evenness. Superscript *e* indicates such *evenness* forms of the indices.

3.1. The Good generalized index (and its family)

The Good generalized index of diversity is expressed as:

(2) $$H(\alpha,\beta) = \sum_{i=1} p_i^{\alpha}[-\log(p_i)]^{\beta}$$

where (α, β) are integer. Baczkowski *et al* (1998) further generalized Good's index so that (α, β) take value in the real plane R^2 and they determine the range of values for which H(α, β) satisfies the Pielou properties. In particular, they show that H(α, β) satisfies the Pielou properties in a closed region within the quadrant $0<\alpha \leq 1$ and $\beta \geq 1$. In the region $\alpha>0$ and $\beta<0$, the index varies inversely with diversity and therefore its complement, inverse or opposite should be used (see below the discussion of the Simpson index).

Setting (α, β) at appropriate values, the H(α, β) index yields a number of indices widely used in literature.

Richness index [H(0,0)]

When $(\alpha=0, \beta=0)$, then :

(3) $$H(0,0) = \sum_{i=1}^{S} 1 = S\ .$$

In this case, the Good index is equivalent to the simple counting of types (richness).

Shannon index [H(1,1)]

When $(\alpha=1, \beta=1)$, then :

(4) $$H(1,1) = -\sum_{i=1} p_i \ \log_2(p_i) \ \ = SH$$

The index was firstly introduced by Shannon (1948) to measure the information content (entropia) of a message. Borrowing from physics, Shannon (1948) calculates that the information content (i.e., the number of bits necessary to describe it) of an outcome is equal to the inverse log (base 2) of the probability of the outcome: the higher its probability, the less its informative content (MacKay 1983). The index was then applied to biological studies on the assumption that the diversity content of a natural system can be measured in way similar to the information content of a message (Good 1953). Because of its origin in information theory, the log_2 basis is normally used (as a consequence the index measures bits of binary codes), but there are not compelling reasons for that. The Shannon Index takes into account both the richness and evenness dimensions of diversity. Stirling and Wilsey (2001) compare the results of various empirical analyses in literature to study the relative impact of the two dimensions on the value of the index.[2]

Shannon relative index

The Shannon index takes value between *0* (when all individuals are of one type) and log_2S (when all types have the same relative abundance). The relative form of the index is therefore:

(5) $$SH^r = \frac{SH}{\log_2(S)}$$

The index was firstly introduced by Pielou (1975) to measure the 'evenness' dimension of diversity only.[3] In fact, its value is independent on richness only up to approximately S=25 (Smith and Wilson 1996).

Other transformations of the index

[2] The index is also referred to as Shannon-Weaver index, a misunderstanding that arose because the original formula was published in a book by Shannon and Weaver (1949).

[3] The same objective can be obtained by using $\log_S$ rather than $\log_2$ in the calculation of SH.

Heip (1974) proposes an alternative measures which move forward towards the objective of independence from richness pursued by Pielou. The Heip Index is expressed as:

(6) $$SH^{e}{}_{Heip} = \frac{(e^{SH} - 1)}{S - 1}$$

Theil (1967) proposes a transformation to measure income inequality. The index is known as Theil index. Its opposite can be used to measure diversity and it is expressed as:

(7) $$SH_{\ Theil} = \frac{1}{S}\sum_{i=1}^{i=S}\frac{n_i}{\overline{n}}\log_2\frac{n_i}{\overline{n}} = SH - \log(S)$$

where $\overline{n} = N / S$ is the average abundance.

Buzas and Hayek (1998) and Hayek and Buzas (1997) note that a simple evenness measure SH^{e}_{BH} can be obtained by dividing e^{SH} by the number of types S. The Shannon index can therefore be decomposed into the sum of its components:

(8) $$SH = \log(S) + \log(SH^{e}_{BH})$$

where $SH^{e}_{BH} = e^{SH} / S$.

Simpson index [1-H(2,0)]

When *(α=2, β=0),* then :

(9) $$H(2,0) = \sum_{i=1} p_i^2$$

The index has a simple interpretation in terms of the probability that any two individuals drawn at random from an infinitely large community belong to the same type.[4] As such, *H(2,0)* in fact increases with un-evenness. Its complement can therefore be used to calculate the probability that any two individuals drawn at random from an infinitely large community belong to different types: [5]

(10) $$SI = 1 - H(2,0)$$

Its simple interpretation can easily explain why this index has been used in very different strands of research. In biology, it is used to measure biodiversity and it is referred to as *Simpson index* from the name of the author who firstly introduced the index (Simpson 1949). In genetics, it occurs under the name of *heterozygosity index*

[4] The correct formula for a finite community is $\sum \frac{n_i(n_i - 1)}{N_i(N - 1)}$, where n_i is the number of individuals in the *i-th* species (see Magurran, 2004, p 115).

[5] In fact, the reciprocal is the most used in the biological literature (see Magurran 2004). Lande (1996) observes that the overall diversity of a set of communities measured as *(1/SI)* may be less than the average diversity of each community (a notion that is 'intuitively intriguing', see Magurran, 2004, p 115) and suggests the use of *(1-SI)* (which is actually widely used in the economic literature - see for example Alesina et al 2003). Rosenzweig (1995) recommends *–ln(SI)* (a transformation firstly used by Pielou 1975). Rosenzweig notes that in this transformation the index better reflects underlying diversity.

(Sham 1998; Svensson 2002). It also called the *Yule index* as a similar index was used by Yule to characterize the vocabulary used by different authors (Magurran 2004). Sociologists and economists use it to measure the degree of ethnic and cultural *fractionalisation* of countries and regions (Alesina et al 2003; Alesina e La Ferrara 2005; Ottaviano and Peri 2005, 2006). In economics, it is also used in its direct form H(2,0) to calculate the degree of market concentration (with p_i equal to the market shares of firms). In such a context, it is normally referred to *Herfindal index of market concentration* (Herfindal 1950).

Simpson relative index

Simpson index of diversity takes value between *0* and ($(S-1)/S$). The relative index is therefore:

(10')
$$SI^r = \frac{S}{S-1} SI$$

Other transformations of the index

With respect to the Shannon Index, the Simpson Index weights less the rare types and it is therefore less dependent on richness. Despite this fact, biologists have tried to obtain pure evenness measures. Smith and Wilson (1996) and Krebs (1999) propose the following measures (based on the reciprocal transformation of H(2,0)):

(11)
$$SI_{SW}^e = \frac{1/H(2,0)}{S}$$

Smith and Wilson (1996) construct this index to decompose the Simpson Index into its richness and evenness dimensions:

(11')
$$1/SI = SI_{SW}^e * S$$

3.2. Indices derived from the inequality literature

This section collates two sets of indices that were originally developed to measure income inequality. The first set of indices is derived from the Lorenz curve (Lorenz 1905). The second set is derived from the statistical concept of variance.

Measures based on the Lorenz curve: the Gini coefficient

The Gini coefficient (Gini 1912) measures the area between the Lorenz curve and the line of equal distribution. Let Θ_i be the cumulative share of individuals up to type *i*:

$$\Theta_i = \frac{1}{N}\sum_{j=1}^{i} p_j$$

the area between the Lorenz curve and the line of equal distribution is given by:

$$S = \frac{1}{2} - \frac{1}{2}\sum_{i=1}^{S}\frac{(i-(i-1))}{S}(\Theta_{1+i} + \Theta_i)$$

Gini proposes to divide this value by the area of the triangle below the line of equidistribution (equal to ½). The Gini index is therefore a *relative* index and it is expressed by:

(12) $$G = 1 - \sum_{i=1}^{S}\frac{(i-(i-1))}{S}(\Theta_{1+i} + \Theta_i)$$

As such, the Gini coefficient is a measure of un-evenness. Its complement *(1-G)* can be used as a measure of diversity and corresponds to the area below the Lorenz curve (relative to its maximum). It can be shown that the Gini coefficient is equivalent to the average linear difference between any two types abundance (in its relative form - Leti 1983, pp. 464-466).

Measures based on the variance: the Coefficient of variation and the Smith and Wilson evenness index

The variance is a standard measure of variability of a distribution. Applied in this context, it measures the average square deviation of abundances from mean abundance.

(13) $$\sigma^2 = \frac{1}{S}\sum_{i=1}^{S}[n_i - \bar{n}]^2$$

where $\bar{n} = N/S$ is the mean abundance. As such, the variance depends on the unit of measurement and the scale of *n*, and increases with the un-evenness of the distribution. We therefore need a twofold transformation in order to make it apt to measure the evenness dimension of diversity.

First, the variance assumes values in the interval between *0* (all abundances are equal to the mean) and $\bar{n}^2(S-1)$. Its *relative* form is therefore:

(13’) $$\sigma^{2^r} = \frac{\sigma^2}{(S-1)\bar{n}^2} = \frac{1}{S(S-1)}\sum_i[\frac{n_i - \bar{n}}{\bar{n}}]^2$$

It can be shown that σ^{2^r} is equivalent to the (relative) *H(2,0)* and to the (relative) squared differences of abundances (Leti 1983, pp 367-468). By dividing the variance by (squared) mean abundance we obtain the Coefficient of Variation, which is independent on the unit of measurement and *n*:

(13’’) $$CV = \frac{1}{S}\sum_i[\frac{n_i - \bar{n}}{\bar{n}}]^2 = \frac{1}{(S-1)}\sigma^{2^r}$$

The Coefficient of Variation can be expressed in its relative form as:

(13’’’) $$CV^r = \frac{CV}{CV_{mas}}\text{, where } CV_{\max} = S(S-1)$$

Second, relative variance and the coefficient of variation are measure of un-evenness. In the empirical application, we will use the complement to 1. It can be shown that the complement of relative variance is equivalent to SI^r and to the (relative) mean of squared differences of abundances (while the Gini coefficient is equal to the relative mean of linear differences. Leti 1983, pp 367-468).

Smith and Wilson (1996) have recently proposed an alternative standardisation of the index, by dividing the variance over log variance to give proportional difference and to make the index independent of the units of measurement. The *-2/π* provides the adequate scaling for the index to vary in the interval [0,1]:

(14) $$SW^e = 1 - [\frac{2}{\pi \arctan[\sum_{i=1}^{S}(\log n_i - \sum_{j=1}^{S}\log n_j / S)^2 / S]}]$$

3.3. Other indices of diversity

A variety of other indices are available. They are developed on conceptually different bases:

- McIntosh index is based on the concept of Euclidean distance in a hyperplane;
- Nee, Harvey and Cotgreave index is based on the slope of the rank/abundance plot (see below);
- Log series α and log normal λ are based on ad-hoc assumption concerning the functional form of the types abundance;

McIntosh measure of diversity

The index (McIntosh 1967) is based on the assumption that a population can be represented as a point in a S-dimensional hypervolume. The Euclidean distance from its origin can then be used as a measure of diversity:

(15) $$U = \sqrt{\sum_{i=1}^{S} n^2{}_i}$$

As such the index depends on N (and S). The following two transformations make the index independent from N and S, respectively:

(15’) $$U' = \frac{N-U}{N-\sqrt{N}}$$

(15’’) $$U^e = \frac{N-U}{N-N/\sqrt{S}}$$

Nee, Harvey and Cotgreave’s evenness measures

The index (Nee *et al,* 1992) is given by the slope *(b)* of the rank-abundance plot. The plot is widely used in biology (see Magurran 2004, p.21) and it is also called Whittaker plot, named after the inventor (Whittaker 1965). It maps relative abundances against the rank of types (from the most to the least abundant): the steeper the curve the more diverse is the community.

(16) $$NHC = b$$

As such, the index heavily depends on the number of types (richness) and takes value in the interval (--∞, 0), with 0 equivalent to max of evenness. A measure of evenness independent from richness and varying in the interval [0,1] is obtained by the following transformation (Smith and Wilson 1996):

(16') $$NHC^{e} = -2/\pi \arctan(b/S),$$

A similar index is proposed by Kempton and Taylor (1976), who's *Q-Statistic* represents the slope of the cumulative types abundance curve in the interquartile interval (to exclude the rarest and most abundant types).

Log series α and log normal λ

The diversity index α is based on the assumption that the distribution of types abundance follows a log series model, where:

(17) $$n_i = \frac{\alpha x^{i}}{i}, i = 1.......S$$

and (as *x* approximate 1) α is approximately equal to the number of types represented by a single individual (Magurran 2004).

The diversity index λ is based on the assumption that the distribution of types abundance follows a log normal model, and:

(18) $$\lambda = S*/\sigma,$$

where σ is the standard deviation of the log normal distribution.

Camargo index

Camargo (1993) introduced a new measure of diversity based on pairwise comparison of types abundances (similarly to the index proposed by Alesina *et al* 2003 to measure polarisation. More details in Section 3.4). For a critical review see Mouillot and Lepetre (1999). The index is expressed as follows:

(19) $$C = 1 - [\sum_{i=1}^{S} \sum_{j=i+1}^{S} (\frac{p_i - p_j}{S})]$$

3.4. Indices measuring dominance and polarisation

This Section discusses two additional features of types distribution (dominance and polarisation) which cannot be categorized in the 'evenness' dimension of diversity as they imply a violation of the Pigou-Dalton principle.

Dominance

The concept of *dominance* is widely used in different disciplines. It refers to the dominant position (by alternative dimensions: proportion of income, population, industry's turnover etc.) of one group or individual over all the other groups or individuals. Dominance increases as transfers take place in favour of the dominant group. Dominance is not affected by transfers involving any other groups.

The Berger-Parker index (Berger and Parker 1970; May 1975) is a simple measure of *dominance* (see Section 2). It is based on the proportional abundance of the most abundant type:

(20) $$BP = N_{MAX} / N$$

It is easy to calculate and it has 'high biological significance' (Magurran 2004). Collier (2001) uses this measure to explore the effect of this dimension of diversity on economic and political outcome of developed and developing countries and finds that it performs significantly differently from the measure of diversity (referred to as *fractionalisation*) based on the Simpson index.

Polarisation

The concept of polarisation was developed when scholars realised that measures of inequality traditionally used neglect the population frequency in each category and therefore disregards information on how population is distributed across different income categories. Yet, such information may be relevant to socio-economic outcomes. Consider for example two populations. The first is uniformly distributed in ten income classes. The second shows a two-spike distribution concentrated on two points. Such polarisation could cause social tensions and conflicts in the second population. However, under any Lorenz-consistent inequality measure, its inequality is lower than in the first (Esteban and Ray 1994; Wolfson 1994).[6]

The key difference between inequality and polarisation is with regard to the Pigou-Dalton principle. When measuring inequality, the effect of a transfer depends on the direction of the transfer. On the contrary, when measuring polarisation, the effect of a transfer depends on the *relative size* of the groups involved. Esteban and Ray (1994) propose the following index of *polarisation*:

(21) $$ER = K \sum_{i=1}^{S} \sum_{j=1}^{S} p_i^{1+\alpha} p_j \left| d_{ij} \right|$$

[6] This is an *ad-hoc* example. Polarisation and inequality do not always conflict.

Short of data on cultural distances, Alesina *et al* (2003) impose an equal 'distance' across all types and obtain the following index:

(21')
$$ER' = K' \sum_{i=1}^{S} \sum_{j=1}^{S} p_i^{1+\alpha} p_j$$

They use (21') to measure cultural *polarisation* as opposed to cultural diversity (which they measure by the Simpson index). They find that ER' index is highly correlated with the Simpson index.

Montalvo and Reynald-Querol (2005) develop an alternative index defined by:

(22)
$$RQ = 1 - \sum_{i=1}^{S} \left(\frac{1/2 - p_i}{1/2} \right)^2 p_i$$

In (22) the square term within the sum captures the deviation of the share of each group from the share it would have in a completely polarized population (1/2).

3.5. Desired and actual properties of indices

The indices reviewed do not necessarily provide the same diversity ranking of populations (ie, population A can be more or less diverse than population B depending on the index adopted). Therefore, diversity cannot be understood as an *absolute* concept, but only *relatively* to the index chosen. The choice of the index is therefore crucial. This Section discusses a lit of desired properties of the indices. The choice of the appropriate index should then be made with respect to these properties.

In biological studies, Smith and Wilson (1996) provide a list of 4 'essential' requirements and 10 'additional' features. In the context of income inequality studies, Subramanian (2004) lists 4 'basic' and 6 'additional' properties (based on Shorrocks 1988 and Anand 1983). In what follows we discuss the properties that are in common to the two strands and provide an evaluation of the indices with respect to those properties. For terminology we draw on both sources, balancing rigour and intuitive appeal.

The following properties are in common:

- *Symmetry*. The index is invariant to permutation of types. It implies that 'all types are equal'. The same principle holds in inequality measurement: all individuals should be considered equal (Subramanian 2004);
- *Independence of richness.* The index does not depend on the number of types. It corresponds to the *replication invariance* property of income inequality indices (the index should be invariant with respect to any *k*-fold population replication of the distribution);
- *Independence on the number of individuals*. It corresponds to the *scale invariance* property of income inequality indices;

- *Transfer* property. The principle states that inequality should increase for any transfer of income from a poorer to a richer individual (and vice-versa). In economics, it is called the Pigou-Dalton principle (which can be expressed in mathematical terms by the strict Schur-convexity of the index functional form), Similarly, Smith and Wilson (1996) require that 'the measure will decrease if the abundance of the least abundance is reduced' and 'the measure will decrease if a very rare type is added to the community'.
- *Normalisation property.* The property concerns the range of values taken by the index. It is required that the index is independent from the unit of measurement and that the index takes a maximum value of 1 when types abundances are equal. Smith and Wilson (1996) also suggest that 'the minimum value is 0'; and that 'the minimum is achieved when abundances are as unequal as possible' (which should be achievable 'for any possible number of types');
- *Easy interpretation.* This principle is difficult to state in a formal way. It is explicit in Subramanian (2004) and it requires that the index has some intuitive appeal. Smith and Wilson (1996) explicit three characteristics: 'the index would respond in a intuitive way to changes in evenness'; 'the index would return an intermediate value for communities that would be intuitively considered of intermediate evenness' and that 'the index is close to its minimum when evenness is as low as it likely to occur in a natural community'.

Up to now the two approaches are in complete agreement. There are some differences with respect to the desirable sensitivity of the index to transfer at the upper and lower end of the distribution:

- *Transfer sensitivity.* Shorrocks and Foster (1987) require that an inequality index be more responsive to income transfer at the lower than at the upper end of an income distribution. On the contrary, Smith and Wilson (1996) prefer the measure to be symmetric with regard to rare and common types.

Two final properties concern the relationships between subgroup inequality and overall inequality. They are explicitly mentioned in Subramanian (2004). It is not included in the Smith and Wilson (1996) list. In the same field of biodiversity measurement, they emerge as a desired property from Magurran (2004):

- *Subgroup consistency.* It requires that, other things being equal, an increase in the evenness of a subgroup does not make the overall evenness to decrease;
- *Decomposability.* It is satisfied when the index is decomposable into a *within-groups* and a *between-group* component.

A final important property concerns the possibility of carrying out hypotheses testing on the indices. Indeed, we know the asymptotical distributions of some of them, which would allow hypothesis testing. This is an important feature of indices, whose discussion goes however beyond the scope of this paper.

Table 1 below summarises the indices and their properties. We only discuss here the first six properties. Additional insights on the transfer sensitivity properties of the indices will be discussed within the context of the artificial example provided in Section 3.6.

Table 1: Properties and indices

Indices			Properties					
			Symmetry	Independence from S	Independence from N	Transfer property	Normalisation property	Easy interpretation
Richness	S=H(0,0)	(3)	√		√			√
Dominance	BP	(20)	√		√		√	√
Polarisation	ER'	(21')	√		√			
Shannon	SH=H(1,1)	(4)	√		√	√		
Simpson	1-SI=1-H(2,0)	(10)	√		√	√	√	√
Gini	(1-G)	(12)	√		√	√		√
C. Variation	$(1\text{-}CV^r)$	(13''')	√		√	√		
McIntosh	U	(15)	√			√		
Camargo	C	(19)	√		√	√		
Shannon-e	Sh^e	(5)	√		√	√	√	
Simpson-e	SI^e	(11)	√	√	√	√	√	
Smith-Wilson	SW^e	(15'')	√	√	√	√	√	
McIntosh-e	U^e	(14)	√		√	√	√	

3.6. An example

In order to illustrate the properties of the indices reviewed, this Section provides an artificial example in which indices are calculated and compared for five alternative ‘standard’ artificial distributions (from the most even to the least): equidistribution, broken-stick model, geometric series, and two distributions very close to the one-gets-all situation. Figure 1 below shows the Lorenz curves of the distribution. The figure confirms that diversity (unevenness) decreases (increases) as we move farther from the equidistribution. Because Lorenz curves do not intersect, the ranking is complete.

Figure 1: Lorenz curves for five standard distributions

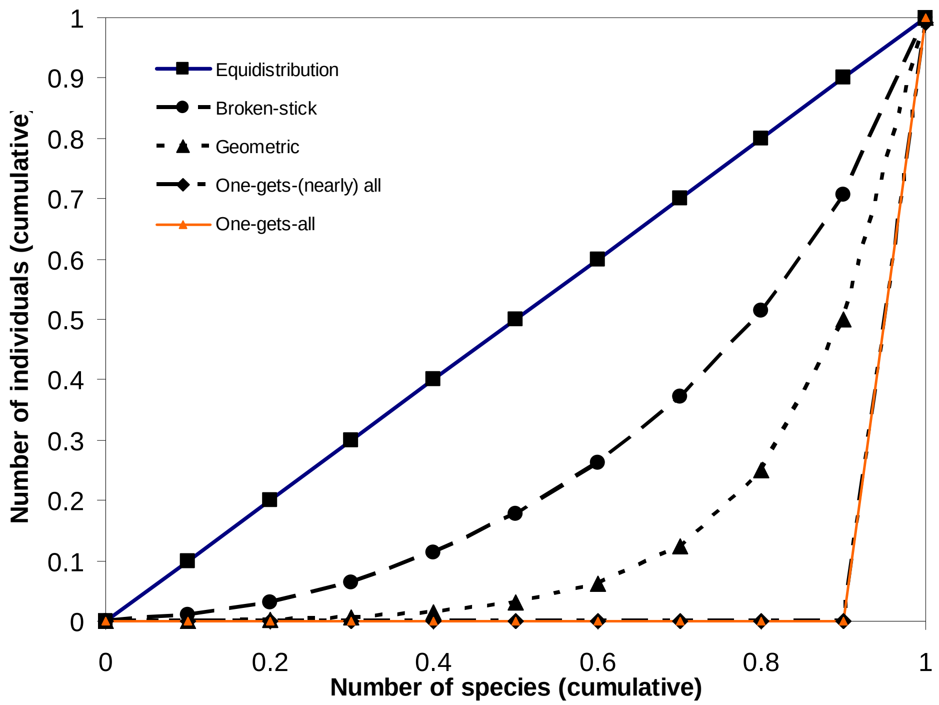


Table 2 shows the relative abundances and the values of the indices. The lower part of the table shows the percentage changes of indices when passing from a distribution to another.

Before discussing the table in details, some considerations concerning the indices reported are needed. *Richness* (the first index) is always set to ten types. This is why we do not report the '*only evenness*' version of the indices. The only exception is the McIntosh index. Here, the normalisation by the number of individuals and the number of types allow to obtain a value of the index within the [0,1] interval (instead of very large numbers). The Shannon index is calculated using $\log_{10}$. As the base is equal to the number of types, the maximum value of the index is one. Finally, the index based on the coefficient of variation is calculated according to equation (13''').

The second and third indices measure respectively dominance and polarisation. They increase as we move farther from the equidistribution. This is because polarisation and dominance are more closely related to the lack of diversity, rather than to diversity. In theory, a higher degree of dominance could co-exist with higher values of diversity indices, but this would require Lorenz-curves to cross at some point (and this is not the case with the simple distributions we are using here – see Figure 1). Short of distances between types, the index of polarisation simply mirrors the behaviour of diversity indices.

Consistently with the ranking provided by the Lorenz curves in Figure 1, the value of all indicators decreases as we move farther from the equidistribution. Nevertheless, important differences emerge when one considers the sensitivity of the indices to changes at the extremes of the distributions. Simpson, Shannon, coefficient of variation and McIntosh indices record the highest percentage changes ((see the lower part Table 2) when passing from a very uneven distribution to a (marginally) more even distribution. Gini and Camargo record the highest changes when moving from the equidistribution to a (marginally) more uneven distribution.

The appropriateness of the use of indices depends therefore on the context. If the issue is the conservation of rare types, the first group of indices is more appropriate (as reflected in their wide use in biological studies). If the issue is the deviation from equal distribution of income, the second group seems more appropriate (and the Gini coefficient is in fact more often employed in studies concerning income inequality).

Table 2: Relative abundances and the values of indices

			Species abundances				
	Species		Equidistribution	Broken-stick	Geometric	One-gets-(nearly) all	One-gets-all
	A		1,000	100	10	10	0
	B		1,000	211	20	10	0
	C		1,000	336	39	10	0
	D		1,000	479	78	10	0
	E		1,000	646	156	10	0
	F		1,000	846	313	10	0
	G		1,000	1,096	626	10	0
	H		1,000	1,429	1,251	10	0
	I		1,000	1,929	2,502	10	0
	L		1,000	2,929	5,005	9,910	10,000
Number of individuals			10,000	10,000	10,000	10,000	10,000
Mean abundance			1,000	1,000	1,000	1,000	1,000
Indexes							
Richness	S=H(0,0)	(3)	10.00	10.00	10.00	10.00	10.00
Dominance	BP	(20)	0.10	0.29	0.50	0.99	1.00
Polarisation	ER'	(21')	0.10	0.17	0.33	0.98	1.00
Shannon	SH=H(1,1)	(4)	1.00	0.86	0.60	0.03	0.00
Simpson	1-SI=1-H(2,0)	(10)	0.90	0.83	0.67	0.02	0.00
Gini	(1-G)	(12)	0.99	0.55	0.30	0.11	0.10
Coefficient of variation	(1-CVr)	(13''')	1.00	0.92	0.74	0.02	0.00
McIntosh	U	(15)	3,162	4,132	5,779	9,910	10,000.00
Camargo	C	(19)	1.00	0.55	0.30	0.11	0.10
McIntosh - evenness	Ue	(15'')	1.00	0.86	0.62	0.01	0.00
Percentage increase from previous distribution (in absolute value)							
Richness	S=H(0,0)	(3)		0	0	0	0
Dominance	BP	(20)		193	71	98	1
Polarisation	ER'	(21')		71	96	194	2
Shannon	SH=H(1,1)	(4)		14	30	95	100
Simpson	1-SI=1-H(2,0)	(10)		8	20	97	100
Gini	(1-G)	(12)		45	46	63	8
Coefficient of variation	(1-CVr)	(13''')		8	20	97	100
McIntosh	U	(15)		31	40	71	1
Camargo	C	(19)		45	46	63	8
McIntosh - evenness	Ue	(15'')		14	28	98	100

4. Dealing with types differences: introducing 'distance'

In this Section we discuss those indices that take into account of the 'distance' (a metric of difference) between types as one of the dimensions of diversity. In the example in Section 1, population C comprising 10 English, 10 Italian, and 10 Japanese speakers was considered *more diverse* than population A constituted by 10 English, 10 Italian, and 10 French speakers. This was because Japanese (in any language taxonomy we can think of) is more different than French from English and Italian.

Pairwise type distances as those considered above ('Japanese to English' or 'French to English') can be derived in a number of ways. In biology, distances are usually derived from phylogenetic information. If one assumes a perfect knowledge of the evolutionary process, the distance between two types can be measured in terms of the temporal distance from the nearest common ancestor. Similarly, language differences could be traced to some form of taxonomic trees (as in Fearon 2003, borrowing from Grimes and Grimes 1996). An alternative possibility is to derive distances by comparing types along a set of micro-characteristics. For example, language differences can be measured by the number of noncognate/cognate words (Weitzman 1992; Kruskal *et al* 1992). In architecture (Weitzman 1992), the distance between different types of building can be measured in terms of micro-characteristics such as the number of floors, the age of construction, the style, the use, the location. More in general, this approach can be used to deal with multidimensional cultural differences (by defining distances, for example, as weighted averages of differences in terms of language spoken at home, religion, type of employment).

However, our focus will not be on whether and how such distances can be measured. We will assume that pairwise distances d_{ij} exist. Our focus is on how to develop an index of collective dissimilarity, *given* the types and the pairwise distances between them.

We will start by discussing those measures that represent types by the present/absent binary choice (ie, indices that consider the *richness* and *distance* dimensions and disregard *evenness*).We will then review the attempts to take into consideration also the relative abundance of types.

4.1. Capturing richness and distance

Weitzman (1992, 1993, 1998) provides a framework for developing an index of diversity that takes distances into consideration.

Let define V(•) the value of the diversity function and let Q be a non-empty proper subset of our community Ω such that:

$$0 \subset Q \subset \Omega \tag{23}$$

Let j be any element belonging to Q but not to Ω:

(23’) $j \in \Omega / Q$

The standard definition of the distance from the point *j* to the set Q is given by:

(24) $d(j,Q) \equiv \min_{i \in Q} d(j,i)$

The measure *d(j, Q)* is a measure of diversity between type *j* and the collective set *Q*. When *d(j, Q)* is small, little diversity should be added to the diversity of *Q*. Correspondingly, for larger *d(j,Q)*, the increase of diversity should be higher. The distance *d(j,Q)* can be interpreted as a derivative or first difference of the value of the diversity function V(•).

(25) $\mathrm{V}(Q \cup \mathrm{j}) - V(Q) = d(j,Q)$

Equation (25) defines then a recursive algorithm to calculate the diversity of the population Ω. It starts by from an arbitrarily value assigned to a subset Q belonging to Ω, including only one species (arbitrarily chosen). The value of the diversity function of an enlarged Q’ is then calculated by bringing an additional species (arbitrarily chosen) in the set and so on. When a type is added, the value of the diversity function is increased by the distance between the new species and the closest already in the subset Q’.

In the general case, the outcome is path dependent. However, Weitzman shows that if it exists a function V(•) such that (24) is satisfied for all *j* and all *Q* and this is unique, then V(•) is *the* diversity function (up to a constant).

The perfect taxonomy

Weitzman shows that (24) is always satisfied and unique *only* in the case of perfect taxonomy (which is equivalent to say that distances are ultrametric: given a reference type, the two most distant types are at an equal distance from it – Figure 3 provides a formalisation of the condition). Ultrametric distances provide something like a integrability condition: ‘when any type becomes extinct, the loss of diversity equals the type distance from its closest relative, and this myopic formula can be repeated indefinitely over any extinction patterns, because any sub-evolutionary tree is also an evolutionary tree’ (Weitzman 1992, p. 370). The intuition is simple: the extinction of a type is equivalent to a branch being cut out of the tree. The length of the branch measures the diversity loss.

Therefore, in the case of perfect taxonomy, the diversity function (up to an additive constant) is given by the *length* of the evolutionary tree, calculated by the sum of lengths of all its vertical branches:

(26) $W = \sum_{i=2}^{S} d_i$, where $d_i = \min(d_{ij})$ for j=1…S and $j \neq i$.

Similar indices have been developed in biological studies (Faith 1992, 1994).

The general case

In the general case, where differences cannot be structured in a taxonomic tree, condition (24) is not satisfied for all Q and all j. Weitzman provides an algorithm (based on dynamic programming) that reduces the actual distances into a taxonomic tree. The algorithm proceeds by comparing actual distances between types and clustering the two types with minimum distances. In this way it produces a set of pairwise ultrametric distances that can be represented with a tree. The tree is just an approximation, but it can be shown that it is the 'most likely' approximation. The approximation bears a cost; while in the case of perfect taxonomy, the diversity of a subset of types can be inferred from the total diversity (as total length of the branch of the sub-tree), this is not possible in the general case (the artificial tree will re-arrange if a type is eliminated from the original tree).

Clarke and Warwick (Warwick and Clarke 1995, 1998 and 2001; Clarke and Warwick 1998 and 1999) develop an alternative approach that measures the average distance between two randomly selected types:

$$(27) \qquad CW = \frac{1}{S^2}\frac{\sum_{i=1}^{S}\sum_{j=1}^{S} d_{ij}}{2}$$

In a further generalisation, Bossert *et al* (2006) calculate distances across several dimensions to construct an index of diversity that simultaneously consider multiple individuals' characteristics (such as income, language, type of employment, sex).

Nehring and Puppe (2002) provide an index that abstracts from the concept of distance and use directly the micro-characteristics of individual observation as basic information. Suppose there exists a set of micro-characteristics $F = \{f_j, j = 1....F\}$. Each type s_i *(i=1…S)* is characterised by a subset of F. Each characteristic is assigned a weight λ_f. The index is then calculated as:

$$(28) \qquad NP = \sum_{j=1}^{F}\sum_{i=1}^{S} \begin{cases} \lambda_{f_j} & \text{if species } s_i \text{ has characteristic } f_i \\ 0 & \text{if species } s_i \text{ does not have characteristic } f_i \end{cases}$$

4.2. Capturing richness, distance and evenness

Clarke and Warwick (1998) and, independently Webb (2000), introduce also types abundance in their measure of diversity (see equation (27)). They propose two alternative forms of the index: in the first form (called 'taxonomic *diversity*') the index measures the average path length (that can be interpreted as 'expected distance') between two randomly chosen individuals (which may belong to the same type):

$$(29) \qquad F = \sum_{i=1}^{S}\sum_{j=1}^{S} d_{ij} p_i p_j$$

The index was actually firstly proposed by the linguist Greenberg (1956) to gauge linguistic diversity in a region (with a very minor difference: the index was based on cultural *resemblance,* rather than diversity. See Fearon 2003). Rao (1982, 1984) and Rao and Nayak (1985) arrive to the same index (called *Quadratic Entropy)* and provide some axiomatisation. Fearon (2003) uses this index to compare cultural and ethnic diversity across countries.

The second form (called 'taxonomic *distinctness*') represents the special case when all individuals are drawn from different types (it is obtained by dividing the first measure by the value it would take if all types belonged to the same hierarchical level in the tree). The two forms collapse in (25) if types can only be characterised in terms of presence/absence.

Desmet *et al* (2005) propose an index that use language distances to measure what they call Peripherality Diversity. The index depends on the relative distance between the dominant group and the minorities. As such it addresses the dominance and polarisation dimensions (rather than evenness). The index can be expressed as follows:

(30) $$PD = \sum_{i=1}^{S} p_i^{1+\alpha} d_{i0} + p_i p_0^{\;1+\alpha} d_{i0}\,,$$

where p_i is the share of type *i, i*=0 identifies the dominant type, α is an exponent between 0 and 1 and d_{i0} is the distance between type *i* and the most abundant type in the community.

4.3. Towards a unified approach

F measures the average distance between two randomly chosen individuals. This Section will explore the meaning and implications of F in two special cases: when pairwise distances are unknown, or types can be represented on a line.

Linear distances

If distances can be represented on a line (as it is the case for income classes), then we can assume that:

(31) $$d_{ij} = (y_j - y_i)\text{, with } j>i$$

where y_i and y_j are respectively the representative income level of type *i* and type *j*.

In this case:

(32) $$F = \sum_{i=1}^{S}\sum_{j=1}^{S}(y_j - y_i)p_i p_j = 2\sum_{i=1}^{S-1} y_i P_i (1-P_i) = 2\mu G^{7}$$

where:

[7] Proofs in Leti (1983, pp. 453-454 (formula De Finetti-Paciello) and pp. 464-466 (equivalence between Gini and average difference).

$P_i = \sum_{h=1}^{i} p_i$ is the cumulative relative abundances up to type (income class) *i;*

$\mu = \sum_{i=1}^{S} y_i\, p_i$ is the average income in the population; and

G is the Gini index of income concentration discussed in Section 3.1.

This application suggests that inequality can be understood as a special case of diversity, where distances are linearly organised.

Distances are unknown

This is the case of all indices discussed in Section 3. As distances between types are unknown, we can assume that:

$d_{ij} = \overline{d}$ for $i \neq j$

$d_{ij} = 0$ for $i = j$

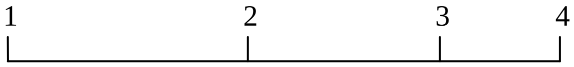
In this case, we have:

$$F = \overline{d}\sum_{i=1}^{S}\sum_{j=1}^{S} p_i\, p_j = \overline{d}(1 - \sum_{i=1}^{S} p_i^{\ 2}) , \tag{33}$$

which collapse into the Simpson index for $\overline{d} = 1$. Therefore, in this case, F measures the probability that any two individuals drawn randomly from an infinitely large community belong to the different types.

Figure 3: Representing Differences

Linear distances

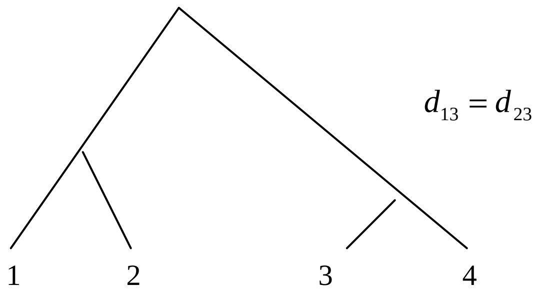


$$d_{13} = d_{12} + d_{23}$$

In the general case, the condition is expressed as:

$$d_{i(i+4)} = d_{i(i+1)} + d_{(i+1)(i+2)} + d_{(i+2)(i+4)}$$

Taxonomic tree



$$d_{13} = d_{23}$$

In the general case, ultrametric distances require (van Roovij 1978):

$$d_{ij} \leq MAX\{(d_{iz}),(d_{zj})\} \quad \forall i, j, z$$

5. An application

This Section provides a first application of the indices discussed in Sections 3 and 4. Values are calculated using the data from United States Current Population Survey for *language* spoken at home (29 language groups). Individual data are grouped by SMSA, Standard Metropolitan Statistical Area (160 SMSA). Data refer to the year 1990.

Table 4 reports a selection of indices for the 10 most diverse and 10 least diverse SMSA in the United States (ranked by the Simpson index). First, we report indices measuring richness, dominance and polarisation. Differently from the example in Table 2 richness varies across cities. It is generally higher in the top group of cities (confirming that richness is an important dimension of diversity). As in Table 2, dominance and polarisation are inversely related to diversity. In many cases dominance exceeds 90%. However, in the top cities dominance is at around 50% (implying that nearly half of the population do not speak English at home).

Second, we report the indices measuring both richness and evenness (Shannon, Simpson, McIntosh, Gini, coefficient of variation, and Camargo).[8] They are consistently high for the top group of cities. The Simpson index in those cities is generally above 0.50, which implies that their inhabitants have a probability above 50% to meet somebody who speaks a different language. The Camargo and Gini have the same values. The Camargo index is in fact the recalculation of the Gini as average difference between any two types abundances (see discussion in Section 3.2).

The third group includes all the indices measuring evenness only (they are indicated by the suffix *–e* in the table). Shannon, Simpson and McIntosh are calculated by using the transformations described in the text. [9] The Smith and Wilson index is itself a measure of evenness and does not need any transformation. The values of the (transformed) Simpson and the Smith and Wilson are not always higher for the top cities than for the bottom ones. This is because these indices measure only the evenness dimension of diversity, i.e. the extent to which the distribution of population across languages actually represented is close to an equidistribution: a population distributed evenly in two types is ranked higher than a population distributed unevenly in ten types. On contrary, the Shannon index is much higher for the top cities than for the bottom ones. This is because the transformation in equation (5) provides only an incomplete correction for richness.

The last two columns report the values of the indices taking into account also of the distances between languages. Distances between Indo-European languages are from Kruskal, Black and Dyen (1992) and are measured by the separation time between a pair of speech varieties. For the purposes of this simple exercise, distances between non Indo-European languages are set arbitrarily to three times the maximum distance between Indo-European languages. Two indices are reported. The Clarke and Warwick

[8] Equation (4), (10), (15), (12), (13''') and (19), respectively. In the calculation of Gini, Coefficient of Variation and Camargo we set to zero the abundance of types that are not represented in the city. For this reason, the indices vary with both evenness and richness.

[9] Equation (5), (11) and (15''), respectively.

and the F.[10] As expected, they are both higher for top group of cities than for the bottom ones.

In order to explore further the broad features of the groups of indices just discussed, Table 5 below shows the correlations between indices calculated across the 160 SMSA in the database.

The following features emerge. First, the indices measuring dominance and polarisation, and those measuring both richness and evenness move very closely together (pairwise correlation coefficients are in absolute value around or above 0.9 - apart for the McIntosh index, which is heavily affected by the size of population) and are strongly correlated with richness (with correlation coefficients often above 0.5). The index of Simpson and the coefficient of variation show perfect correlation (although their values are different – see Table 4). The group also includes the (transformed) Shannon.[11] This is because the transformation in (7) provides only an incomplete correction for richness. In Table 5, relevant cross-indices correlations are marked in bold.

Second, the (transformed) Simpson, the (transformed) McIntosh and the Smith-Wilson tend to move very closely (correlation coefficients generally above 0.7) and are negatively correlated with richness.[12] As a result they move quite independently from the indices in the first group. In Table 5, relevant cross-indices correlations are marked in bold.

Finally, among the indices that take distances into consideration, the Clarke and Warwick moves quite independently from most of indices but shows relatively high correlation with richness.[13] The F index shows correlation coefficients above 0.8 with most of the indices in the first group (and for this reason it is also in bold in Table 5). [14]

[10] Equation (27) and (29), respectively.
[11] Equation (7) and (15”), respectively.
[12] Equation (11), (15’’), (14), respectively
[13] Equation (27).
[14] Equation (29).

Table 4: Linguistic diversity in US cities

Index	Richness	Dominance	Polarisation	Shannon	Simpson	Gini	C. Variation	McIntosh	Camargo	Shannon-e	Simpson-e	McIntosh-e	Smith-Wilson-e	Clarke-Warwick	Fearon
City	S = H(0,0)	BP	ER'	SH = H(1,1)	SI = 1-H(2,0)	(1-G)	(1-CV^r)	U	C	SH^e	SI^e	U^e	SW^e	CW	F
	(3)	(20)	(21')	(4)	(10)	(12)	(13''')	(15)	(19)	(5)	(11)	(15'')	(14)	(27)	(29)
Jersey City	28	0.51	0.44	1.38	0.62	0.12	0.65	3962	0.12	0.41	0.09	0.48	0.54	318	100
Los Angeles	29	0.57	0.48	1.26	0.58	0.11	0.60	65339	0.11	0.37	0.08	0.43	0.54	14	122
Miami	27	0.51	0.51	1.01	0.56	0.09	0.58	14364	0.09	0.31	0.08	0.42	0.54	116	57
New York	28	0.64	0.50	1.38	0.55	0.13	0.57	53022	0.13	0.41	0.08	0.41	0.53	297	105
San Francisco	27	0.68	0.53	1.26	0.52	0.11	0.54	13462	0.11	0.38	0.08	0.38	0.54	79	166
San Antonio	26	0.58	0.56	0.84	0.51	0.07	0.52	8531	0.07	0.26	0.08	0.37	0.54	167	53
Corpus Christi	11	0.56	0.57	0.76	0.50	0.07	0.52	2008	0.07	0.32	0.18	0.42	0.57	425	48
San Jose	27	0.70	0.56	1.21	0.49	0.11	0.51	14534	0.11	0.37	0.07	0.35	0.54	130	139
Salinas	22	0.70	0.59	1.03	0.47	0.08	0.48	2287	0.08	0.33	0.09	0.34	0.54	97	85
El Paso	19	0.66	0.60	0.77	0.47	0.06	0.48	3858	0.06	0.26	0.10	0.35	0.56	386	49
Lima	10	0.97	0.95	0.18	0.06	0.04	0.06	1259	0.04	0.08	0.11	0.04	0.57	231	8
York	14	0.97	0.95	0.19	0.06	0.04	0.06	4384	0.04	0.07	0.08	0.04	0.56	366	9
Monroe	8	0.97	0.95	0.17	0.06	0.04	0.06	1417	0.04	0.08	0.13	0.04	0.58	1895	9
Green Bay	12	0.97	0.95	0.18	0.05	0.04	0.06	2089	0.04	0.07	0.09	0.04	0.56	1821	7
Chattanooga	16	0.97	0.95	0.17	0.05	0.04	0.06	4349	0.04	0.06	0.07	0.04	0.55	675	8
Macon	9	0.97	0.95	0.16	0.05	0.04	0.05	1509	0.04	0.07	0.12	0.04	0.57	637	8
Johnstown	9	0.97	0.95	0.17	0.05	0.04	0.05	3256	0.04	0.08	0.12	0.04	0.58	233	5
Muncie	12	0.97	0.95	0.17	0.05	0.04	0.05	1420	0.04	0.07	0.09	0.04	0.55	961	8
Springfield	14	0.97	0.95	0.17	0.05	0.04	0.05	2335	0.04	0.06	0.08	0.03	0.55	241	8
Altoona	12	0.98	0.96	0.14	0.04	0.04	0.04	1489	0.04	0.06	0.09	0.03	0.55	1363	6

Table 5: Correlation between indices

Index	Equation																
Richness	S = H(0,0)	(3)	1														
Dominance	BP	(20)	-0.32	1													
Polarisation	ER'	(21')	-0.38	**0.98**	1												
Shannon	SH = H(1,1)	(4)	0.50	**-0.92**	**-0.97**	1											
Simpson	SI = 1-H(2,0)	(10)	0.37	**-0.98**	**-1.00**	**0.97**	1										
Gini	(1-G)	(12)	0.58	**-0.82**	**-0.87**	**0.95**	**0.86**	1									
C. Variation	(1-CV^r)	(13''')	0.37	**-0.98**	**-1.00**	**0.97**	**1.00**	**0.86**	1								
McIntosh	U	(15)	0.65	**-0.29**	**-0.33**	**0.42**	**0.33**	**0.54**	**0.33**	1							
Camargo	C	(19)	0.58	**-0.82**	**-0.87**	**0.95**	**0.86**	**1.00**	**0.86**	0.54	1						
Shannon-e	SH^e	(5)	0.36	**-0.94**	**-0.98**	**0.98**	**0.98**	**0.90**	**0.98**	0.33	**0.90**	1					
Simpson-e	SI^e	(11)	-0.74	-0.32	-0.25	0.10	0.27	-0.03	0.27	-0.34	-0.03	0.26	1				
McIntosh-e	U^e	(15'')	-0.90	-0.07	-0.01	-0.11	0.02	-0.23	0.02	-0.57	-0.23	0.02	0.84	1			
Smith-Wilson-e	SW^e	(14)	-0.79	0.02	0.07	-0.20	-0.06	-0.34	-0.06	-0.39	-0.34	-0.07	0.76	0.78	1		
Clarke-Warwick	CW	(27)	0.53	-0.16	-0.18	0.23	0.17	0.26	0.17	0.27	0.26	0.16	-0.44	-0.48	-0.41	1	
F	F	(29)	0.42	-0.78	**-0.85**	**0.89**	**0.84**	**0.84**	**0.84**	**0.36**	**0.84**	**0.87**	0.09	-0.07	-0.11	0.22	1.00
			Richness	Dominance	Polarisation	Shannon	Simpson	Gini	C. Variation	McIntosh	Camargo	Shannon-e	Simpson-e	McIntosh-e	Smith Wilson	CW	F

6. Conclusions

The number of available diversity measures is so large that their comparison a daunting task. The reason is that the various indices have been developed in parallel by different scholars in different disciplines with different purposes.

We have proposed a systematization of the main statistical indicators of diversity beyond the different names that different disciplines often give to very similar measures. In so doing, we have clarified that the crucial distinctions between indices arises from the specific components of diversity they aim at capturing: richness, evenness or distance, or combinations of the three.

We have shown that, when targeted at the same component(s) of diversity, different indices yields very similar results. Most naturally, differences emerge only when the components of diversity addressed are in fact different. In particular, the indices measuring only evenness only might differ substantially from those measuring richness only or richness and evenness together. By showing how many indices are indeed closely related, our results provide a framework for comparing the methodology and the results of existing studies on diversity across disciplines. To future studies, our results also provide a toolbox that greatly simplifies the choice of the correct diversity measures. As discussed in Section 3.6, the choice of index will depend on the context and objectives of the research (Baumgärtner 2006). Nevertheless, our toolbox allows discarding complex and unintuitive indices where simple and intuitive ones provide comparable information. Take, for example, the case of distance between groups. At least in the context of linguistic diversity, our analysis suggests that most pieces of information contained in the sophisticated Clarke-Warwick and Fearon indices are already captured by simpler indices based on richness only or richness-plus-evenness respectively.

References

Anand S. (1983), *Inequality and Poverty in Malaysia: Measurement and Decomposition*, Oxford University Press, New York.

Alesina A., Devleeschauwer. A., Easterly W., Kurlat S. and Wacziarg R. (2003), Fractionalization, *Journal of Economic Growth* 8 (2), pp. 155-194.

Alesina A., E. La Ferrara (2004), Ethnic Diversity and Economic Performance, *Journal of Economic Literature*, 43, pp. 762-800.

Baczokwski A.J., Joanes D. and Shamia G. (1998), Range of Validity of α and β for a Generalised Index of Diversity $H(\alpha, \beta)$ due to Good, *Mathematical Biosciences* 148, pp. 115-128.

Baumgärtner S. (2006), Measuring the Diversity of What? And for What Purpose? A Conceptual Comparison of Ecological and Economic Biodiversity Indices, Working Paper, University of Heidelberg.

Berker W.H. and Parker F. L (1970), Diversity of Planktonic Foraminifera in Deep See Sediments, *Science* 169, 1345-1347.

Bossert W., C. D'Ambrosio and E. La Ferrara (2006), A Generalized Index of Fractionalization, Innocenzo Gasparini Institute for Economic Research, Working Paper 313.

Buzas M.A. and Hayek L.-A.C. (1998), SHE Analysis for Biofacies Identification, *J. Foraminiferal Res*. 28, pp. 233-239.

Camargo J.A. (1993), Must Dominance Increase with the Number of Subordinate Species in Competitive Interactions?, *Journal of Theoretical Biology* 161, pp. 537-542.

Clarke K. R. and Warwick R. M. (1998), A Taxonomic Distinctness Index and Its Statistical Properties, *Journal Applied Ecology* 35, pp. 523-531.

Clarke K. R. and Warwick R. M. (1999), The Taxonomic Distances Measure of Biodiversity: Weighting of Step Lengths Between Hierarchical Level, *Mar. Ecol. Prog. Ser.* 184, pp. 21-29..

Collier P. (2001), Implication of Ethnic Diversity, *Economic Policy* 32, pp. 129-166.

Desmet K, Ortuño I. and Weber (2005), Peripheral Diversity and Redistribution, CEPR Discussion Paper 5112.

Esteban, J-M and Ray. D (1994), On the Measurement of Polarization, *Econometria* 62, No. 4, p. 819-851.

Faith D.P. (1992), Conservation Evaluation and Phylogenetic Diversity, *Biological Conservation* 61, 1-10.

Faith D.P. (1994), Phylogenetic Pattern and the Quantification of Organismal Biodiversity, Phil. Trans. R. Soc. Lond B 345, pp. 45-58.

Fearon J.D. (2003), Ethnic and Culture Diversity by Country, *Journal of Economic Growth* 8 (2), pp. 195-222.

Greenberg J. (1956), The Measurement of Linguistic Diversity, *Language* 32, pp. 105-109.

Grimes J. and Grimes B. (1996), *Ethnologue: Languages of the World,* 13edn Summer Institute of Linguistics, Dallas, TX.

Gini C. (1912), Variabilità e mutabilità, *Studi Economico-Giuridici della Facoltà di Giurisprudenza dell'Università di Cagliari III, parte II.*

Good I. (1953), The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika* 40, p 237.

Hayek L.-A.C. and Buzas M. A. (1997), *Surveying Natural Populations*, Columbia University Press, New York.

Heip C. (1974), A New Index Measuring Evenness, *Journal of Maritime Biological Association UK* 54, pp. 555-557.

Herfindal O.C. (1950), *Concentration in the Steel Industry,* PhD Thesis, Columbia University.

Kempton R.A. and Taylor L. R. (1976), Models and Statistics for Species Diversity, *Nature* 262, 818-820.

Krebs C. J. (1999) *Ecological Methodology,* Harper and Row, New York.

Kruskal J. B., Black P. and Dyen I. (1992), An Indo-European Classification: A Lexicostatistical Experiment, *Transaction of the American Philosophical Society* 82, part 5.

Lande R. (1996), Statistics and Partitioning of Species Diversity, and Similarity among Multiple Communities, *Oikos,* 76, pp. 5-13.

Leti G. (1983), *Statistica descrittiva,* Il Mulino, Bologna.

Lorenz, M. O., (1905), Methods for Measuring the Concentration of Wealth, *American Statistical Association*, 9, pp. 209-219.

MacKay D. J. C. (2003), *Information Theory, Inference and Learning Algorithms,* Cambridge, Cambridge University Press.

Magurran A. E. (2004), *Measuring Biological Diversity,* Oxford, Blackwell Publisher.

May R.M. (1975), Patterns of Species Abundances and Diversity, in *Ecology and Evolution of Communities,* M. L. Cody and J. M. Diamond, pp. 81-120, Harvard University Press, Cambridge, MA.

McIntosh R.P. (1967), An Index of Diversity and the Relation of Certain Concepts to Diversity, *Ecology* 48, 392-404.

Montalvo J.C. and M. Reynald-Querol (2005), Ethnic Polarization, Potential Conflicts and Civil Wars, *American Economic Review*, 95, pp. 796-816.

Mouillot D. and Lepetre A. (1999), Introduction of Relative Abundance Distribution (RAD) Indices, Estimated from the Rank-Frequency Diagrams (RFD) to Assess Changes in Community Diversity, *Environmental Monitoring Assessment* 63, pp. 279-295.

Nee S., Harvey P. H. and Cotgreave P. (1992), Population Persistence and the Natural Relationship Between Body Size and Abundance, in *Conservation of Biodiversity for Sustainable Development,* O.T. Sandlund, K. Hindar and A.H.D. Brown, Scandinavian University Press, Oslo.

Nehring K. and C. Puppe (2002), A Theory of Diversity, *Econometrica,*70, pp. 1155-1198.

Ottaviano G.I.P. and Peri G. (2005), Cities and Cultures, *Journal of Urban Economics*, 58, pp. 304-337.

Ottaviano G.I.P. and Peri G. (2006), The Economic Value of Cultural Diversity: Evidence from US Cities, *Journal of Economic Geography*, 6, pp. 9-44.

Pielou E. C. (1975), *Ecological Diversity,* Wiley InterScience, New York.

Rao R. C. (1982), Diversity: Its Measurement, Decomposition, Apportionment and Analysis, *Sankhyā,* 44, A, pp. 1-22.

Rao R. C. (1984), Convexity Properties of Entropy Functions and Analysis of Diversity, in *Inequalities in statistics and Probability,* U.K. Tong Ed, IMS Lecture Notes, 5, pp. 68-77.

Rao R. C. and T. K. Nayak (1985), Cross Entropy, Dissimilarity Measures and Characterization of Quadratic Entropy, *IEEE Transactions on Information Theory,* IT-31, 5, pp. 589-593.

Rosenzweig M. L. (1995) *Species Diversity in Space and Time,* Cambridge University Press, Cambridge.

Sham P. (1998), *Statistics in Human Genetics*, Arnold, London.

Shannon C. E. (1948), A Mathematical Theory of Communication, *Bell. Sys. Technology Journal* 27, pp 379-423; 623-656.

Shannon C. E: and Weaver W. (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.

Shorrocks A. F. (1988) Aggregation Issues in Inequality Measurement, in W. Eichorn (ed.) *Measurement in Economics: Theory and Applications in Economic Indices,* Physica Verlag, Heidelberg.

Shorrocks A. F. and J. Foster (1987), Transfer Sensitive Inequality Measures, *Review of Economic Studies*, 54(1), pp. 485-497.207-231

Simpson E. H. (1949), Measurement of Diversity, *Nature* 163, 688.

Smith B. and Wilson J. B. (1996), A Consumer’s Guide to Evenness Measures, *Oikos* 76, pp. 70-82.

Stirling G. and Wilsey B. (2001), Empirical Relationships between Species Richness, Evenness and Proportional Diversity, *The American Naturalist* 158 n. 3, pp. 286-299.

Svensson A. (2002), *Diversity Indices for Infectious Strains*, Mathematical Statistics, Stockholm University, Research Report 2002:5.

Subramanian S. (2004), *Indicators of Inequality and Poverty*, UNU-WIDER Research Paper 2004/25.

Theil H. (1967), *Economics and Information Theory*, North Holland, Amsterdam.

van Roovij A. C. M. (1978), *Non Archimedean Functional Analysis*, Marcel Dekker, New York.

Warwick R. M. and Clarke K. R. (1995), New Biodiversity Measures Reveal a Decrease in Taxonomic Distinctness with Increasgin Stress, *Mar. Ecol. Prog. Ser.* 129, 301-305.

Warwick R. M. and Clarke K. R. (1998), Taxonomic Distinctness and Environmental Assessment, *Journal of Applied Ecology* 35, pp. 532-543.

Warwick R. M. and Clarke K. R. (2001), Practical Measures of Marine Biodiversity Based on Relatedness of Species, *Oceanogr. Mar. Biol. Ann. Rev.* 39, pp. 207-231.

Webb C. O. (2000), Exploring the Phylogenetic Structure of Ecological Communities: An Example from Rain Forest Trees, *Am Nat.* 156, 145-155.

Weitzman M. L. (1992), On Diversity, *Quarterly Journal of Economics*, 107, pp. 363-405..

Weitzman M. L. (1993), What to Preserve? An Application of Diversity Theory to Crane Conservation, *Quarterly Journal of Economics*, 108, pp. 157-182.

Weitzman M. L. (1998), The Noah Ark Problem, *Econometrica*, 66, pp. 1279-1298.

Whittaker, R. H. (1965), Dominance and Diversity in Land Plant Communities, *Science* 147, 250-260.

Whittaker, R.H. (1972), Evolution and Measurement of Species Diversity, *Taxon* 21 pp 213-251.

Wolfson, M. (1994), Conceptual issues in Normative Measurement: When Inequality Diverge, AEA Papers and Proceedings.

Annex: The data

Indices are calculated using the data from United States Current Population Survey for *language* spoken at home (29 language groups). Individual data are grouped by SMSA, Standard Metropolitan Statistical Area (160 SMSA). Data refer to the year 1990.

Distances between Indo-European languages are from Kruskal, Black and Dyen (1992) and are measured by the separation time between a pair of speech varieties. For the purposes of this simple exercise, distances between non Indo-European languages are set arbitrarily to three times the maximum distance between Indo-European languages.

Tables A.1 and A.2 list language groups and SMSAs.

Table A.1 The 29 language groups

English/native
Scandinavian
Dutch
French
Celtic
German
Polish
Czech
Slovac, and other Balto-Slavic
African languages
Russian, Ukrainian, Ruthenian
Other Indoeuropean
Hungarian
Rumanian
Yiddish, Jewish
Greek
Italian
Spanish
Portuguese
Chinese, Tibetan
Arabic, Syriac, Aramaic
Albanian
Persian
Hindi
Hebrew, Israeli
East-Southeast Asian, Indonesian, Malaya
Filipino, Miconesian, Polynesian
American Indian
Other, not listed, not reported

Table A.2: The 160 US SMSA

40	Abilene
80	Akron
160	Albany-Schenectady-Troy
200	Albuquerque
240	Allentown-Bethlehem-Easton
280	Altoona
320	Amarillo
460	Appleton-Oshkosh-Neenah
520	Atlanta
560	Atlantic-Cape May
600	Augusta-Aiken
640	Austin-San Marcos
680	Bakersfield
720	Baltimore
760	Baton Rouge
840	Beaumont-Port ArthuR
880	Billings
920	Biloxi-Gulfport-Pascagoula
960	Binghamton
1000	Birmingham
1040	Bloomington-Normal
1080	Boise City
1240	Brownsville-Harlingen-San Benito
1280	Buffalo-Niagara Falls
1320	Canton-Massillon
1360	Cedar Rapids
1400	Champaign-Urbana
1440	Charleston-North Charleston
1520	Charlotte-Gastonia-Rock Hill
1560	Chattanooga
1600	Chicago
1640	Cincinnati
1680	Cleveland-Lorain-Elyria
1720	Colorado Springs
1740	Columbia
1760	Columbia
1840	Columbus
1880	Corpus Christi
1920	Dallas
1960	Davenport-Moline-Rock Island
2000	Dayton-Springfield
2040	Decatur
2080	Denver
2120	Des Moines
2160	Detroit

2240	Duluth-Superior
2320	El Paso
2360	Erie
2400	Eugene-Springfield
2560	Fayetteville
2640	Flint
2680	Fort Lauderdale
2760	Fort Wayne
2840	Fresno
2900	Gainesville
2960	Gary
3000	Grand Rapids-Muskegon-Holland
3080	Green Bay
3120	Greensboro--Winston-Salem--High Point
3160	Greenville-Spartanburg-Anderson
3200	Hamilton-Middletown
3240	Harrisburg-Lebanon-Carlisle
3320	Honolulu
3360	Houston
3400	Huntington-Ashland
3480	Indianapolis
3520	Jackson
3560	Jackson
3600	Jacksonville
3640	Jersey City
3680	Johnstown
3720	Kalamazoo-Battle Creek
3760	Kansas City
3800	Kenosha
3840	Knoxville
3880	Lafayette
3920	Lafayette
4000	Lancaster
4040	Lansing-East Lansing
4120	Las Vegas
4280	Lexington
4320	Lima
4360	Lincoln
4400	Little Rock-North Little Rock
4480	Los Angeles-Long Beach
4520	Louisville
4600	Lubbock
4680	Macon
4720	Madison
4800	Mansfield
4920	Memphis
5000	Miami
5080	Milwaukee-Waukesha
5120	Minneapolis-St. Paul
5170	Modesto
5200	Monroe

5240	Montgomery
5280	Muncie
5360	Nashville
5560	New Orleans
5600	New York
5640	Newark
5720	Norfolk-Virginia Beach-Newport News
5800	Odessa-Midland
5880	Oklahoma City
5920	Omaha
5960	Orlando
6080	Pensacola
6120	Peoria-Pekin
6160	Philadelphia
6200	Phoenix-Mesa
6280	Pittsburgh
6440	Portland-Vancouver
6640	Raleigh-Durham-Chapel Hill
6680	Reading
6720	Reno
6760	Richmond-Petersburg
6780	Riverside-San Bernardino
6800	Roanoke
6840	Rochester
6880	Rockford
6920	Sacramento
6960	Saginaw-Bay City-Midland
7040	St. Louis
7080	Salem
7120	Salinas
7160	Salt Lake City-Ogden
7240	San Antonio
7320	San Diego
7360	San Francisco
7400	San Jose
7480	Santa Barbara-Santa Maria-Lompoc
7500	Santa Rosa
7600	Seattle-Bellevue-Everet
7680	Shreveport-Bossier City
7800	South Bend
7840	Spokane
7920	Springfield
8120	Stockton-Lodi
8160	Syracuse
8200	Tacoma
8280	Tampa-St. Petersburg-Clearwater
8320	Terre Haute
8400	Toledo
8480	Trenton

8520	Tucson
8560	Tulsa
8600	Tuscaloosa
8640	Tyler
8680	Utica-Rome
8720	Vallejo-Fairfield-Napa
8800	Waco
8840	Washington
8920	Waterloo-Cedar Falls
8960	West Palm Beach-Boca Raton
9040	Wichita
9160	Wilmington-Newark
9200	Wilmington
9280	York
9320	Youngstown-Warren

CHAPTER 4

DIVERSITY AND PRODUCTIVITY: EVIDENCE FROM EUROPEAN CITIES

1. Introduction

Growing international flows in goods, factors and knowledge are fostering the global interactions among a rising and increasingly diversified number of people. At the European level, this phenomenon is reinforced by the twin processes of deeper integration and enlargement. As a consequence, 'diversity' is more and more at the core of public debates and a central issue for policy-making in the EU.

The debate is 'double faced'. On the one hand, the official rhetoric looks at diversity as a main asset for development and human welfare. At the global level, the 2001 *Universal Declaration on Cultural Diversity* of the United Nations Educational Scientific and Cultural Organisation (UNESCO) states that "cultural diversity is as necessary for humankind as biodiversity is for nature" (Art. 1). Similarly, at the EU level diversity is seen as the core concept of European identity (and *United in Diversity* is the motto that was proposed in the European Constitution). On the other hand, the general public perceives issues such as migration and enlargement as very problematic. The relevance of the 'Polish plumber' in the French debate on the European Constitution and the calls for restrictions to migration in several European countries are two of the main examples.

From an economic point of view, the key question is whether a culturally homogenous society is more efficient than a culturally diversified one. The answer is not obvious and equally 'double faced'. On the one hand, cultural diversity generates potential costs as it may entail racism and prejudices resulting in open clashes and riots (Abadie and Gardeazabal 2003), as well as conflicts of preferences, leading to a suboptimal provisions of public goods (Alesina, Baqir and Easterly 1999; Alesina, Baqir and Hoxby 2004). On the other hand, cultural diversity creates potential benefits by increasing the variety of goods, services and skills available for consumption, production and innovation (Lazear 1999; O'Reilly Williams and Barsade 1998; Ottaviano and Peri 2006; Berliant and Fujita 2004).

This paper is co-authored with Elena Bellini, Fondazione Eni Enrico Mattei (FEEM) and Gianmarco I. P. Ottaviano, University of Bologna, FEEM and CEPR. We thank Giovanni Peri, Andrea Ichino and all participants to the doctoral seminar of 20 December 2006 in Bologna for extremely useful suggestions. An early version of the paper was presented at the final ENGIME Conference on 'Multicultural Cities: Diversity, Growth and Sustainable Development', Rome, November 2004. We thank all the participants for their comments.

Recent evidence on US data show that richer diversity is indeed associated with higher wages and productivity of natives with causation from the former to the latter (Ottaviano and Peri 2005 a,b; Ottaviano and Peri 2006). Using a new regional database for Europe, we takes this research agenda forward and look for the first time at the relationship between diversity and productivity across European regions. We find results that are broadly consistent with those on US cities as also in EU regions richer diversity is associated with higher productivity. In particular, we provide evidence that causation again runs from the former to the latter.

The rest of the paper is organised as follows. Section 2 summarises the economic literature on diversity and places our contribution into context. Section 3 describes the data and presents some stylized facts. Section 4 introduces the theoretical model and Section 5 discusses the empirical results. Section 6 concludes.

2. The literature on diversity

The link between cultural diversity and economic performance has attracted considerable attention over the last decade. Using cross-country regressions, an early paper by Easterly and Levine (1997) shows that richer diversity is associated with slower economic growth.[1] Despite strong criticism (see for example Arcand *et al* 2000), the Easterly and Levine results have been confirmed by a number of studies. In particular, Alesina and La Ferrara (2005) find that going from perfect homogeneity to complete heterogeneity (i.e., the index of fractionalisation going from 0 – there is just one group – to 1 – each individual forms a different group) would reduce a country yearly growth performance by 2 per cent. Angrist and Kugler (2002) find a small but significant negative impact of migration on employment levels in the EU. La Porta et al (1999) and Alesina et al (2003) argue that higher levels of diversity might result in suboptimal decisions on public good provisions, consequently damaging the growth performance in the long-run. They show that diversity is negatively correlated with measures of infrastructure quality, illiteracy and school attainment, and positively correlated with infant mortality. Similarly, Alesina, Glaeser and Sacerdote (2001) find that higher diversity is associated with lower levels of social spending and social transfers by the government. The interpretation is that ‘redistributive policies’ are less valued in ethnically fragmented societies.

However, the conclusion that diversity has a negative effect on the economy need to be further qualified. Collier (2001) argues that diversity has negative effects on productivity and growth only in non-democratic regimes. Alesina and La Ferrara (2005) find that diversity has a more negative effect at lower levels of income (implying that poorer countries suffer more from ethnic fragmentation). Easterly (2001) constructs an index of institutional quality aggregating data from Knack and Keefer (1995) on contract repudiation, expropriation, rule of law and bureaucratic quality. He finds that the negative effect of ethnic diversity is significantly mitigated by ‘good’ institutions.

[1] Easterly and Levine (1997) use a fractionalisation index of diversity calculated from the *Midas Atlas* database.

Moreover, a number of studies relating diversity to urban agglomeration suggest that diversity can have also positive economic consequences. Jacobs (1961) sees diversity as the key factor of success of a city: the variety of commercial activities, cultural occasions, aspects, inhabitants, visitors as well as the variety of tastes, abilities, needs and even obsessions are the engine of urban development (Jacobs, 1961, p 137). Sassen (1994) studies ‘global cities’ - such as London, Paris, New York and Tokyo – and their strategic role in the development of activities that are central to world economic growth and innovation, such as finance and specialised services. A key characteristic of ‘global cities’ is the cultural diversity of their population. Bairoch (1985) sees cities and their diversity as the engine of economic growth. More recently, Florida (2002) argues that diversity contributes to attract knowledge workers thereby increasing the creative capital of cities and the long-term prospect of knowledge-based growth (Gertler, Florida, Gates and Vinodrai 2002).

Cross-country comparisons may not therefore be the correct tool to identify the possible positive effect of diversity. Finer spatial units, such as cities, where differences more easily interact, seem more appropriate laboratories. The focus on cities also allows one to partial out differences in institutional quality and stage of development.

Glaeser, Scheinkman and Shleifer (1995) examine the relationship between a variety of urban characteristics in 1960 and urban growth (income and population) between 1960 and 1990 across US cities. They find that racial composition and segregation are basically uncorrelated with urban growth. However, segregation seems to positively influence growth in cities with large non-white communities. Alesina and La Ferrara (2005) use the basic specification of Glaeser, Scheinkman and Shleifer (1995) to estimate population growth equations across US counties over 1970-2000. Consistently with their result at the country level discussed above, they find that diversity has a negative effect on population growth in initially poor counties and a less negative (or positive) effect for initially richer counties.

Following Roback (1982), Ottaviano and Peri (2006) develop a model of a multicultural system of open cities that allows them to use the observed variations of wages and rents of US-born workers to identify the impact of cultural diversity on productivity. They find that on average, US-born citizens are more productive in a culturally diversified environment (the result is robust to the use of IV implying a causal relationship from diversity to productivity). This main result is qualified in two specific respects. Firstly, local diversity has a negative effect on the provision of public goods (consistently with findings at the national level). Second, the positive effects are stronger when only second and third generation immigrants are considered, which suggests that the positive effects are reaped only when some degree of interaction between communities takes place.

These results somehow contrast with earlier findings in the economic literature showing a negative impact of immigrants on the wages of natives and a positive impact on returns on capital (Borjas 1995 and 2003). However, Ottaviano and Peri (2005b) notice that those results rely on the key assumptions of perfect substitution between natives and foreigners and fixed capital assets. Allowing for imperfect substitutability between natives and foreigners as well as endogenous capital accumulation, Ottaviano and Peri (2005b) find that the effects of immigration on the average wages of natives turn positive and rather large. Moreover, they find that the effect is particularly strong for the most educated (college graduates) and negative for the least educated (high-school drop-outs). The latter result is consistent with previous results showing a negative impact on the relative wages of less educated workers (Borjas 1994, 1999, 2003; Borjas, Freeman and Katz 1997; and to a minor extent Butcher and Card 1991; Card 1990 and 2001; Friedberg 2001; Lewis 2003).

The economic literature discussed so far is based either on cross-country analyses or focuses primarily on the US. This is not only because diversity is one of the hallmarks of US society, but also for the pragmatic reason that the richness and the quality of data readily available in the US make micro-analyses feasible. In this paper we use a newly constructed database to (partially) overcome the latter constraint in the case of Europe. Contrary to the US, in Europe cultural differences are historically inherited and are largely enshrined in national states (with established regional minorities either recognised or challenged by the national states). The migration flows over the last two centuries (from southern to northern Europe and from the colonies to colonial powers) have not dramatically altered this situation and simply led to the establishment of relatively stable ethnic communities in some European states. This situation is changing now as an increasing flow of people is crossing the EU national borders from inside and outside of the EU thereby creating a fluid landscape that resembles more closely the US situation. Indeed, this is at the basis of a current vivid public debate. For these reasons, we believe that our European focus represents an important complement to the existing studies from both an academic and a policy points of view.

3. The dataset

The dataset[2] includes demographic, economic and geographical data for over 500 European regions from 11 countries of the EU15 (Austria, Belgium, Denmark, France, Ireland, Italy, the Netherlands, Portugal, Spain, Sweden and the United Kingdom). Data are collected at NUTS 3 level (equivalent to *county* in the UK, *province* in Italy or *département* in France) and refer to two different points in time: 1991 (1990 for Finland and the Netherlands) and 2001 (2000 for Finland and the Netherlands; 1999 for France). The choice of reference years is constrained by the availability of Census data in each country (more on this below).

Economic data include GDP, employment (3-sector level), unemployment, active population and hotel and restaurant prices (more on this below). GDP, employment, unemployment, and active population are from Eurostat's Cronos REGIO database. When data are not available at NUTS 3 level, they are interpolated by using NUTS 2 data (kindly provided by Cambridge Econometrics). Geographical data include the areas (in square Km^2) of the region (from the Eurostat's REGIO database) and a travel time

[2] The dataset has been developed at Fondazione Eni Enrico Mattei with support from the European Commission, 6th RTD Framework Programme, Contract n° SSP1-CT-2003-502491 (PICTURE).

matrix (kindly provided by the European Commission DG Regio). Geographical data are used to calculate the density of population and the market potential of each region. The market potential of a region is calculated as the weighted average of the GDP of that region and the GDP's of the surrounding regions, with weights inversely related to the travel time (by car) between the regions.

Hotel and restaurant prices are used to proxy for local price indexes that are unavailable at NUTS 3 level. They have been chosen because typically they are highly correlated with the prices of non-tradables, in particular of land, which have been used by Ottaviano and Peri (2005a, 2006) to disentangle the productivity and the amenity effects of diversity. Hotel and restaurant prices are derived from the Michelin Guides of each country for the reference years. By exploiting the rating system of Michelin we have constructed price indexes that refer to restaurants and hotels of comparable quality across countries and cities. In particular, the hotel (restaurant) price for each region is calculated by averaging across the prices of all *two-houses* hotels (*two forchettes* restaurants) reported in the guide for that region. Hotel prices are for a two-bed room with no breakfast included. Restaurant prices exclude fixed-price menus.

Demographic data are constructed from the National Statistical Institutes of each country (mostly from national Census Surveys or Registry data) and cover population by gender, age (0-14; 15-39; 40-64; 65 or more), marital status (unmarried, married, divorced, widow) and level of education (basic or not educated, secondary school, degree or higher education - harmonized using the ISCED classification of the OECD) and citizenship (country of birth for the UK and Ireland) grouped by main area of provenience to achieve a common classification (autochthonous, other UE countries, other European countries, Africa, America, Asia, Oceania, unknown).

4. Measuring diversity

Measuring the diversity of a population requires two steps.[3] First, it is necessary to find one or more criteria to distinguish 'cultural groups' within the population. In ethnology the 'right list' of groups (Fearon 2003) would be based on a process of 'self-categorisation' where people recognize the distinction of groups and anticipate that significant actions are or could be conditioned on belonging or not to a group. A direct approach to the identification would involve carrying out worldwide surveys. Because of the costs involved, no such experiment has been carried out and indirect approaches have been used in literature. Indirect approaches require the choice of one or more 'identity markers' as a basis for the identification of the groups. Extra & Yağmur (2004) compare the theoretical strengths and weaknesses of four possible 'identity markers' (nationality, country of birth, language spoken at home and self-categorisation). Table 1 summarises their results.

[3] Whittaker (1972) distinguishes α–diversity (the diversity of a given population, or inventory diversity), and β–diversity (the variation of diversity across different populations, or differentiation diversity). Here, we will only use α–diversity measures, as we only refer to diversity within regions.

Table 1: Criteria for the definition and identification of population groups in a multicultural society (P/F/M = person/father/mother) (source: Extra & Yağmur 2004:31)

Criterion	Advantages	Disadvantages
Citizenship (CIT)	• objective • relatively easy to establish	• (intergenerational) erosion through naturalisation or double CIT • CIT not always indicative of ethnicity/identity • some (e.g., ex-colonial) groups have CIT of immigration country
Country-of-birth (CoB)	• objective • relatively easy to establish	• intergenerational erosion through births in immigration country • CoB not always indicative of ethnicity/identity • invariable/deterministic: does not take account of dynamics in society (in contrast of all other criteria)
Self-categorisation (SC)	• touches the heart of the matter • emancipatory: SC takes account of person's own conception of ethnicity/identity	• subjective by definition: also determined by language/ethnicity of interviewer and by spirit of times • multiple SC possible • historically charged, especially by World War II experiences
Home language (HL)	• HL is the most significant criterion of ethnicity in communication processes • HL data are prerequisite for government policy in areas such as public information or education	• complex criterion: who speaks what language to whom and when? • language is not always core value of ethnicity/identity • useless in one-person households

At national level, the best known and most widely used effort to distinguish 'cultural groups' within countries was carried out by a team of Soviet ethnographers in the early 1960s and published as *Atlas Narodov Mira.* The Soviet team mainly used *language* to define groups, but sometimes included groups that seem to be distinguished by some notion of *race* rather than language, and quite often used *national origin* (Fearon 2003). In the attempt of clearing from potential sources of arbitrariness (why should one use language alone in one case, language and race in a second one and language and national origin in a third one?) Alesina *et al* (2003) develop separate measures based on linguistic and religious groups (as well as ethnic groups, as a combination of the two) in a sample of about 190 countries.

At regional and urban level, data are much more scattered. For European regions, the only identity marker available is 'citizenship' ('country of birth' for the UK and Ireland), which is subject to intergenerational erosion. For the US, Ottaviano and Peri (2005a, 2006) compare measures of urban diversity based on country-of-birth, language-spoken-at-home, citizenship and race and find that such measures are highly correlated across cities (this is true to a lesser extent also for religion). The bias introduced by the citizenship-based measure of diversity may therefore not be too large.

In Europe, however, the problem of intergenerational erosion is reinforced by the fact that Member States have different citizenship laws and therefore different naturalisation rates. We will discuss in Section 7 the implications for the econometric analysis and how we deal with them.

The second step towards diversity measurement involves the construction of a synthetic index. A plethora of indexes have been proposed in various strands of literature (from biology to economics) that can serve this objective (a full review is proposed in the fourth paper of this dissertation). Here we will adopt two of the most used indexes in the relevant economic literature. The first is simply the share of foreigners in the whole resident population. The second is the *fractionalisation* index. Given a population of L_c individuals divided in *i=1…M* cultural groups, the fractionalisation index can be calculated as:.

$$(1)\ d_c = 1 - \sum_{i=1}^{i=M} \left(\frac{L_{ci}}{L_c} \right)^2$$

where L_{ci} is the number of individuals that in city *c* belong to group *i*. The index is widely used in biology (known as the Simpson index of diversity) and corresponds to the complement to one of the Herfindal index of concentration across groups. It measures the probability that two individuals randomly extracted belong to different groups. The index varies between 0 and 1 and increases with both the number of groups and the evenness of the distribution of individuals across groups.

5. Diversity in European regions

We can now use the database presented in Section 3 to discuss the main features of the European landscape of diversity and how this has changed over the period 1991 to 2001.

Figure 1 shows the percentage of foreigners in European regions in 1991.[4] At that time, diversity characterized only regions in the core of Europe: France around Paris and Lyon, Belgium, the Netherlands and the south of the UK. Regions of Spain, Italy, Austria and Nordic countries were fairly homogenous. In Italy and Spain the percentage of residents with foreign citizenship was below 2% everywhere. The situation has rapidly changed over the 1990s. In 2001 (see Figure 2) most of Austrian regions have reached a percentage of foreigners higher than 8% and the percentage of foreigners in most regions of Italy and Spain is between 4 and 8%. Overall, the share of foreigners increased from 4.8% in 1991 to 6.1% in 2001 (an increase of nearly 30% in absolute terms).

[4] Here and in what follows, we will refer to 'foreigner' as 'foreign-born' in the UK and Ireland, and 'with foreign citizenship' elsewhere. For the sake of illustration, we present the data using NUTS 2 regions. As explained in Section 3, data are collected at NUTS 3 level.

The data also allow for some analysis in terms of migrants’ provenience. On average, the largest group of foreign population is represented by migrants from other EU15 countries (representing around 1.9% of population in 1991), but this group has not significantly increased over the decade. Migrants from Africa represent the second largest group (1.5% of population in 2001) followed by Asian and other European (both groups amounted to around 1% of population in 2001). Contrary to migrants from the EU, the number of migrants from those three groups has been growing very fast with an increase of over a third during the decade.

Figure 3 and Figure 4 show the percentage of foreigners respectively from inside and outside the EU15. Figure 3 shows a geographical pattern that is very similar to the one shown in Figure 1 with the highest shares in the core regions of Europe and very little outside. Hence, internal migration flows tend to reproduce old core-periphery patterns. Figure 4 is more similar to Figure 2 with relatively high shares also in the regions of Austria, Italy and Spain. Contrary to migrants from the EU, recent migration flows from outside seem to affect to a greater extent the regions of more recent immigration, particularly those that are close to the Mediterranean and the Eastern border in Southern Europe (the lack of data for Germany and Finland makes it difficult to analyze the influence of migration from the northern part of the Eastern border)

Figure 1. Shares of foreigners in European regions, 1991

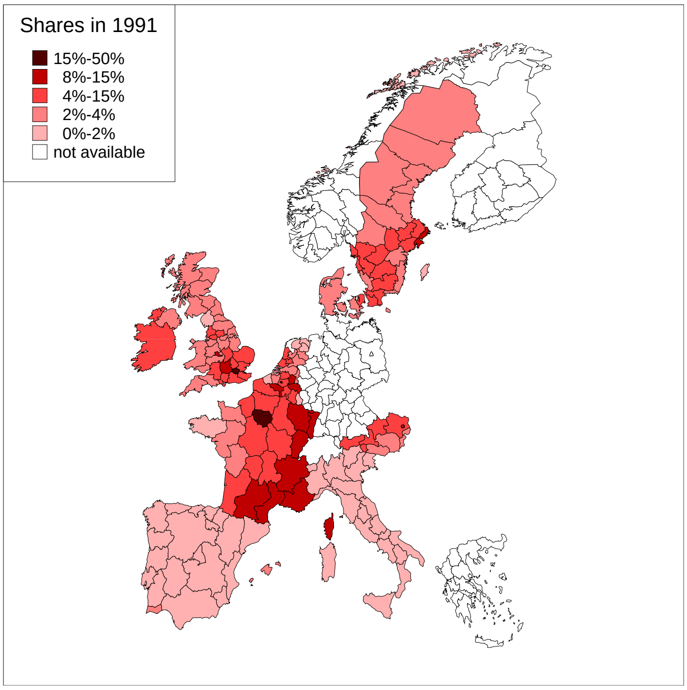


Figure 2. Shares of foreigners in European regions, 2001

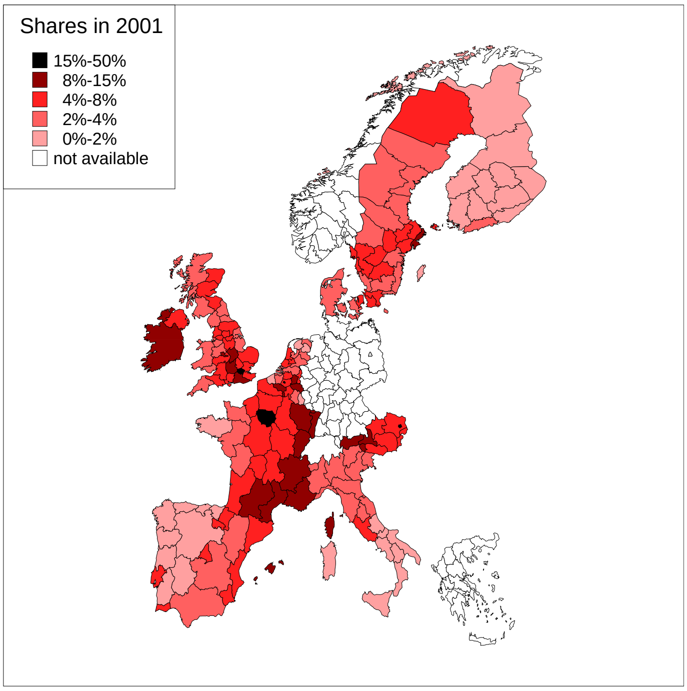


Figure 3: Share of foreigners from within the EU15, 2001

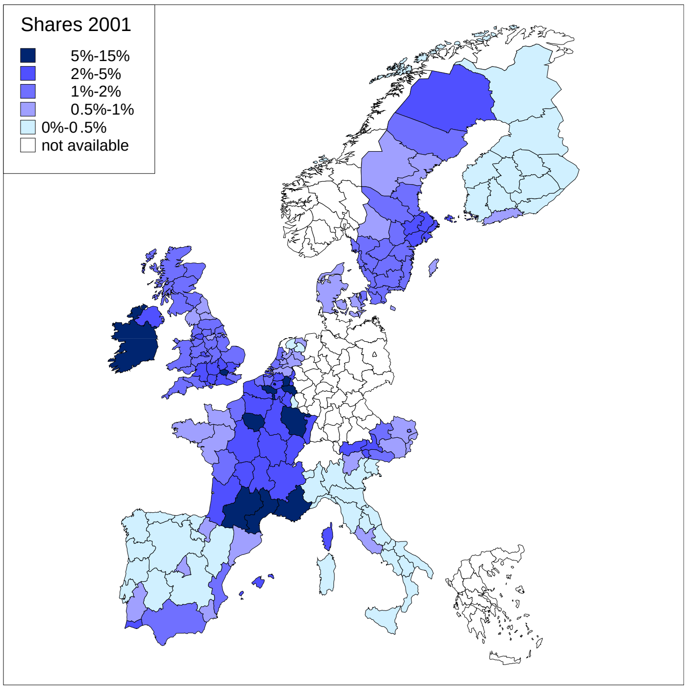


Figure 4: Share of foreigners from outside the EU15, 2001

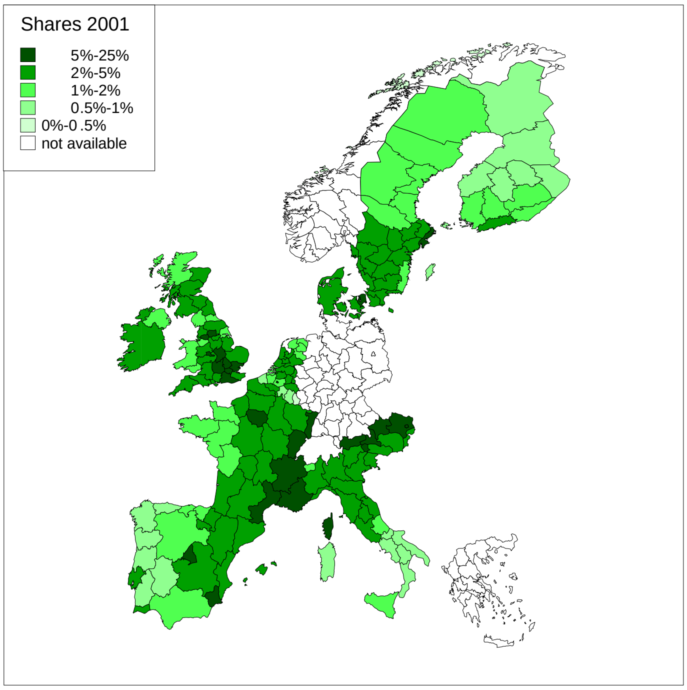


Table 2 shows the most and the least diverse EU regions in 1991 and 2001 ranked according to the Simpson index of diversity (fractionalization) discussed in Section 4. The share of foreigners in total population is also reported. Urban regions are at the top of the ranking both in 1991 and 2001. French and UK regions reach the highest score in both cases, joined in 2001 by Bruxelles and surroundings. Interesting features emerge comparing the distribution of diversity in and around Paris and London. While in Paris diversity is more concentrated in the *banlieu* (Seine-Saint-Denis being more diverse than Paris), the opposite is true for London where diversity is more concentrated in the *core* (Inner London being more diverse than Outer London). Vienna appears in the top ten only in 2001, following the immigrant inflows from Eastern Europe after 1989. Rural regions are at the bottom of the ranking both in 1991 and 2001. In 1991, the group of regions at the bottom end shows nearly no diversity and includes only rural Italian and Spanish regions. The picture is different in 2001. Some degree of diversity also characterises the most homogenous regions and some of the Italian and Spanish regions have been replaced by rural regions in France and Belgium in terms of lack of diversity.

Table 2: Most and least diverse European regions, 1991 and 2001

Most diverse	1991			2001	
	Simpson	Share of foreigners		Simpson	Share of foreigners
Inner London (UK)	0.334	27.8%	Inner London (UK)	0.409	33.6%
Seine-Saint-Denis (FR)	0.261	24.1%	Seine-Saint-Denis (FR)	0.315	27.9%
Outer London (UK)	0.230	18.0%	Outer London (UK)	0.304	22.9%
Paris (FR)	0.228	21.7%	Paris (FR)	0.243	21.9%
Bruxelles (BE)	0.223	28.6%	Hauts-de-Seine (FR)	0.208	18.1%
Hauts-de-Seine (FR)	0.190	17.4%	Val-de-Marne (FR)	0.203	19.4%
Val-de-Marne (FR)	0.166	17.6%	Val-d'Oise (FR)	0.191	17.8%
Val-d'Oise (FR)	0.162	15.7%	Bruxelles (BE)	0.182	27.1%
Rhône (FR)	0.136	13.8%	Wien (AT)	0.181	16.4%
Leicestershire (UK)	0.136	9.1%	Berkshire (UK)	0.175	13.1%
Least diverse	**1991**			**2001**	
	Simpson	Share of foreigners		Simpson	Share of foreigners
Taranto (IT)	0.001	0.1%	Benevento (IT)	0.005	0.4%
Terni (IT)	0.001	0.1%	Vandée (FR)	0.005	0.4%
Albacete (ES)	0.001	0.1%	Taranto (IT)	0.004	0.6%
Badajoz (ES)	0.001	0.1%	Oristano (IT)	0.004	0.3%
Jaen (ES)	0.001	0.1%	Ypres (BE)	0.004	0.3%
Ciudad Real (ES)	0.001	0.1%	Enna (IT)	0.004	0.4%
Zamora (ES)	0.001	0.1%	Tâmega (PT)	0.004	0.5%
Isernia (IT)	0.001	0.1%	Brindisi (IT)	0.004	0.4%
Campobasso (IT)	0.001	0.1%	Eeklo (BE)	0.004	0.2%
Chieti (IT)	0.000	0.0%	Dixmude (BE)	0.002	0.6%

Source: Authors' calculation based on national Censuses data for population by *country of birth* for Ireland and the UK and *citizenship* for the other countries (see Section 3).

Notes: Data are for 1991 and 2001 except for the Netherlands (1990 and 2000) and France (1991 and 1999).

Finnish regions are excluded (1991 data are not available).

It is common sense to believe that US cities are very diverse ‘melting pots’, while European cities are generally considered more homogenous both within (low α-diversity, in the classification of Whittaker 1972) and between themselves (low β-diversity, following the same classification). Although a direct comparison is not possible, useful indications concerning the validity of this statement can be drawn by comparing Table 1 with the data presented by Ottaviano and Peri (2005a, Table 2) for US cities.[5] A more complex picture seems to appear. The most diverse US cities are Los Angeles and New York with a share of foreign born in total population of respectively 37% and 31% in 1990 (corresponding to diversity indexes in the range of 0.5 to 0.6). The percentage is not dramatically different from the percentage of foreign population in the most diverse European regions in 2001 (Inner London reached 33% in 2001). Differences are apparently larger at the bottom. The least diverse European regions have a share of foreigners in total population that is smaller than 0.5% whereas their counterparts in the US (such as Cincinnati and Pittsburgh) reach a share of 2.3%. Nevertheless, European regions have levels of α-diversity that are comparable with those of US cities and span a range of diversity (β-diversity) that is not significantly smaller than the range of diversity spanned by US cities.

6. Theoretical model

To structure the empirical analysis, we use the theoretical framework developed by Ottaviano and Peri (2006), who model an open system of cities in which ‘diversity’ affects both the productivity of firms and the satisfaction of consumers through localized external effects. Both the model and the identification procedure of the impact of diversity on city dwellers build on Roback (1982).

The framework considers a system of a large number N of regions, indexed by $c=1,\dots,N$. There are two factors of production, labour (perfectly mobile) and land (fixed). The total amount of land is exogenously allocated to regions and H_c denotes the amount land in region c. To ensure that the rental income of workers, if any, is independent of residence and therefore does not affect migration choices, land is assumed to be owned by locally resident landlords.

Total supply of labour is L and each worker inelastically supplies one unit of work. L_c denotes the number of workers living and working in region c. In order to rule out commuting, intraregional commuting costs are zero and interregional commuting costs are prohibitive, so we can focus on the interregional allocation of workers.

Workers are identical in terms of attributes that are relevant for market interactions, but they differ in terms of non-market attributes, which exogenously classifies them into M different groups (‘cultural identities’) indexed by $i=1,\dots,M$. The diversity of regional population is measured by d_c (calculated as in (1)). Diversity affects both production and

[5] Ottaviano and Peri (2006) use ‘country of birth’ as identity marker. Data are therefore directly comparable with our data for the UK and Ireland but not for the rest of the regions (for which we use ‘citizenship’). The higher values of the Simpson index for US cities also depends on the larger number of ‘cultural’ groups used by Ottaviano and Peri (as the Simpson index varies with both the number and relative size of groups).

consumption as an externality that can be either positive or negative. The objective is to identify the dominant externality (consumption or production) and its sign.

As a result of those assumptions, the interregional allocation of land is exogenously given while the interregional allocation of labour will be endogenously determined in equilibrium. Similarly, the degree of cultural diversity for the system is exogenously given, while intraregional diversity is endogenously determined by the entry decisions of firms and the migration decision of workers.

Preferences are defined over the consumption of land *H* and a homogenous good *Y* that is freely traded among regions. The utility of a typical worker of group *i* in region *c* is given by:

(2) $U_{ic} = A_U(d_c)H_{ic}^{1-\mu}Y_{ic}^{\mu}$, where $0<\mu<1$.

In **(2)**, H_{ic} and Y_{ic} are land and good consumption, while $A_U(d_c)$ captures the consumption externality associated with local diversity d_c. If the first derivative $A_U'(d_c)$ is positive, then diversity has a positive effect on workers utility (i.e., an amenity effect). If the first derivative $A_U'(d_c)$ is negative, then diversity has a negative affect on workers utility (i.e., a disamenity effect). Workers move to the region that offers them the highest utility. Given **(2)** and utility maximisation, the indirect utility function is given by:

$$V_{ic} = (1-\mu)^{1-\mu}\mu^{\mu}A_U(d_c)\frac{E_{ic}}{r_c^{1-\mu}p_c^{\mu}}$$

(3)

where E_{ic} is workers expenditures. Given our assumption about land ownership, E_{ic} will consist of wage only: $E_{ic}=w_c$.

As to production, good Y is supplied by perfectly competitive firms using both land and labour as input. The typical firm in a region *c* produces according to the following technology:

(4) $Y_{jc} = A_Y(d_c)H_{jc}^{1-\alpha}L_{jc}^{\alpha}$, where $0<\alpha<1$.

In (4), H_{ic} and L_{ic} are land and labour inputs, while $A_Y(d_c)$ captures the productivity externality associated with local diversity d_c. If the first derivative $A_Y'(d_c)$ is positive, then diversity has a positive effect on firms' productivity (i.e., a positive productivity effect). If the first derivative $A_Y'(d_c)$ is negative, then diversity has a negative affect on firms productivity (i.e., a negative productivity effect). Given (4) and profit maximisation, it is possible to solve for the marginal cost pricing condition:

(5) $$p_c = \frac{r_c^{1-\alpha}w_c^{\alpha}}{(1-\alpha)^{1-\alpha}\alpha^{\alpha}A_Y(d_c)}$$

As *Y* is freely traded, its price will be the same everywhere and we can choose it as numeraire, i.e. $p_c=1$. [6]

[6] With reference to the empirical analysis, it is important to note that by imposing $p_c=1$, we are *de facto* requiring that the law-of-one-price holds for tradable goods and that land rents are a reasonable approximation of non-tradable goods prices (in the model, as land is the only fixed factor, differences in local prices are entirely driven by land rents).

We can now determine the spatial equilibrium. This is identified by a set of prices for labour and land (w_c, r_c) with $c=1,\ldots,N$ such that in all regions workers and landlords maximise their utilities given their budget constraints, firms maximise profits given their technological constraints, factor and product markets clear. At the equilibrium, no worker has an incentive to move. For an interior equilibrium to exist (i.e., $L_c>0$ for any $c=1,\ldots,N$), workers must be indifferent between locations, i.e. their indirect utility is equalised across regions:

$$(6) \qquad V_{ic} = V_{ik} \forall k, c = 0...N$$

In what follows, we will refer to (6) as the '*free migration condition*'. Similarly, in equilibrium no firm has an incentive to exit or enter the market. This is ensured by the marginal cost pricing condition that, given the choice of numeraire, can be re-written as:

$$(7) \qquad r_c^{1-\alpha} w_c^{\alpha} = (1-\alpha)^{1-\alpha} \alpha^{\alpha} A_Y(d_c)$$

In what follows, we will refer to (7) as the '*free entry condition*'.[7] In order to use the model for the empirical investigation, it is necessary to solve for the rent and wage levels at the equilibrium allocation. This requires solving together the free migration condition (6) and the free entry condition (7) while taking account of (3). The result is the '*wage equation'*:

$$(8) \qquad \ln w_c = \frac{(1-\mu)\eta_Y - (1-\alpha)\eta_U}{1-\alpha\mu} + \frac{1}{1-\alpha\mu} \ln\left(\frac{[A_Y(d_c)]^{1-\mu}}{[A_U(d_c)]^{1-\alpha}}\right)$$

and the '*rent equation'*:

$$(9) \qquad \ln r_c = \frac{\eta_Y + \alpha\eta_U}{1-\alpha} + \frac{1}{1-\alpha\mu} \ln(A_Y(d_c)[A_Y(d_c)]^{\alpha})$$

where $\eta_Y \equiv (1-\alpha)^{1-\alpha}\alpha^{\alpha}$, $\eta_Y \equiv (1-\mu)^{1-\mu}\mu^{\mu}/\nu$ and ν is the value of the indirect utility function at the equilibrium (the same across all regions).

Equations (8) and (9) give the relation between diversity and factors prices and represent the theoretical foundation of our empirical investigation. In the wake of Roback (1982), they must be estimated together as the estimation of only one of them would run into an identification problem. To see this, consider estimating equation (9). A positive correlation between diversity and wages would be consistent with both a disamenity effect (A_U' $(d_c)<0$) and a positive productivity effect (A_Y'$(d_c)>0$). Analogously, a positive correlation between diversity and rents would be consistent with both an amenity effect (A_U' $(d_c)>0$) and a negative productivity effect (A_Y' $(d_c)<0$). Only the joint estimation of (8) and (9) will allow the identification of the dominant effect. Specifically:

[7] The free migration and the free entry conditions can then be solved to determine the spatial allocation of workers. A complete discussion is given in Ottaviano and Peri (2006).

(10)

$$\frac{\partial r_c}{\partial d_c} > 0 \quad \text{and} \quad \frac{\partial w_c}{\partial d_c} > 0 \quad \text{iff dominant positive productivity effect} \quad A_Y'(d_c)>0$$

$$\frac{\partial r_c}{\partial d_c} > 0 \quad \text{and} \quad \frac{\partial w_c}{\partial d_c} < 0 \quad \text{iff dominant consumption amenity} \quad A_U'(d_c)>0$$

$$\frac{\partial r_c}{\partial d_c} < 0 \quad \text{and} \quad \frac{\partial w_c}{\partial d_c} < 0 \quad \text{iff dominant negative productivity effect} \quad A_Y'(d_c)>0$$

$$\frac{\partial r_c}{\partial d_c} > 0 \quad \text{and} \quad \frac{\partial w_c}{\partial d_c} > 0 \quad \text{iff dominant consumption disamenity} \quad A_Y'(d_c)>0$$

Figure 5 provides a graphical representation of the spatial equilibrium and the associated identification problem. Regional nominal wages (*w*) are measured along the vertical axis and regional land rents (*r*) along the horizontal one. Downward sloping lines depict the '*free entry condition'*, i.e. the combination of rents and wages that make firms indifferent across locations. Their downward slope reflects the fact that firms can earn the same profit in different regions provided that higher wages correspond to lower rents and vice-versa. Upward sloping lines depict the '*free migration condition'*, i.e. the combination of rents and wages that make workers indifferent across locations. Their upward slope reflects the fact that workers can achieve the same utility ('real wage') in different regions provided that higher rents correspond to higher wages and vice-versa. The intersection between the two curves gives the wage and rent equilibrium.

Local diversity d_c acts as a shift parameter on the two curves. A positive shock to diversity shifts the *free entry condition* upward (downward) if diversity has a positive (negative) productivity effect. It shifts the *free migration condition* downward (upward) if diversity has a consumption amenity (disamenity) effect. We can therefore identify the dominant effect of diversity by looking at the impacts of shocks on the equilibrium factor prices.

Suppose A represents the initial equilibrium at factor prices (*r*,*w*). Suppose also that there is a shock to diversity and we observe higher wages (*w'*>*w*) after the shock.

Figure 5 shows that in principle this could be associate either with a upward shift of the free entry condition (point B) indicating a positive productivity effect; or with an upward shift of the free migration condition (point C) indicating a negative effect on workers quality of life (or consumption disamenity). To distinguish whether higher wages signal higher productivity or worse quality of life, additional information is needed. In Figure 5 that is provided by rents: whereas higher productivity is associated with higher wages and higher land rents (point B), worse quality of life is associated with higher wages but lower land rents (point C). By symmetry the foregoing arguments can be applied to downward shifts of the firm and worker indifference lines. A reduction in productivity shifts the firm line downward, which reduces both wages and land rents (point D). An improvement in the quality of life shifts the worker line downward, thus decreasing wages and increasing land rents (point E).

Figure 5: The spatial equilibrium

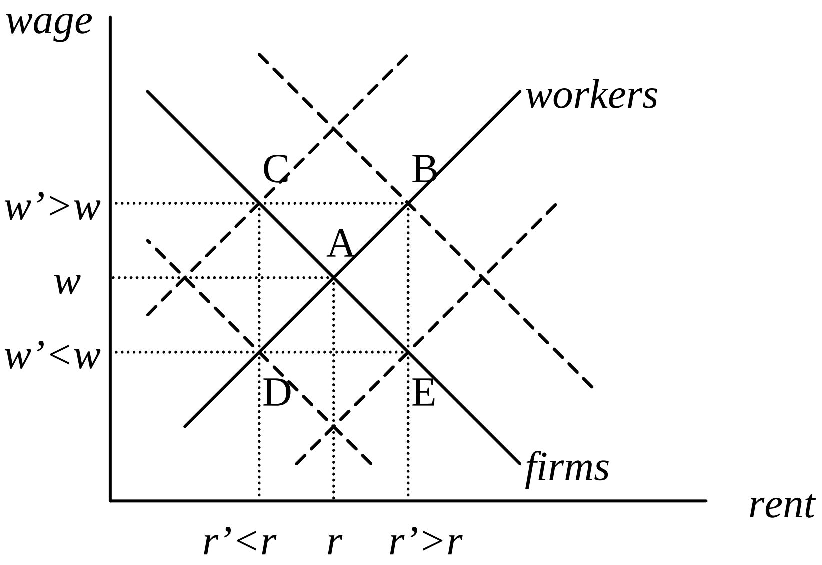


Table 3 summarizes the overall identification procedure that will be used in Section 7 to assess whether and to what extent diversity affects productivity across EU regions.

Table 3: Identification strategy

		Rent variation	
		Positive	Negative
Wage variation	Positive	Positive productivity effect	Disamenity effect
	Negative	Amenity effect	Negative productivity effect

Before moving to the empirical results, it is however important to discuss the consequences of Europe's low labour mobility for the empirical implementation. Consider the extreme case of no labour mobility. In such case, the '*free migration condition*' becomes vertical and wage differentials measure productivity differentials. If this were the case for Europe, we could simply estimate the wage equation and identify wage responses to diversity shocks as productivity effects. Since labour mobility in Europe is low but it is not absent (particularly among migrants), we will nevertheless estimate the rent regressions in order to rule out any possibility that higher wages reflect the disamenity effects of diversity.

7. Empirical results

We now present the results of the empirical analysis, which is carried out in four steps. First, following the identification strategy set out in Section 6, we estimate the wage equations. As wage data for European regions and cities are scattered and not available at NUT3 level, we use GDP per capita as a proxy.[8] Under the model assumption of free firm mobility the two measures are equivalent, as profits are equalised across regions and income differentials are entirely driven by wage differentials.

Second, we estimate the rent equations. EU-wide comparable data for land rents at city level are not available (and data for a close proxy such as house prices are only available for a restricted number of major cities). However, in our theoretical model, rents *de facto* capture non-tradeable good prices (see footnote 6), which we proxy by the average prices (in logs) of *two-forchettes* restaurants as detailed in Section 3.[9]

Third, with respect to Roback (1982) we face an additional problem. While she estimates the effect of exogenous factors (such as climate) on productivity and the quality of life, our independent variable (diversity) is endogenous and therefore we cannot be sure that any correlation found reveals a causal link from diversity to local incomes and prices. We use instrumental variables (IV) to tackle such an endogeneity problem.

Fourth and last, we carry out some robustness checks. In particular, we adjust the measures of diversity to account for differences in citizenship laws.

First step: Income regressions

The basic equation is the following:

$$(11) \quad \ln y_{ct} = \beta_o + \beta_1\, div_{ct} + \beta_2 edu_{ct} + \beta_3 agri_{ct} + \beta_4 \ln(dens)_{ct} + \beta_5 \ln(mpot)_{ct} + D_t + D_c + e_{ct}$$

where *c* indexes the city and *t* the time. As discussed, the dependent variable *(ln y_{ct})* is GDP per capita (in logs). The key regressor is the city's diversity (div_{ct}). We use two measures of diversity: the Simpson index (see Section 4) and the simple share of foreigners in total population.[10] We include standard control variables (see Temple 1999 for a review of the recent literature on income and growth regressions) such as the share of agriculture in total employment ($agri_{ct}$) to control for differences in industrial

[8] REGIO also contains data for 'Compensation of employees' but scattered and only available at NUTS 2 level.

[9] Where data availability makes computation possible, the correlation between restaurant prices and house prices is typically large and positive. For example, in a sample of 12 major Italian cities such correlation was roughly 70 per cent in 2001.

[10] As from Section 3 population is classified by citizenship in all countries apart from the UK and Ireland for which we use the 'country of birth' .

structure and the share of inhabitants with at least secondary education (*edu_{ct}*) to control for differences in human capital endowments. The density of population *$ln(dens)_{ct}$* is introduced (in logs) to control for those 'non-pecuniary' externalities that derive from sheer proximity of economic actors.[11] Market potential (*$ln(mpot)_{ct}$*) controls for the 'pecuniary' externalities that derive from the agglomeration of economic activities, as highlighted by the new economic geography literature (see Redding and Venables 2004, Ottaviano and Pinelli 2006). In all regressions, we introduce region and time fixed effects. Region fixed effects (D_c=1 for the NUTS 3; 0 otherwise) control for those characteristics, such as institutions and geographical location, that do not change over time. When region fixed effects are introduced, only the time variation of data is left to be explained, and the resulting regression in levels is equivalent to a differences-on-differences regression. The fixed effects then capture time-invariant differences in local diversity deriving from the identity marker used (country of birth or citizenship) and differences in national citizenship laws. Lastly, the time fixed effect D_t controls for Europe-wide trends.

Table 4 shows the results of the basic first-step regressions. Robust standard errors (heteroskedasticity often characterises cross-regional analyses), are reported in brackets. The first three columns report the results of regressions without market potential. The control variables are correctly signed. The share of agriculture has a negative and significant coefficient, consistently with most findings in literature (see, for example, Bivand and Brundstad 2003). The human capital variable has a positive and significant coefficient, consistent with the growth literature (Temple 2001). The density of population has a negative coefficient suggesting that negative congestion effects prevail (similar results are found by Ottaviano and Pinelli 2006 across Finnish communes).[12] Turning to our key variables, both measures of diversity have positive and significant coefficients. When market potential is added to the set of regressors (Table 4, last two columns), the coefficients on the diversity measures remain significant and become even larger. This suggests that the positive relationship between diversity and incomes is not simply due to the fact that migrants move towards regions where economic activities are agglomerated. Market potential has a positive and significant coefficient, consistently with theoretical predictions and recent empirical findings (Head and Mayer 2004; Redding and Venables 2004; Ottaviano and Pinelli 2006).

Under the realistic assumption of no labour mobility, the results would point out to a positive effect of diversity on firms' productivity. Nevertheless, in order to rule out the possibility that the higher wages simply reflect aversion to diversity (rather than a genuine effect on productivity), we study below the relationship between diversity and local prices.

[11] Local external effects can be positive, due to easier non-market interactions leading to technological externalities (see Ciccone 2002; Ciccone and Hall 1996) or negative, due to higher congestion and consequent waste of resources that make interactions difficult.

[12] Introducing fixed-effects at the NUTS 2 (rather than NUTS 3) level, we obtain a positive coefficient on density, which is consistent with previous findings that densely populated areas have an economic advantage over scarcely populated areas within the same region (see for example Ciccone and Hall 1996; Ciccone 2002).

Table 4: GDP per capita regressions - basic regressions

Dependent variable *Independent variables*	*GDP per capita*					
Share of agriculture	-2,358*** (0,482)	-2,38*** (0,483)	-2,162*** (0,448)	-2,344*** (0,455)	-2,336*** (0,437)	-2,091*** (0,406)
Density	-0,668*** (0,141)	-0,705*** (0,134)	-0,74*** (0,123)	-0,698*** (0,127)	-0,754*** (0,108)	-0,781*** (0,1)
Human capital	0,795*** (0,121)	0,588*** (0,124)	0,487*** (0,127)	0,503*** (0,178)	0,092 (0,190)	0,068 (0,179)
Market Potential				0,845*** (0,315)	1,193*** (0,311)	1,086*** (0,268)
Simpson Diversity Index		2,528*** (0,685)			3,423*** (0,729)	
Share of foreigners			4,524*** (0,98)			5,074*** (0,939)
N.	268	268	268	268	268	268
R^2	43%	49%	53%	50%	59%	62%

Notes:
*** = significant at 1%; **= significant at 5%; *=significant at 10%. Robust standard errors in parentheses.

Second step: Price regressions

The basic regression is the following:

$$(12) \quad \ln p_{ct} = \gamma_0 + \gamma_1 \, div_{ct} + \gamma_2 edu_{ct} + \gamma_3 agri_{ct} + \gamma_4 \ln(dens)_{ct} + \gamma_5 \ln(mpot)_{ct} + D_t + D_c + e_{ct}$$

The dependent variable (*ln* p_{ct}) is average restaurant price in the region. As before, the key regressor is the regional diversity (div_{ct}.). Standard control variables are included together with region and time fixed effects.

Table 5 shows the results of the prices regressions following the same structure of Table 4. All regression have very low explanatory power (R^2 is between 0.10 and 0.13 compared to 0.50-0.60 of the income regressions) supporting the hypothesis of low labour mobility and thus a vertical *free migration* condition. Nevertheless, control variables are correctly signed, as in the first step. Market potential and human capital variables are positively signed (although human capital is no longer significant when market potential is included), consistently with the theoretical prediction of NEG models and the recent literature on human capital. The coefficient on the share of agriculture is significant and negative in all regressions, confirming that a higher specialisation in agriculture is negatively associated with productivity. Concerning our key variable, both measures of diversity have a positive (but not significant) coefficient in all regressions. Following our identification strategy, this rules out the possibility of diversity being a consumption disamenity and points out a positive correlation between diversity and productivity.

Table 5: Restaurant prices regressions – basic regression

Dependent variable *Independent variables*	*Restaurant prices*					
Share of agriculture	-1,388** (0,72)	-1,386** (0,722)	-1,379* (0,731)	-1,351** (0,692)	-1,361** (0,691)	-1,313* (0,699)
Density	0,031 (0,93)	0,032 (0,094)	0,028 (0,095)	0,013 (0,098)	0,004 (0,102)	0,000 (0,102)
Human capital	0,614*** (0,168)	0,620*** (0,184)	0,599*** (0,190)	0,299 (0,204)	0,215 (0,229)	0,230 (0,230)
Market Potential				0,792*** (0,335)	0,863*** (0,347)	0,832*** (0,343)
Simpson Diversity Index		-0,076 (0,783)			0,632 (0,813)	
Share of foreigners			0,196 (1,006)			0,728 (1,041)
N. R^2	223 10%	233 10%	223 10%	223 13%	223 13%	223 13%

Notes:
*** = significant at 1%; **= significant at 5%; *=significant at 10%.Robust standard errors in parentheses.

Third step: Instrumental variables

Short of a randomized experiment, we cannot be sure that the positive correlation found between diversity and wages (and hence productivity) reveals a causal link from diversity to productivity. For this reason we use instrumental variables (IV) to tackle the endogeneity problem and analyse the direction of causality. A set of good instruments should be correlated with the change in diversity of regions from 1991 to 2001, and not otherwise correlated with the residuals in the structural equations (11) and (12). Previous literature has proposed two approaches to construct such instruments. The first uses the 'shift-share methodology' firstly applied by Card (2001) and, more recently, by Saiz (2003) and Ottaviano and Peri (2006). The key idea is that migrants tend to settle close to where migrants of the same provenience already reside. Therefore, the predicted end-of-period composition of a region's population can be computed on the basis of its beginning-of-period composition by attributing to each group in the region its average growth rate for the EU as a whole. However, Section 5 shows that recent migration waves into Europe are settling in regions and cities that were previously rather homogenous, which makes this approach not applicable to our case.

The second approach looks more promising. The key idea here is that migrants enter through 'gateways' and tend to settle in their proximity due to the presence of costs of travelling and spreading information as well as the existence of ethnic networks (Ottaviano and Peri 2006). In this case, the distance from such 'gateways' is presumably highly correlated with diversity and exogenous to income and local prices. Section 5 shows that over 1991-2001 the main migration shocks came from Eastern Europe (following the fall of the Iron Curtain in 1989 and the Balkans wars of the 1990s) and from Africa. Accordingly, we construct two instrumental variables: the distance from the Eastern border (*lneast*) and the distance from the Mediterranean coast (*lnmed*).[13] The distance from the Eastern border is calculated as the region's minimum distance

[13] A similar approach is used by Angrist and Kugler (2002) that exploit the Balkan war as an exogenous shock by using the distance of national capitals from Pristina and Sarajevo to instrument countries' shares of migrants in total population.

from the Austrian and Italian borders with Hungary, Czech Republic and Slovenia as well as from the main ports on the Adriatic (Trieste, Brindisi and Taranto).[14] The distance from the Mediterranean is calculated as the region's minimum distance from one of the main ports on the Mediterranean coast (Genoa, Cagliari, Palermo, Leghorn, Naples, Marseille, Algeciras, Barcelona and Valencia). We also construct a third instrumental variable (*lnmain*) using the region's minimum distance from the largest ports by freight (Rotterdam, Antwerp, Le Havre) or passengers (Dover, Calais) not considered in the construction of previous variables.[15] However, the latter variable is more subject to endogeneity as income may enter the determination of such a Europe-wide hierarchy of ports. For this reason, we use only the first two (geography-based) instrumental variables in all regressions (the only exception being the four regressions in the fourth step using the share of foreigners as measure of diversity).

Results are shown in Table 6. The first four columns report the results for the *income* regressions (with and without market potential). The F and Hansen-J tests indicate that the choice of instruments is correct. The F-test of exclusion of instruments from the first stage regression is always above 10 (the value normally taken as reference value) showing that the instruments are strongly correlated with the endogenous variable. The Hansen-J is generally low and the null hypothesis of exogeneity of the instruments cannot be rejected in three out of four cases. The control variables are correctly signed. Both diversity measures bear significant and positive coefficients. They are larger than those reported in Table 4, providing evidence of an attenuation bias in the OLS estimates.

The last four columns report the results of the *price* regressions. The coefficients on the control variables are similar to the corresponding OLS coefficients in Table 5. The coefficients on diversity measures are positive and not significant when market potential is not included, as it was the case for OLS. When market potential is instead included, the coefficients on the diversity terms are significant and much larger than in the OLS regressions. The latter effect may be explained by the low F-test, which implies that the estimates are less precise.

Overall, IV results do not contradict the OLS estimates and point at a positive causal relationship from diversity to productivity.

[14] In order to select the 'main' ports we have proceeded as follows. Firstly, we have taken the first 3 seaports in each country by yearly freight tonnage as published by the European Seaport Organisation (ESPO) in its Factual Report on the European Port Sector 2004-2005 (data cover the period 2000-2003). Secondly, we have added to the resulting set the ports those appearing in the top fifteen by passengers (all passenger) traffic according to the REGIO database (average 1991-2001). REGIO contains data only for NUTS 2. We have identified the relevant NUTS 3 within the NUTS 2 using the list of seaports provided on the ESPO website. For example, when identifying the main ports on the Mediterranean, Cagliari, Palermo and Naples have been included in the set because Sardinia, Sicily and Campania are among the top fifteen regions among European NUTS 2 ranked by sea passenger traffic.

[15] Data on freight are taken from European Seaport Organisation (ESPO). Data on passengers are taken from from REGIO (at NUTS 2 level).

Table 6: Instrumental variable regressions

Dependent variable *Independent variables*	*GDP per head*				*Restaurant prices*			
Share of agriculture	-2.439*** (0.404)	-2.135*** (0.386)	-1.596*** (0.218)	-1.397*** (0.180)	-1.419** (0.608)	-1.294** (0.612)	-1.555* (0.627)	-0.562 (0.710)
Density	-0.709*** (0.109)	-0.750*** -0.099	-0.850*** (0.082)	-0.835*** (0.077)	0.015 (0.084)	-0.003 (0.098)	-0.165 (0.198)	-0.250 (0.229)
Human capital	0.563*** (0.148)	0.444*** -0.171			0.503* (0.269)	0.468** (0.318)	-1.386 (1.015)	-1.124 (0 .819)
Economic potential			0.922*** (0.123)	0.907*** (0.129)			2.233** (0.954)	1.609** (0.674)
Simpson Diversity Index	2.843*** (1.113)		6.818*** (1.230)		1.289 (2.115)		12.88* (6.970)	
Share of foreigners		5.157*** (2.048)		5.134** (1.856)		2.014 (3.327)		15.22* (7.617)
N.	268	268	467	467	223	223	220	220
R^2	49%	53%	33%	45%	8%	8%	na	na
Hansen-J	2.56	3.34	1.61	21.65	6.19	6.14	2.83	2.20
F-test on instruments	23.76	10.39	23	16.23	13.58	8.51	4.56	5.28
Instrumental variables	lneast lnmed	lneast lnmed	lneast lnmed	lneast lnmed	lneast lnmed	lneast lnmed	lneast lnmed	lneast lnmed

Notes:
*** = significant at 1%; **= significant at 5%; *=significant at 10%. Robust standard errors in parentheses.

Fourth step: Correcting for differences in citizenship laws

As discussed in Section 4, our measures of diversity are based on 'citizenship' and therefore subject to the problem of intergenerational erosion through naturalisation or double citizenship. As there is not a common approach to citizenship at European level, the issue of naturalization is regulated in different ways by Member States thereby introducing a potential important bias in our measures of diversity. The problem is further complicated by the fact that data for the UK and Ireland refer to 'country-of-birth' rather than citizenship. In Step 1 to 3 we have dealt only partially with this issue by introducing region-specific dummies. These should control for time-invariant differences in diversity resulting both from the existence of different citizenship laws and from the use of a different identity marker (country-of-birth instead of citizenship).

In a further step to eliminate the bias, we use the OECD data on annual naturalisation rates (i.e., shares of foreign residents acquiring citizenship every year) in each member country. We regress the two measures of diversity (in first differences) on the average naturalisation rate for the period of reference. We then use the residuals as alternative explanatory variables in difference-on-difference regressions. We drop the UK and Ireland from the regression in order to eliminate the potential distortion deriving from the different identity marker used.

Table 7 shows the results of OLS regressions (income and price). The results are very similar to those obtained in Table 4 and Table 5 (before correcting the diversity measures): the coefficients on diversity measures are positive and significant in the income regression (and similar in size to those in Table 4) and not significant in the price regressions.

Table 8 shows the results of IV regressions. The first four columns report the results of *income* regressions. When market potential is excluded (first and second column), the F-test on the exclusion of instruments is very low, which explains the extreme variability of coefficients on diversity measures. On the contrary, when market potential is included (third and fourth column), the F-test is significant and the results obtained in Table 6 with uncorrected variables are fully confirmed.[16] The coefficients on both diversity measures are positive and significant. Their size is remarkably similar to that reported in Table 6. The last four columns report the results of the *price* regressions. Column 5 and 6 reports the result of the regression with the basic set of control variables. There are no significant variables, which points out to some collinearity among the variables (adding market potential does not improve the results of the regressions). For this reason, we report in columns 7 and 8 the results of a more parsimonious specification. The coefficients on diversity measures are positive and significant. Given collinearity, such results may simply be the outcome of the exclusion of control variables. In any case, the results of price regressions rule out a negative effect of diversity on the quality of life, thereby confirming the positive effect of diversity on productivity.

Table 7: Corrected variables: OLS regressions

Dependent variable *Independent variables*	*GDP per head*				*Restaurant prices*			
Share of agriculture	-2.230*** (0.358)	-2.113*** (0.348)	-1.343*** (0.188)	-1.299*** (0.181)	-1.851*** (0.705)	-1.746** (0.711)	-1.690** (0.695)	-1.623** (0.702)
Density	-0.348* (0.208)	-0.382* (0.185)	-0.566*** (0.147)	-0.512*** (0.147)	0.348 (0.415)	0.279 (0.434)	0.431 (0.425)	0.406 (0.443)
Human capital	0.322 (0.211)	0.180 (0.221)			-0.070 (0.310)	-0.139 (0.335)	0.241 (0.321)	0.211 (0.349)
Economic potential			0.0635 (0.128)	-0.040 (0.126)			2.816** (1.333)	2.626* (1.325)
Simpson Diversity Index	2.063*** (0.712)		2.521*** (0.335)		1.261 (1.017)		1.168 (0.988)	
Share of foreigners		3.470*** (0.829)		2.168*** (0.386)		2.073 (1.392)		1.573 (1.349)
N. R^2	207 42%	207 45%	384 33%	384 31%	164 11%	164 11%	161 13%	161 13%

Notes:
*** = significant at 1%; **= significant at 5%; *=significant at 10%. Robust standard errors in parentheses.

[16] In order to strengthen the instruments, the regression using the simple share of foreigners is estimated employing all three instrumental variables (and not just two, as in all previous regressions).

Table 8: Corrected variables: IV regressions

Dependent variable	*GDP per head*				*Restaurant prices*			
Independent variables								
Share of agriculture	-2.093***	0.396	-1.448***	-1.330***	-1.475	1.040	-1.663***	-1.637***
	(0.428)	(2.770)	(0.227)	(0.189)	(1.652)	(4.931)	(0.413)	(0.430)
Density	-1.214**	-4.571	-0.847***	-0.661***	-5.434	-5.504		
	(0.604)	(4.198)	(0.175)	(0.148)	(10.34)	(9.460)		
Human capital	-1.068	-7.984			-7.990	-7.434		
	(0.880)	(7.965)			(14.25)	(12.18)		
Economic potential			0.537**	0.197				
			(0.272)	(0.204)				
Simpson Diversity Index	10.21**		5.945***		44.066		8.653***	
	(5.134)		(1.383)		(76.36)		(1.870)	
Share of foreigners		60.74		4.263***		52.12		13.76***
		(54.07)		(1.301)		(81.90)		(2.378)
N.	207	207	384	384	164	164	308	308
R^2	na	na	16%	24%	na	na	na	na
Hansen-J	22.73	0.43	0.39	28.56	0.18	0.19	15.22	5.71
F-test on instruments	3.39	0.54	11.35	13.46	0.37	0.78	19.92	14.37
Instrumental variables	lneast lnmed	lneast lnmed lnmain	lneast lnmed	lneast lnmed lnmain	lneast lnmed	lneast lnmed lnmain	lneast lnmed	lneast lnmed lnmain

Notes:
*** = significant at 1%; **= significant at 5%; *=significant at 10%. Robust standard errors in parentheses.

8. Conclusions

In this paper we have studied the impact of cultural diversity on productivity across European city-regions. We have based our empirical analysis on Ottaviano and Peri (2006), who model a system of open cities in which cultural diversity affects both productivity and consumption through an externality (which can be positive or negative). Building on this model, we have developed an empirical strategy based on the estimation of price and income equations in order to identify the dominant channel of externality (consumption or production) and its sign. We have estimated price and income equations using a newly developed database including demographic, economic and geographical variables for more than 500 NUTS3 regions in 11 countries of the EU15. Data refer to two different points in time, 1991 and 2001. We have constructed two measures of diversity and in both cases we have used ‘citizenship’ as (the only available) identity marker (country-of-birth for Ireland and the UK).

We have found that diversity is positively correlated with income. Under the realistic assumption of no labour mobility, such positive correlation would indicate that richer diversity is associated with higher productivity. However, if labour were mobile, higher wages could simply reflect the wage premium that workers require if averse to diversity. As the latter would imply a negative correlation between diversity and local prices, we have estimated a price equation using average regional restaurant prices as proxies for local prices. We have found nil or positive correlation between the two variables. We have therefore concluded that richer diversity is associated with higher productivity.

Furthermore, using the distances from the Eastern border and from the Mediterranean coast as instruments for diversity, we have provided evidence of causation running from diversity to productivity. These results have been shown to be robust to different measures of diversity, to the exclusion of the UK and Ireland (for which country-of-birth is used as identity marker) and to the use of measures of diversity that correct for differences in naturalisation rates across Member States.

Our results are consistent with those in Ottaviano and Peri (2006), who find that urban diversity has a positive effect on natives' wage and productivity levels across US cities. They could be consistent with those in Alesina and La Ferrara (2005), who find that urban diversity is positively associated with population growth across rich US counties. However, this would require that the EU regions included in our dataset are sufficiently rich. Future work should further investigate this issue

Our results are not consistent with previous cross-country studies, which tend to find a negative association between diversity and economic outcomes. There are two main explanations. First, the focus on Europe (and the US) clears the results from the effects of different institutional and development scenarios that may affect cross-country studies. In fact, Collier (2001) argues that diversity has a negative effect on productivity and growth only in non-democratic regimes; and similarly, Easterly (2001) finds that the negative effect of ethnic diversity is significantly mitigated by 'good' institutions'. Second, regions and *a fortiori* cities, rather than countries, are likely to be the appropriate laboratory to observe diversity at work, as differences interact more easily and positive externalities can be tapped.

Additional robustness tests are still needed. These may include the use of alternative measures of diversity (taking into account EU vs. non-EU resident foreigners), or regressions by regional sub-groups using parameters that can be geographic (e.g., coastal vs. landlocked regions) or economic (e.g., rich vs. poor regions). Although IV fully supports OLS conclusions and our instruments stand to statistical testing, further effort is needed to investigate the endogeneity issue and the direction of causality. A possible direction is the use of 'control' regions having similar economic structure and recent history but differently exposed to diversity shocks.

References

Abadie A. and J. Gardeazabal (2003), The Economic Costs of Conflict: A Case Study for the Basque Country, *The American Economic Review*, 93 (1), pp. 113-132.

Alesina A., R. Baqir and W. Easterly (1999), Public Goods and Ethnic Division, *Quarterly Journal of Economics*, 111 (4), pp. 1243-1284.

Alesina A., R. Baqir and C. Hoxby (2004), Political Jurisdictions in Heterogenous Communities, *Journal of Political Economy,* 112 (2), pp. 384-396.

Alesina A., A. Devleschawuer, W. Easterly, S. Kurlat and R. Wacziarg (2003), Fractionalization, *Journal of Economic Growth,* 8, pp. 155-194.

Alesina A., E. Glaeser and B. Sacerdote (2001), Why Doesn't the US have a European Style Welfare State?, *Brooking Paper on Economic Activity,* Fall.

Alesina A., E. La Ferrara (2005), Ethnic Diversity and Economic Performance, *Journal of Economic Literature*, 43, pp. 762-800.

Angrist J. D., A. D. Kugler (2002), Protective or Counter-Productive? Labour Market Institutions and the Effect of Immigration on EU Natives, IZA Discussion Paper 433.

Arcand, J-L., P. Guillaumont and S. Guillaumont Jeanneney (2000), How to Make a Tragedy: on the Alleged Effect of Ethnicity on Growth, *Journal of International Development,* 12, pp. 925-938.

Bairoch P. (1985), *De Jéricho à Mexico: Villes et economie dans l'histoire*, Paris, Editions Gallimard.

Berliant M. and M. Fujita (2004), Knowledge Creation as a Square Dance on the Hilbert Cube, Discussion Paper, Kyoto University.

Bivand R. S. and R. J. Brunstad (2004), Regional growth in Western Europe: an Empirical Exploration of Interactions with Agriculture and Agricultural Policy, in *European Regional Growth*, B. Fingleton (ed.), Springer, New York, pp. 351-375.

Borjas G.J. (1994), The Economics of Immigration, *Journal of Economic Literature,* 32, pp. 1667-1717.

Borjas G.J. (1995), The Economic Benefits of Immigration, *Journal of Economic Perspectives,* 9, pp. 3-22.

Borjas G.J. (1999), *Heaven's Door,* Princeton, Princeton University Press.

Borjas, G. (2003) The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market, *Quarterly Journal of Economics*, CXVIII (4), pp. 1335-1374.

Borjas G. J., R. Freeman and L. Katz (1997), How Much do Immigration and Trade Affect Labor Market Outcomes?, *Brookings Papers on Economic Activity*, 1, pp. 1-90.

Butcher K. C. and D. Card (1991), Immigration and Wages: Evidence from the 1980s, *American Economic Review*, Papers and Proceedings, 81 (2), pp. 292-296.

Card D. (1990), The Impact of the Mariel Boatlift on the Miami Labor Market, *Industrial and Labor Relation Review*, XLIII, pp. 245-257.

Card D. (2001), Immigrant Inflows, Native Outflows, and the Local labor Market Impacts of Higher Immigration, *Journal of Labor Economics*, XIX, pp. 22-64.

Ciccone A. (2002), Agglomeration Effects in Europe, *European Economic Review*, 46, pp. 213-227.

Ciccone A. and R. Hall (1996), Productivity and the Density of Economic Activity, *American Economic Review*, n. 87, pp. 54-70.

Collier P. (2001), Implication of Ethnic Diversity, *Economic Policy*, 32, pp. 129-166.

Easterly W. (2001), Can Institutions Resolve Ethnic Conflict?, *Economic Development and Cultural Change*, Vol. 49, No. 4, pp 687-706.

Easterly W. and R. Levine (1997), Africa’s Growth Tragedy: Policies and Ethnic Division, *Quarterly Journal of Economics*, 111 (4), pp 1203-1250.

Extra G. and K. Yağmur (eds) (2004), *Urban multilingualism in Europe. Immigrant minority languages at home and school*, Clevedon, Multilingual Matters.

Fearon J. D. (2003), Ethnic and Culture Diversity by Country, *Journal of Economic Growth*, 8 (2), pp. 195-222.

Florida R. (2002), *The Rise of the Creative Class* (tr. it., L’ascesa della nuova classe creativa, Milano, ed. Arnoldo Mondadori Editore SpA).

Friedberg R. (2001), The Impact of Mass Migration on the Israeli Labor Market, *Quarterly Journal of Economics*, vol. 116(4), pp. 1373-1408.

Gertler M. S., R. Florida, G. Gates, T. Vinodrai (2002), *Competiting on Creativity: Placing Ontario's Cities in North American Context*, Institute for Competitiveness and Prosperity, Ontario Ministry of Enterprise.

Glaeser E.L., Scheinkman J.A., Shleifer A. (1995), Economic Growth in a Cross-Section of Cities, *Journal on Monetary Economics*, 36(1), pp. 117-144

Head K. and T. Mayer (2004), The Empirics of Agglomeration and Trade, in Henderson V and J-F Thisse (eds), *Handbook of Regional and Urban Economics*, vol. 4, Elsevier, Amsterdam.

Jacobs J. (1961), *The Death and Life of Great American Cities,* New York, Vintage.

Knack S. and P. Keefer (1995), Institutions and Economic Performance: Cross-country Tests using Alternative Institutional Measures, *Economics and Politics*, 7, pp. 207-227.

La Porta R., F. Lopez de Silanes, Shleifer A. and R. Vishny (1999), The Quality of Government, *Journal of Law, Economics and Organisation*, 15(1), pp. 222-279.

Lazear E. P. (1999), Globalisation and the Market for Team-Mates, *The Economic Journal*,109, C15-C40.

Lewis E. (2003), Local Open Economies within the US. How do Industries respond to Immigration?, Federal Reserve Bank of Philadelphia, Working Paper 04.

O'Reilly C., K. Williams and S. Barsade (1998), Group Democracy and Innovation: Does Diversity Help?, *in Research and Managing Groups and Teams*, D Gruenfeld et al (eds) JAI Press.

Ottaviano G.I.P. and Peri G. (2005a), Cities and Cultures, *Journal of Urban Economics*, 58, pp. 304-337.

Ottaviano G.I.P. and Peri G. (2005b), Rethinking the Gains from Immigration. Theory and Evidence from the US, CEPR Discussion Paper, n. 5226.

Ottaviano G.I.P. and Peri G. (2006), The Economic Value of Cultural Diversity: Evidence from US Cities, *Journal of Economic Geography*, 6, pp. 9-44.

Ottaviano G.I.P. and Pinelli D. (2006), Market Potential and Productivity: Evidence from Finnish Regions, *Regional Science and Urban Economics*, 36, pp. 636-657.

Redding S. and A. Venables (2004), Economic Geography and International Inequality, *Journal of International Economics* 62, pp 53-82.

Roback J. (1982), Wages, Rents and the Quality of Life, *Journal of Political Economy*, 90, pp. 1275-1278.

Saiz A. (2003) Immigration and Housing Rents in American Cities, Federal Reserve Bank of Philadelphia, Working Paper 03-12.

Sassen S. (1994), *Cities in a World Economy*, Pine Forge Press, Thousand Oaks, US.

Temple J. (1999), The New Growth Evidence, *Journal of Economic Literature,* 37, pp. 112-156.

Temple J. (2001), Generalization that aren't, Evidence on Education and Growth, *European Economic Review,* 45, pp. 905-918.

Whittaker, R.H. (1972), Evolution and Measurement of Species Diversity, *Taxon* 21 pp 213-251.