WP3/12 SEARCH WORKING PAPER

Skilled labour mobility: Tracing its spatial distribution

Ernest Miguélez and Rosina Moreno

May 2013

SEARCH
SHARING KNOWLEDGE ASSETS: INTERREGIONALLY COHESIVE NEIGHBORHOODS

EUROPEAN COMMISSION
European Research Area
SEVENTH FRAMEWORK PROGRAMME
Funded under Socio-economic Sciences & Humanities

The research leading to these results has received funding from the European Community's
Seventh Framework Programme (FP7/2010-2.2-1) under grant agreement n° 266834

Skilled labour mobility: Tracing its spatial distribution

Ernest Miguélez, Rosina Moreno

AQR-IREA. Department of Econometrics, Statistics and Spanish Economy. University of Barcelona, Av. Diagonal 690, 08034 Barcelona, Spain. E-mails: emiguelez@ub.edu; rmoreno@ub.edu

Abstract
This paper aims to provide new insights into the well-studied phenomenon of knowledge flows. We study one of the main mechanisms through which these flows occur, that is, the mobility of highly-skilled individuals. In contrast to earlier studies, we focus on the geographical mobility of inventors across European regions. Thus, patent data are used to trace the pattern of inventors' mobility across European regions, to track down focuses of attraction of talent throughout the continent, and to study their distribution across the space. To do so, we first gather information from PCT patent documents and match the names which seemed to belong to the same inventor using name matching algorithms; second, we create a new algorithm to decide whether each patent applied for under each name belongs to the same inventor, according to set of predetermined characteristics.

Key words: inventors' mobility, knowledge flows, name matching algorithms, exploratory data analysis

JEL: C8, J61, O31, O33, R0

1. Introduction

Knowledge flows occur through a variety of mechanisms, such as market transactions, collaborative research networks, monitoring of competitors, foreign direct investment, spin-offs, or pure externalities (Trippl and Maier 2007, Döring and Schnellenbach 2006). Among these mechanisms, the mobility of highly-skilled personnel – across firms, between academia and the business sector, across geographic locations – is a key source of knowledge transmission. This mobility is the main concern of this paper. In the age of modern technology and the knowledge-based economy, the transmission of information, or codified knowledge, is becoming increasingly important for the creation of innovations. However, "tacit" knowledge (Polanyi 1966), which cannot be transferred into documentation such as papers, patents, and so on, is playing an ever greater role in the creation of new knowledge and valuable goods. Face-to-face interactions between talented individuals are the mechanism through which this knowledge diffuses. Thus, the mobility of skilled individuals like inventors is intrinsically related to the aforementioned tacit knowledge diffusion.

We focus our analysis on mobility among these particularly highly-skilled workers, that is to say, inventors[1]. Their movement is important since they are carriers of knowledge – not only codified knowledge, but also tacit knowledge. The analysis of this phenomenon is intrinsically interesting from a policy viewpoint: information on the patterns of movement of these researchers and the effect on their productivity and on potential positive (or negative) social externalities resulting from their movement, may help policy makers to design suitable frameworks able to exploit this phenomenon for collective purposes.

As stressed in the literature, the accumulation of human capital and skilled individuals is critical for growth and regional development (Lucas, 1988, Florida, 2002). In such a setting, countries, regions, and cities compete each other to attract these people (OECD, 2008), and therefore those areas receiving talent are better positioned to take advantage of individuals' embodied knowledge (Maier et al., 2007) and their spatial human capital externalities (Lucas, 1988; Moretti, 2004). To the best of our knowledge, the identification and characterisation of these magnetic focuses is an important issue not completely addressed in the literature. Thus, our main objective is to perform a detailed exploratory spatial analysis for the focuses of attraction of talent throughout the European geography, looking for the *agglomeration centres for knowledge flows*.

Second, we would like to test whether these focuses are randomly distributed across the space or, on the contrary, they follow certain spatial pattern. We hypothesize (and test) that these focuses of attraction of talent are indeed spatially correlated because of at least three major reasons: (1), the attractiveness of a certain region –job opportunities, R&D facilities, amenities,…- spills over that region to nearby ones; (2), certain regions experience congestion which translates to crowding-out effects, favouring the location of, say, labs, universities or firms, outside the main region but in neighbouring regions in order to take advantage of the existence of agglomeration economies; and (3), the existence of country specific features, like industrial tradition, research prestige or wage premium, makes all the regions of a given country attractive for inventors from abroad. The existence of this spatial correlation would mean that not only region-specific

[1] Our analysis in this paper is exclusively focused on inventors who have applied for patents. We use, however, the term highly skilled workers or highly skilled individuals, among others, as synonyms in this study, although we acknowledge that these latter terms encompass broader concepts.

endowments are important, but also that location and geography are still valuable. That is why this spatial distribution should also be a matter of concern. Third, and finally, we use several descriptive statistics which point to the fact that geographical movements of inventors is a phenomenon bounded in the space, country specific, and very concentrated. However, we acknowledge that this hypothesis deserves further empirical analysis.

From a methodological viewpoint, our study contributes to the existing literature in several ways. Firstly, it will shed some light on the geographical mobility of inventors – as opposed to the organisational mobility examined by the greater part of the literature[2] (see section 2 below). This is an important issue, because, as is well known, highly-skilled workers are not only a source of innovation, growth and well-being for the firm for which they are working but also a major source of knowledge and human capital spillovers for the whole region. This means that their geographical movement is an important concern for regional development. We strongly believe that this shift-focus is significant, because the literature on the geography of innovation and knowledge flows, and spatial economics, or spatial econometrics, has not studied so far the exact mechanisms through which knowledge flows from one region to another one. Second, our study will be performed at a very detailed level of regional aggregation – as opposed to the literature on star-scientists' mobility (see section 2 below), which is carried out at country level. Although the number of studies of labour mobility among skilled workers and inventors have increased in recent years, there has been little analysis of their geographical mobility, and, to the best of our knowledge, this geographical mobility has not been examined at such a disaggregated level as the one we use in our empirical analysis – the NUTS 3 level. As is well known, knowledge flows are a localised phenomenon (Jaffe et al. 1993, Bottazzi and Peri 2003) and their analysis should therefore be carried out at a highly disaggregated level in terms of spatial units. Our analysis is therefore made at regional level. Again, this is important since low levels of regional disaggregation are appropriate levels of analysis for the study of innovation and knowledge diffusion phenomena (Anselin et al. 1997, Acs et al. 2002). Furthermore, unlike other recent studies in this field, our study will cover the whole of Europe.

Finally, as we explain in section 3, we propose a useful methodology for identifying the mobility patterns of inventors using information contained in their patent documents and computerised algorithms to be able to do this on a large scale (the whole of Europe). By looking at the names that appear in the patent documents relating to their inventions, our approach is divided in two stages: first name matching algorithms are used in order to group possible similar names, and then an algorithm is designed to establish computationally whether inventors with the same or similar names are actually the same person, on the basis of features reported in the patent document – self-citations, the applicant, the region from where the inventor makes the application, or its technological class. Although this is a secondary objective in the present study, few attempts are found in the literature to do so, and it should be considered therefore one important contribution of the paper.

The outline of the paper is as follows: section 2 reviews the literature on inventors' mobility; section 3 describes the methodology used to match each record to

[2] As pointed out by Laforgia and Lissoni (2009), in the absence of information on acquisition and merger activities of firms, it is difficult to know whether the mobile inventors identified (inter-organizational mobility) have actually changed employers, or if it is just a case of an absorption or merger between two applicants. These authors also note the existence of free lance inventors, as well as patented inventions with multi-applicant firms. This is why inter-firm mobility patterns should be treated with caution.

the correct inventor; section 4 presents an exploratory spatial analysis of the resulting final dataset and several additional statistics; and section 5 concludes.

2. Theoretical and empirical background

The rationale behind our investigation is that mobility of highly-skilled within the local labour market and across firms is important for regional development because they are carriers of knowledge (Trippl and Maier 2007). However, their geographical mobility across regions is also important for the inter-regional and international diffusion of knowledge, and its analysis will be, therefore, one major contribution of the paper. While moving, inventors and scientists in general take their knowledge to other places and share it with their new colleagues; they acquire new knowledge from these colleagues, they set up new links and social networks based on trust for future collaborations[3] and, in general, promote new combinations of knowledge (Op. Cit). It is also argued that knowledge diffusion through mobility is far from being unidirectional (Ackers 2005); it is actually multi-directional, leading to what Saxenian (2005) called "brain circulation". The literature also lays emphasis on the return phenomenon (temporary migration) and circular migration, which are also considered important processes (OECD 2008).

An important strand in the literature has analysed these phenomena using data from different datasets containing the CV of a given researcher. Zucker, Darby and colleagues (1998ab, 2002, 2006) have undertaken an extensive research program on the effects of star-scientist movements from academia to industry in the field of biotechnology. Their research shows that this movement promotes success and also that it is these star-scientists, more than their disembodied knowledge, that represent the main determinant of firm location in the sector, and subsequently in the formation and transformation of high-tech industries. Less is known about geographical mobility at the empirical level. However, this star-scientist literature has also been descriptively analysed in Maier et al. (2007), who used data from *ISIHighlyCited.com* to map spatial distribution and mobility patterns. Their analysis shows that the US is the nation that received by far the highest number of star scientists, whilst Western Europe and the UK are the regions that lose the majority of these scientists. In spite of the narrow definition of mobile scientists (those whose country of birth differs from their current country) and their level of aggregation, their results are revealing. These patterns are particularly accentuated in the fields of physics, computer sciences, and engineering.

Most studies on the mobility of highly-skilled workers focus on patent data to trace the pattern of inventors' movements. These data and their information about inventors and applicants have been used widely to trace inter-firm mobility, and normally use patent citations to examine knowledge flows[4]. However, the vast majority of these studies are restricted to relatively small samples due to the difficulties involved in obtaining large, reliable datasets on inventors' mobility. For instance, Almeida and Kogut (1999) analyse the inter-firm mobility of engineers in the US semiconductor

[3] The literature empirically examining the influence of mobility on the structure of interpersonal networks within an across firms is extensive (Singh, 2005; Fleming et al, 2007; Breschi and Lissoni, 2009). Note, however, that social research networks can boost mobility of inventors at the same time –so the other way around. In any case, the relationship between mobility and networks deserves further empirical research. We would like to thank the anonymous referees for pointing out the possible existence of such phenomenon.

[4] Earlier research has used patent citations as an indicator of knowledge flows. We consider that inventors' mobility itself can be identified as a knowledge transfer.

industry using patent data, tracing their mobility through their names and through interviews. Their study concludes that this phenomenon undoubtedly influences the local/regional transfer of knowledge. Next, Song et al. (2003) shed some light on the determinants of mobility across US engineers who moved from US to non-US firms, finally suggesting that the knowledge acquired by hiring engineers from other firms is useful for innovation. Using patent data and patent citations data for the semiconductor industry, these authors identified the mobility patterns of those inventors through coincidences in the names and surnames that appeared in patent documents and using manual checks. In the European case, Crespi et al. (2007) is one of the best-known attempts to empirically analyse the phenomenon of inventors' mobility. Using data from the PatVal-EU database[5], they investigated the mobility patterns of inventors who applied for one of 9,000 selected EPO[6] patents in the mid-nineties across six large European countries, focusing their attention on the movement from university to industry. Their findings suggest that hiring the inventor of a patent from academia gives the employer access to her tacit knowledge, and that the cumulative knowledge of the inventor and the value of her patents are significant factors in firms' decisions regarding recruitment. The PatVal database is also used in the studies by Hoisl (2007, 2009), which show that mobility has a positive and significant impact on inventors' productivity – for 3,049 German inventors – and Lenzi (2009), who focuses on job-to-job mobility for a group of Italian inventors and the determinants of their movements. Finally, Agrawal et al. (2006) test the hypothesis that when an inventor leaves, she does not break her links with her former colleagues. In their study, inventors' mobility is traced using the same exact name and surname and the technological class of the patents.

In recent years, several authors have suggested interesting methodologies for using patent data to trace the pattern of movements for large samples. Our work is closely related to their proposals. One of the pioneering studies was by Trajtenberg et al. (2006), who undertook a huge research program to design computerised algorithms able to identify inventors using their names and information contained in the patent document. Interesting follow-up studies are Trajtenberg and Shiff (2008), who examined the mobility patterns of a subset of Israeli inventors (both across assignees and within and outside Israel), and Shalem and Trajtenberg (2008) for a sample of Israeli software inventors. Kim et al. (2006) also use a similar methodology to the one suggested in Trajtenberg et al. (2006) to investigate the mobility patterns of US semiconductor and pharmaceutical industry inventors. Even more recently, Lissoni et al. (2006), Lissoni et al. (2008) and Lissoni (2008) have also developed a methodology to automatically trace inventors' movements with computerised algorithms and using complete information from the patent documents through data from the European Patents Office for Swedish, French and Italian academic inventors (the KEINS database). Breschi and Lissoni (2009) use data on US inventors applying to the EPO, who are matched when the name, surname and address coincide[7]. When the address is not exactly the same, a program is used to give different scores to every pair of the same name and surname with different addresses depending on the coincidence of

[5] The PatVal-EU database was a project created under the sponsorship of the European Commission, involving research groups from six European universities. Inventors from six major European countries who applied for European patents were surveyed. An outline of the project and first descriptive results can be found in Giuri et al. (2007).

[6] EPO stands for European Patent Office.

[7] Although Breschi and Lissoni (2009) carry out a clean up process of the dataset before using the computerized algorithms, unlike Trajtenberg et al. (2006) or Kim et al. (2006) they do not deal with name similarity or spelling problems.

information such as the technological class of the patent, the patent applicant, the location of the inventor or networks of co-inventors. The most interesting result of the authors' analysis of invention, mobility and co-inventors' networks was that "mobile inventors and short social chains of co-inventors are largely responsible for the localisation of knowledge flows" (Breschi and Lissoni, 2009; p. 27). Thus, an important conclusion is that knowledge flows are localised to the extent that inventors' mobility and networks are also localised.

To recap, the review of the empirical literature shows that inventors on the move are a source of new knowledge for the firm hiring them and for their former employer, but also for the geographical locations involved in the movement. This point, as we mentioned above, has not been studied in depth. Nonetheless, we are convinced of the importance of tracking down those locations attracting talent.

3. Data

As stressed in the first section, our main purpose is to explore the geographical mobility of inventors across European regions at a very fine level. To do so, we first need to define who can be considered as inventor moving across European regions. Instead of gathering data directly from the EPO, the USPTO, or the JPO[8], we consider patent applications filled at the Patent Cooperation Treaty (PCT) at the international phase which has been designed at the EPO. The aim of this section is to describe the dataset used in this study and the methodology to identify the mobility patterns of inventors.

3.1 Who is who?

The PCT database comprises patent applications filed to patent offices throughout all its 139 contracting countries. The dataset is actually made up of innovations patented at the three major offices worldwide. In this study, we will use only PCT data filed to the EPO. This procedure, which is an intermediate step between applying in the priority year (when a patent is filed for the first time worldwide) and filing for patent protection abroad, has been used increasingly in recent years (Usai 2008)[9]. Since PCT procedures are costly (at least more costly than merely filing the patent to the EPO), we assume that, by compiling these data, we have covered the most valuable inventions, and thus include the inventors who take more knowledge from one region to another when they move.

The choice of Europe as our focus of analysis is not a matter of chance. As we noted in the previous section, most large-scale studies have focused their analysis on the US (or use the US as a reference point), while for Europe studies have concentrated on country-specific cases or survey-based information. We rule out information regarding the inventors who have applied either from countries outside Europe or those who have applied to the USPTO or the JPO. There are many reasons for our choice. First, as already stressed, the US and Japan are the most productive countries in terms of innovation and patents (Maraut et al. 2008, Usai 2008). The fact that Europe is some way behind – especially some of its southernmost regions – could mean that our analysis would be biased to American or Japanese particularities, whereas our aim is to scan Europe's distinctive features. Second, and in relation to the first point, we are

[8] USPTO stands for United States Patent and Trademark Office, and JPO stands for Japan Patent Office.

[9] The PCT procedure allows to seek for patent protection in different countries at the same time using a unique application although only the offices like the EPO or the USPTO are able to grant the patent.

seeking a certain amount of homogeneity in our data, both administratively and geographically. The administrative problem is caused by bureaucratic idiosyncrasies in the European Office when a patent is filed, and leads us to rule out information from other offices. And since our main target is the analysis of geographical mobility across regions we need a certain amount of geographical homogeneity in relation to regional features in terms of size, statistical representation, complementary variables, and so on, and this is why we omit inventors living in countries outside Europe. We acknowledge that in this way we are missing important information, especially concerning the exclusion of the US – the literature has identified the US as one of the most important countries in terms of talent attraction (Trajtenberg et al. 2006, Maier et al. 2007), and some countries like the UK and Germany have a deep bilateral deficit of talent flows with the US. However, its inclusion would hide information regarding local/regional particularities which we are especially interested in identifying – even acknowledging that the most important patents worldwide tend to be filed to the USPTO, which provides extremely useful data (see Trajtenberg et al. 2006 and Kim et al. 2006).

The raw data for our study were collected from the OECD REGPAT database (OECD, May 2008 edition). This dataset links the addresses of the inventor(s) and applicant(s) for each patent to more than 2,000 regions throughout the OECD countries – see Maraut et al. (2006) for a methodological note. Thanks to their fruitful work, we can identify the region from which each inventor works when she applies for a patent. Basically, they are concerned with the process of regionalisation of patent data at very low levels of disaggregation, which they assess using the addresses of the inventor documented in the patent – the ZIP code, or, in its absence, the town name. This regionalisation procedure provides researchers with a complete dataset of patents applied for under the PCT procedures which contains a wealth of information for each patent, i.e., the publication number, the priority year (that is to say, the year when a patent was filed for the first time), information about the name, address, region code and country code of the inventor(s) and applicant(s) of each patent, the share of the patent that corresponds to each inventor or applicant, in order to take account of co-authorships and multi-applicants, and finally the technological class(es) to which each patent corresponds. For the purpose of our research we aim to identify those inventors who have applied for more than one patent, and who have done so while living in different regions. Here, therefore, we face the first problem, since the information contained in the PCT dataset does not include an ID for each inventor corresponding to that inventor and to no-one else, and which would be the same even if this inventor moved to another region. To overcome this problem we use the inventor's name (actually, her name, surname and middle name, if stated) to trace the movements of inventors who have moved region (this is the procedure used in Agrawal et al. 2006, Kim et al. 2006, and Trajtenberg et al. 2006, among others).

Unfortunately, two main problems arise in dealing with the strategy sketched above, which could be summarised as the "who is who" problem (Trajtenberg et al. 2006). The first occurs when the name (and surname) of the same inventor is spelled differently on different occasions (Trippl *versus* Tripl; Ericsson *versus* Eriksson; Smith *versus* Schmyt; and so on). The second concern is known in the literature as "the John Smith problem": i.e. when two inventors with exactly the same name are not actually the same inventor.

Given that our dataset contains almost 1,200,000 records[10], any manual procedure would be extremely time-consuming and, although relatively reliable, would

[10] A record is a unique combination of name and surname of the inventor, the location from which she applies, patent numbers, and the share of the patent which corresponds to that inventor.

not be immune from human error. Thus, as in Trajtenberg et al. (2006) or in Kim et al. (2006), for instance, our methodology is divided into two stages. The first one deals with name dissimilarity using name matching algorithms; the second stage seeks to identify each inventor (even if two records share exactly the same name) using several features linked to each record – the address of the inventor, her assignee, the technological class in which she is working, and so on. However, the methodology of the authors mentioned above must be adapted to our particularities and the caveats of the OECD REGPAT dataset, which was created only recently. A detailed explanation on how we have built our two-stages algorithm can be found in Appendix 2.

As a result, our methodology has identified 102,652 unique inventors who have applied for patents at least twice or more[11]. In the following table it is shown the distribution of our inventors among the top twenty regions hosting them. It can be seen that their distribution across European regions is far from being balanced. Thus, as expected, the majority of identified inventors are in the core European regions, with Germany having eleven regions within the top twenty.

[Insert Table 1 about here]

Moreover, among these identified inventors, 10,403 have crossed at least once one of our regional borders (we have defined 698 regions, see section 3.2 for an explanation) throughout the whole period of analysis (1990-2006), meaning the 10,13% of them.

3.2 Dataset and variable construction

Since we aim to reflect the geographical mobility of inventors, tracking down those regions attracting talent, we now describe the construction of the variable that is used to obtain this information. We are interested in the areas which attract talented personnel, and so we consider the sum of the number of inventors in each combination of region and year who already applied for patents from another region in a previous year (inflows of inventors). Our period of analysis covers a large range of years, from 1990 up to 2006 – although movements originating in the sending region between 1980 and 1990 that fall into the range of receiving regions in the period 1990-2006 are also considered.

The geography of innovation and knowledge spillovers on a large scale is usually analysed in large regions (NUTS 1 or NUTS 2 for the case of Europe, where Bottazzi and Peri 2003, Peri 2005, Moreno et al. 2005, Miguélez et al. 2008, are some examples) or even at the level of countries or US states due to data constraints. However, in this paper, we try to deal with this drawback by carrying out our study at a lower level of regional disaggregation. Ideally, bearing in mind the phenomenon we are studying, the ZIP code level would be interesting, because otherwise we may well not take account of a number of movements within larger regions – larger than the ZIP code – and we may underestimate the extent of this phenomenon. However, we should also bear in mind that, by doing so, we might identify movements of inventors who apply for patents from different workplaces within the same firm or research institution – so we would be overestimating the number of geographical movements. In any case, we prefer to

[11] We only include identified inventors with at least two patents because only them are those potentially moving inventors.

underestimate movement counts rather than overestimate them, so as not to affect possible future econometric estimations[12].

Given all the above arguments, we have chosen NUTS 3 level as the spatial unit for our analysis[13]. However, the size and scope of this administrative division in Europe varies greatly, and that is why we proceed as follows. First, we calculate the average area of NUTS 3 regions for the whole of Europe; we do the same for NUTS 0, 1, 2, and 3 regions in each country; and then we choose the level of NUTS for each country which is closest to the average European NUTS 3 size obtained in the first stage. Therefore, in our case this process obtains a sample of NUTS 3 regions for the majority of countries except for the case of Belgium, Germany, the Netherlands, Switzerland, and United Kingdom[14], where NUTS 2 regions will be considered. Moreover, because Eurostat has undertaken several country-specific reorganisations of these regions in recent years, we consider NUTS 2 regions for the case of Poland, and NUTS 0 for Denmark[15]. Our final sample covers 698 regions in 29 European countries.

Further practicalities: in spatial analysis (especially when depicting data on a map) data for areas come in the form of counts, like in the present study –counts of inflows. However, areas differ in size in terms of the population able to apply for a patent (the number of inventors of each region). Therefore, data need to be adjusted so that population size effects do not distort comparisons (Haining, 2004). Most cross-section studies of innovation issues, for example, tend to use population as denominator – considering, thus, patents per capita instead of the absolute number of patents in each region. However, total population does not represent the relevant population for our purposes, since potential movers must be involved in innovation. To deal with this issue, we compile data on Human Resources in Science and Technology (HRST) from Eurostat[16]. Specifically, we compile data on the percentage of HRST over total active population. We then multiply these percentages by the total population of an area for the whole period (these data are also compiled from Eurostat) and thus obtain our measure of the relevant population in each region[17]. Moreover, we multiply these data by the percentage of patents applied from each region (using not only patents applied at the PCT phase, but all EPO patents), in order to take account of the tendency of certain

[12] We are also aware of the existence of the “Modifiable Areal Unit Problem” (MAUP). In spatial statistics and econometrics, results – especially concerning spatial association statistics – may well change radically depending on the spatial scale of the analysis, so our results should be considered, as usual, with caution.

[13] In this sense, it is worth mentioning a recent study by Ponds et al. (2007), who studied the phenomenon of knowledge flows at NUTS3 level as well. However, unlike our work, the main focus of that paper as a mechanism for knowledge flows is not the mobility of inventors, but networks of scientific collaborations.

[14] The whole list of countries considered and the number of regions in each one can be found in the Appendix 1.

[15] We also consider the island of Sardinia as a whole NUTS 2 (instead of NUTS 3) and the German *Land* of *Sachsen-Anhalt* as a single NUTS 1 region, and we have omitted the regions of Las Palmas de Gran Canaria, Tenerife, Ceuta, Melilla, Madeira, Açores, Guadeloupe, Martinique, Guyane and Reunion due to their distance from continental Europe.

[16] HRST are defined by Eurostat as people who fulfil at least one of two conditions: either successfully completed tertiary education, or are not formally qualified but are employed in an S&T occupation where the mentioned qualifications are normally required. In order to restrict the definition of those potentially movers in our study, we consider only those people who meet both requirements, and which are labelled as the CORE of HRST.

[17] We are aware, however, that data on HRST from Eurostat are only disaggregated at NUTS 2 level, so, when necessary, we use the same percentage for all NUTS 3 regions within a given NUTS 2 region.

regions to further innovate[18]. Here we give an example to clarify this point: let's say that Berlin had in 1990 about 41,072 people considered HRST. Imagine that Berlin hold 21% of the total amount of patents applied to the EPO throughout European regions in 1990. To calculate Berlin's IMR for 1990, we would divide the number of Inflows over 41,072*(1+0.21) –we will also divide this latter number over 1,000 to consider as IMR the number of Inflows over *thousands* of HRST.

$$IMR = \frac{I}{[(HRST \cdot (1+R))/1{,}000]} \qquad (1)$$

Thus, IMR will be the Inward Migration Rate, our main variable, where I stands for the inflows (already described), HRST stands for Human Resources in Science and Technology, and R will be the aforementioned correcting factor.

4. A first insight into the spatial distribution of scientists' mobility

This section aims to provide a preliminary idea of the spatial distribution of geographical knowledge flows driven by the geographical mobility of inventors, using the data presented in the above section. Applying a set of exploratory analysis and spatial statistics tools (ESDA), the objectives of our analysis are divided into three groups. First, by describing and visualising in maps the distribution of our variable, we aim to identify the foci of attraction of talent among European regions, these *agglomeration centres for knowledge flows*. As already stressed in the introductory section, this is an important issue to be addressed for its implications for growth, development and innovation capacity at regional level. Second, using these ESDA tools, we aim to assess whether there is some kind of spatial pattern in the geographical distribution of this phenomenon— specifically, whether they present a significant spatial concentration, or whether their distributions are characterised by any significant local regime. Basically, we are interested in elucidating why these movements could be concentrated in space – if in fact they are – or what the relationship might be between geographical inflows of inventors in one area and that of its neighbours. As already sketched in the introduction of this text, we have the hypothesis that these centres of attraction are not randomly distributed across the space, but they present certain pattern of spatial correlation –so those region with high values of the IMR are located nearby other high-value regions. The theoretical rationale behind our hypothesis is three-fold. First, the attractive regions are located nearby on the space because the attractive characteristics of a given region –amenities, job opportunities, social networks, research

[18] We acknowledge that the ideal solution will involve the use of the inventors identified through our algorithms (therefore all inventors not only those moving) as a denominator. Doing so will come about with two problems. First, although we can identify inventors throughout European regions and also their movements, we cannot know if they stop patenting at some point (in case they retire or die, for instance). Thus, at each point in time it is not possible to know the exact number of inventors applying from a given region unless we adopt strong assumptions. From our point of view, any possible heuristic might be misleading, and that is why we opt for gathering secondary data on inventors and scientists. The second problem is related to identification. We will see it through an easy (and extreme) example. Consider the Greek Western region of Corfu (Kerkyra), where we have identified only one inventor during the whole period, who patented there after being hired by a local firm from her former work, located in Athens. Not being identified any other inventor in the region (since only applied patents are used to identify them), Corfur will be among the top receiving regions in our sample, since its I.M.R. would be maximum (will be 1). Again, this problem leads us to gather secondary data to adjust our variable for the size effect.

facilities, multinational firms, contacts with the academia, and the like- may well spill over its administrative boundaries. Second, some of these attractive regions, especially urban ones, may suffer some kind of congestion effects, due to high land prices, traffic jumps, or pollution, which would favour the location of the research agents outside that region, but nearby it at the same time in order to take advantage of the possible existence of agglomeration economies. Finally, we believe that certain European countries deserve an aura of attractiveness thanks to their research prestige, their wage premium, or their industrial tradition, that makes all the regions of these countries attractive for the inventors from abroad. Finally, the last point of our analysis will show that the movements of inventors are geographically mediated, that is, they are bounded in the space, and also quite a lot a-spatially concentrated, although further econometric analysis would be necessary to confirm this point.

4.1 Spatial distribution of inflows of inventors

We now analyse the spatial distribution of the movements of inventors in Europe. To do so, we examine the spatial patterns of the IMR of inventors across NUTS3 European regions. Since data of this kind may exhibit lumpiness from year to year, we use the average of the 17 years under consideration for our purposes. Before visualising the data, however, in the following table can be found the top twenty regions in terms of Inflows, and the top twenty regions in terms of the IMR, both as average of the whole period. Few points to be highlighted: nine over the twenty regions are from Germany in terms of average inflows, although France, Switzerland, and the United Kingdom have also representation. The leader, however, is the Dutch region of North-Brabant. When taking into account the size effect, however, the ranking fills up of Austrian, Finish, Italian, and Swedish regions also, and the overall picture is slightly different, being the Austrian region of Wiener Umland/Südteil located at the top of the ranking.

[Insert Table 2 about here]

In map 1, the Inward Migration Rate is sketched using a quintile distribution, showing the regions that attract talent. A clear distinction appears between regions in countries with high values of the variable, including Belgium, the Netherlands, Germany, Ireland, United Kingdom, Luxembourg, Switzerland, Italy, Austria, Denmark, Norway, Sweden, Finland, and, to a lesser extent, Slovenia, and countries with low values through the majority of their regions, including Spain, Portugal, Greece, Malta, Cyprus, and the eastern countries of Romania, Bulgaria, Hungary, Czech Republic, Slovak Republic, Poland, Latvia, Lithuania, and Estonia. Therefore, by including eastern countries we identify a clear “Core-Periphery” division rather than a “North-South” segregation, with Nordic countries being inside the Core as well. This pattern is also observed for the great majority of economic variables, especially those most related to innovation and knowledge spillovers. It is worth highlighting some particular cases within this general pattern. For the case of Germany, for instance, nearly all regions belong to the highest quintiles, especially those in the western and southern part of the country, although more central regions like Hannover and Braunschweig also show high levels of the IMR. Certain regions in the north-west of the country are also especially interesting, like Düsseldorf, Münster, Arnsberg, and Köln (where the city of Aachen is located), which show some of the highest values in our sample and seem to form a high value cluster with the Dutch region~~s~~ of North Brabant (which includes Eindhoven) and, to a lesser extent, with the Belgian regions of Wallonia, Leuven, and Brussels. Meanwhile, the regions located in the south and west of Germany show high values as well, which seem to be partially correlated with Swiss regions, Austrian regions, and the French regions within the NUTS2 regions of Alsace and Lorraine. Some high-valued French regions are also located near the administrative boundaries with Switzerland and Italy, on the Mediterranean coast, Ille-et-Vilaine in Brittany (whose capital is Rennes), and around Paris (including the capital) – especially to the south of the capital (like Hauts-de-Seine). For Italy, the picture is slightly more random, although of course the higher values are concentrated in the north (large part of the NUTS3 regions within Lombardia –where Lodi is within the top-twenty regions of Europe in terms of IMR- and Emilia-Romagna), with the surrounding regions of Rome – Rieti, L'Aquila, Pescara, and Chieti – located in the fourth and fifth quintile of the distribution. For its part, the highest values in the United Kingdom are recorded in the south-east of the country. It is worth mentioning that the NUTS 2 regions of London (Inner and Outer London) are located in the fourth quintile of the distribution, while all their surrounding regions are located in the upper quintile – which seems to indicate some kind of congestion effect in the capital. Likewise, some central and northern regions like Cheshire, Derbyshire and Nottinghamshire, North Yorkshire (which includes Leeds), Tees Valley and Durham, and Northumberland and Tyne and Wear, are also located in the upper quintile. In the case of Nordic countries, all Swedish regions are located in the fifth quintile of the distribution (especially Uppsala and Skane), and also some Finnish regions (especially on the south coast of the country), while only the Norwegian regions of Telemark, Oslo and Sor-Trondelag are located in this quintile. Austria shows high values for the majority of its regions as well, especially for Vienna and its neighbours and the regions bordering with Germany and Switzerland, while some regions in Slovenia show high values as well, though only the region of Koroska, bordering with the Austrian region of West- und Südsteiermark (which is in the top-twenty), is located in the upper quintile – pointing to the existence of some kind of cross-national spillover effects. As stressed, the case of Austria is paradigmatic. Although none of its regions are within the top-twenty receiving more inflows, it has 5 regions within the top-twenty IMR ranking, and especially for those surrounding the capital. Finally, Denmark is located is the third quintile, while the region across the border in Sweden and Germany are in the upper quintile. However, we should bear in mind that Denmark is not divided into different regions in our study and, given that it seems at first sight that the great majority of movements are within each country, this result is as expected.

In the case of the countries with low values of inward migration, some exceptions should be highlighted. These exceptions are not a real focus of attraction compared with the whole sample, but they are if compared with their surrounding areas. This is the case, for instance, of the Hungarian regions of Budapest and Pest, and some Spanish regions in the north-east, the Mediterranean coastal regions, León and Lugo in the north-west and those areas immediately surrounding Madrid and the capital itself – which nonetheless seems to experience a degree of congestion and crowding-out.

[Insert Map 1 about here]

From Table 2 and Map 1, an initial conclusion arises: only a subsample of countries, and a subsample of regions therefore, are benefiting from the immigration of talented individuals and, as a consequence, these regions are those *potentially* benefiting from knowledge flows and human capital spatial externalities. This is in line with the concluding ideas suggested in Maier et al. (2007), although their work is not focused on regions and inventors, but on countries and star-scientists. All in all, from this first analysis it is important to bear in mind that (1), even when controlling for innovation potential and patenting bias, skilled individuals' attraction is specially reserved for few countries and regions, whilst this phenomenon is very poor or inexistent in other countries; (2), large cities and capital cities register high values of our IMR most of the times, even in poor performing countries in terms of inflows –supporting the theses about the importance of urban agglomerations; and (3), in some cases, the regions surrounding these large or capital cities are even more magnetic, pointing to the existence of spillovers of attractive features and/or crowding-out effects. These ideas will be partially addressed in the following section.

4.2 Spatial patterns of association of inventors' movements

The next step consists in dealing with significant spatial effects, atypical allocations, outliers, and the like. Our idea is to test the hypothesis that the IMR is correlated across the space indeed, due to spillovers of regional attractive features, crowding-out effects due to congestion, and country specific characteristics. Before addressing this issue, we need to define a measure of "neighbouring", which will be summarised in a nxn matrix of spatial weights, n being the number of regions, where $W = \{w_{ij}\}$. The most usual definition of neighbouring is first-order physical contiguity, that is, if two regions share the same administrative border $w_{ij} = 1$, and $w_{ij} = 0$ otherwise. However, the first-order contiguity matrix for Europe, in which there are a number of islands, would induce a matrix with rows and columns with only zeros, which would change the sample size and the interpretation of statistical inference. Other contiguity criteria have been defined in the literature, such as commercial exchanges (Cabrer-Borràs and Serrano-Domingo 2007) and technological proximity (Moreno et al. 2005), although some endogeneity problems may well arise. The appropriate weight matrix should, then, be chosen with care. In this regard, we consider appropriate a distance-based matrix with a fixed number of neighbours – see Le Gallo and Ertur (2003) for methodological concerns regarding these matrices. When fixing an equal number of neighbours for all regions, we avoid certain methodological problems that may occur when the number of neighbours is allowed to vary –when contiguity or distance-based matrices without a fixed number of regions are used (see Le Gallo and Ertur 2003). Given that the average number of neighbours for our sample using first-order contiguity matrices is 4.87, and the median is located between five and six neighbours, we will assign five fixed neighbours in our matrix. Nonetheless, we will also check the robustness of our analysis using distance-based matrices with 10, 15, and 25 neighbours, given that we are working with relatively small areas (which differ somewhat in terms of size)[19].

What we would like to know is whether there exists a relationship between the immigration rate of inventors in one region and in the neighbouring regions. In spite of the intuitive conclusions arising from the visualisation of our maps, we must use some statistical analyses to verify the existence of a spatial structure of migration data. To shed some light on the possible existence of global spatial autocorrelation (SAC) in our sample, we use the Moran's I and Geary's c statistics. The results leave no doubt (see table 3). There exists a strong, positive spatial autocorrelation in the IMR, with no differences in relation to the definition of the contiguity criteria or the statistic used. This positive spatial autocorrelation implies that regions with a high value of the Inward Migration Rate are the neighbours of other high-performing regions. In contrast, low registers tend to be located in regions next to other poor performers (a negative,

[19] All the results using alternative matrices will be provided by the authors upon request.

significant autocorrelation would suggest that there are clusters of regions with high levels of the IMR surrounded by a set of low-IMR regions – and vice-versa). Therefore, we interpret these results as evidence of the existence of certain features which favour the spatial correlation of our immigration rate. Among other potential reasons behind this spatial correlation in the distribution of inventors' inflows, we suggest that the attractive characteristics of certain regions might be spilling over to neighbouring regions, or alternatively that these regions could be experiencing crowding-out effects, or finally that country-specific features also matter. However, the exact mechanisms underlying the spatial correlation encountered are not known given our approach, and in any case its investigation exceeds the purpose of this study. A follow-up interpretation of this preliminary evidence would stress the importance of geography and location for attracting talent. However, again, further confirmatory analyses would corroborate the relative importance of both within regional attractive endowments and cross-border effects (geography).

[Insert Table 3 about here]

Several other questions are also of interest. Are there any local geographical patterns driving the positive global SAC? Put another way, which regions contribute most to the global SAC? Are there local clusters of migration rates? Can they be identified as spatial regimes? (If so, spatial non-stationarity should be considered aside from the SAC). Are there atypical allocations? To partially answer these questions, first of all we use the Moran scatterplot. This spatial tool (Figure 1) plots the value of IMR ~~GMR~~ for each observation against its spatial lag. It is worth computing the percentage of regions in each quadrant. As observed, the bulk of regions (75%) are located in the upper-right (HH) and the lower-left (LL) quadrants of the scatterplot, where regions with high (low) values of our variable are surrounded by regions with high (low) values as well. Among this vast majority, however, the distribution is uneven, since 26% are located in the HH quadrant and 49% in the LL one. In principle, the regions located in those quadrants would form clusters of high and low values respectively, and therefore different spatial regimes (spatial non-stationary) would be identified, although this extreme would need to be confirmed using LISA tests and Moran's scatter maps. Therefore, these local clusters are assumed to be driving the local forces towards global SAC. The remaining quadrants (upper-left, LH; lower-right, HL) show those atypical allocations, that is to say, those regions with high levels of the IMR surrounded by regions with low values (and vice-versa), which deviate from the global pattern of positive SAC.

[Insert Figure 1 about here]

Although this figure is quite revealing, we cannot say anything about the significance of these spatial regimes and atypical allocations without performing a local Moran's I statistic. We perform this test using our main weighting matrix (distance-based matrix with 5 fixed neighbours) and also using matrices with 10, 15, and 25 neighbours. Significant local clusters[20] are located in the southern, central, and eastern regions of the Iberian peninsula, the extreme south of Italy and Sicily, Malta, Cyprus, Greece, Romania, Bulgaria, Poland, Lithuania, Latvia, and partially Estonia, southern Sweden and Denmark, the region of Paris and its surrounding areas, the south-east of

[20] The results are not shown in order to save space but will be provided upon request.

the United Kingdom, and a large cluster in the core of Europe covering a large part of German regions, some Swiss and Austrian regions, some northern Italian regions, and several Belgian regions. So these regions present a certain spatial dependence which stands out from the average spatial autocorrelation of the sample.

By combining the information from the Moran scatterplot and the local Moran's I statistic, we obtain the Moran scattermap (Map 4). This map shows the regions which display significant local spatial autocorrelation, which are the same as those in the previous list of regions. Furthermore, the regions are encoded according to the quadrant (HH, red; LL, blue; HL, light red; LH, light blue) allocated in the Moran scatterplot. Two spatial regimes of low values are clearly identifiable in the south and east of Europe, covering some of the regions of the Iberian Peninsula (except the north-east), the extreme south of Italy and Sicily, Malta, Cyprus, Greece, Romania, Bulgaria, Poland, Lithuania, Latvia, and partially Estonia. A large spatial regime of high values can also be identified in the core of Europe, Paris and its surrounding regions, some regions of Scandinavia and part of the United Kingdom. Besides, the local Moran's I statistic identifies a number of atypical areas, i.e. those that exhibit negative spatial correlation with their neighbours: atypical low (high) levels of the IMR areas on the periphery of the high (low) levels of flows regimes. Some interesting findings emerge: first, only two regions (Toledo and Leon, from Spain) located in the HL quadrant (high values surrounded by low values) are actually significant. In contrast, several regions allocated in the LH quadrant are significant, and mainly correspond to regions located near the high-value spatial regime which do not follow their neighbours in terms of in- and out-movements of skilled personnel. These regions are four regions in Austria and one in Slovakia, one Belgian region, one North Italian region, one Dutch region, a few Finnish regions, Seine-Saint-Denis, in France, next to Paris, and finally Denmark – confirming our suspicions about the Danish results. All in all, this subset of regions can be identified as significant atypical allocations – clusters of dissimilar values, that is to say, regions with high (low) levels of in- and out-flows of inventors surrounded by regions with low (high) levels – which would lead to a negative SAC if they predominated in our sample, which obviously is not the case.

[Insert Map 4 about here]

Finally, it is also worth analysing the origin-destination flows of inventors across European regions. In this respect, we would like to show a double fact. First, that the movements of inventors across European regions are bounded in the space; and second, we would like to confirm that only a subsample of these regions are making the most of this phenomenon, and therefore only this subsample of regions are potentially benefiting from the flow of knowledge which these movements may well imply[21]. To do so, we come back to the inflows measure for the whole period instead of the IMR.

[21] In this sense, it is essential to attract the attention to an important point. As already stressed, we have focused the analysis on geographical mobility (which occurs when an inventor crosses an administrative boundary). However, it is important to bear in mind that part of these geographical movements will occur within the same firm (a plant localisation change, a temporal change of the destination of the employee, and so on). In front of this situation, several scholars argue that since the employee has not changed her firm, there is not a real flow of knowledge because all her knowledge and skills remain within the firm. On the contrary, part of the literature also argues that even in front of this situation, the mere location change of the employee implies a knowledge flow because she will interact with other individuals different than before, even though only outside the plant (during her leisure time, in her relationships with other customers and suppliers, and so on). However, the mechanism through which knowledge diffuses between individuals is an open debate.

We have also calculated the outflows measure, to check its a-spatial concentration behaviour. We show several figures in the table above. As can be seen in the first row of the table, a large part of the inflows (44%) throughout the whole period (1990-2006) come from regions located within the 10 nearest neighbours of a given region. What is more, more than 30% of them come from the 5 nearest neighbours. However, the striking fact is that more than 76% of those inflows come from a region located within the same country. All in all, it seems clear to us that the migration movements of the inventors are localized phenomena, in other words, geographically mediated. Gini indexes are shown in the second row of the table. Both inflows and outflows are quite a lot concentrated during the whole period, since both are near to 1 -the Gini index ranges from 0 (perfect equality) to 1 (perfect inequality). Finally, we can confirm this a-spatial concentration from the last row of the table, since, as shown, more than 40% of the inflows during the whole period are concentrated only in 20 regions. The same applies for the other side of the coin, that is, the outflows of inventors. In this sense, it is important to notice that 17 regions are in both top rankings, corroborating the fact that only a subsample of regions are participating of this phenomena. This conclusion is in line with what we saw at the end of section 4.1. Furthermore, we believe that the features of these movements (localized and reserved for relatively few regions) are also contributing to the existence of global positive spatial correlation. We acknowledge, however, that these ideas and hypotheses deserve further investigation and econometric techniques should be used, although the figures presented are already quite revealing.

[Insert Table 4 about here]

5. Conclusions and lines of future research

The main goal of this paper is to determine which European regions are foci of attraction of talent. Thus, our research has extended the existing literature on inventors' mobility by focusing on the geographical aspect of this phenomenon, rather than on job-to-job mobility. The analyses so far have shown that regions with high levels of the IMR are located in the core of continental Europe (German regions, Switzerland, Austrian regions, eastern French regions and those around Paris, northern Italy, and so on), Nordic countries and the United Kingdom, forming a spatial cluster of high values. As stressed all over the present text, these areas are those potentially receiving larger amounts of knowledge flows, through inventors' mobility. Moreover, simple figures at the end of section 4.2 partially showed that the inflows (not the IMR) and outflows phenomena are quite a lot concentrated in some regions. These encountered facts are in line with recent research on the spatial distribution and movements of highly skilled individuals (see the analysis of star-scientists by Maier et al., 2007). Further, we would like to know whether regions with high (low) levels of the IMR of inventors are located near other regions with high (low) levels of flows. Thus, we had the hypothesis that global spatial correlation may well exist due to spatial spillovers of attractive characteristics across regions, crowding-out effects, and country-specific features. To do this, we conduct an ESDA. The empirical approach confirms the spatial correlation of our data, although the reasons underlying this phenomenon are not possible to disentangle yet given our descriptive methodology. In any case, we can confirm the importance of being located nearby those leading regions and cities in terms of economic performance, innovation activity and inventors' stocks and inflows, so co-location and geography matter for attracting talent. We also had the hypothesis that this phenomenon is geographically bounded and reserved only to a set of regions which are making the most of these movements. Our exploratory approach confirms this extreme, although further empirical analysis should be carried out in the future. Our conclusion is that, even acknowledging that those regions have a high tendency towards innovation, they are more likely to receive inflows of inventors. Thus, following the suggestions of Breschi and Lissoni (2009), a conclusion should be borne in mind, i.e., knowledge flows are localised to the extent that inventors' mobility is also localised.

In this regard, our plans for future research will be concerned with depicting the relationships between regions through inventors' mobility, and with the factors influencing this phenomenon, i.e., whether geographical distance is the main driving force behind it, or whether other forces such as technological similarity, income level, socio-cultural similarity, national boundaries, and so on, are also involved. The estimation of origin-destination flow models (gravity models) will be used to assess the importance of geography and other regional features on explaining the inventors' inter-exchange phenomenon across European regions[22].

In order to carry out our proposals, we have put forward a methodology for tracing the mobility patterns of inventors who have applied for patents to the EPO under the Patent Cooperation Treaty over a long period (1990-2006) across European regions (mainly NUTS 3 regions, and some NUTS 2 regions when necessary). To do so, our suggestion is divided in two stages. The first one deals with name similarities using Soundex, a well-known name matching algorithm. In the second stage, several features linked to each patent and inventor – the technological class of the patent, the applicant, the region from which the inventor makes her application, the name of the inventor, and so on – are used to test whether each encoded inventor's name belongs to the same person or not.

Finally, we should mention the main limitations of our research study. The first one is related to the raw data. The OECD regularly launches new editions of its dataset, which is continuously updated by users who report possible mistakes, and therefore the use of new editions will also improve on our matching procedure and our final dataset. Next, although we sought to homogenise the raw data prior to the study, these data should be more thoroughly monitored, due to a tendency of the OECD REGPAT database to mix records with the full name and middle name with other records with only the initials of both names, and other practices that may cause mistakes. Finally, once the raw data are improved, the algorithms could also be refined, for instance by changing the number and type of tests, the scores or the thresholds. Actually, one of our lines of future research involves the design of optimisation algorithms able to decide the score of each test and the thresholds by themselves. Despite these limitations, we think that the analysis performed in this study is reliable and the conclusions are not affected by the problems we have described.

[22] In this regard, we would like to thank the anonymous referees for suggesting us the estimation of such a model.

Acknowledgements
We would like to thank Ismael Gómez Miguélez for his invaluable help. Ernest Miguélez and Rosina Moreno acknowledge financial support from the *Ministerio de Ciencia e Innovación*, ECO2008-05314, and Jordi Suriñach also from the *Ministerio de Ciencia e Innovación*, ECO2008-02291. The authors would also like to share their appreciation of helpful comments received from the participants of the III World Spatial Econometrics Conference (Barcelona, July 8th-10th, 2009), the participants of 9th European Summer School on Industrial Dynamics "Agglomeration, Industries and Innovation" (Barcelona, August 31st-October 4th, 2009), the DIME Workshop on Technology, Skills and Geography at SPRU (Sussex University, September 11th-12th, 2009), the editors, and three anonymous referees. However, any mistake or omission remains ours.

References

Acs Z, Anselin L, Varga A (2002) Patents and innovation counts as measures of regional production of new knowledge. *Research Policy* 31: 1069–1085

Agrawal A, Cockburn I, McHale J (2006) Gone but not forgotten: labour flows, knowledge spillovers, and enduring social capital. *Journal of Economic Geography* 6: 571-591

Ackers L (2005) Moving people and knowledge: scientific mobility in the European Union. *International Migration* 43(5): 99-131

Almeida P, Kogut B (1999) Localisation of knowledge and the mobility of engineers in regional networks. *Management Science* 45: 905-917

Anselin L, Varga A, Acs Z (1997) Local Geographic Spillovers between University Research and High Technology Innovations. *Journal of Urban Economics* 42: 422-448

Bottazzi L, Peri G (2003) Innovation and spillovers in regions: Evidence from European patent data. *European Economic Review* 47: 687 – 710

Branting LK (2003) A comparative evaluation of name-matching algorithms, International Conference on Artificial Intelligence and Law

Breschi S, Lissoni F (2009) Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, pp. 1-30

Cabrer-Borrás B, Serrano-Domingo G (2007) Innovation and R&D spillover effects in Spanish regions: A spatial approach. *Research Policy* 36: 1357–1371

Corredoira RA, Rosenkopf L (2006) Learning from those who left: the reverse transfer of knowledge through mobility ties, Management Department Working Paper

Crespi G, Geuna A, Nesta L (2007) The mobility of university inventors in Europe *Journal of Technology Transfer* 32(3): 195-215

Döring T, Schnellenbach J (2006) What do we know about geographical knowledge spillovers and regional growth?: A survey of the literature. *Regional Studies* 40.3: 375-395

Fleming L, King C, Juda A (2007) Small worlds and innovation. *Organization Science* 14(5): 375-393

Florida, R (2002) *The rise of the creative class: and how it's transforming work, leisure, community and everyday life*. Basic books, cop. New York

Giuri P, Mariani M, Brusoni S, Grespi G, Francoz D, Gambardella A, Garcia-Fontes W, Geuna A, Gonzales R, Harhoff D, Hoisl K, Le Bas C, Luzzi A, Magazzini L, Nesta L, Nomaler Ö, Palomeras N, Patel P, Romanelli M, Verspagen B (2007) Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy* 36: 1107-1127

Haining, R (2004) *Spatial Data Analysis. Theory and Practise*. Cambridge University Press, Cambridge

Hoisl K (2009) Tracing mobile inventors: The causality between inventor mobility and inventor productivity. *Research Policy* 36(5): 615-636

Hoisl K (2007) Does mobility increase the productivity of inventors? *Journal of Technology Transfer* 34: 212-225

Jaffe AB, Trajtenberg M, Henderson R (1993) Geographic localisation of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577-598

Kim J, Lee SJ, Marschke G (2006) International knowledge flows: Evidence from an inventor-firm matched dataset. NBER Working Paper 12692

Laforgia F, Lissoni F (2009) What do you mean by ‘mobile’? Multi-applicant inventors in the European Biotechnology Industry. In; Malerba F., Vonortas N. (eds.) *Innovation Networks in Industries*, Edward Elgar (forthcoming)

Le Gallo J, Ertur C (2003) Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980-1995. *Papers in Regional Science* 82: 175-201

Lenzi C (2009) Patterns and determinants of skilled workers’ mobility: evidence from a survey of Italian inventors. *Economics of Innovation and New Technology* 18(2): 161-179

Lissoni F (2008) Academic inventors as brokers: An exploratory analysis of the KEINS database *CESPRI Working Paper*, 213

Lissoni F, Llerena P, McKelvey M, Sanditov B (2008) Academic patenting in Europe: new evidence from the KEINS database. *Research Evaluation* 16: 87–102

Lissoni F, Sanditov B, Tarasconi G (2006) The Keins database on academic inventors: methodology and contents *CESPRI Working Paper*, 181

Lucas, RE (1988) On the mechanics of economic development. *Journal of Monetary Economics* 22: 3-42

Maier G, Kurka B, Trippl M (2007) Knowledge spillover agents and regional development: spatial distribution and mobility of star scientists, *DYNREG* Working Papers 17/2007

Maraut S, Dernis H, Webb C, Spiezia V, Guellec D (2008) The OECD REGPAT Database: A presentation *STI Working Paper* 2008/2

Miguélez E, Moreno R, Artís M (2008) Does social capital reinforce technological inputs in the creation of knowledge? Evidence from the Spanish regions, *IREA Working Papers 2008/13*, forthcoming in *Regional Studies*

Moreno R, Paci R, Usai S (2005) Geographical and sectoral clusters of innovation in Europe. *Annals of Regional Science* 39: 715–739

Moretti, E (2004) Human capital externalities in cities. In: V. Henderson and J. Thisse (eds) *Handbook of Urban and Regional Economics*, vol. 4.

OECD (2008) *The Global Competition for talent. Mobility of the highly skilled*. Organisation for Economic Co-operation and Development

Peri G (2005) Determinants of Knowledge Flows and Their Effect on Innovation. *The Review of Economics and Statistics* 87(2): 308-322

Polanyi M (1966) *The tacit dimension*. Routledge & Kegan Paul, cop., London

Ponds R, van Oort F G, Frenken K (2007) The geographical and institutional proximity of research collaboration. *Papers in Regional Science* 86: 423-443

Raffo J, Lhuillery S (2007) How to play the “Names Game”: Patent retrieval comparing different heuristics, *Research Policy*, In Press: doi:10.1016/j.respol.2009.08.001

Saxenian A (2005) From brain drain to brain circulation: transnational communities and regional upgrading in India and China. *Studies in Comparative International Development* 40: 35-61

Snae C (2007) A comparison and analysis of name matching algorithms. *Proceedings of World Academy of Science, Engineering and Technology* 21: 252-257

Shalem R, Trajtenberg M (2008) Software patents, inventors and mobility, Working Paper

Singh J (2005) Collaborative networks as determinants of knowledge diffusion patterns. *Management Science* 51(5): 764-786

Song J, Almeida P, Wu G (2003) Learning-by-hiring: When is mobility more to facilitate interfirm knowledge transfer? *Management Science* 49(4): 351-365

Thoma G, Torrisi, S (2007) Creating Powerful Indicators fo Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases. *CESPRI WP* n. 211

Trajtenberg M, Shiff G, Melamed R (2006) The “names game”: harnessing inventors’ patent data for economic research, *NBER working paper 12479*

Trajtenberg M, Shiff G (2008) Identification and mobility of Israeli patenting inventors, The Pinhas Sapir Center for Development, Tel Aviv University DP No. 5-2008

Trippl M, Maier G (2007) Knowledge spillover agents and regional development, *SRE-Discussion 2007/01*

Usai S (2008) The geography of inventive activities in OECD regions, STI Working Paper 2008/3

Zucker LG, Darby MR, Armstrong J (1998a) Geographically localized knowledge: Spillovers or markets? *Economic Inquiry* 36: 65-86

Zucker LG, Darby MR, Brewer MB (1998b) Intellectual human capital and the birth of U.S. biotechnology enterprises. *American Economic Review* 88(1): 209-306

Zucker LG, Darby MR, Torero M (2002) Labor Mobility from Academe to Commerce. *Journal of Labor Economics* 20(3): 629-660

Zucker LG, Darby MR (2006) Movement of star scientists and engineers and high-tech firm entry, *NBER* Working Paper 12172

Table 1. Number of identified inventors. Top-twenty regions (1990-2006)

Germany	Oberbayern (Munich)	5254	Sweden	Stockholm	2184
Germany	Stuttgart	5029	Germany	Freiburg	2048
Germany	Darmstadt	3825	Germany	Berlin	1891
The Netherlands	North-Brabant (Eindhoven)	3506	United Kingdom	East Anglia (Cambridgeshire)	1838
Germany	Düsseldorf	3494	Switzerland	Eastern Switzerland	1787
Germany	Karlsruhe	3301	Germany	Tübingen	1723
Germany	Köln (Aachen, Bonn)	3287	Finland	Uusimaa (Helsinki)	1702
Denmark	Denmark	2981	United Kingdom	Berkshire, Buckinghamshire and Oxforshire	1622
Germany	Rheinhessen-Pfalz (Kaiserslautern)	2777	France	Hauts-de-Seine	1613
France	Paris	2272	Germany	Mittelfranken	1594

Note: In parenthesis can be found the most known city(ies) of each region, if any. The table counts all the inventors identified through our matching-sppliting algorithm who have applied for patents to the EPO (under the PCT phase) from each of the top-20 regions.

Table 2. Ranking of most receiving regions. Inflows and IMR. Average 1990-2006

(i)			(ii)		
Average 1990-2006 inflows			Average 1990-2006 I.M.R.		
The Netherlands	North-Brabant (Eindhoven)	35.29	Austria	Wiener Umland/Südteil	1.65
Germany	Karlsruhe	32.06	Germany	Rheinhessen-Pfalz (Kaiserslautern)	1.60
Germany	Oberbayern (Munich)	28.35	Switzerland	Eastern Switzerland	1.57
Germany	Darmstadt	28.29	Sweden	Uppsala	1.24
Germany	Rheinhessen-Pfalz (Kaiserslautern)	28.29	The Netherlands	North-Brabant (Eindhoven)	1.20
Germany	Köln (Aachen, Bonn)	27.18	Switzerland	Ticino	1.13
France	Paris	26.94	Germany	Karlsruhe	1.13
Germany	Stuttgart	24.88	Finland	Uusimaa (Helsinki)	1.09
Germany	Düsseldorf	24.41	Austria	Wiener Umland/Nordteil	1.06
United Kingdom	Surrey, East and West Sussex	23.00	Finland	Pirkanmaa (Nokia)	1.04
France	Hauts-de-Seine	17.82	Sweden	Västmanland	0.97
Sweden	Stockholm	15.88	Germany	Oberpfalz	0.90
Germany	Freiburg	15.00	Austria	West- und Südsteiermark	0.88
Switzerland	Espace Mittelland	14.82	Austria	Linz-Wels	0.88
Switzerland	Eastern Switzerland	14.29	France	Paris	0.81
United Kingdom	Berkshire, Buckinghamshire and Oxforshire	13.94	France	Hauts-de-Seine	0.79
Germany	Tübingen	13.00	Italy	Lodi	0.78
United Kingdom	East Anglia (Cambridgeshire)	13.00	Switzerland	Espace Mittelland	0.78
United Kingdom	Bedfordshire and Hertfordshire	12.35	Austria	Bludenz-Bregenzer Wald	0.77
United Kingdom	Inner London	12.29	Sweden	Skane	0.76

Note: In parenthesis can be found the most(s) known city of each region, if any. Column (i) is the counts of inflows as an average for the seventeen years. Column (ii) is the counts of inflows over thousands of (corrected) HRST, as an average for the seventeen years.

Table 3. Global spatial autocorrelation tests (Moran's test). Average 1990-2006

	W1	W2	W3	W4
Moran's I (z-statistic)	22.16	27.46	29.71	34.591
p-value	*0.000*	*0.000*	*0.000*	*0.000*
Geary's c (z-statistic)	-15.02	-16.38	-15.86	-15.09
p-value	*0.000*	*0.000*	*0.000*	*0.000*

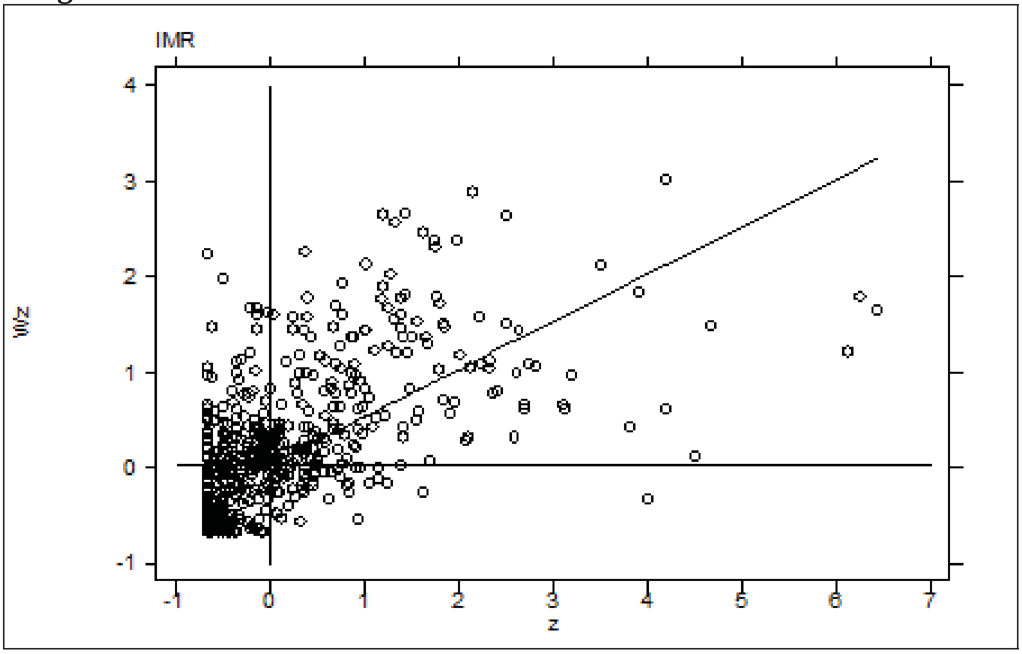
Notes: W1: main matrix, distance-based matrix with 5 neighbours; W2: distance-based matrix with 10 neighbours; W3: distance-based matrix with 15 neighbours; W4: distance-based matrix with 25 neighbours. All matrices are row-standardized.

Table 4. Spatial and a-spatial concentration of movements

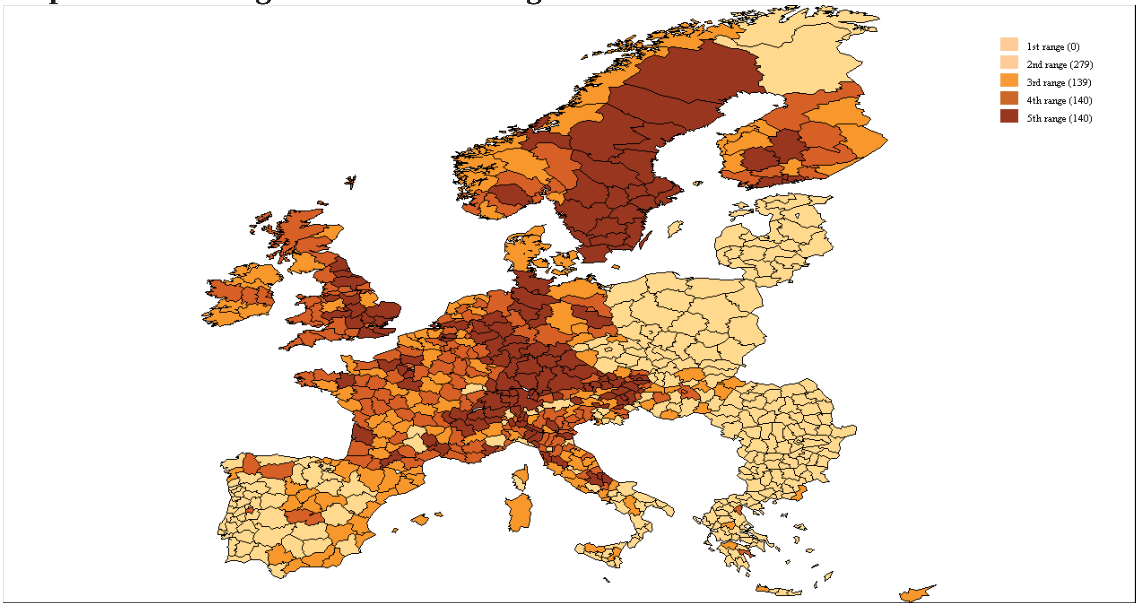
10 nearest neighbours	5 nearest neighbours	National Movements
44.33%	30.79%	76.18%
Inflows Gini Index		Outflows Gini Index
0.83		0.84
Inflows Cumulative Sum (20/698)	Outflows Cumulative Sum (20/698)	Regions in both rankings
41%	43%	17/20

Notes: The first row shows, respectively, the percentage of counts of inflows coming from the 10 nearest regions over the overall counts of inflows, the percentage counts of inflows coming from the 5 nearest regions over the overall counts of inflows, and the percentage counts of inflows coming from the same country. The second row shows the Gini Index for the inflows and outflows measures, both pointing at the existence of a high level of concentration of these variables (for the Gini index, 0 means perfect equality and 1 means perfect inequality). Finally, the last row computes the percentage of the cumulative sum for those 20 regions with more inflows and more outflows during the whole period, and the number of regions in both rankings.

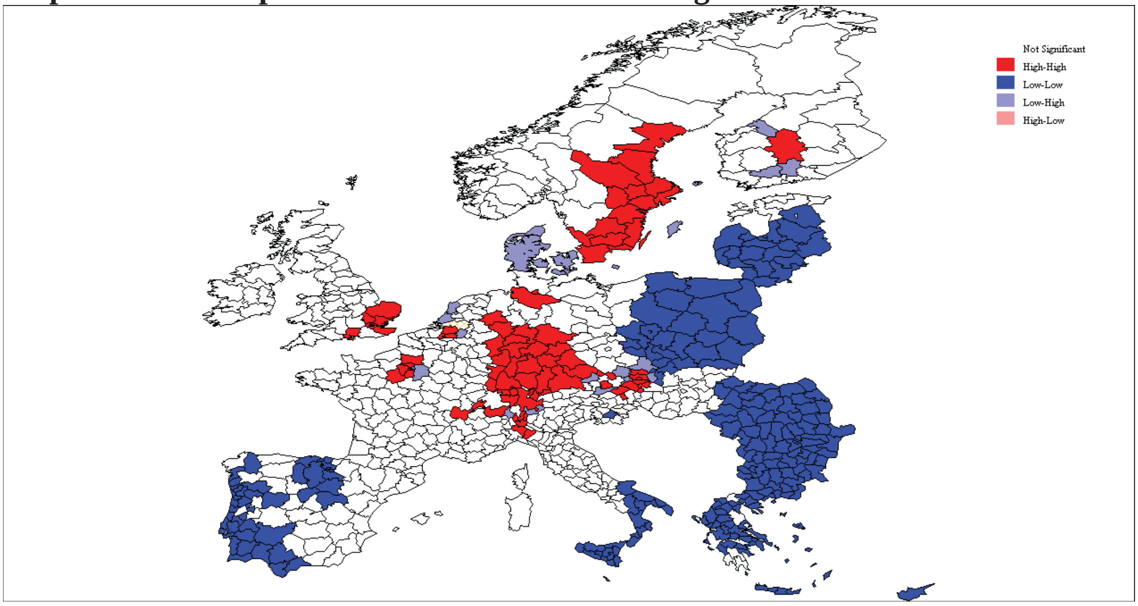
Figure 1. Moran Scatterplot of the IMR, average 1990-2006, 5 nearest neighbours



Map 1. Inward Migration Rate. Average 1990-2006



Map 4. Cluster map. IMR 1990-2006. 5 nearest neighbours



Appendix 1

List of countries (and number of regions in each one) considered in our sample
Austria –AT- (35), Belgium –BE- (11), Bulgaria –BG- (28), Switzerland –CH- (7), Cyprus –CY- (1), Czech Republic –CZ- (14), Germany –DE- (29), Denmark –DK- (1), Estonia –EE- (5), Spain –ES- (48), Finland –FI- (21), France –FR- (96), Greece –GR- (51), Hungary –HU- (20), Ireland –IE- (8), Italy –IT- (100), Lithuania –LT- (10), Luxemburg –LU- (1), Latvia –LV- (6), Malta –MT- (2), the Netherlands –NL- (12), Norway –NO- (19), Poland –PL- (16), Portugal –PT- (28), Romania –RO- (42), Sweden –SE- (21), Slovenia –SI- (12), Slovak Republic –SK- (8), United Kingdom –UK- (37).

Appendix 2

This section describes the two algorithms used to identify inventors using the patent documents. Our approach is based on up to date literature (Trajtenberg et al., 2006; Kim et al., 2006; Raffo and Lhuillery, 2007; Thoma and Torrisi, 2007; and Lissoni et al., 2008). Those authors have tried to contribute to the correct identification of unique inventors using basically their names, several patent characteristics, and different ad-hoc heuristics, in what they called "the Names Game" (Trajtenberg et al, 2006; Raffo and Lhuillery, 2007). Our suggestions within this appendix strongly feed from this former literature, and try to contribute to enrich it at the same time.

The name matching algorithm

Name matching algorithms are designed to solve spelling problems like the ones described in section 3.1. Actually, name variation takes many forms. As reviewed in the literature (Branting, 2003; Snae, 2007) the sources of mistakes might refer to character variations, including capitalisation (Trippl *versus* trippl), punctuation (López Bazo *versus* López-Bazo), spacing (ERNESTMIGUELEZ *versus* ERNEST MIGUELEZ), or qualifiers (Rosina Moreno *versus* Prof. Dr. Rosina Moreno). It might refer to spelling variation, including insertion (McCann *versus* MacCann), omission (Iammarino *versus* Iamarino), substitution (Maier *versus* Mayer), or transposition (Fingelton *versus* Fingleton). And finally it might refer to phonetic variations (Cooper in English would be spelled Cuper in German).

A name matching system must deal not only with spelling and phonetic concerns, but also with cultural aspects (Snae, 2007). For instance, there exist spelling analysis-based algorithms (like the Guth and Levenshtein alogarithms), based on sequences and character strings. There are also phonetics-based algorithms (like Soundex, Metaphone or Phonex), and some composite (ISG) or hybrid (LIG) examples.

Given the features of our dataset (with a predominance of English and German-origin names), phonetic algorithms seem to be the most suitable. Among them, the Soundex algorithm is one of the most widely used. Although it was initially designed for English names, it has since been extended to other languages. It is the name matching algorithm used in Trajtenberg et al. (2006) and Kim et al. (2006) as well, and, as the authors recognise, the algorithm is quite reliable except for Asian names (whose presence in our dataset, we suspect, will be nominal, and stronger in datasets derived from the USPTO).

Soundex was developed in the 1930s by the US Census Bureau and used to list all the individuals in the US census records starting from 1880. It encodes the first letter of each string followed by a number of digits (usually three) representing the phonetic categories of the next consonants. The vowels and the consonants H, W and Y are ignored, and adjacent letters from the same category are encoded with a single digit. The coding is as follows: (1) for B, P, F, V; (2) for C, S, K, G, J, Q, X, Z; (3) for D, T; (4) for L; (5) for M, N; (6) for R. The 0 is used when the string has finished before using the whole number of digits.

Before using the algorithm we have (1) written the field with the inventor's name and surname in capital letters; (2) dropped punctuation symbols, slashes, apostrophes, bar marks, numbers, commas, periods, spaces between words, and the like; and (3) separated the name and surname in different fields. Afterwards, we encode the surname with the first letter of the string and six additional digits, and encode the name of the inventor using the initial letter and three additional digits. Combining the surname code and the name code we build what Trajtenberg et al. (2006) called the *p-sets* (potentially the same inventor)[23]. Each different *p-set* is therefore identified as a different, unique inventor. In this way, we encode, with the same Soundex code, the strings that differ slightly but actually belong to the same person (like those of the former examples). Notwithstanding, this procedure might induce another important error: that is, when two records which actually belong to different inventors are matched under the same *p-set*. Thus, "François De La Poype", "Frank De Wolf", and "Francis Dell'Ova" will share the same *p-set* code, D410000-F652 – although obviously they are not the same person. Of course, Soundex will encode two researchers named "John Smith" with the same code, even though they do not belong to the same person. To solve these two types of error, we need to go on to the second stage of our methodology.

The splitting algorithm

Here, we describe our method for distinguishing whether each pair of records encoded under the same *p-set* belongs to the same inventor or not. This is actually the most difficult task in the project. To do so, we will compare every pair of records contained within each *p-set* according to several features of each record. Following Trajtenberg et al.'s (2006) suggestions, we will give different scores to each comparison made between each pair of records within the same *p-set*, which we will call test, although we will adapt the number and types of tests, and the scores, to our particularities. This is not the procedure followed in Kim et al. (2006), since those authors preferred to decide whether two records belong to the same inventor when several conditions are reached – that is, without giving scores. As these authors argue, in this way they avoid the arbitrary decision to give a predetermined score to each test. However, by not doing this an important aspect stressed by Trajtenberg et al. (2006) is missed. Consider two records within the same *p-set* with the name, say, "John Smith", both working for "Phillips" and living in London. Now, consider another pair of records within another *p-set*, this time with the name "Camilla Rönnqvist", working for both "Pliva Hrvatska D.D.O." and living in the Croatian peninsula of Pula. Which pair is more likely to belong to the same inventor? Obviously, the second one. Thus, following Trajtenberg et al.'s (2006) suggestions, we will weight each of the scores given to every single pair-wise comparison of records. Unlike Trajtenberg et al., however, we will divide the distribution of the variables used to do the tests into eight frequencies, and will weight each test with the standardised inverse of these frequencies. Thus, the scoring scheme let us to weight the scores given to each pair-wise comparison depending on the frequency of occurrence of a given feature of the test performed. So, these "two Croatian inventors" of the example sharing the same name will be said to be the same person more likely than those inventors whose names and associates characteristics are quite common within our dataset. Moreover, the use of different frequencies to weight the scores will let to smooth the jumps done between frequency and rareness of the names and characteristics[24].

[23] In this study, the full name and surname are encoded using Soundex with the initial letter of both followed by six additional digits. Meanwhile, our *p-sets* only take account of the first letter of the name and three additional digits – aside from the surname with six additional digits. This more relaxed definition of *p-set* obliges us to check all the movers identified in our dataset one by one, in order to avoid incorrect matching throughout our procedure – that is, to make sure that an inventor with a different name is not considered the same person.

[24] Trajtenberg et al. (2006) give different scores if the name is rare or frequent, if the city is large or small, if the patent class is large or not, and so on, so deciding a threshold up to which the score given to a single comparison could be very different if it is situated in the upper or lower part of the distribution. In our

We now turn to the detailed description of our splitting methodology. For each pair of records within every *p-set*, we run five comparisons (tests): we compare the Soundex-code (with the initial letter and six additional digits) of the name of each one, the NUTS3 region from where the patent application is made, the technological class to which each patent is associated[25], the Soundex-code (the initial letter and six additional digits) of the name of the applicant, and finally we check whether these two records cite one another[26] – as noted in the literature, the probability of self-citation is higher than the probability of citing someone else. The election of the tests is arbitrary, but most of them have been already used in the literature (Trajtenberg et al., 2006; Lissoni et al., 2008). In any case, we have run as much tests as the raw data permit, squeezing all the information linked to each patent in order to optimise the identification procedure. Next, we give a score to each of these five tests, which will be properly weighted. Further, in contrast to the dataset used in Trajtenberg et al. (2006), each patent belongs to a different number of technological classes and associated applicants – usually (but not always) more than one. To deal with this particularity, we will give a score for each matching within each test, weighted by the possible matches which can be made in each test. Thus, for instance, for a given comparison between two records, say A and B with two and three technological classes respectively, we give a single score for every pair of matches (six in this case) which will be weighted by the number of positive matches over six. The same applies for the applicants' Soundex-code. The following table displays the scores used for each criterion.

[Insert Table A.1 about here]

Once each test has been performed and properly weighted, we add up total scores for every pair-wise comparison within each *p-set* and weight this result again with the inverse of the frequency of each *p-set* itself[27]. Afterwards, we compare it with a pre-determined numerical threshold – up to which we decide if two records belong to the same inventor or not[28] – set at 99.

After doing this, transitivity must be imposed also for logical reasons. It is done in the sense that, although two inventors, say A and C, are not considered to be the same person – i.e., their total score derived from their multiple comparisons does not reach

case, we use several intervals of rareness and frequency instead of just one threshold for each test (specifically, we use eight intervals). Thus, we weight those tests related to the applicant's name, the region, the technological class of the patent, and the Soundex-code of the inventor's name. The self-citation test is not weighted. Moreover, the results of the weighting process are again weighted depending on whether the built *p-set* is rare or frequent (in this case, we have divided our distribution into four intervals).

[25] We have used the IPC classification to do this, establishing that a pair of patents belong to the same technological class if they share the same two first digits of this encoded classification.

[26] Data for citations is gathered from the OECD citations database.

[27] This is done again because of the fact that two very common Soundex-codes of the surname-name strings are less likely to be the same person than other rare ones.

[28] The values of the scores and thresholds are assigned through a trial and error mechanism, comparing the results of every trial made with a subsample of inventors. Specifically, we use the Spanish inventors' subsample – since we are more familiar both with the Spanish innovation system and with Spanish names, surnames, regions, and the like – to make these comparisons, by assessing the goodness of the splitting algorithm record by record. Thus, we depart from an arbitrary scoring scheme, and we run iteratively our algorithm doing marginal changes on the scores and the threshold each time. When a marginal change improves the goodness-of-fit of the matching process, we keep this change. We do it until we are not able to improve the performance of the algorithm.

the minimum threshold – we impose that they are the same person if A is the same person as B and B is the same as C.

Table A.1.1 Scores of each test

Test	Score
Self-citation	140
Same technological class	120
Same applicant Soundex-code	120
Same inventor’s name Soundex-code (with 6 digits)	110
Same NUTS-3 region	90