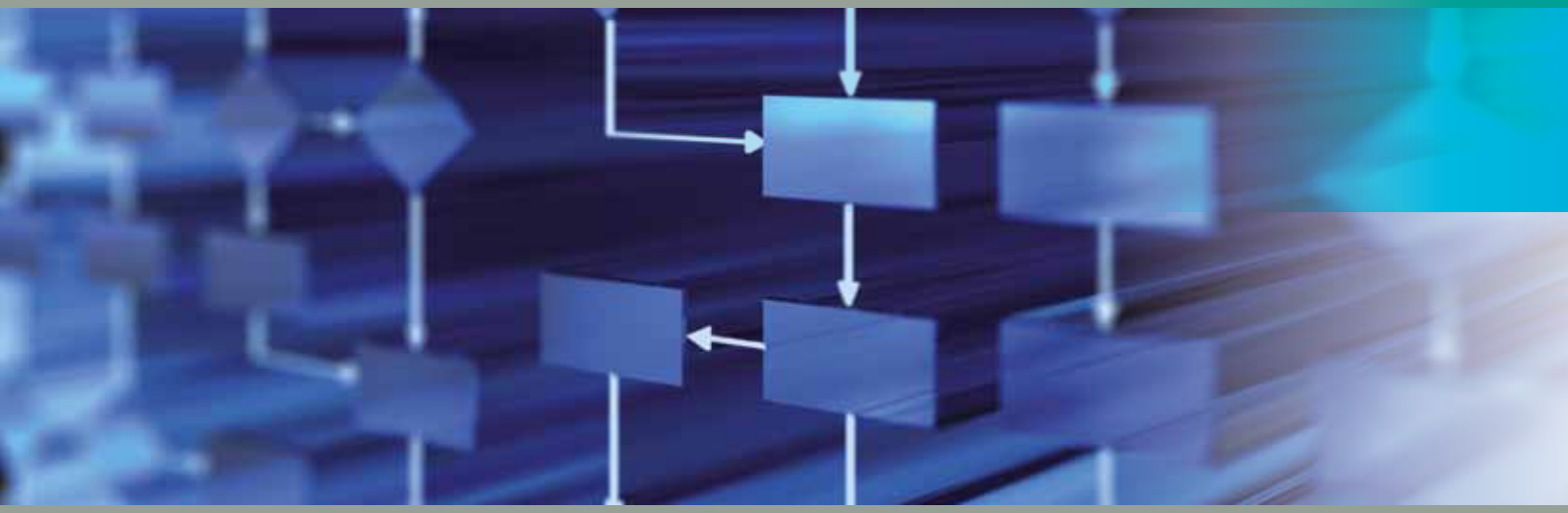


WIPO Economics & Statistics Series

January

2014

Economic Research Working Paper No. 17

Inventor Data for Research on Migration and Innovation: A Survey and a Pilot

Stefano Breschi
Francesco Lissoni
Gianluca Tarasconi

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Inventor Data for Research on Migration & Innovation: A Survey and a Pilot

Stefano Breschi[1], Francesco Lissoni[2], Gianluca Tarasconi[3]

Abstract

This paper discusses the existing literature on migration and innovation, with special emphasis on empirical studies based on patent and inventor data. Other sources of micro-data are examined, too, for comparative purposes. A pilot database, based on patent filings at the European Patent Office is presented. It contains information on individual inventors, including their country of residence and of origin. Preliminary evidence suggests that immigrant inventors contribute to innovation not only in the US, but also in selected European countries, where they often rank among the most productive individuals. Data on returnee inventors to selected countries of origin suggest the phenomenon to be of limited scale, and highly subject to errors of measurement.

Keywords: immigration, innovation, inventor data, patent data

JEL codes: F22, O15, O31

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the views of the World Intellectual Property Organization or its member states.

Acknowledgments

This paper benefitted from comments by the participants to the WIPO Experts Meeting on "Intellectual Property, the International Mobility of Knowledge Workers and the Brain Drain" (April 2013, Geneva), at which it was first presented. In particular, Bronwyn Hall produced an extensive review and pointed out several issues in the first version (which we could remedy only in part). Ernest Miguelez made available to us the WIPO-PCT data we used for comparison with our Ethnic-INV ones. Curt Baginski assisted us in the exploration of IBM-GNR©'s potential and technical details.

[1] CRIOS – Università Bocconi, Milan
[2] GREThA – Université Montesquieu, Bordeaux IV; CRIOS – Università Bocconi, Milan; e-mail: francesco.lissoni@u-bordeaux4.fr
[3] CRIOS – Università Bocconi, Milan

1. Introduction

Migration and innovation are two phenomena whose ties date back a long time in history, well before the emergence of professional science and engineering (S&E). David's (1993) historical excursus on the birth of modern intellectual property rights (IPRs) reminds us that the latter originate from the *privilegi* granted by Italian states of the 14th-15th century to foreign craftsmen, in order to lure them away from their home countries (or rival states) and inject new techniques in the local industry. In the same years, Tudor England was engaging actively in «the negotiation … of secret agreements designed to attract skilled foreign artisans into [the Crown's] service. German armorers, Italian shipwrights and glass-makers, French ironworkers were enticed to cross the Channel in this fashion» [David, 1993; pp. 47-48]. Coming to more recent times, Moser et al. (2011) show how Jewish scientists seeking refuge from Nazi Germany were responsible for a significant growth in US patenting activity in several fields, while Halary (1994) discusses the globalisation of the academic world over the course of the second half of the 20th century.

What makes the study of migration and innovation a hot research topic nowadays is the steady increase in the global flows of scientists and engineers (S&Es) observed over the past 20 years, both in absolute terms and as a percentage of total migration flows (Freeman, 2010; Docquier and Rapoport, 2012). These flows have been fed by an increasing number of countries, most notably China, India, and the former soviet-block countries. This raises a number of questions on the role these migrants play in the innovation process in both their destination and origin countries.

The most common questions asked with reference to destination countries, most notably the United States, can be summarized as follows: Are foreign S&Es complements or substitutes of local ones? In other words, do they increase their destination country's innovation potential, or do they simply displace the local workforce (Borjas, 2004; Chellaraj et al., 2008; Hunt and Gauthier-Loiselle, 2010)? Are destination countries increasingly dependent on the immigration of S&Es (including graduate students) to maintain their present technological leadership? Does such dependence require the implementation of dedicated immigration policies (Chaloff and Lemaitre , 2009)?

As for origin countries, the key research questions concern the extent of their loss of human capital ("brain drain") and the nature and effectiveness of potential compensating mechanisms, such as knowledge spillovers from destination countries or the contribution to local innovation by returnee S&Es and entrepreneurs (Agrawal et al., 2011; Kerr, 2008). In this respect, some debate exists on the role of intellectual property, most notably in the aftermath of many origin countries' subscription of TRIPs, the Trade Related Intellectual Property Agreements that come with the adhesion to the World Trade Organization (Fink and Maskus, 2005).

While rich in questions, this emerging literature is still poor in answers. One important limitation concerns the empirical side, and the lack of extensive and detailed data for micro-econometric analysis. Another important limitation concerns its almost exclusive focus on one destination country, the US, and the origin countries which have recently become its top providers of foreign talents, namely India, China, and other East Asian countries. US-centrism is not peculiar to this field of studies, but here it bears the additional disadvantage of reducing a multi-polar phenomenon, one in which several countries act both as source and destination of migration flows, to a set of binary relationships between the US and a limited set of origin countries.

In this paper we explore the potential of patent data as a source of information capable to address both limitations. We do so both by reviewing the existing literature and by experimenting with a "pilot" database (to which we will refer with the working name of *Ethnic-Inv*). The database is built following Kerr's (2007) seminal methodology, which consists in combining inventor-based data from the USPTO with extensive information on the ethnic origin of names and surnames. Differently from Kerr, however, we try explicitly to focus on Europe, both by making use of European Patent Office data and by exploiting a different names-and-surname database, one containing finer-grained information on countries of origin. At this stage, our aim is mainly methodological and consists in:

i. discussing the main benefits and drawbacks of analysing the migration and innovation phenomenon through the lens of ethnic inventor data, both in general and with reference to our specific approach;

ii. providing a first illustration of the importance of ethnic inventors in Europe, so as to attract to the phenomenon the attention it deserves;
iii. discussing the specificities of inventors' migration to Europe, especially with reference to countries of origin and their role in destination countries.

In what follows, we first survey the existing literature on migration and innovation, with an exclusive focus on quantitative studies (section 2). We then discuss the key methodological problems one faces when using inventor-based patent databases, with special reference to migration issues (section 3). Finally, we provide some descriptive statistics and simple econometric exercise based upon the *Ethnic-Inv* database in order to discuss the latter's reliability and potential (section 4). Section 5 concludes.

2. Quantitative studies of migration and innovation

In this section we review existing quantitative studies that address, either directly or indirectly, the relationship between migration and innovation. We first consider general studies on the growing phenomenon of highly skilled (tertiary educated) migration, which makes use of macro information from national census and labour force survey data. We compare them to reports (by OECD and other organisations) on the international mobility of doctoral holders and academic scientists, who represent the most mobile category among the highly skilled. Finally, we move on to examine the methodology and results of micro-econometric studies based upon *ad hoc* collections of data on scientists, college graduates, and inventors.

2.1 Highly skilled and scientific migration

Studies on highly skilled (*hs*) migration belong to a long standing tradition of research on migration and development (for a survey of economic studies, see Docquier and Rapoport, 2012; for a cross-disciplinary survey, see de Haas, 2010). The core phenomenon under study is that of "brain drain", namely the emigration of highly educated men and women from less developed countries, and the associated risk of possible depletion of local human capital.

Recent efforts aimed at quantifying the extent of the phenomenon have produced data and statistical evidence of great interest also for studies more directly focussed on innovation. In this respect, two contributions stand out: the dataset produced by Docquier and Marfouk (2006), to which we will refer as DM06; and DIOC, the Database on Immigrants in OECD countries, produced by the OECD.[4]

The two datasets have been collected with similar methodologies, and DIOC can be considered a more extensive and up-to-date version of DM06. Both of them contain figures for the stock of foreign born residents in OECD countries in given years (1990 and 2000 for DM06; 2000 and 2005/6 for DIOC)[5], disaggregated by migrants' origin country, age class, gender, and level of educational attainment. These figures come either from census data or (in a minority of cases, which include however a few big countries) labour force surveys, with the latter generally returning lower figures. Data on the number of residents in origin countries (around 195 for DM06, 230 in DIOC) are also included and broken down by age, gender, and education.

Based either on DMO06 or DIOC one can build matrixes having destination countries on rows (*i*), origin countries on columns (*j*), and $foreign_{ij}$ in cells, the latter being the stock of foreign born residents in country *i* coming from country *j*. To the extent that the OECD includes all major destination countries, summing up over destination countries provides reasonable approximations of total emigration from any single origin country ($foreign_j = \sum_i foreign_{ij}$).[6] In a similar fashion, one can compute the total stock of foreign born residents in any destination country *i* ($foreign_i = \sum_j foreign_{ij}$). We indicate stocks of highly skilled migrants with *hs_foreign*.

[4] DM06 comes in various releases, all available from the author's website (http://perso.uclouvain.be/frederic.docquier/oxlight.htm). The most recent release is used in Docquier et al. (2009). For DIOC methodology see Widmaier and Dumont (2011), while data are downloadable from: http://www.oecd.org/els/mig/dioc.htm.
[5] Widmaier and Dumont (2011) warn that DIOC 2000 is not entirely comparable to DIOC 2005/06. DIOC 2000 also comes in an extended version, which includes, among destination countries, also around 70 non-OECD countries.
[6] In DIOC 2005/06 no data are available for emigration to South Korea, as well as Estonia, Hungary, Iceland, Slovenia, Slovakia, and Turkey.

On such basis, the brain drain rate for each origin country is calculated as:

$BrainDrain_j = hs\_foreign_j/(hs\_foreign_j+hs\_origin_j)$,

which is the ratio of all *hs* emigrants from the same origin country over the sum of such emigrants and the origin country' residents with the same educational level (hs_origin_j). In a similar fashion, one can calculate any destination country's reliance on *hs* immigrants (let's call it "brain intake") as:

$BrainIntake_i = hs\_foreign_i/hs\_residents_i$,

where $hs\_residents_i$ are the total *hs* residents in the destination country.[7]

Although extremely valuable, these type of data are not entirely exempt from limitations. We point out only three of them, which serve as an introduction to our discussion of other data sources (for more details, see Docquier et al., 2009). First, there are some difficulties in defining foreign born individuals. In principle, these should be those whose country of birth differs from that of residence. However, the concept of foreign born is not homogeneous across all countries, some of which restrict it to foreign citizens born abroad. At the same time, data sources for a few countries report data only for foreign citizens, who are generally fewer than foreign born, but may include individuals born in the country of residence from foreign born parents.[8] Besides limiting the cross-country comparability of data, uncertainty on the definition of "foreign born" individuals makes it difficult to identify *hs* returnees to origin country. This category of migrants is an all-too-important one for innovation, to the extent that they may carry with them key knowledge assets acquired in their countries of destination. Second, information is not available on where foreign born individuals received their tertiary education, which makes it difficult to distinguish *hs* emigrants from both individuals who left their origin country at a young age and received their education abroad, and international students who decide to stay in the country where they obtain their degree. In this respect, information on age of entry in the destination country may help, but this is not always available or reliable. Third, migrants are assigned to the *hs* category on the basis of their educational attainments (tertiary education), but it is often the case that they accept jobs for which they are overqualified. This is more likely for those who completed their education abroad and meet difficulties in having their academic title officially validated or properly appreciated by employers. For what concerns the innovation process, it may be the case that foreign scientists and engineers, while figuring in *hs* immigration statistics, are neither employed in R&D nor contribute to other innovation input such as design, production management and the likes.

As for substantive information one can get from datasets on *hs* migration, a joint reading of empirical analyses based upon DM06 and DIOC data suggests that, at least since the 1990s, *hs* migration has grown in stock (from around 13 million units worldwide in 1990, according to DM06, to 26 millions 2005/06, according to DIOC) and as a share of total migration (from 30% in 1990 to almost 40% in 2005/06). Besides, migration rates for the tertiary educated are higher than those for the non-tertiary educated (around 5.5% on average vs. 1.3%, according to Docquier and Rapoport, 2012).

For what concerns the stock of *hs* migrants, origin countries are inevitably the largest ones, especially those whose languages are internationally diffused, regardless of their development level. Among the top 30 origin countries worldwide we find several European ones, starting with the UK (always top of the list, with over 1 million *hs* emigrants, regardless the database considered), followed by Germany (almost one million), Poland, Italy, France, Russia, the Netherlands, Ukraine, Romania, Greece, and Serbia. Notice that the US figures in this top30 list too, as well as several large emerging economies such as China, India, and Vietnam. This suggests that, when studying the migrants' contribution to innovation, a very important role may be played by emigrants from industrialized, R&D intensive countries. It also suggests that Europe contributes decisively to the *hs* migration stock worldwide, both with its more advanced countries and with its less advanced ones. In a few cases, this translates into rather high brain drain rates (16% in Poland, and over 11% in the UK in 2005/06) or at least in above-world-average ones (7.2% for Germany, around 6% for both Italy and France).

[7] Notice that the only residents (both in origin and in destination countries) one should consider in these calculation should be those aged over 25, who are those old enough to be tertiary educated. This rule is followed in the various publications by Docquier and co-authors based upon DM06, but not by Widmaier and Dumont (2011), who makes use of DIOC.

[8] Docquier and Marzouk (2006) point out that country of birth is a better indicator of migrant status than citizenship, as it is time invariant, while citizenship may change due to naturalization.

At the same time, European countries within the OECD have the lowest brain intake rates. This is mainly due to their immigration policies, which usually do not select by skill and are dominated by family reunions or, in several countries, humanitarian reasons.[9] The net result of *hs* emigration and immigration is nonetheless positive for several large European countries (including the UK, France, and Germany), with the main exception of Italy, which suffers of a net loss. In terms of innovation studies, these figures suggest that OECD European countries may face difficulties due to the emigration of their S&Es, which they may find difficult to compensate with immigrant S&Es of the same quality.

As for non-OECD European countries, most of them suffer both of net losses, and high brain drain rates (Bulgaria, 11%; Hungary, 9%; Romania, 18%; all figures from DIOC 2005/06), the main exception being Russia, with only 1%. Notice that non-European largest contributors to *hs* migration, such as India and China, do not suffer of high brain drain rates (respectively 4% and 2%), due to the size of their population and their generally high (and increasing) education level.

An important category of *hs* migrants are those holding doctoral degrees, especially in scientific and technical fields. DM06 and DIOC do not include separate figures for them, but some information can be obtained from the survey on the Careers of Doctorate Holders (CDH), conducted jointly by the OECD and UNESCO in 2007 and covering 25 OECD countries (plus a seven-country pilot project in 2003; see Auriol, 2007 and 2010). Although not explicitly targeted at migration, and even less so at innovation, the CDH dataset contains useful, complementary information to *hs* migration statistics. First, we learn that "the labour market of doctorate holders is … more internationalized than that of other tertiary-level graduates" [Auriol, 2010; p.19] and that, in Europe, 15% to 30% of native doctorate holders can be considered as returnees having lived, in the 10 years before the survey, in at least one different country. Auriol (2007) shows that, in 2003, around 13% of doctorate holders in Germany were foreign born, almost double the DM06 figures for all tertiary educated (7%, 2000 data; 42% vs. 32% in Switzerland, and 26% vs. 11% in the US). Second, as far as Europe is concerned, most of the mobility takes place within the continent (over 60% of total mobility). Last, France, Germany and the UK emerge as the important destination countries along with the US, the latter being however the top destination for all doctorate holders from East Asia and India (who make 57% of foreign doctorate holders in the US, as opposed to only 27% of Europeans).

When it comes to contributing more directly to the migration and innovation topic, however, the CDH data suffer of several drawbacks. First, doctoral graduates represent only from 1% to 3% of all tertiary graduates in most countries (the maximum is 4.5% in Switzerland). Second, industrial researchers are less likely to hold a doctorate degree than academic ones. Therefore, most doctorate holders end up contributing to innovation in a decisive, but in a rather indirect way. In particular, they tend to undertake academic careers and contribute mainly to education and scientific advancement via publication of their research results.[10]

In this respect, data from the *GlobSci* publication-based survey confirm the exceptional degree of globalization achieved by the academic labour market (Franzoni et al., 2012; Scellato et al., 2012). The *GlobSci* survey concerns authors of papers published in high quality scientific journals in 2009, in the fields of biology, chemistry, environmental science, and materials, active in the 16 top countries for authors' affiliation (70% of published articles, the only large country excluded from the survey being China). Early results show that foreign-born authors (defined as those who entered the country of affiliation after the 18th year of age) are more than half of all authors in Switzerland (57%) and around a third in the US (38%) and in between a third and a fifth in several European countries (38% in Sweden, 33% in the UK, 28% in the Netherlands, 22% in Denmark, 23% in Germany, 18% in Belgium, and 17% in France). The only top countries with limited contributions of foreign-born scientists are Spain (7%), Japan (5%), and Italy (3%).

[9] Selective immigration policies are those that target specifically hs migrants. They mainly consists in reserving quotas for university graduates or specific professional figures and, in a few cases, for allowing hs perspective immigrants to enter the country even before having found an occupation. The most notable cases are those of Australia, New Zealand, and Canada, whose immigration flows now are largely dominated by selective policies, followed by the US, with its H1-B visas. Germany has also recently implemented a fast track visa granting procedure for IT specialists, which however has not met the expected success. For a comprehensive discussion, see Chaloff and Lemaitre (2009).

[10] These limitations affect even more severely another potential source of information on migration and innovation, namely the MORE survey on the mobility of European researchers (MORE, 2010). Here the main focus is on academic researchers (data for industrial researchers are based on a non representative sample) and no questions are asked with a direct relevance for the innovation process.

GlobSci also confirms that migration in Europe is mainly intra-continental and driven by proximity and language effects (for example, Italians are the principal foreign group in bordering France, while Germans make almost 40% of foreign scientists in Switzerland and are the top group also in Belgium, Denmark, and the Netherlands). On the contrary, the US is confirmed to be the main attractor of Chinese and Indian nationals (which are the most represented among foreign-born authors, with shares respectively of 17% and 12%).

Another interesting piece of evidence from the *GlobSci* survey concerns the foreign-born scientists' propensity to engage in collaboration with colleagues from home countries and fellow expatriates (both of them very high, with 40% of collaborations being with home country and as much with fellow expatriates, in the same or different affiliation countries). This suggests that, at least within academic science, an effective "ethnic" network is at work, with the potential of delivering knowledge spillovers to origin countries.

2.2 Migrants' impact on innovation

In very recent years, various attempts have been made to exploit archival data for retrieving information on the impact of *hs* immigrant on their destination country, almost all of them centred on the US. As for the impact of emigration on innovation in origin countries, this has been studied almost exclusively on the basis of patent and inventor data, yet again with an almost exclusive US focus. We examine the two streams of literature separately.

2.2.1 Migrants' contribution to innovation in destination countries

The United States research system (both industrial and academic) has been an historical destination for foreign born scientists and engineers. Universities have been playing a key role in encouraging the inflow of foreign students and postdocs, the former now making around 45% of graduate students enrolled in S&E programmes, and around 60% of postdocs (2006 data, as reported by Black and Stephan, 2010). A debate is ongoing in both the US academic and non-academic press on the extent of foreign researchers' contribution to scientific advancement and innovation, and the related visa policies to undertake. Negative concerns have been expressed by migration scholars already sceptical on immigrants' general contribution to US economic growth, such as Borjas (2009). These concerns address mainly the possibility of local S&E students and workers being crowded out by the inflow of foreign competitors. Evidence in favour of this thesis is the dramatic drop of US citizens' enrolment in S&E university programmes. In addition, it has been noticed that more recent cohorts of foreign-born academic researchers in the US tend to concentrate in more peripheral and less productive universities and departments, which do not offer attractive career prospects to native students (chs.7-8 in Stephan, 2012; Su, 2012). And yet, such evidence could simply prove the existence of a natural division of labour, with US citizens entering professions for which mastering the local language and culture, as well as having a larger social capital, matters more than having acquired specific scientific or technical skills. Much research has therefore focussed on testing the hypothesis that immigrant S&Es may self-select on the basis of superior skills and contribute more than natives with similar education levels or jobs to innovation (this would imply that immigrants and natives are not perfect substitutes, and that the former do not merely displace the latter by accepting lower wages).

A pioneer empirical effort in this direction is that by Stephan and Levin (2001), who focus on the presence of foreign-born and foreign-educated among eminent scientists and innovators active in the US in 1980 and 1990. The authors assemble a sample of about 5000 highly productive or distinguished S&Es, which include members of the National Academy of Science (NAS), the National Academy of Engineering (NAE), the authors of highly cited scientific papers (from early ISI databases, now part of the Web of Science published by Thomson), a selection of academic entrepreneurs in the life sciences, and a small number (around 180) of inventors of highly cited USPTO patents. Countries of birth and education of sample members are ascertained by exploiting the biographical information made available by NAS, NAE, and various directories of scientific and medical societies. The share of foreign born and that of foreign educated in each category of eminent scientists and innovators is then compared to US S&E labour force's shares of foreign born and foreign educated, the latter being calculated on the basis of NSCG data (National Survey of College Graduates). Two-tail Chi-square tests prove that in all cases but one the foreign born are over-represented in the eminent scientist and innovator group.

In a few cases, a cohort effect is detected, with foreign born entered in the US before 1945 being particularly productive (this is not the case, however, for top inventors and academic entrepreneurs). Finally, the foreign-educated are found to contribute disproportionately to these results, which suggests both that the US benefit of positive externalities generated by foreign countries and that immigrant S&Es are self-selected on the basis of skills.

Stephan's and Levin's results on the foreign-born's contribution to entrepreneurship are confirmed for more recent years by other surveys, most notably those conducted by Wadhwa et al. (2007a-c) and No and Walsh (2010). The former find that around 25% of all engineering and technology companies established in the U.S. between 1995 and 2005 were founded or co-founded by at least one foreign-born. The percentage increases remarkably in high-tech clusters such as the Silicon Valley (52%) or New York City (44%). These foreign entrepreneurs are mostly found to hold doctoral degrees in S&E, and to be better educated than control groups of natives. At the same time, it is found that most of these entrepreneurs first entered the US as students, and not with the specific purposes of setting up a new company. As for No and Walsh (2010), they survey 1900 US-based inventors of "triadic patents" (patents filed in the US, Japan, and Europe), asking them, among other things, to self-evaluate their inventions' technological impact and economic value. Both measures are found to be higher for inventions by foreign born inventor after controlling for the patents' technology class, the inventors' education level, and a number of characteristics of both the patent applicants and the inventive projects. The immigrants' contribution to patenting has been further investigated by Hunt and Gauthier-Loiselle (2010). The two authors exploit the 2003 edition of the NCSG, which contains a question on the number of patents filed by respondents, starting from 1998. Descriptive statistics show that the foreign born graduates are more likely than the natives to have filed one or several patents. However, this depends chiefly on a composition effect, the foreign-born graduates being more likely to belong to S&E disciplines. Cross-sectional econometric evidence shows that, at the state level, an increase in the number of foreign-born college graduates generate more innovation (measured by patents per head) than an equivalent increase of local graduates. Finally, it is found that any increase in the foreign born share of graduates translates into an equivalent increase in the ratio of college educated workers (foreign born plus natives) over the total workforce, after controlling for changes in the population age structure. This suggests that *hs* immigration adds to, and does not displace the native *hs* workforce. Hunt (2009, 2013) provides interesting extensions of these results. Still based on NSCG 2003 data, Hunt (2009) compares the contribution to innovation of foreign-born and native college graduates, not only in terms of patents, but also in terms of publications (papers in refereed professional journals, conference proceedings, and books). In addition, a distinction is drawn between foreign-born graduates who entered the US with student visas, postdoc visas, and all others (the latter being, most likely, children reunited to their families at an early age, who then shared the same school background of natives). Two results deserve comments. First, the advantage of foreign born over natives with respect to publications is higher than that for patents, and it is not entirely explained by the same composition effect. This suggests that self-selection of highly skilled immigrants is particularly strong when it concerns the academic labour market, so that not all findings on academic scientists can be immediately extended to inventors in non-science based technologies. Second, the only foreign born graduates who hold any advantage over natives are postdocs, which suggests that highly productive foreign S&Es enter the US via the academic labour market, rather than as undergraduate or graduate students (nor they emerge from immigrants' families via the US education system).

As for Hunt (2013), the NSCG data are used to prove that a correlation exists between graduates' patent productivity and wages, so that the author can go on arguing that the latter can be taken as proxy of engineers' innovativeness (most patents coming from engineering graduates). Data from the American Community Survey are then used to prove that:

(i) Foreign-born working as engineers are over-represented among top wage earners, other things being equal
(ii) Foreign-born qualified as engineers, on the contrary, are under-represented

The contrast between (i) and (ii) is explained by the over-qualification (or underemployment) of engineering graduates immigrated from less developed countries (including those arrived at an early age). The latter may face difficulties in getting an engineering job or in reaching managerial positions, being impeded by lack of language skills or social capital. On the contrary, immigrants from richer countries and India are more common among the foreign-born actually working as engineers.

Hunt (2010) finds similar results for computer scientists. One important policy conclusion concerns the viability of selective visa policies, which apparently run the risk of attracting overqualified workers from a subgroup of origin countries.

Chellaraj et al. (2008) make use of a production function approach to estimate the impact of both foreign-born high skilled workers and international students on innovation in the US. The authors propose a time series regression, in which the dependent variable is the number of patents filed at the USPTO by US companies, as a percentage of the US labour force. The regressors of interest (inputs in the innovation production function, with a 5-year lag) are the foreign-born share of graduate students, the highly skilled foreign-born share of total workforce, the foreign-born holders of a doctoral degree in S&E as a percentage of total workforce, and the percentage R&D/workforce ratio. Controls include the depreciated stock of patents filed over 5 years before the focus year. Further exercises alternatively use as dependent variables the patents filed by business companies as opposed to those filed by universities.
The elasticity of patents to the presence of skilled immigrants is found to be positive and significant, and even more so the elasticity with respect to foreign graduate students. This difference can be explained with the composition effect highlighted by Hunt and Gauthier-Loiselle (2010): while highly skilled immigrants comprise many professions, foreign graduate students are concentrated in S&E, and therefore have a much more direct impact on innovation. The same composition effect may explain the superior estimated impact of foreign graduate students with respect to local ones.[11]

A partial exception to the US-centrism of the literature is the study by Ozgen et al. (2011), also based on an innovation production function approach. The study concern 170 NUTS2 regions in Europe, observed over two periods (late 1990s and early 2000s). The study makes no use of data on classes of highly skilled immigrants directly relevant for innovation (scientists and engineers, inventors, or graduate students), but only of regional figures on the share of foreign-born residents, the average skill of immigrants (proxied by the income level of origin countries), and the heterogeneity of countries of origin, plus controls. In this sense, the study is closer to the tradition of studies on the value of cultural diversity on innovation and growth (Ottaviano and Peri, 2006; Bellini et al., 2013), than to a direct evaluation of migration's impact on innovation. In a similar vein, Niebuhr (2010) focus on the foreign-born contribution to cultural diversity in R&D employment, as opposed to total employment, as well as in other professions classified as highly skilled. She then investigates the effect of cultural diversity on the patenting rate of 95 German regions over two years (1995 and 1997), finding a positive association.

2.2.2 Migrants' contribution to innovation in origin countries

A longstanding tradition of emigration studies has consisted in evaluating the type and extent of positive returns from emigration for origin countries. Early studies placed special emphasis on emigrants' remittances and the role they might play in capital formation in less favoured countries and regions. More recently, due to the increasing importance of *hs* migration, more attention has been paid to emigrants' contribution to knowledge formation and innovation. This may come in two, none mutually exclusive forms, namely:

(i) *"Ethnic-bound" knowledge spillovers.* Emigrant scientists and engineers may retain social contacts with former fellow students or educational institutions in their home countries, and transmit them the scientific and technical skills they have acquired abroad (either on a friendly or contractual basis, through visiting professor programmes, research collaborations, or firm consultancy)
(ii) *Returnees' direct contribution.* Emigrant scientists and engineers who have worked as academic or industrial researchers, may decide to move back to their origin countries and continue their activities over there. In the case of entrepreneurs, they may keep base in the destination

[11] In a related paper, Stuen et al. (2012) examine the impact of foreign-born (by origin country) vs. native students on the scientific publications (number and citations received) by 2300 US university departments. Panel data on publications come from the ISI Web of Science in years 1973-2001, while data on students come from the Survey of Earned Doctorates (SED), for years 1960-1997. Endogeneity problems in estimating the impact of foreign-born students are tackled by instrumenting their arrival rates with shocks in both their origin country and the US (GDP fluctuations, introduction or lift of restrictions to study abroad, visa policy changes). Foreign-born and local students are found to impact similarly on their departments' publication activity and quality, which goes in the direction of suggesting their substitutability.

countries, but set up new or subsidiary companies in their home country (Meyer, 2001; Wadhva, 2009a,b; Kenney et al., 2013, and references therein).

While case studies on these phenomena abound, large scale quantitative evidence is scant, and entirely based on patent data.

The most comprehensive enquiry has been conducted by William Kerr, in a series of papers (some with co-authors) that exploit two sources of information:

- the NBER Patent Data File, compiled by Hall et al. (2001), which includes information on name, surnames, and addresses of inventors
- the Melissa ethnic-name database, a commercial repository of names and surnames of US residents, classified by likely country of origin, mainly used for direct-mail advertisements.

Names and surnames from the two sources are matched in order to assign an "ethnic affiliation" to each inventor. "Ethnicity" here identifies populations coming from groups of countries with linguistic or cultural affinities. The latter are measured according to a US-centric perspective, with more details for Asian countries. As a result, only nine groups are identified: English (which includes US natives), European (which includes all Europe, with the exception of Russia and Spain), Russian (former Russian-speaking USSR countries), Hispanic-Filipino (all Spanish-speaking countries), Chinese (including Taiwanese and Singaporean), Indian (including Pakistani and Bengali), Japanese, Korean, and Vietnamese. To the extent that recent waves of highly skilled migration into the US have come from China, India, and other Asian countries, this classification, albeit incomplete, suffices to explore a large part of the phenomenon of interest.

Another limitation of the "ethnic" method for approximating the phenomenon of migration is the impossibility to distinguish between foreign-born individuals (who include all important categories of the foreign-educated and international students) and second generation immigrants, or members of longstanding ethnic minorities. At least for the US, this limitation is mitigated by working on longitudinal data, to the extent that increments of non-native ethnic groups of interest are mainly due to immigration (witness the highly skilled immigration statistics we reviewed in section 2.1).

Descriptive analysis by Kerr (2007) reveals several stylized facts, most of which are coherent with those concerning *hs* and scientific migration:

(i) The ethnic inventors' share of all US-resident inventors grows remarkably over time, from around 17% in the late 1970s to 29% in the early 2000s. Notice that the latter figure is in the same order of magnitude of CDH estimates of the foreign-born share of doctoral holders in 2003 (26%) but much larger than that for *hs* migration from DIOC 2005/06 (around 16%; see section 2.1).[12]

(ii) The fastest growing ethnic inventor groups are the Chinese and Indian ones, while the overall growth appears to be stronger in science-based and high technologies

(iii) When distinguishing patents according to the institutional nature of the applicant (academic vs. business) one observes the growth of ethnic inventorship to occur early on in universities, with firms catching up later (in coincidence with the rise of the phenomenon of ethnic entrepreneurship described in the previous section)

(iv) Ethnic inventors appear to cluster in metropolitan areas (with a correlation between city size and percentage of ethnic patents), thus contributing to the growing spatial concentration of inventive activity observed in the US over the past 20 years (this evidence is reprised in detail by Kerr, 2009).

[12] These comparisons cannot be but suggestive, as the figures for doctorate holders and highly-skilled migrants refer to stocks, while Kerr's figures are better seen as flows.

The most important applications of the ethnic inventor database concern the theme of knowledge spillovers. In Kerr (2008) these are measured by citations running from patents filed at USPTO from foreign residents (in years 1985-97) to patents filed up to ten years before by local residents.[13] Citations are grouped according to four criteria (inventor's ethnicity and technological class of the citing patent, plus inventor's ethnicity and technological class of the cited patent). A negative binomial regression is then run, with citation groups as observations, the number of citations in each group as the dependent variables (which is often zero), and a series of dummies as regressors. Among the latter, the "co-ethnicity" dummy is of particular interest, as it indicates whether the ethnicities of inventors in the cited and citing patents in the citation group are the same. This allows estimating that co-ethnic citation groups are on average 50% more numerous than mixed ethnic ones. This basic finding is interpreted by recalling the vast economic and sociological literature on the tacit nature of technical knowledge and the roles of social ties (see footnote 13). In this interpretation, ethnicity represents a social bond between inventors, which favours the transfer of tacit knowledge assets not comprised in the patent description, but necessary to the understanding and development of the invention.

Kerr (2008) further uses patent data as regressors in a first-difference panel data econometric exercise concerning origin countries of immigration into the US. Here the dependent variables are alternative measures of economic growth (growth of manufacturing output or employment, or of labour productivity), while the number of ethnic patents in the US (with ethnicity coinciding with that of the origin country) is the focal regressors. He finds that a one percentage point increase of ethnic patents in the US is associated to a 10% to 30% increase of the country of origin's output measures. The result weakens, but resists, when excluding China from the origin country set, or Computer and Drugs from the technologies considered. This suggests that ethnic-mediated spillovers, while having a stronger impact in high technologies and in one particular economy, are not irrelevant for a more general set of countries and technological fields.

Reverse causality problems are also discounted, by excluding from the regressions Western European countries of origin, plus Japan: in the case of such advanced economies, it could be the case that ethnic patents in the US grow as a consequence (and not as the cause) of home technical progress, with the country of origin's multinationals finally expanding into the US and localizing there a few of their own inventors.

Foley and Kerr (2011) exploit the same database to investigate the specific role of ethnic inventors in relation to multinational companies' activities in origin countries. In particular, they find that US multinationals with a high share or quantity of ethnic patents invest and innovate more in their ethnic inventors' origin countries, while at the same time relying less on joint ventures with local companies for doing so. This suggests that ethnic inventors may not only channel back to their origin countries some key economic and innovation activities, but also act of substitutes of local intermediaries, thus diminishing their companies' costs of engaging into foreign direct investments.

Patent inventor data have been exploited by a few other studies on immigrants' knowledge feedbacks to origin country, all of them dealing with the case of India. Agrawal et al. (2011) propose a theoretical model which compares the importance of spatial proximity between inventors (co-location in the same country, region or city) to that of social proximity (same ethnic origin) in facilitating the transmission of technical knowledge. To the extent that spatial proximity facilitates knowledge transmission more than social proximity does, the origin country of inventors stand to lose from migration (emigrant inventors will engage more in knowledge exchanges within their destination countries, than with their home countries). In order to determine the relative importance of spatial and social (ethnic) proximity, they reformulate Jaffe et al.'s (1993) classic exercise on the determinants of patent citations (see again footnote 13).

[13] Only first inventors and their addresses are considered, and self-citations at the company level are excluded, so to avoid counting knowledge transfers internal to multinational companies. The use of patent citations to measure knowledge flows is both widespread and controversial. It originates with Jaffe et al. (1993)'s application to the theme of spatial concentration of knowledge spillovers, where it is proved that citations are more likely to occur between patents by co-localized inventors, after controlling for the spatial of concentration of patents, by technological classes. This exercise has been criticized for methodological reasons by Thompson and Fox-Kean (2005). Breschi and Lissoni (2005, 2009) and Agrawal et al. (2006) prove that other types of distance between inventors, in particular social distance, matter as much or more than spatial distance. Technical issues are reviewed by Breschi and Lissoni (2004). For a general critique of the use of patent citations in innovation studies see Roach and Cohen (2013).

First, on the basis of an Indian surname database, they identify ethnic Indian inventors of USPTO patents both from India and from the US (1981-2000).
Second, citations running from patents by ethnic Indian inventors from India are investigated. It is shown that, when controlling for the technological class of these patents, most citations are directed to other patents from India, rather than from the US; and that ethnic ties between inventors in India and in the US do not increase the probability to observe a citation link between those inventors' patents. The only exceptions are citations in Electronics, and those mediated by multinational firms, both of which are (weakly) mediated by ethnic ties. Overall, these results go in the direction of suggesting that pure knowledge spillovers (that is, knowledge flows not mediated by contractual norms) are not a major source of feedbacks from emigrated inventors to origin countries. However, ethnic ties are found to matter within the US (albeit less than co-location ties), as proved by an earlier paper by the same authors, based on the same methodology (Agrawal et al., 2008; see also Almeida et al., 2010). Quite interestingly, Agrawal et al. (2011) also try to identify Indian returnee inventors, but find only very few of them, who are responsible of just 18 patents. Similarly, Alnuaimi et al. (2012) examine around 3500 USPTO patents assigned to over 500 India-located patentees (local firms, subsidiaries of foreign companies, and universities) in between 1985 and 2004, and find very few inventors once active in subsidiaries of foreign companies who then move to local firms. This suggests that, as far India is concerned and inventors are examined, returnees and multinational employees in origin countries are not a direct source of knowledge transfer. Therefore, Foley's and Kerr's (2011) results can be only explained by indirect activities by ethnic inventors, not captured by patents, such as reference, advice, and cultural mediation.

A more recent contribution by Miguelez (2013) exploits the information on inventors' nationality contained in PCT patent applications (see 3.1 below). The author estimates the impact of foreign inventors on the extent of international technological collaborations between origin and destination countries, as measured by co-patenting activity. Findings suggest a positive and significant impact for all countries of origin, that is, not only for the largest ones, such as China and India.

3. Migration, innovation and patent data: methodology and potential

In the previous section, we have examined a number of potential sources of information on the phenomenon of migration and innovation, including information on inventors from patent data. We discuss here more in depth the latter's potential, as well as the methodological problems they pose. As for their potential, this appears very large with respect to:

1) Direct measurement of migrants' contribution to innovation in their destination countries. This applies in particular to science-based and advanced technologies (electronics, ICT, drugs, biotech, and scientific instruments), for which patent data better capture inventive activity (compared to more traditional technologies). Proper classification of inventors by name ethnicity, along with regular updating, may deliver systematic information on the weight of foreign inventors in terms of patent shares, and shares of highly cited patents.
2) Patent citations may be exploited to track knowledge flows among inventors from the same origin country, either within the same destination country or back towards the country of origin.
3) By concentrating on ethnic inventors with more than one patent (a necessary condition for having the possibility to observe two different countries of residence), one may also hope to track returnee inventors, despite existing studies suggest figures to be very low.

Notice that the technologies for which patent data are informative are the same in which universities all over the world are very active either directly, that is through patenting (see Lissoni, 2012), or indirectly, by educating future inventors. To the extent that universities are a key point of entry for migrant S&Es into destination countries, this reinforces patent information's potential for producing large enough figures amenable of statistical analysis.

In order to fulfil this potential, however, a number of technical challenges have to be tackled. We examine them in turn.

3.1 Ethnic identity vs. migrant status

First, a systematic effort must be undertaken in order to uncover all exploitable sources of information on the ethnic origin of names and surnames and improve the quality of ethnic classification. Notice that assigning a name or surname to a country of origin is an exercise whose outcome depends heavily on the country of residence considered and its immigration or geopolitical history. For example, an ethnic Italian name in France may indicate either a recent Italian immigrant, a descendant of immigrants from the late 19th or 20th century, or a member of an ethnic minority like the Corsicans; but the same name in Japan would undoubtedly points at a recent immigrant. Similarly, Turkish names in Germany may point to grandchildren of unskilled immigrants from the 1950s, or to recently arrived doctoral students. A large literature exists outside economics which both discusses classification problems and provides untapped ethnic name repositories (Cheshire et al., 2011; Mateos et al., 2011). In the next section we explore the potential of a commercial repository (namely, the IBM Global Name Recognition system), but in the future we plan to explore a few alternatives, such as the Onomap system, produced by the Department of Geography at University College London. In addition, a plethora of smaller datasets have been assembled by geneticists engaged in isonymic studies[14] (two classic references being Lasker, 1977; and Piazza et al., 1987), or by public health specialists who study the access of immigrants and minorities to medical care and/or their exposure to specific diseases (e.g. Razum, 2001).

Such a variety of sources should also help going beyond a major limitation of Kerr's (2007) pioneer effort, and of the migration-and-innovation literature in general, namely its US-centrism. This is all too necessary when we recall, from the discussion conducted in section 2.1, that highly skilled immigration is a relevant phenomenon in Europe. There we observed both a set of non-European origin countries that does not entirely overlap with that of migration into the US; and an important phenomenon of brain circulation between advanced European countries, which is of great interest to local policy-makers and stakeholders. In order to explore these phenomena, we cannot clearly content ourselves of concentrating only on inventors from China or India.

Alternatively, one could use information on the country of birth or nationality of the inventors. The former is generally available from census data, but they are not easily linkable to inventor data (for an exception, see Zhang and Ejermo, 2013); or from inventor surveys, but in this case country-of-destination subsamples may be too small to produce significant statistical evidence. As for the nationality of inventors, this is available on patent applications submitted to PCT (Patent Cooperation Treaty), with a request of extension into in the US, for a number of years up to 2010 (Miguelez and Fink, 2013). These data, which we will make use of below and refer to as the WIPO-PCT dataset, are extremely valuable to the extent that they document with unprecedented details the migration flows of inventors over the past 10 years or so.[15] In particular, these flows are highly correlated with those of *hs* migrants, but more extreme, with the US standing out even more noticeably as the top destination country, and India and China as the top origin ones. Unfortunately, the time series one can obtain from WIPO-PCT stop in 2010 and will not be updated in the future. In addition, measuring migration through nationality may lead to an under-estimation of the former, to the extent that long-term immigrants (especially highly skilled ones) often end up getting the nationality of their country of residence.

3.2 Inventor disambiguation

A second methodological problem to be tackled is that of name disambiguation. Once ignored by economists making use of patent data, the increasing exploitation of information on inventors has now attracted more attention to the issue. By "name disambiguation" (or, in information technology jargon, "entity resolution") we mean the identification of two or more inventors listed on several patents as the same person, based on their homonymy or quasi-homonymy (identity or similarity of names and surnames). This operation hides a number of traps, which may result in non-random measurement errors, and bias estimates of the phenomena under interest. In particular, type I errors (false positives) are generated whenever two inventors are presumed to be the same person, when in fact they are not;

[14] Isonomic studies exploit the co-occurrence of surnames within a population (isonomy) to calculate the latter's inbreeding rate, which in turn raises the probability to observe some genetic traits of interest

[15] As explained by Miguelez and Fink (2013), the WIPO-PCT dataset covers in principle all years from 1978 to 2011, but it is significantly populated only from 1991.

while type II errors (false negatives) are generated whenever two inventors who are indeed the same person are not identified as such.
In jargon, a disambiguation exercise that produces a small number of false positives is said to ensure a high precision rate, while a low number of false negative is said to ensure a high recall rate.[16]

Raffo and Lhuillery (2009) discuss at length the implication of measurement errors (low precision or low recall) in a number of applications of inventor data. Here we limit ourselves to point out a technical problem that inevitably occurs when dealing with ethnic names and surnames, and its implications for studies on migration and innovation.

A key element of name disambiguation algorithms consists in measuring the edit or phonetic distance between similar names and surnames, and setting some distance thresholds under which different names and surnames are considered the same. The objective here is to avoid treating as two different persons all individuals whose names or surnames have been misspelled (e.g. "Francesco" turned into "Francisco") or transliterated between alphabets in different ways (e.g. "Mao Ze Dong" being also written as "Mao Tse Tung"). At the same time, one must avoid substantive differences between names to be treated as misspellings or transliteration variants (e.g. treating "Joe" and "John" as the same).[17]

Unfortunately, when applied to inventors from different countries of origin, general rules return different results in terms of precision and recall, depending on the orthographic rules and the frequency of common names and surnames typical of each country. Chinese, Korean, and Vietnamese surnames, for example, are both short (which makes it arduous to tell them apart on the sole basis of edit distances) and heavily concentrated on a few, very common ones (such as Wang, Kim, or Nguyen). The opposite holds for Russian surnames. Similar examples can be made for names.

An immediate implication is that the same algorithm may return a low precision rate (many false positives) when applied, say, to East Asian names and surnames, and a high one when applied to Russian ones. Alternatively, it may be that the algorithm returns a high recall rate (few false negatives) for East Asians, and a low one for Russian ones. In both cases, it will often be the case that many pairs or groups of East Asian inventors will be treated as the same person, as opposed to only a few Russian ones, and introduce three possible biases:

1) Over-estimating the average and/or maximum productivity of inventors in the low precision/high recall group (the East-Asians) and under-estimating those in the high precision/low recall one (the Russians). Indeed, while several inventors named Wang or Wong will be collapsed into one, and then treated as a highly productive individual, the same will apply only to a few Russian inventors.
2) Over-estimating the number of returnee inventors in the low precision/high recall group (the East-Asians), and under-estimating it in the high precision/low recall (the Russians). While several Wang *aka* Wong inventor will be found to be active both in the US and China, the same will apply only to a few Russian inventors.
3) Under-estimating the rate of ethnic citations for the low precision/high recall group (the East-Asians), and over-estimating it in the high precision/low recall (the Russians). Understanding this point requires bearing in mind that the citations of interest (those representing knowledge spillovers) are those that run between different individuals; that is, self-citations at the individual level are excluded from the computation. This implies that many citations between, say, East Asian inventors may be excluded, to the extent that the citing and cited inventor (e.g., Wang and Wong, respectively) are more likely to be treated as the same person; while many citations between Russian inventors may be retained, as the pair of inventors involved appear to be two different persons.

[16] Precision and recall rates are measured as follows:

$Precision = \frac{tp}{tp+fp}$; $Recall = \frac{tp}{tp+fn}$

where

$$tp = number\ of\ true\ positives$$
$$tn = number\ of\ true\ negatives$$
$$fp = number\ of\ false\ positives$$
$$fn = number\ of\ false\ negatives$$

[17] In the case of inventor data, most algorithm supplement the edit and phonetic distance investigation with contextual information (whether the two inventors with the same or similar names have addresses in the same location, patent in the same class or for the same company etc.). See Pezzoni et al. (2012).

Two complementary strategies may help tackling the problems just exposed. One consists in calibrating the disambiguation algorithm by taking into account the specificities of each linguistic group. The other consists in making the best possible use of the contextual information contained in patents (see footnote 16). Disambiguation algorithms specific for inventor data, which follow both strategies, have been developed, among others, by Lai et al. (2011) and Pezzoni et al. (2012). Two public available inventor-data produced with such algorithms are the EP-INV dataset, originally developed for the identification of academic inventors, but comprising all inventors of patent applications filed at the European Patent Office from 1978 to around 2010; and the US Patent Inventor Database, developed by Lee Fleming and associates, which contains USPTO data.[18]

Applications of disambiguated inventor data to the study of inventor mobility and networks or university patenting has nowadays become very common (see respectively Marx et al., 2009; Breschi and Lissoni, 2009; and Lissoni, 2012). The same cannot yet be said of applications to the study of migration and innovation.

Kerr (2007) and extensions make use a non-disambiguated inventor data set (the NBER dataset), as no attempts are made to estimate the productivity of inventors, nor the number of returnees (and citations are treated at a relatively aggregate level).[19]

As for studies that deal with citations and/or returnee figures, Agrawal et al. (2008, 2011) and Almeida et al. (2010) do not provide details on the disambiguation techniques they have used, while Alnuaimi et al. (2012) apply a "perfect matching" technique, by which only inventors with exactly the same name and surname are considered as the same person, without further checks (which in principle works as a high precision algorithm, but still can suffer of a false negative problem, due to the presence of homonyms). In what follows, we illustrate the problems just discussed with a few examples from the Ethnic-Inv database, all of them also of substantive interest for the migration and innovation issue.

3.3 Returnee tracking

As discussed in detail by Hall's comments to an early version of the present paper (Hall, 2013), patent data provide only partial information on inventors' mobility in space. This is because mobility can be observed only for inventors having signed at least two patents over the observation period, who constitute a minority of all inventors (the majority having signed just one patent). More precisely, in order to define an inventor as "mobile", we need to observe his two or more patents having been taken in different cities, regions or states. In case of inter-state migration a mobile inventor qualifies as a migrant, so it is possible, in principle, to estimate migration rates as part of overall mobility rates.

According to Hall's simulations, however, any indicator of migration based exclusively on mobility data will certainly underestimate the real extent of inventors' migration, even by a half or more, depending on inventors' propensity to emigrate, and to whether this is related to productivity (number of patents filed).[20]

To the extent that the studies we reported above do not rely just on within-the-patent information to measure migration, but infer it from the ethnic origin of the inventors' names and surnames, they are less affected by this problem. This is because migration rate estimates are based on samples that include not only the inventors with more than one patent, but also those with one patent only.

However, the problem re-appears as soon as one attempts to track returnee inventors: in this case, one goes back to the need of observing at least two patents per migrant inventor, in order to define the latter as a returnee if the any patent following the one(s) reporting a foreign address do report a home country one.

[18] To download the APE-INV inventor database: http://www.esf-ape-inv.eu/index.php?page=3#EP-INV. APE-INV serves as the basis for the Ethnic-Inv database we present in section 4. To access the Lai et al.'s database: http://dvn.iq.harvard.edu/dvn/dv/patent

[19] Nevertheless, one may presume disambiguation algorithms have been used to produce the Melissa name database used for identifying ethnic inventors.

[20] Hall's simulations are based on a realistic frequency distribution of inventors by number of patents (as observed in the literature) and several hypothetical parameters representing their probability to emigrate.

As suggested by Hall, one partial remedy to the problem may consist in using the number of observed returnees as inputs to a simulation exercise, whose objective should consist in providing estimates of the returnee phenomenon, based on realistic assumptions of migrant inventors' productivity and propensity to go back to their home countries. Formulating such assumptions requires however conducting an inventor survey, which goes beyond the scope of this paper.

4. The *Ethnic-inv* database: a first look

The *Ethnic-inv* database, still at its pilot stage, results from matching names and surnames of inventors in the APE-INV inventor database with information on their countries of origin obtained by Global Name Recognition, a name search technology produced by IBM (from now on, IBM-GNR).

4.1 Methodology

4.1.1 Disambiguated inventor data: the EP-INV dataset

The EP-INV inventor database contains information on 2,806,516 inventors with different names and/or addresses listed on the patent applications filed at the European Patent Office (EPO), from its year of opening in 1978 to around 2009 (filing data from 2009 onward not being complete, due to publication delays). Raw data come from the October 2011 version of PatStat, the Worldwide Patent Statistical Database published regularly by the European Patent Office.[21] Disambiguation is performed by making use of Massacrator 2.0, a 3-step algorithm working as follows (for details, see Pezzoni et al., 2012):

Disambiguation Step 1. *Cleaning & Parsing*: the relevant text strings (those containing information on name, surname and address of the inventor) are purged of typographical errors, while all characters are converted to a standard set. The string containing the inventors' complete name is parsed into tokens, each of which containing a given name or a surname (out of all given names and surnames the inventor may have); tokens containing professional titles or name qualifications ("jr", "II" etc) are dropped. The address is parsed, too.

Disambiguation Step 2. *Matching*: the algorithm selects pairs of inventors, from different patents, who are likely candidates to be the same person, due to homonymy or similarity of names. The matching procedure compares all tokens produced at step 1, sort them alphabetically, and assign them to groups, based on 2-gram distances between subsequent tokens in the alphabetical list. Any pair of inventors whose complete name strings are composed of tokens belonging to the same group are matched. In case the matched strings are composed by a different number of tokens, the minimum common number of tokens is considered.

Disambiguation Step 3. *Filtering*: the matched inventor pairs are filtered according to additional information retrieved either from the patent documentation or external sources. Typical information from within the patent documentation are the address (e.g. quasi-homonyms sharing the same address are believed to be the same person) or some characteristics of the patent, such as the applicant's name (e.g. homonyms whose patents are owned by the same company may be presumed to be the same person) or its technological contents (as derived from the patent classification system or patent citations).

Massacrator 2.0 is a general tool, one that can be calibrated to maximize precision (minimize false positive) or recall (minimize false negatives), or to achieve the best possible combination of the two (that is, to strike a balance between different types of errors). The calibration takes place at step 3, by assigning different weights to filtering criteria. In what follows, unless otherwise stated, we will make use of the "balanced" version of the database, one that, when tested against a benchmark sample of French academic inventors, returned a precision rate of 88%, and a recall rate of 68%. However, some

[21] Access and methodological information for PatStat at: http://forums.epo.org/epo-worldwide-patent-statistical-database/ - last visited: 4/4/2013. See also the unofficial blog: http://rawpatentdata.blogspot.com.

exercises in section 4.2.3 will be conducted also on a "recall-oriented" database, with a recall rate of 93%, but a low precision (56% only).
Notice that the number of unique individuals in the "balanced" database are 2,366,520 (-16% with respect to the number of inventors in the raw data), while the same figure in the "recall-oriented" one is 1,697,976 (-39%).[22]

4.1.2 Inventors' country of origin: the Ethnic-INV dataset

Our basic source of information on inventors' country of origin is the database feeding the *IBM-GNR system*, a commercial product which performs various name disambiguation tasks. Among such tasks, the one of interest here is the association of names and surnames to one or (more often) several countries of likely origin. This association originates from a database produced by US immigration authorities in the first half of the 1990s, which registered all names and surnames of all foreign citizens entering the US, along with their nationality, for a total of around 750,000 full names. In addition, variants of registered names and surnames are considered, according to country-sensitive orthographic and abbreviation rules.[23]

When fed with either a name or a surname or both, IBM-GNR returns a list of "countries of association" (from now on: CoAs) and three scores:

- "frequency", which indicates to which percentile of the frequency distribution of names or surnames the name or surname belongs to, for each CoA (e.g. an extremely common Vietnamese surname such as Nguyen will be associated both to Vietnam and to France, which hosts a significant Vietnamese minority; but in Vietnam it will get a frequency value of 90, while it France it will get only, say, 50, the Vietnamese being just a small percentage of the population);
- "significance", which approximates the frequency distribution of the name or surname across all CoA (continuing with the previous example, the highest percentage of inventors names Nguyen will be found in Vietnam, followed by the US and France)
- "confidence", which indicates the reliability of the frequency and significance scores for each CoA, as a function of the number of observations available for the latter in the database (with 90 indicating maximum reliability).

The IBM-GNR list of CoAs associated to each inventor is too long for being immediately reduced to a unique country of origin for each inventor in our database. This operation requires filtering a large amount of information through an *ad hoc* algorithm, one that compares the frequency and significance of the two lists of CoAs associated, respectively, to the inventor's name and surname to the inventor's "country of residence" at the moment of the patent filing (which we obtain from the inventor's address in the EP-INV dataset). Figure 1 illustrates the type of information provided by IBM-GNR, the position of our algorithm in the information processing flow, and the final outcome. Notice that we refer to "country of association" (CoA) when considering the raw information from IBM-GNR, and to "country of origin" when considering the final association between the inventor and one of the many CoAs proposed by IBM-GNR (or one of our "meta-countries" based on linguistic association). The full description of our algorithm (to which we will refer with the working name of *Ethnic-INV* algorithm) is available on request.[24] Here it suffices to point out its basic principles:

[22] Recall and precision rates were estimated against two datasets of academic inventors from France and Switzerland, which implies that the same rates could be lower or higher when considering inventors from different countries of origin. Future versions of the APE-INV inventor database will try to correct for this.
[23] Information on IBM-GNR reported here comes from IBM online documentation (http://pic.dhe.ibm.com/infocenter/gnrgnmv/v4r2m0/index.jsp?topic=%2Fcom.ibm.iis.gnm.overview.doc%2Ftopics%2Fgnr_gnm_con_gnmoverview.html; last visit: 4/4/2013) as well as: Patman (2010) and Nerenberg and Williams (2012). E-mail and phone exchanges with IBM staff were also decisive to facilitate our understanding. Still, being IBM-GNR a commercial product partly covered by trade secrets, we did not have entire access to its algorithms and we had to reconstruct through deduction.
[24] Notice that the *Ethnic-INV* algorithm and the *Massacrator 2.0* disambiguation algorithm described above are totally independent. Although we use them in combination, nothing prevents perspectives users from combining inventor data disambiguated by means of *Massacrator 2.0* to other sources of information than IBM-GNR, or to apply the *Ethnic-INV* algorithm to other sources of inventor disambiguated data.

Ethnic-INV Step 1. The algorithm considers first whether the maximum-significance CoAs associated respectively to the name and surname of the inventor are the same (or whether the product of significance of first name and surname exceeds a critical threshold value), in which case it creates a country of origin identical to such CoA, irrespective of the country of residence. In the example of figure 1, this is the case of Francesco Lissoni, whose name and surname are both associated with maximum significance to Italy.

Ethnic-INV Step 2. In case this simple rule does not apply, the top significance CoA associated to the inventor's surname is considered, provided its value exceeds a certain threshold. Various threshold values are specified, for as many variants of the algorithm (sixteen, at this stage of our research).

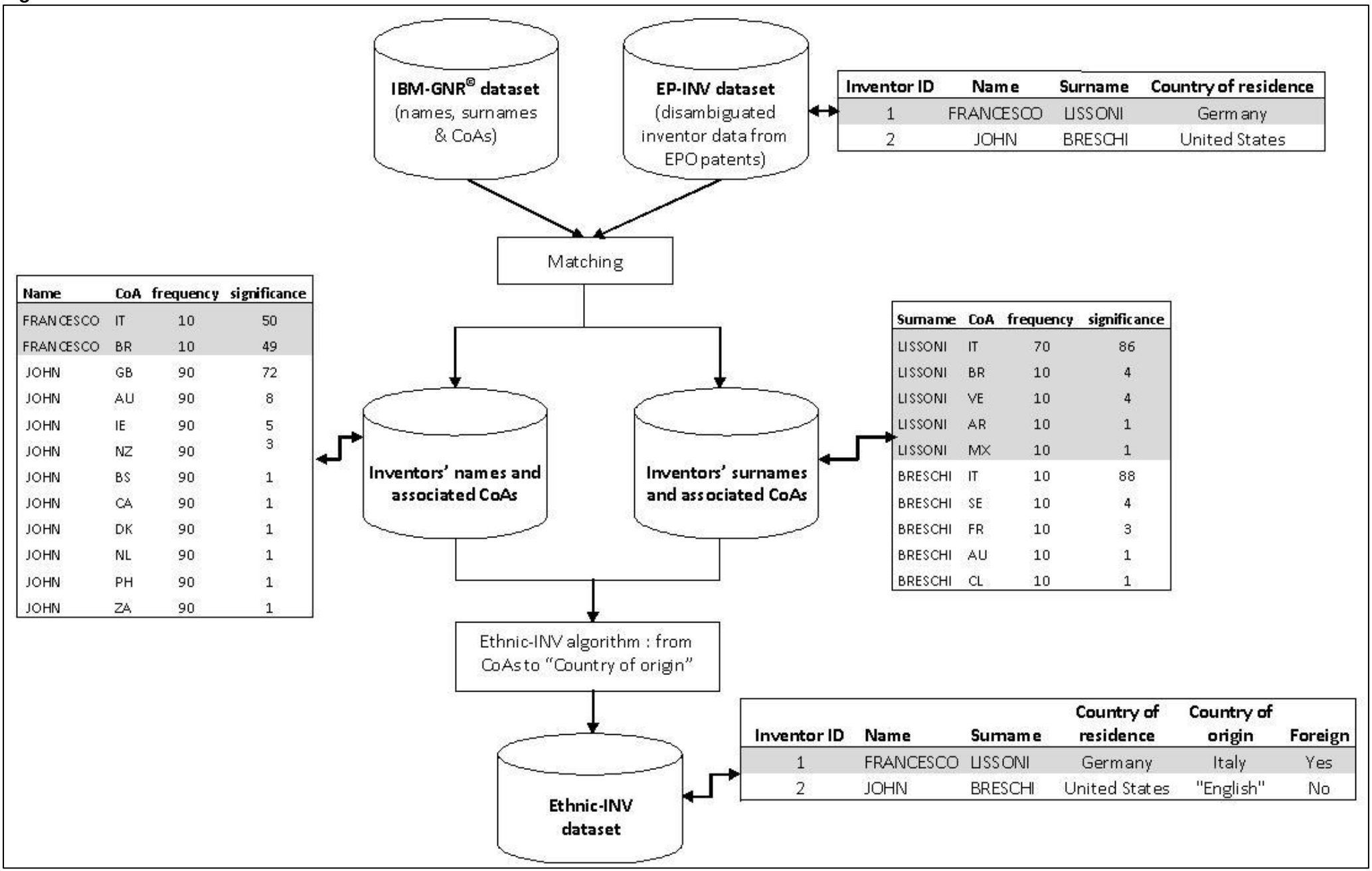
Ethnic-INV Step 3. In case none of the two previous sets of rules applies, the inventor's country of origin is presumed to be the same as that of residence one. Similarly, in order to correct for the presence of ethnic minorities in the country of residence, the latter is chosen as country of origin if either the inventor's name or her surname appears to be very frequent (frequency > 90) there. In the example of figure 1, this is the case of John Breschi, who has a very frequent name in his country of residence, the US, despite a very uncommon surname (one that would push the algorithm to treat him as Italian).

Ethnic-INV Step 4. After completing this process, the algorithm produces a *Foreign* dummy variable, which takes value 1 (Yes) whenever the country of origin and residence differs (or the country of residence is not comprised within the meta-country of origin). We will refer to inventors with *Foreign* =1 as IFOs (Inventors of Foreign Origin). In the example of Figure 1, Francesco Lissoni is considered an IFO (his country of origin being Italy, while that of residence is Germany), while John Breschi is not (his country of residence, the US, is comprised within his "English" meta-country of origin).

Ethnic-INV Step 5. Notice that Step 2, by requiring to fix threshold values of frequency and significance, allows for calibrating the algorithm in order to orient it towards more precision or more recall (or a combination of the two). The latter are defined as in footnote 12, although by "positives" ("negatives") we now mean the cases of inventors whose country of origin differ from (is the same as) that of residence at the priority date of the patent. In order to identify true vs. false positives and negatives we would need a benchmark datasets reporting the true country of origin (ideally, the country of birth) of a representative sample of individuals. At this stage of our research, we have not yet managed to assemble a similar database, so we rely as benchmark on the WIPO-PCT dataset (Miguelez and Fink, 2013). This is a less-than-satisfactory choice to the extent that the latter does not report the inventors' country of birth, but their nationality.[25]

[25] By now, the WIPO-PCT database is the best benchmark at hand, as it is very similar to the EP-INV database in terms of size and information contents (albeit without inventor disambiguation). This allows us to define as a true positive every inventor that our algorithms classifies as of foreign origin and is of foreign nationality according to WIPO-PCT (similarly, a false positive will be any inventor classified as of foreign origin, who turn out not to be a foreign national; and so on for the false negatives). Still, it is an imperfect benchmark, to the extent that our definition of "country of origin" and that of nationality differ in many ways. First, nationality of the country of residence can be acquired over time, and this applies especially to prolific immigrant inventors, who may have resided a long time in a foreign country (indeed, a close look at WIPO-PCT data reveal several cases of double nationality). Second, the concept of country of origin points at some ethnic ties that can be maintained after the change of nationality or, in some cases, over more than one generation of immigrants, including those who were born abroad and automatically acquired the nationality of the country of residence (according to some form of *ius soli* legislation). For these reasons, we cannot expect very high precision rates when using nationality as a benchmark, nor should we interpret too strictly any false positive as an error in our country of origin classification.

Figure 1. From inventor data to the Ethnic-INV database



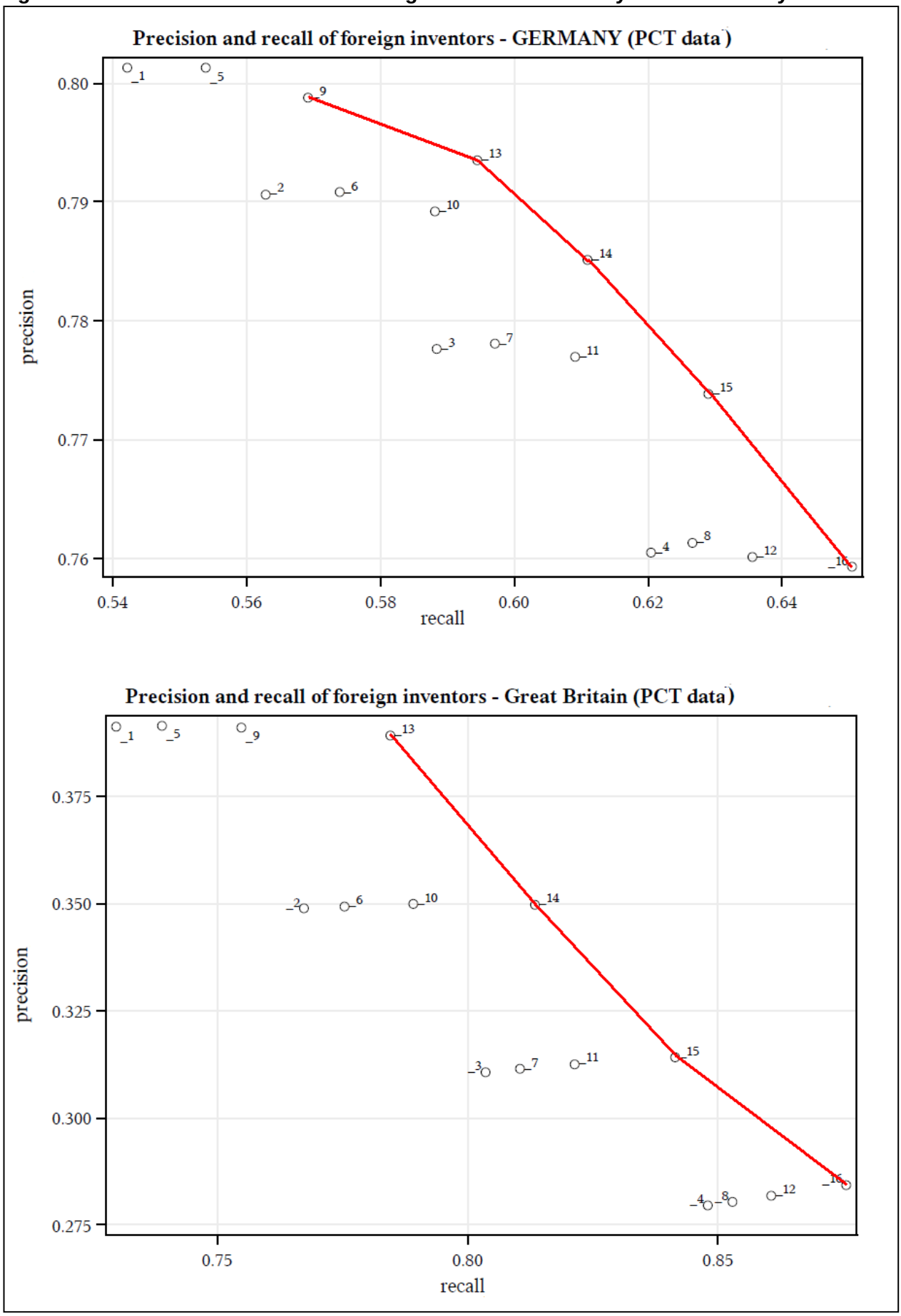
Inventor ID	Name	Surname	Country of residence
1	FRANCESCO	LISSONI	Germany
2	JOHN	BRESCHI	United States

Name	CoA	frequency	significance
FRANCESCO	IT	10	50
FRANCESCO	BR	10	49
JOHN	GB	90	72
JOHN	AU	90	8
JOHN	IE	90	5
JOHN	NZ	90	3
JOHN	BS	90	1
JOHN	CA	90	1
JOHN	DK	90	1
JOHN	NL	90	1
JOHN	PH	90	1
JOHN	ZA	90	1

Surname	CoA	frequency	significance
LISSONI	IT	70	86
LISSONI	BR	10	4
LISSONI	VE	10	4
LISSONI	AR	10	1
LISSONI	MX	10	1
BRESCHI	IT	10	88
BRESCHI	SE	10	4
BRESCHI	FR	10	3
BRESCHI	AU	10	1
BRESCHI	CL	10	1

Inventor ID	Name	Surname	Country of residence	Country of origin	Foreign
1	FRANCESCO	LISSONI	Germany	Italy	Yes
2	JOHN	BRESCHI	United States	"English"	No

Figure 2. Calibration of the Ethnic-INV algorithm at the country level: Germany vs. Great Britain



(1) Dots represent precision-and-recall outcomes of different calibrations of the algorithm; see section 4.1.2 for details

(2) In red: efficient calibrations (precision cannot be improved without reducing recall, and viceversa)

It is important to stress that the calibration occurs at the level of the inventor's country of residence. Figure 2 reports the results of the calibration exercise conducted for two countries of interest, namely Great Britain and Germany. Each dot on the graphs is the result of a different calibration; dots on the red line represent efficient calibration results, whose outcome in terms of precision cannot be improved without losing out significantly in terms of recall (and vice versa). We first notice that the range of precision and recall values we obtain is different for the two countries (precision: 27.5%-40% for Britain vs. 76%-60% for Germany; recall: 70%-90% for Britain; 54%-66% for Germany). Second, we observe a different distribution of calibration results, with efficient dots being just four for Britain, and five for German. Finally, while for both countries the maximum-recall calibration is represented by dot 16, the maximum precision one is dot 13 for Britain and dot 9 for Germany (as for "balanced" calibration, the choice for Britain is between dot 13 and dot 14, while for Germany the choice includes dot 15, too).

As a source of information on countries of origin, the most important limitation of IBM-GNR, for our purposes, is the absence of the US among the list of available CoAs, as US citizens never entered the original database. When manipulating the information obtained by IBM-GNR we manage, by now, to assign as presumed US origin to inventors residents in the US [see below], but cannot to do the same for inventors outside the US. Compounded with other problems (chiefly, the difficulty to distinguish between CoAs whose dominant language is English), this forces us to assign US inventors in a generic "English" meta-country of origin, which also include British, Australian, and Irish inventors.

"Neighbourhood effects" problems are present, too. Most Spanish-speaking travellers and immigrants to the US come from Mexico, followed by other Latin American countries. When applying IBM-GNR to Europe, this leads to over-estimating the number of inventors coming from such countries, and under-estimating those from Spain. The same applies for Brazil and Portugal. Similar problems affect Chinese-, German-, and Russian–speaking countries, in this case because one large country (respectively, China, Germany and Russia) overwhelm all others in terms of significance scores. Once again, at present, our solution for these problems consists in creating meta-countries of origin (such as "Chinese", "German", "Russian", or "Spanish", with reference to the linguistic group). Still, these meta-countries of origin are more detailed and less US-biased than Kerr's (2007) ethnic groups.[26] Another important limitation is the potential obsolescence of the reference database, which may lead to some errors in the assignment of inventors to countries of origin. For example, during the first half of the 1990s, entries in the US from Eastern European countries were still limited, so that IBM-GNR may not be able to assign a CoA to inventors from such countries. They will then be confounded with local inventors, thus leading to an underestimation of their presence abroad.

Despite these limitations, the IBM-GNR has the potential to identify immigrant inventors from a variety of countries of great interest from a European perspective. Its accuracy may vary across origin and destination countries, and should be tested against at least one benchmark dataset for country. Besides, nothing prevents us from combining it, in the near future, with other, similar resources.

In what follows we will use just a subset of the Ethnic-Inv data, namely those data concerning inventors active in the 12 European countries with the largest number of patent filings at the EPO since 1978 (namely: Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Spain, Sweden, Switzerland, United Kingdom), plus the 3 most important non-European countries, according to the same criterion (namely: US, Japan, and South Korea). As shown in table 1, this amount to over 1,700,000 inventors (with inventors active in n>1 countries, a tiny minority indeed, being counted n times). Almost one third of these inventors were active in the US, and almost as many in Japan, which leaves around a third of them in Europe. Here it is Germany that dominates, with around 14% of total observations in the sample.

[26] The full list of meta-countries of origin is:
- "Arabic": Egypt, Algeria, Kuwait, Lebanon, Syria, Tunisia
- "Chinese": China, Taiwan, Hong Kong
- "English": UK, Australia, Ireland (it includes also the US, albeit not listed as a CoA)
- "Former Czech-Slovakia": Czech Republic, Slovakia
- "Indian": India, Bangladesh, Nepal, Pakistan, Sri Lanka
- "Portuguese": Portugal, Brazil
- "Russian": Russia, Belarus, Bulgaria, Azerbaijan, Kazakhstan, Serbia & Montenegro, Ukraine, Uzbekistan
- "Spanish": Spain, Mexico, Colombia, Costa Rica, Cuba, Venezuela, Uruguay

Table 1: Inventors* in the Ethnic-Inv database, by country of residence (selected countries only)

	Nr	%
Austria	16,608	0.9
Belgium	20,499	1.2
Denmark	14,103	0.8
Finland	17,433	1.0
France	114,254	6.4
Germany	252,823	14.3
Great Britain	86,219	4.9
Italy	47,318	2.7
Netherlands	46,943	2.7
Spain	17,100	1.0
Sweden	31,617	1.8
Switzerland	35,510	2.0
Japan	504,431	28.4
South Korea	42,690	2.4
US	526,850	29.7
Total	*1,774,398*	*100*

** Inventors active in n>1 countries are counted n times*

4.2 Applications

In what follows we discuss a few applications of the *Ethnic-inv* database, in its present version. They are both of substantive interest and allow some further methodological reflection.

4.2.1 A comparison with highly skilled immigration and WIPO-PCT data

Table 2 provides estimates of IFOs' share of resident inventors in several countries, according to Ethnic-Inv data. Columns 3 to 5 report data based on patents filed between 1985-95, while columns 6 to 8 refer to 1995-2005. Columns within each set differ according to the calibration of the Massacrator 2.0 algorithm (respectively: maximum precision, maximum recall, and "balanced"). The table also provides a comparison of the IFO-related figures with similar figures for inventors of foreign nationality from the WIPO-PCT database, as well as with *hs* immigration shares from the DIOC 2005/06 datasets (respectively, in columns 1 and 2). This comparison serves the purpose of checking whether differences between inventor and highly skilled immigration figures differ by an order of magnitude too high to be credible, therefore suggesting the possibility of some gross estimation error.

The shaded cells in columns (6) to (8) indicate the estimates of IFOs' incidence that are closest to equivalent estimates for foreign national inventors in WIPO-PCT data. We observe that:

(i) For some countries of residence, the most reliable algorithm appears to be the one that maximises precision, in others the one that maximises recall or the "balanced" one. This confirms that the best algorithm for assigning the countries of origin varies according to the country of residence.
(ii) The Ethnic-Inv database does quite a poor job in capturing IFO in the US, due to the limitation of the IBM-GNR system: the distance between the WIPO-PCT estimate (around 16%) and ours (from a minimum of 25% to over 40%) is too high to be explained only by the conceptual difference between "nationality" and "country of origin" (see footnote 16). We clearly over-estimate the overall share of foreign inventors in the US.

(iii) The share of IFO (whether estimated on the basis of the Ethnic-Inv database, or the WIPO-PCT ones) and that of highly skilled workers are quite similar in terms of ranking (with Switzerland and the US on top, and Italy and the Japan at the bottom) but quite different in terms of values. This suggests that demand factors (such as the R&D intensiveness of the local system of innovation) and institutional factors (such as the attractiveness of the education system) may interfere with more general economic forces and immigration policies when it comes to the immigration of a very specialized category of high skilled workers, such as the scientists and engineers who form the bulk of the inventors' community, as discussed in section 2.

Table 3 reports a set of figures similar to those in table 1, but for a selected set of countries of origin, namely: Arabic, Chinese, Indian and Russian meta-countries (see footnote 22), as well as Iran, Poland, Romania, Turkey, Vietnam. These countries are at the same time among the top providers of *hs* migrants, and those whose languages clearly differ from any language spoken in the countries of destination of our interest. Such linguistic differences ought to help reducing the number of false positives, thus increasing the precision of our estimates. Indeed, a comparison of column (1) with columns (2) to (4) suggests that for these countries of origin the Ethnic-Inv estimates that best approximate the WIPO-PCT ones are those based on the maximum precision algorithm, with the exception of estimates for Japan and Spain (best algorithm is the balanced one) and South Korea (maximum recall). Still, we have over-estimation problems for the US.

Table 2: Inventors of foreign origin as % of resident inventors: estimates from Ethnic-Inv and comparison with estimates from other data sources

	Foreign-*born* as % of highly skilled residents (DIOC 2005/06)	Foreign *nationals* as % of resident inventors, (WIPO-PCT, 1991-2010)	Foreign *origin* inventors as % of residents (Ethnic-Inv, 1985-1995) ; by calibration of the Ethnic-INV algorithm [§]			Foreign *origin* inventors as % of residents (Ethnic-Inv, 1995-2005) ; by calibration of the Ethnic-INV algorithm		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Countries of residence :								
Austria	17.83	13.41	6.52	10.12	8.29	8.68	12.78	10.66
Belgium	12.46	17.84	10.54	14.36	12.40	12.61	17.15	14.85
Denmark	9.88	7.89	4.64	6.47	5.58	5.78	8.42	7.13
Finland	3.00	6.11	4.43	7.16	5.89	6.31	8.91	7.74
France	12.84	7.79	5.02	10.05	7.45	7.13	12.64	9.79
Germany	11.70	6.50	4.16	9.55	6.77	5.56	10.98	8.24
Great Britain	17.17	11.70	4.38	7.24	5.86	9.79	14.00	12.00
Italy	6.94	3.78	2.55	3.59	3.08	3.43	4.75	4.10
Japan	1.09	1.61	0.48	0.60	0.55	0.78	0.93	0.87
Netherlands	10.51	18.12	10.82	15.51	13.24	15.25	21.38	18.35
South Korea	n.a.	1.84	1.01	1.50	1.26	0.87	1.07	0.98
Spain	12.64	6.45	4.00	5.91	4.97	4.34	5.68	5.13
Sweden	15.46	7.75	6.80	10.79	8.89	9.80	14.38	12.19
Switzerland	29.10	36.10	18.19	25.98	22.00	22.14	31.07	26.44
United States	15.81	15.93	18.68	34.76	26.68	24.84	40.59	32.66

[§] Columns (3) and (6): max precision ; Columns (4) and (7): max recall ; Columns (5) and (8): "balanced" algorithm [see section 4.1.2 for details]

Table 3: Inventors of foreign origin as % of resident inventors: estimates from WIPO-PCT and Ethnic-Inv (selected countries of origin*)

	Foreign *nationals* as % of resident inventors (WIPO-PCT, 1991-2010)	Foreign *origin* inventors as % of residents (Ethnic-Inv, 1985-2005) ; by calibration of the Ethnic-INV algorithm §		
	(1)	(2)	(3)	(4)
Austria	1.18	1.77	2.18	1.98
Belgium	1.56	1.98	2.55	2.27
Denmark	1.19	1.41	1.71	1.55
Finland	2.22	2.32	2.72	2.54
France	1.71	2.26	2.92	2.58
Germany	1.54	2.10	2.52	2.31
Great Britain	2.21	3.34	3.97	3.67
Italy	0.49	0.53	0.64	0.59
Japan	0.75	0.73	0.81	0.76
Netherlands	2.89	3.47	3.94	3.68
South Korea	0.96	0.72	0.89	0.82
Spain	0.70	0.60	0.80	0.70
Sweden	1.69	2.82	3.44	3.16
Switzerland	2.24	2.73	3.37	3.03
United States	7.02	12.43	15.24	13.79

* Selected (meta-)countries of origins: Arabic (meta), Chinese (meta), Indian (meta), Iran, Poland, Romania, Russian (meta), Turkey, Vietnam
§ Column (2): max precision ; Column (3) : balanced; Column (3) : max recall [see section 4.1.2 for details]

4.2.2 Immigrant inventors' contribution to destination countries

For each inventor in the database we produce two productivity measures:

i. An "outstanding productivity" binary variable, based upon the frequency distribution of inventors by total number of patent applications filed at the EPO and year of entry into the database (entry = first patent application filed at EPO by the inventor): an inventor is said to exhibit outstanding productivity if he/she falls into the top 95% of the distribution.
ii. The patent h-index (number of patent receiving a number of citations equal to or higher than *h*).[27]

We then conduct two simple regression exercises (to be interpreted as exploratory partial correlations), both of them limited to the set of 12 European countries with the largest patent stock (see section 3) and the US, for comparison.

The first exercise consists of a Logit regression, where "outstanding productivity" is the dependent variable, and regressors are simple dummies indicating:

- whether the inventor is of foreign origin; in particular, we experiment with two sets of dummies:
  - the first set (*foreign by entry cohort*) consists of four dummies taking value 1 if the inventor has a country of origin different than that of residence and belongs to one out of four cohorts of entry (1985-90, 1991-95, 1996-00, 2001-05); the distinction by cohort of entry is meant to capture any potential change in the composition (by skills or countries of origin) of IFO over time
  - the second set of dummies indicate separately each selected countries of origin

  We experiment with both high precision and high recall estimates of the foreign origin on inventors. In particular, we consider both high precision and high recall *foreign by entry cohort* dummies (while for dummies representing selected countries of origin we rely exclusively on high precision estimates);
- dummies for the inventor's entry year, irrespective of whether the inventor is of foreign origin or native, which are meant to capture any calendar effect on the distribution of inventors by productivity;
- technology dummies, which take value 1 for each technological class in which the inventor has at least one patent (multiple classifications for the same patent are possible). They control for cross-technology differences in the average number of patents per invention (due to differences in patent scope and in the relative proportion of occasional vs. professional inventors).[28]

The second econometric exercise we conduct concerns the h-index measure of contribution to inventive activity. Being it a count variable (with a maximum around 25) we run a Poisson regression, where the explanatory variables are the same set of dummies of the previous exercise. While Poisson regressions may return negatively biased estimates of zero values, they perform as well as other count regression techniques (while saving computational time) for large count values, which are the ones in which we are most interested into (Long, 1997).

Table 4 reports the key summary statistics, separately for Europe (selected countries) and the US. All variables in the table (but h-index) are dummies, so their (min,max) values are always (0,1) (for the h-index the minimum is always zero, while the maximum is 23 for the US and 27 for Europe).

[27] The h-index, first proposed by Hirsch (2005), is a synthetic indicator of scientific productivity commonly used to rank individual scientists in a large number of disciplines (see also: Bornmann and Daniel, 2005 and 2007). The number of citations received by a scientific paper indicates the quality of the latter, in terms of potential for further research. While the total or average number of citations received by a scientific author may depend on just one or a very few highly cited items, the h-index better captures the overall quality of the authors total production. Being a stock measure, however, it is positively correlated with the author's age. In a similar fashion, citations to patent documents are interpreted as indicators of the economic or technological impact of the patent (Breschi and Lissoni, 2004; Hall et al., 2005).

[28] We consider seven technological classes (*Electrical engineering & Electronics ; Scientific & Control Instruments; Chemicals & Materials; Pharmaceuticals & Biotechnology; Industrial processes; Mechanical Engineering, Machine & Transport; Consumer good; .Civil engineering)* , based of the original IPC (International Patent Classification) codes assigned to each patent and produced by OST (2008)

Table 4: Summary statistics

	US			Europe		
Variable	Obs	Mean	Std Dev	Obs	Mean	Std Dev
top5	526850	0.047	0.212	700427	0.049	0.049
h index	526850	0.970	1.062	700427	0.890	0.890
Foreign (1985-90 cohort) – max precision	526850	0.029	0.167	700427	0.010	0.010
Foreign (1991-95 cohort) – max precision	526850	0.035	0.185	700427	0.011	0.011
Foreign (1996-00 cohort) – max precision	526850	0.064	0.245	700427	0.021	0.021
Foreign (2001-05 cohort) – max precision	526850	0.101	0.301	700427	0.032	0.032
Foreign (1985-90 cohort) – max recall	526850	0.056	0.230	700427	0.018	0.018
Foreign (1991-95 cohort) – max recall	526850	0.064	0.245	700427	0.019	0.019
Foreign (1996-00 cohort) – max recall	526850	0.107	0.309	700427	0.033	0.033
Foreign (2001-05 cohort) – max recall	526850	0.161	0.367	700427	0.049	0.049
Chinese	526850	0.053	0.223	700427	0.004	0.004
Iran	526850	0.004	0.066	700427	0.001	0.001
Poland	526850	0.010	0.100	700427	0.003	0.003
Romania	526850	0.002	0.040	700427	0.001	0.001
Russian	526850	0.011	0.102	700427	0.004	0.004
Turkey	526850	0.002	0.049	700427	0.001	0.001
Indian	526850	0.046	0.209	700427	0.004	0.004
Arabic	526850	0.004	0.060	700427	0.002	0.002
Other_foreign	526850	0.257	0.437	700427	0.100	0.100

[§] Europe= Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Spain, Sweden, Switzerland, United Kingdom

The top section of table 5 (which define IFOs on the basis of high precision algorithms) shows that both in Europe and in the US IFOs have a higher-than-average probability of exhibiting outstanding productivity (columns 1 and 2). Odds ratios are pretty similar for the two regions and well over one for all IFO cohorts. Notice however that, due to our over-estimation of the number of IFOs in the US, the latter are more likely to include false positives. Under the hypothesis that IFOs are more productive, on average, than native inventors, this should introduce a negative bias in our estimates of IFOs' productivity in the US.

When we break down the data by European country, the results cease to hold for a few cohorts, which vary across country, albeit the odds ratios remain always positive, with three exceptions (the 2001-05 cohorts in France, Netherlands, and Italy). The results are particularly weak for the countries with low absolute number of foreign inventors such as Italy, whose IFO's share of total inventors is low; or Sweden and Switzerland, which have higher IFO's share, but have relatively few inventors (in the case of Switzerland, there is also the additional complication that the inventors of German or Austrian origin were counted as IFO, due to the high number of false positives in the Ethnic-Inv database).

The bottom section of table 5 illustrates the problem we may encounter when changing algorithm to define IFO. There we make use of high recall algorithms, which reduce the number of false negatives (foreign inventors mistaken for locals) at the cost of increasing the number of false positives (local inventors mistaken for foreign). Results in columns 1 and 2 resist (which is somewhat reassuring in terms of robustness of our analysis) but the value of the odds ratios is closer to one, compared to the top section. This is generally true of all columns, too. We also observe a higher number of non-significant estimates. The explanation is as follows: by "mixing up" too many local inventors with foreign ones, we dilute the relationship between foreign origin (as defined by the algorithm) and productivity. The only exception to this change is the United Kingdom. We do not have yet a clear explanation for this exception, but we observe that the United Kingdom is a very special case, in which several immigrants from the US, Ireland, and former Commonwealth countries get confounded with locals. It may well be that the high precision algorithm leaves out too many of them (and indeed it is a very different algorithm than the one used for other countries), so that the high recall algorithm may do a better job in identifying highly productive IFO.

Table 6 reports the results of the same exercise of the bottom block of table 5, but for specific countries of origins, those whose languages (in particular, the languages used for names and surnames) do not coincide with any language used in the countries of destination of our interest. In the future, we plan to refine this analysis by considering also other countries of origins, including some western European ones (such as Italy or Greece, whose *hs* emigration within Europe is remarkable) as well as several Central and Eastern European ones now excluded (most notably, the Czech Republic, joint with Slovakia, as well as Hungary, Albania, and Bulgaria).

The results we obtain suggest a few interesting patterns, possibly as result of different dyadic relationships between countries of origin and countries of destination in the US and Europe (and across European countries). The US seems to attract highly productive inventors from all the countries explicitly considered in the table, with the partial exception of Romania. Europe fails to attract this type of IFO from Iran and Turkey and, partially, from Arabic countries. Intuitively, this result can be explained by that fact that Europe hosts large groups of second generation immigrants from these countries, who enter the S&E labour market along with locals, instead of self-selecting them on the basis of skill as first generation *hs* immigrants do (they enter the countries of destination after having received all or a large part of their education in their home country or abroad; see section 2). A similar explanation may apply to the results for Indians in the UK. At the same time, this result is in line with the observation that *hs* skilled migration has a preference for the US, and that the large presence of home country minorities in Europe fails to compensate for the US' attraction power (the best example is that of Turkey, whose *hs* emigrants have a well-documented preference for the US compared to Germany, despite the large Turkish minority in the latter).

Table 5: Foreign origin and outstanding productivity: Logit regression on cohorts of immigrants (dep. variable: inventor's probability to fall in top 5% of the distribution by nr of patents; odds ratios reported)

	(1) US	(2) Europe[§]	(3) Germany	(4) France	(5) UK	(6) Netherl.	(7) Italy	(8) Sweden	(9) Switzerl.
High precision definition of foreign origin									
Foreign (1985-90 cohort)	1.253***	1.303***	1.316***	1.387***	1.260	1.533***	1.329	1.289	1.014
	(0.0453)	(0.0652)	(0.124)	(0.174)	(0.193)	(0.239)	(0.369)	(0.327)	(0.148)
Foreign (1991-95 cohort)	1.320***	1.344***	1.090	1.251*	1.251	2.018***	2.061***	1.523**	1.401**
	(0.0456)	(0.0655)	(0.109)	(0.158)	(0.172)	(0.325)	(0.480)	(0.309)	(0.200)
Foreign (1996-00 cohort)	1.504***	1.499***	1.382***	1.440***	1.617***	1.735***	1.457	1.161	1.067
	(0.0448)	(0.0560)	(0.108)	(0.159)	(0.159)	(0.170)	(0.338)	(0.169)	(0.129)
Foreign (2001-05 cohort)	1.381***	1.318***	1.181**	0.980	1.849***	0.999	0.732	1.509***	1.273**
	(0.0447)	(0.0517)	(0.0969)	(0.117)	(0.190)	(0.124)	(0.227)	(0.223)	(0.145)
Technology controls	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*Y*
Year controls	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*Y*
Constant	0.00506***	0.00445***	0.00455***	0.00391***	0.00385***	0.00333***	0.00613***	0.00347***	0.00353***
	(0.000213)	(0.000160)	(0.000266)	(0.000344)	(0.000370)	(0.000517)	(0.000788)	(0.000638)	(0.000574)
High recall definition of foreign origin									
Foreign (1985-90 cohort)	1.181***	1.170***	1.066	1.243**	1.273*	1.431**	0.884	1.075	1.152
	(0.0357)	(0.0477)	(0.0766)	(0.119)	(0.167)	(0.200)	(0.223)	(0.228)	(0.144)
Foreign (1991-95 cohort)	1.200***	1.233***	1.127	1.166	1.274**	1.832***	1.575**	1.244	1.232
	(0.0363)	(0.0496)	(0.0821)	(0.115)	(0.153)	(0.267)	(0.345)	(0.221)	(0.162)
Foreign (1996-00 cohort)	1.382***	1.355***	1.179***	1.269***	1.533***	1.715***	1.427*	1.136	1.049
	(0.0385)	(0.0434)	(0.0725)	(0.115)	(0.139)	(0.155)	(0.287)	(0.141)	(0.115)
Foreign (2001-05 cohort)	1.275***	1.254***	1.180***	1.046	1.635***	0.886	0.922	1.346**	1.111
	(0.0392)	(0.0425)	(0.0757)	(0.0978)	(0.162)	(0.0995)	(0.226)	(0.180)	(0.120)
Technology controls	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*Y*
Year controls	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*Y*
Constant	0.00497***	0.00446***	0.00459***	0.00389***	0.00383***	0.00329***	0.00620***	0.00353***	0.00340***
	(0.000213)	(0.000161)	(0.000269)	(0.000343)	(0.000368)	(0.000513)	(0.000796)	(0.000649)	(0.000558)
Observations	526,411	699,944	252,644	114,193	86,178	46,908	47,291	31,578	35,498

se in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

[§] Europe= Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Spain, Sweden, Switzerland, United Kingdom

Table 6: Foreign origin and outstanding productivity: Logit regression on specific countries of origin of immigrants – high recall definition of foreign origin (dep. variable: inventor's probability to fall in top 5% of the distribution by nr of patents; odds ratios reported)

	(1) US	(2) Europe[§]	(3) Germany	(4) France	(5) UK	(6) Netherl.	(7) Italy	(8) Sweden	(9) Switzerl.
Chinese	1.520***	1.418***	1.271	1.089	1.525**	1.673**	0.664	1.725**	1.856*
	(0.0452)	(0.131)	(0.270)	(0.327)	(0.278)	(0.381)	(0.673)	(0.458)	(0.603)
Iran	1.514***	1.208	1.093	0.827	0.934	2.514		1.235	0.995
	(0.149)	(0.194)	(0.324)	(0.390)	(0.350)	(1.411)		(0.560)	(0.758)
Poland	1.187**	1.290**	1.090	0.778	1.456	2.614**		2.127*	2.007
	(0.0811)	(0.137)	(0.182)	(0.235)	(0.400)	(1.031)		(0.824)	(0.885)
Romania	1.367*	1.582**	1.548	0.554	3.209**	5.994***			
	(0.253)	(0.308)	(0.518)	(0.344)	(1.624)	(2.525)			
Russian	1.292***	1.450***	1.829***	1.165	2.020***	2.079***	0.0775**	1.202	0.623
	(0.0865)	(0.136)	(0.272)	(0.346)	(0.547)	(0.559)	(0.0888)	(0.430)	(0.269)
Turkey	1.856***	1.072	1.005	1.603	1.409	1.256		3.904*	0.912
	(0.234)	(0.182)	(0.228)	(0.860)	(0.797)	(0.884)		(3.177)	(0.569)
Indian	1.561***	1.436***	1.335	1.380	1.213	2.586***	0.717	1.525	1.212
	(0.0498)	(0.122)	(0.310)	(0.500)	(0.167)	(0.489)	(0.509)	(0.538)	(0.452)
Arabic	1.617***	1.243*	1.604	0.914	1.178	2.431*		2.065	1.717
	(0.168)	(0.158)	(0.484)	(0.180)	(0.509)	(1.209)		(1.390)	(0.741)
Other_foreign	1.150***	1.246***	1.115***	1.225***	1.498***	1.256***	1.347**	1.139	1.110*
	(0.0198)	(0.0244)	(0.0406)	(0.0615)	(0.0936)	(0.0794)	(0.156)	(0.0956)	(0.0674)
Year controls	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*
Technology controls	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*	*y*
	(0.0198)	(0.0244)	(0.0406)	(0.0615)	(0.0936)	(0.0794)	(0.156)	(0.0956)	(0.0674)
Constant	0.00494***	0.00443***	0.00456***	0.00390***	0.00380***	0.00335***	0.00611***	0.00346***	0.00343***
	(0.000207)	(0.000158)	(0.000266)	(0.000343)	(0.000364)	(0.000515)	(0.000786)	(0.000634)	(0.000552)
Observations	526,411	699,944	252,644	114,193	86,178	46,908	47,203	31,551	35,476

se in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

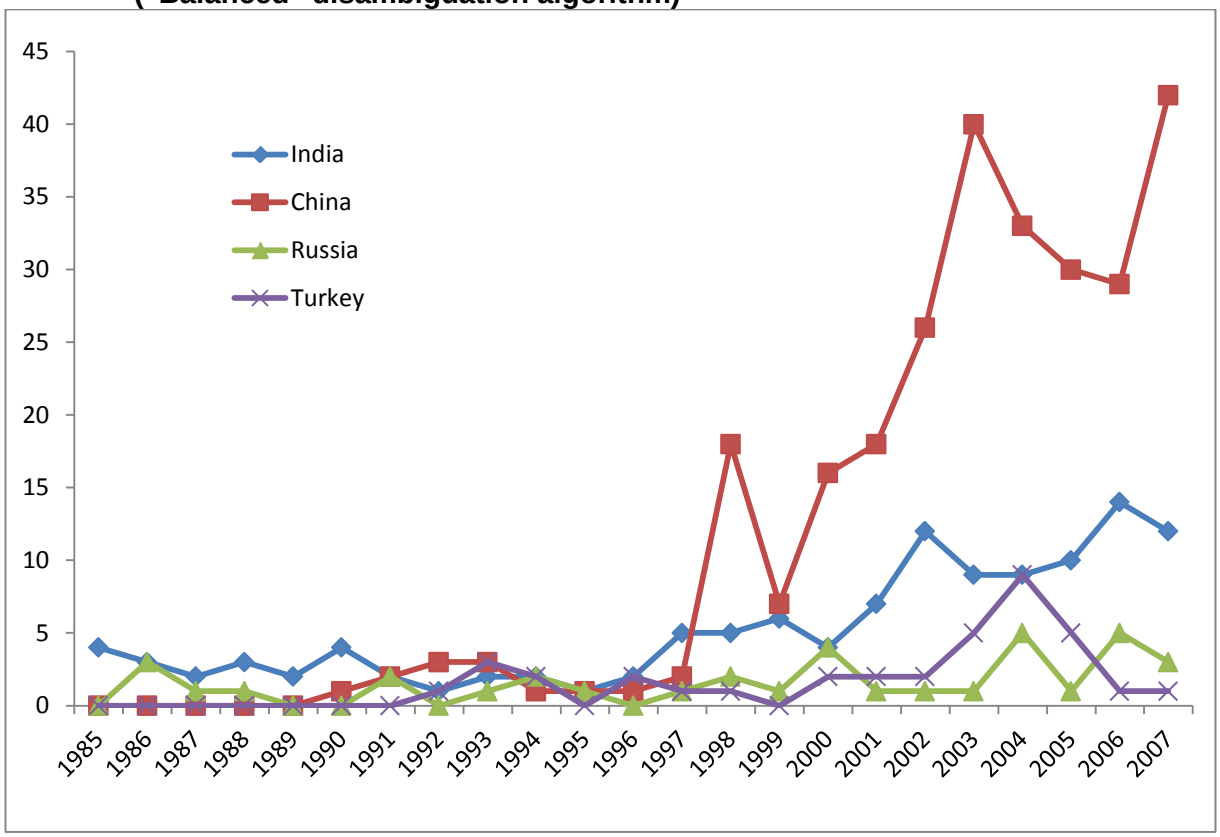
[§] Europe= Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Spain, Sweden, Switzerland, United Kingdom

Among the European countries of origin, the most pervasive presence of highly productive IFO is that of Russians (among the top productive inventors in four countries out of seven). Among the countries of destination, those who make the most of IFO from the countries of origin considered are the Netherlands and the UK. At the opposite end, we find France.

4.2.3 Returnee inventors

While the statistics examined so far concern the impact of immigrant inventors on innovation within their countries of destination, a large part of the literature we reviewed in section 2 concerns their contribution to innovation in their countries of origin. As discussed in section 2, a potential contribution channel consists of returnee S&Es who innovate in their home countries, based on their experience in the former country or countries of destination. Table 7 reports our estimates of the phenomenon, according to two different versions of the Ethnic-Inv database. The former version is the one based on a "balanced" name disambiguation algorithm, the second on the "high recall" one (both make use of a "high precision" algorithm for the identification of foreigners). Returnee inventors are defined as inventors with at least two patents, one of which filed as resident in a country different from that of origin and another filed as resident in the country of origin, the former having been filed not later than the latter. This definition excludes inventors who filed patents first as residents in their country of origin and then abroad (they being emigrants, not returnees), but it includes both inventors who, after having returned to their country of origin, keep patenting as residents of other countries, and inventors who patent at the same time as residents in their country of origin and elsewhere.

Figure 3. Number of returnee inventors per year(1), 1985-2007 – selected countries ("Balanced" disambiguation algorithm)(2)



(1) Only inventors with at least two patents are considered. Returnees are defined as inventors with at least one patent in a country different from the country of origin and one patent in the country of origin (the latter not being filed before the former). Year of return is the year of first patent filing from origin country

(2) Balanced disambiguation: unique inventors identified with an algorithm with estimated 88% precision rate [→ true positives/true+false positives] and 68% recall rate [→ true positives/(true positives + false negatives)

Table 7. Returnee inventors[1], by selected countries of origin and disambiguation algorithm

	"Balanced" disambiguation[2]			"High recall" disambiguation[3]				
	nr returnee inventors	returnees as % of emigrant inv.	returnees as % of resident inv.	nr returnee inventors	returnees as % of emigrant inv.	returnees as % of resident inv.		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(5)/(2)*	*(6)/(3)*
China	220	3.09	0.88	1499	19.15	8.62	6.2	9.8
Iran	11	0.79	5.70	14	1.04	8.19	1.3	1.4
Poland	14	0.37	0.43	41	1.11	1.40	3.0	3.3
Romania	6	0.67	1.11	17	1.95	3.53	2.9	3.2
Russia	32	1.03	0.31	76	2.45	0.81	2.4	2.6
Ukraine	11	5.67	1.52	16	8.16	2.33	1.4	1.5
Turkey	27	2.23	1.21	47	4.02	2.84	1.8	2.3
India	92	0.74	0.65	407	3.41	4.13	4.6	6.4
Pakistan	3	0.56	6.25	3	0.57	6.67	1.0	1.1
Algeria	0	0	0	1	1.28	5.56	-	-
Morocco	6	2.39	5.36	11	4.35	11.00	1.8	2.1
Tunisia	1	0.62	1.10	4	2.38	5.19	3.8	4.7

(1) Only inventors with at least two patents are considered. Returnees are defined as inventors with at least one patent in a country different from the country of origin and one patent in the country of origin (the latter not being filed before the former).

(2) Balanced disambiguation: unique inventors identified with an algorithm with estimated 88% precision rate [→ true positives/true+false positives] and 68% recall rate [→ true positives/(true positives + false negatives)

(3) High recall disambiguation: unique inventors identified with an algorithm with estimated 56% precision rate, and 93% recall rate

The figures for returnee inventors we obtain from the "balanced" database are extremely low. The country of origin with the highest number of returnees is China (220 inventors), which amounts to around 3% of returnee rate (share of emigrant inventors who return) and less than 1% of impact on the country of origin (share of returnee inventors over inventors resident in the country of origin). In India, a much studied case, both rates are less than 1%. In all other cases, the number of returnees is negligible and we observe higher than 1% values only in countries with either few emigrant inventors or few resident ones.

The results change dramatically when using a "high recall" database. China is the country of origin most affected, but also the one for which estimates are most likely to suffer by low precision (as discussed in section 3). Here we move to over 1,400 returnees, with a returnee rate of 19% and an impact on country of origin of over 8%. India is also quite affected by the change of algorithm, much less the European countries, for which the precision-recall trade off may be less severe. While of no substantive interest, these results help clarifying the importance of technical issues concerning name disambiguation when studying the relationship between emigrant inventors and their countries of origin. Low figures for inventors active in such countries mean that any distortion introduced by the choice of one algorithm over another will be magnified, even in the case of large countries such as China or India.

Finally, and quite interestingly, figure 3 shows that the phenomenon of returnee inventorship can be meaningfully studied with the help of patent data only from the second half of the 1990s onward, at least for the case of less developed or developing countries of origin. Before then, none of these countries had adopted a patent legislation: while this did not impede inventor to patent abroad, it may imply some lack in the necessary legal and administrative infrastructure to support extensive patent filing at foreign patent offices. Besides, and more importantly, these same countries did not host, until recently, many R&D-performing companies, so that innovation, if occurred, did not leave any patent track behind it.

5. Conclusions and further research

The relationship between migration and innovation is an important one both for countries of destination and countries of origin. In this paper, we have reviewed the existing empirical evidence on the phenomena, with an eye mainly on methodological issues.

In particular, we have reviewed in detail sources of macro- and micro-data used in the literature and concluded that inventor data, to be extracted from large patent datasets, have a potential for shedding light on several issues. We have then provided a few examples of this potential by producing descriptive statistics and simple econometric exercises based on *Ethnic-Inv*, a pilot dataset based on inventor information extracted from patent filings at the European Patent Office (EPO), and IBM-GNR, a commercial database on ethnic origin of names and surnames. We have discussed at length the technical issue of name disambiguation, which is a crucial one for ensuring the minimum level of data quality, but has not yet received enough attention in applications of patent data to migration studies.

We see our effort to create Ethnic-Inv as complementary to a similar effort conducted by Kerr (2007) on the basis of USPTO data and a different commercial dataset of names. Kerr's data have been instrumental to shed light on the migration-innovation nexus involving the US and Asian countries, but cannot help overcoming what we saw as a main limitation of the existing literature, namely its US-centrism. While the US are certainly the country whose innovation activities have most benefitted (both historically and recently) by the influx of foreign S&Es, general data on highly skilled migration show that the latter matter in Europe, too. However, European countries of destination differ from the US in that most highly skilled migration originates from neighbouring countries, with richest countries serving both as destination and origin. This makes it necessary to identify separately each European country of origin, a task not performed by Kerr's data, which poses several technical challenges.

Despite the many limitations of our pilot database, the results we presented, albeit not yet robust, confirm its potential. First, we have seen that immigrant inventors represent a sizable group in the European countries of destination most active in patenting at the EPO. Figures for shares of immigrant inventors over residents are in the same order of magnitude of figures for comparable shares of highly skilled migration. Second, in several European countries foreign inventors rank high in terms of productivity, very much in line with what found for the US by Stephan and Levin (2001).

These results hold also when examining specific countries of origin. At the same time, though, they do not hold or are weaker for some countries (such as Italy) where highly skilled migration is traditionally low or for countries in which we may not be able to distinguish immigrant inventors from inventors from ethnic minorities (such as Germany, France, and Sweden). Finally, descriptive statistics suggest that the phenomenon of returnee inventorship is very limited, or at least difficult to capture with patent data, due to the high sensitivity to name disambiguation issues.

Besides improving the Ethnic-Inv database, our research plans for the immediate future will be mainly addressed at refining our findings on destination countries, especially European ones. They will be of immediate relevance to policy issues related to recent and less recent efforts to create a truly homogeneous European Research Area, its attractiveness for S&Es from Eastern Europe (compared to the US), and recent signals of a reprise of highly skilled migrations from Southern Mediterranean countries.

References

Agrawal, A., Cockburn, I. & Mchale, J. 2006. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. Journal of Economic Geography, 6, 571-591.

Agrawal, A., Kapur, D. & Mchale, J. 2008. How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data. Journal of Urban Economics, 64, 258-69.

Agrawal, A., Kapur, D., Mchale, J. & Oettl, A. 2011. Brain Drain or Brain Bank? The Impact of Skilled Emigration on Poor-Country Innovation. Journal of Urban Economics, 69, 43-55.

Almeida, P., Phene, A. & Li, S. 2010. Communities, knowledge and innovation: Indian immigrants in the US semiconductor industry. Working paper, 58/2010.

Alnuaimi, T., Opsahl, T. & George, G. 2012. Innovating in the periphery: The impact of local and foreign inventor mobility on the value of Indian patents. Research Policy, 41, 1534-1543.

Auriol, L. 2007. Labour market characteristics and international mobility of doctorate holders: results for seven countries. OECD STI Working Paper. Paris: OECD Publishing.

Auriol, L. 2010. Careers of doctorate holders: employment and mobility patterns. OECD STI Working Paper. Paris: OECD Publishing.

Bellini, E., Ottaviano, G. I. P., Pinelli, D. & Prarolo, G. 2013. Cultural diversity and economic performance: evidence from European regions. Geography, Institutions and Regional Economic Performance. Springer.

Black, G. C. & Stephan, P. E. 2010. The economics of university science and the role of foreign graduate students and postdoctoral scholars. American universities in a global market. University of Chicago Press.

Borjas, G. J. 2004. Do foreign students crowd out native students from graduate programs? National Bureau of Economic Research.

Borjas, G. J. 2009. Immigration in High-Skill Labor Markets: The Impact of Foreign Students on the Earnings of Doctorates. In: FREEMAN, R. B. & GOROFF, D. L. (eds.) Science and Engineering Careers in the United States: An Analysis of Markets and Employment. University of Chicago Press.

Bornmann, L., Daniel, H.D. (2005). Does the h-index for ranking of scientists really work? Scientometrics, 65(3), 391-392.

Bornmann, L., Daniel, H.D. (2007). What do we know about the h index? Journal of the American Society for Information Science and technology,58(9), 1381-1385.

Breschi, S. & Lissoni, F. 2005a. "Cross-Firm" Inventors and Social Networks: Localized Knowledge Spillovers Revisited. Annals of Economics and Statistics / Annales d'Économie et de Statistique, 189-209.

Breschi, S. & Lissoni, F. 2005b. Knowledge networks from patent data. In: MOED, H. F., GLÄNZEL, W. & SCHMOCH, U. (eds.) Handbook of quantitative science and technology research. Berlin: Springer Science+Business Media.

Breschi, S. & Lissoni, F. 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. Journal of Economic Geography, 9, 439-468.

Chaloff, J. & Lemaitre, G. 2009. Managing highly-skilled labour migration: a comparative analysis of migration policies and challenges in OECD countries. OCED Social, Employment and Migration WP. Paris: OECD Publishing.

Chellaraj, G., Maskus, K. E. & Mattoo, A. 2008. The Contribution of International Graduate Students to US Innovation. Review of International Economics, 16, 444-62.

Cheshire, J., Mateos, P. & Longley, P. A. 2011. Delineating Europe's Cultural Regions: Population Structure and Surname Clustering. Human Biology, 83.

David, P. A. 1993. Intellectual property institutions and the Panda's thumb: Patents, copyrights, and trade secrets in economic theory and history, in: WALLERSTEIN, M.B., MOGEE, M-E., SCHOEN, R.A. (eds.).*Global dimensions of intellectual property rights in science and technology*. National Academies Press

De Haas, H. (2010) "Migration and Development: A Theoretical Perspective", International Migration Review 44.1: 227-264.

Docquier, F. & Marfouk, A. 2006. International migration by educational attainment (1990-2000). In: ÖZDEN, Ç. & SCHIFF, M. (eds.) International migration, remittances and the brain drain. New York: The World Bank - Palgrave Macmillan.

Docquier, F. & Rapoport, H. 2012. Globalization, Brain Drain, and Development. Journal of Economic Literature, 50, 681-730.

Docquier, F., Lowell, B. L. & Marfouk, A. 2009. A gendered assessment of highly skilled emigration. Population and Development Review, 35, 297-321.
Fink, C. & Maskus, K. E. 2005. Intellectual property and development: lessons from recent economic research, World Bank Publications.
Foley, C. F. & Kerr, W. R. 2011. Ethnic Innovation and US Multinational Firm Activity. NBER working paper. National Bureau of Economic Research, Inc.
Franzoni, C., Scellato, G. & Stephan, P. 2012. Foreign Born Scientists: Mobility Patterns for Sixteen Countries. National Bureau of Economic Research Working Paper Series, No. 18067.
Freeman, R. B. 2010. Globalization of scientific and engineering talent: international mobility of students, workers, and ideas and the world economy. Economics of Innovation and New Technology, 19, 393-406.
Halary, C. (1994). *Les exilés du savoir: les migrations scientifiques internationales et leurs mobiles*. Harmattan.
Hall, B. H. (2013). Discussion of: Inventor Data for Research on Migration and Innovation, mimeo.
Hall, B. H., Jaffe, A. B. & Trajtenberg, M. 2001. The NBER patent citation data file: Lessons, insights and methodological tools. National Bureau of Economic Research.
Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569–16572.
Hunt, J. 2009. Which Immigrants Are Most Innovative and Entrepreneurial? Distinctions by Entry Visa. *NBER Working Papers.*
Hunt, J. 2013. Are Immigrants the Best and Brightest U.S. Engineers? *National Bureau of Economic Research Working Paper Series,* No. 18696.
Hunt, J. & Gauthier-Loiselle, M. 2010. How Much Does Immigration Boost Innovation? American Economic Journal: Macroeconomics, 2, 31-56.
Jaffe, A. B., Trajtenberg, M. & Henderson, R. 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. The Quarterly Journal of Economics, 108, 577-598.
Kenney, M., Breznitz, D. & Murphree, M. 2013. Coming back home after the sun rises: Returnee entrepreneurs and growth of high tech industries. Research Policy, 42, 391-407.
Kerr, W. 2009. The Agglomeration of US Ethnic Inventors. NBER Working Papers.
Kerr, W. R. 2007. The Ethnic Composition of US Inventors.
Kerr, W. R. 2008. Ethnic Scientific Communities and International Technology Diffusion. Review of Economics and Statistics, 90, 518-537.
Lai, R., D'amour, A., Yu, A., Sun, Y., Torvik, V. & Fleming, L. 2011. Disambiguation and co-authorship networks of the US Patent Inventor Database. Harvard Institute for Quantitative Social Science, Cambridge, MA, 2138.
Lasker, G. W. 1977. A coefficient of relationship by isonymy: a method for estimating the genetic relationship between populations. Human Biology, 49.
Lissoni, F. 2012. Academic patenting in Europe: An overview of recent research and new perspectives. World Patent Information, 34, 197-205.
Lissoni, F., Llerena, P. & Sanditov, B. 2011. Small Worlds in Networks of Inventors and the Role of Science: An Analysis of France. Bureau d'Economie Théorique et Appliquée, UDS, Strasbourg.
Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. SAGE Publications, Thousand Oaks CA.
Marx, M., Strumsky, D. & Fleming, L. 2009. Mobility, skills, and the Michigan non-compete experiment. Management Science, 55, 875-889.
Mateos, P., Longley, P. A. & O'sullivan, D. 2011. Ethnicity and population structure in personal naming networks. PloS one, 6, e22943.
Meyer, J.-B. 2001. Network Approach versus Brain Drain: Lessons from the Diaspora. International Migration, 39, 91-110.
Miguelez E. (2013) Inventor diasporas and the internationalization of technology, paper presented at the "Patent Statistics for Decision Makers" conference, Rio de Janeiro, November 12 & 13, 2013 (http://ernestmiguelez.com/job-market-paper/)
Miguelez, E. & Fink, C. (2013) Measuring the International Mobility of Inventors: A New Database. World Intellectual Property Organization-Economics and Statistics Division
More (2010), Study on mobility patterns and career paths of EU researchers, Final Report to the European Commission – Research Directorate, Brussels (http://ec.europa.eu/euraxess/pdf/research_policies/MORE_final_report_final_version.pdf; last visited: April, 2013)

Moser, P., Voena, A. & Waldinger, F. 2011. German-Jewish émigrés and US invention. mimeo (available at SSRN 1910247).
Nerenberg, S., Williams, K. (2013) The Case for Analytical Name Scoring over Name Variant Expansion, IBM® InfoSphere Global Name Management report, IBM Corporation Armonk, NY
Niebuhr, A. 2010. Migration and Innovation: Does Cultural Diversity Matter for Regional R&D Activity? Papers in Regional Science, 89, 563-85.
No, Y., Walsh, J. P. 2010. The importance of foreign-born talent for US innovation. *Nature biotechnology*, *28*(3), 289-291.
Ottaviano, G. I. P. & Peri, G. 2006. The economic value of cultural diversity: evidence from US cities. Journal of Economic Geography, 6, 9-44.
Ozgen, C., Nijkamp, P. & Poot, J. 2011. Immigration and innovation in European regions. Discussion paper series//Forschungsinstitut zur Zukunft der Arbeit.
Patman, F. (2010) Advanced Global Name Recognition Technology, IBM® InfoSphere Global Name Management report, IBM Corporation Armonk, NY
Pezzoni, M., Lissoni, F. & Tarasconi, G. 2012. How To Kill Inventors: Testing The Massacrator© Algorithm For Inventor Disambiguation. Cahier du GREThA nr. 29. Groupe de Recherche en Economie Théorique et Appliquée – Université Bordeaux IV.
Piazza, A., Rendine, S., Zei, G., Moroni, A. & Cavallisforza, L. L. 1987. MIGRATION RATES OF HUMAN-POPULATIONS FROM SURNAME DISTRIBUTIONS. Nature, 329, 714-716.
Raffo, J. & Lhuillery, S. 2009. How to play the "Names Game": Patent retrieval comparing different heuristics. Research Policy, 38, 1617-1627.
Razum, O., Zeeb, H. & Akgün, S. 2001. How useful is a name-based algorithm in health research among Turkish migrants in Germany? Tropical Medicine & International Health, 6, 654-661.
Roach, M., & Cohen, W. M. (2013). Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research. *Management Science*,*59*(2), 504-525.
Scellato, G., Franzoni, C. & Stephan, P. 2012. Mobile Scientists and International Networks. National Bureau of Economic Research Working Paper Series, No. 18613.
Stephan, P. 2012. How economics shapes science, Harvard University Press.
Stephan, P. E. & Levin, S. G. 2001. Exceptional contributions to US science by the foreign-born and foreign-educated. Population Research and Policy Review, 20, 59-79.
Stuen, E. T., Mobarak, A. M. & Maskus, K. E. 2012. Skilled Immigration and Innovation: Evidence from Enrolment Fluctuations in US Doctoral Programmes*. The Economic Journal, 122, 1143-1176.
Su, X. 2012. International Doctoral Science and Engineering Students: Impact on Cohorts' Career Prospects. Journal of Studies in International Education.
Thompson, P. & Fox-Kean, M. 2005. Patent citations and the geography of knowledge spillovers: A reassessment. American Economic Review, 450-460.
Wadhwa V., Jasso G., Rissing B. A., Gereffi G., Freeman R. B. (2007c) Intellectual Property, the Immigration Backlog, and a Reverse Brain-Drain (America's New Immigrant Entrepreneurs: Part III), mimeo, Duke University
Wadhwa V., Rissing B. A., Saxenian A., Gereffi G. (2007b) Education, Entrepreneurship and Immigration (America's New Immigrant Entrepreneurs: Part II), mimeo, Duke University
Wadhwa V., Saxenian A., Freeman R. B., Gereffi G. (2009a) America's Loss is the World's Gain (America's New Immigrant Entrepreneurs), mimeo, Duke University
Wadhwa V., Saxenian A., Freeman R. B., Salkever A. (2009b) Losing the World's Best and Brightest (America's New Immigrant Entrepreneurs), mimeo, Duke University
Wadhwa V., Saxenian A., Rissing B. A., Gereffi G. (2007a), America's New Immigrant Entrepreneurs: Part I. Duke Science, Technology & Innovation Paper No. 23, Duke University
Widmaier, S. & Dumont, J.-C. 2011. Are recent immigrants different? A new profile of immigrants in the OECD based on DIOC 2005/06, Paris, OECD Publishing
Zheng, Y. & Ejermo, O. (2013), Characteristics of Foreign-born Inventors in Sweden., paper presented at the 8th DRUID Academy, January 16-18, Aalborg (http://druid8.sit.aau.dk/acc_papers/jlg2f36g3171tofl73ae4hteqblr.pdf)