Foreign inventors in the US: Testing for Diaspora and Brain Gain Effects

Stefano Breschi [1], Francesco Lissoni [1/2/§], Ernest Miguelez [2/3/4]

[1] *CRIOS – Università Bocconi, Milan*

[2] *GREThA UMR CNRS 5113 – Université de Bordeaux*

[3] *AQR-IREA (Barcelona, ES)*

[4] *CReAM (London, UK)*

[§] contact author: francesco.lissoni@u-bordeaux.fr

This version: *29 January 2015*

To be presented at: Berkeley Innovation Seminar, Fung Institute for Engineering Leadership, UC Berkeley (March 4, 2015)

Please do not quote or circulate

Abstract

We assess the role of ethnic ties in the diffusion of technical knowledge by means of a large database of EPO patents filed by US-resident inventors of foreign origin ("ethnic" inventors), 1980-2010, which we identify both through linguistic analysis of names and surnames and information on nationality. We consider ten important countries of origin of highly skilled migration to the US, both Asian and European. We test whether ethnic inventors' patents are disproportionately cited by: (i) co-ethnic migrants in the same destination country ("diaspora" effect); and (ii) non-migrant inventors residing in their country of origin ("brain gain" effect). We find evidence of the diaspora effect for China, India, and, to less extent, other Asian countries, but not for European ones, with the exception of Russia. Diaspora effects, however, do not translate necessarily into a brain gain effect, most notably for India. Some evidence exists of an "international diaspora" effect, by which co-ethnic inventors in different destination countries (but not necessarily the country of origin) cite disproportionately each other's patents. It remains to investigate the role of ethnic ties in the formation of inventors' networks.

Keywords: migration, brain gain, diaspora, diffusion, inventors, patents

JEL codes: F22, O15, O31

ACKNOWLEDGEMENTS: Unique identifiers for inventors in the EP-INV database come from the APE-INV project (Academic Patenting in Europe), funded by the European Science Foundation. The pilot project for assigning inventors to ethnic group was funded by the World Intellectual Property Organization (WIPO), which also made available to us the WIPO-PCT dataset. Discussants at the first WIPO *Experts Meeting on IP and the Brain Drain* (Geneva, April 2013) provided useful suggestions. The same apply to participants to the following conferences : MEIDE (Santiago de Chile, November 2013), PATSTAT (Rio de Janeiro, November 2013), EUROLIO (Utrecht, January 2014) and EPIP (Brussels, September 2014); as well to participants to seminars at University College Dublin, LSE, CRIOS-Bocconi, Kassel University, Collegio Carlo Alberto (Turin), IMT (Lucca), LUISS (Rome) and GREThA. Gianluca Tarasconi provided outstanding technical assistance and silly jokes. We owe the discovery of the IBM-GNR© system to Lars Bo Jeppesen, while Curt Baginski assisted us in its implementation.

1. Introduction

The last decade has seen the convergence of two important streams of literature dealing with the diffusion of technical knowledge and the mobility of scientists and engineers. First, research in the geography of innovation has explored the role of social ties in facilitating knowledge diffusion, and in determining its spatial reach. Among such ties, a good deal of attention has been paid to those binding members of scientific and technical "diasporas", namely the communities of migrant scientists and engineers from the same origin country (Agrawal et al., 2008; Kerr and Lincoln, 2010; Saxenian et al., 2002). Second, migration and development scholars have explored to what extent these diasporas contribute to innovation in their home countries, through international knowledge flows (Kapur, 2001; Kuznetsov, 2006; Saxenian et al., 2002). Emerging naming conventions label the social ties in question as "ethnic", a synthetic but imperfect adjective we will also adopt, for want of better alternatives.

The two streams of literature share a common necessity in going beyond anecdotal evidence and success stories. This requires measuring the importance of ethnic ties as vehicles for knowledge diffusion, and assessing the relative weight of their multiple embodiments. The latter comprise multinational firms operating in both the destination and home countries of migrants (Foley and Kerr, 2011), several academic and professional exchange networks (Meyer, 2001; Meyer and Brown, 1999), as well as returnee migration (Alnuaimi et al., 2012; Nanda and Khanna, 2010), often tied to entrepreneurial ventures (Saxenian, 2006; for a skeptical view, see: Kenney et al., 2013).

Patent and inventor data have been increasingly used to address these measurement issues. Migrant inventors are identified as such either by using information on their nationality, available on Patent Cooperation Treaty (PCT) applications (Miguelez, 2014; Wadhwa et al., 2007b); or by linguistic analysis of their names and surnames (Agrawal et al., 2008; Agrawal et al., 2011; Foley and Kerr, 2011; Kerr, 2008; Kerr and Lincoln, 2010).

So far, however, both streams of literature have focussed almost exclusively on the US as a destination country, and on China and India as origin countries of highly skilled migrants in general, and migrant inventors in particular. This overlooks the fact that several European countries are also important sources of highly skilled migrants to the US; and that Europe hosts quite robust flows of intra-continental migration (Docquier and Marfouk, 2006; Widmaier and Dumont, 2011). The focus on China and India is also at the origin of present difficulties in assessing whether the evidence on the role of those countries' diasporas can be generalized to other countries (Pandey et al., 2006).

In this paper we contribute both substantively and methodologically to this emerging field by analysing the forward citation patterns of patents filed by foreign inventors in the US from five Asian countries (China, India, Ira, Japan, and South Korea) and as many European ones (France, Germany, Italy, Poland, and Russia). All our data are novel and come from EP-INV, a database of uniquely identified inventors listed on patent applications filed at the European Patent Office, combined with information from IBM-GNR© (Global Name Recognition system, courtesy of IBM), by means of an original algorithm. Complementary data come from PCT applications, as described by Miguelez and Fink (2013).

We test for the existence of "diaspora" and "brain gain" effects. We state a diaspora effect to exist when foreign inventors of the same ethnic group and active in the same country of destination (in our case, the US) have a higher propensity to cite one another's patents, compared to patents by other inventors, other things being equal. We state a "brain gain" effect to exist when the same foreign inventors' patents are

also disproportionately cited by inventors active in their countries of origin, so that the latter stand to gain from their migrants'" knowledge remittances". We find evidence of the diaspora effects for the Asian countries (especially for China and India), but not for the European ones, with the exception of Russia. Diaspora ties, however, do not appear to of primary importance in conveying knowledge diffusion (social ties established through co-inventorship carry more weight).

As for "brain gain" effects, our evidence is more mixed. Not all countries of origin that exhibit a diaspora effect also stand to gain in terms of absorption of knowledge generated by such diaspora. By contrast, we find that we find that multinationals play an important role. In a few cases, the absence of a brain gain effects stands in contrast with the existence of an "international diaspora" effect exists, by which migrants from the same country of origin dispersed across different destination countries cite disproportionately each other.

In what follows, we first survey the literature on migration, innovation, and knowledge flows, with special emphasis on patent-based studies (section 2). We then present our research questions and data (section 3). In section 4 we report the results of our empirical exercise. Section 5 discusses such results and concludes.

2. Background literature

2.1 Localized knowledge flows and the role of social ties

Localized knowledge flows are a key topic in the geography of innovation (surveys by Breschi and Lissoni, 2001; Breschi, 2011). Under the form of pure externalities, they play a key role in Marshallian and Jacobian location theories (Ellison et al., 2007; Henderson, 1997). Yet, their importance has been questioned both by New Economic Geography models (Krugman, 1991 and 2011) and by evolutionary theories of clustering (Boschma and Frenken, 2011). A key point of contention in the debate has been that of measurement, which is fraught with technical as well as conceptual difficulties.

As for technical difficulties, these were first tackled by Jaffe et al. (1993), who introduced the use of patent citations along with a simple, yet attractive methodology for testing their localization in space (from now on, JTH test). The test makes use of two set of patent pairs. The first one includes a sample of cited patents and all the related citing ones, with exclusion of self-citations at the company level (cited-citing or "case" pairs); the second includes the same sample of cited patents, with citing ones replaced by controls with the same technological classification and priority year (cited-control or "control" pairs). After geo-localising patents on the basis of their inventors' addresses, a simple test of proportions is carried out, one that proves the share of co-localized cases to be significantly higher than the share of co-localized controls (with co-localization specified either at the city, state, or country level). The test can be generalized by means of a regression analysis, with the probability of a citation to occur as the dependent variable, and the stacked sets of cited-citing and cited-control patent pairs as observations (Singh and Marx, 2013).[1]

[1] Technical refinements of the JTH test also concern the level of detail chosen for the technological classification of patents (Henderson et al., 2005) and the origin of patent citations (Alcácer and Gittelman, 2006; Breschi and Lissoni, 2005a; Thompson, 2006)Alcacer, J., Gittelman, M., 2006. Patent citations as a measure of knowledge flows: The influence of examiner citations. The Review of Economics and Statistics, 88(4), 774-779, Breschi, S., Lissoni, F., 2005, Knowledge networks from patent data. In: H.F. Moed, W. Glänzel, U. Schmoch (Eds.). Handbook of quantitative science and technology research. Springer Science+Business Media, Berlin, pp. 613-643, Thompson, P., 2006. Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations. The Review of Economics and Statistics, 88(2), 383-388..

Further research has been devoted to uncovering the actual mechanisms behind localized knowledge flows and their economic characteristics. Breschi and Lissoni (2009, 2005b) show that a large share of localized patent citations are self-citations at the individual level, associated to inventors who move across firms, or act as consultants for different firms in the same location or region[2]. Other localized citations occur between patents signed by socially close inventors, namely inventors located at short geodesic distances on networks of inventors[3]. These results have two implications:

1) They cast some doubts on the pure externality (spillover) interpretation of localized knowledge flows: market mechanisms may be in place, such as labour mobility and/or licensing. It may also be the case that mergers and acquisitions are at work, with some firms' knowledge being incorporated into new entities, along with its inventors (see section 2.3).

2) They suggest that spatial proximity is largely a proxy for social proximity, to be intended here as professional proximity (who worked with whom). Professional proximity may force inventors to share knowledge (as when they work together on the same project) or create strong enough social bonds to induce obligations or opportunities to share. Agrawal et al. (2006) show that professional ties may indeed resist to physical distance, as when new patents by inventors who relocated keep being cited by former co-inventors who did not.

This line of research has evolved in the direction of uncovering other forms of social ties besides the professional ones, and of exploring their relationship with spatial distance. This has led Agrawal et al. (2008) to explore the role of ethnic ties in the US-resident population of Indian inventors, which the literature describes as closely-knit "diaspora" (Kapur, 2001).

First, based on an Indian surname database, the authors identify ethnic Indian inventors of USPTO patents, all of them resident in the US (1981-2000). Second, they apply the JTH methodology and extend it to test not only the extent of knowledge flows' co-location (at the MSA level), but also the importance of ethnic ties[4]. Co-ethnicity of two patents' inventors is found to be associated to an increase in the probability to observe a citation link between such patents. Besides, co-ethnicity and co-location seem to act as substitutes, with Indian inventors in the US activating their ethnic connections to reach knowledge assets located outside their metropolitan area. The sociological or organizational origins of ethnic ties (whether they derive from previous common study or work experiences, by affiliation to formal expat organizations, or are simply due to informal linguistic and cultural bonds) are not explored.

[2] Academic scientists who produce patentable inventions do not always disclose them to their universities' technology transfer offices, either in the attempt to exploit them individually or as a result of contractual arrangements with industry sponsors. In both Europe, the US, and Japan, academic patents are a non-negligible share of total domestic patents, especially in science-based fields (survey by Lissoni, 2012). Academic inventors play a key role in bridging gaps and shortening distances in inventors' networks (Lissoni et al., 2013).

[3] Networks of inventors are obtained by examining co-inventorship patterns. Technically, they are one-mode projections of "two-mode" or "affiliation" networks (Borgatti and Everett, 1997). All co-inventors are at distance one, while inventors who have never worked together on the same patent, but have at least one co-inventor in common, are at distance two; if they have no common co-inventors, but at least two of their co-inventors once worked once together, they are at distance three; and so forth. When examining complex networks, in which any two inventors can be linked one another by several different "chains" of co-inventors (paths), the social distance between any two inventors is usually measured as the length of the shortest path between the two (geodesic distance).

[4] One important difference between Agrawal et al.'s (2008) method and JTH is that the former makes use, as observations, of inventor pairs, rather than patent pairs. Whenever a cited-citing patent pairs include n>1 inventors in the cited patent and m>1 inventors on the citing one, this produces m*n observations. Another difference is that controls are selected on cited patents, rather than citing ones.

Almeida et al. (2014) also rely on an ad hoc collection of surnames to identify Indian inventors in the US (in the semiconductor industry). They find both evidence of intra-ethnic citations, as well as some indications that reliance on such citations is correlated to inventors' productivity.

Agrawal et al. (2011) extend the Agrawal et al.'s (2008) data and methodology to the case of international knowledge flows and find that patents by Indian inventors in the US do not seem to attract a higher-than-average rate of citations from the inventors' home country. The only (weak) exceptions are patents in Electronics, and patents owned by multinational firms. Overall, these results go in the direction of suggesting that the Indian diaspora is not a major source of knowledge feedbacks for its country of origin. As suggested by Singh and Marx (2013), country boundaries seem to provide obstacles to knowledge diffusion that resist to controlling for spatial distance. It is at this point that studies in the geography of innovation tradition blend with research on migration and brain gain.

2.2 Migrants' contribution to innovation in origin countries

Migration studies have traditionally looked for possible positive returns from emigration for origin countries. Early research placed special emphasis on emigrants' remittances and their role in capital formation. More recently, due to the increasing importance of highly skilled migration, more attention has been paid to contributions to knowledge stock and innovation (Bhagwati and Hanson, 2009).

These may come in three, non-mutually exclusive forms, namely:

(i) *"Ethnic-driven" knowledge flows.* Emigrant scientists and engineers may retain social contacts with professional associations and educational institutions in their home countries, and transmit them scientific and technical skills either on a friendly or contractual basis (Meyer and Brown, 1999; Meyer, 2001)

(ii) *Internal transfers by multinational companies* (Blomström and Kokko, 1998; Veugelers and Cassiman, 2004)

(iii) *Returnees' direct contribution.* Migrant scientists, engineers and/or entrepreneurs may decide to move back to their origin countries and continue their activities over there, possibly keeping base in their destination countries (Wadhva, 2007a,b; Kenney et al., 2013, and references therein).[5]

While case studies on these phenomena abound, large-scale quantitative evidence is scant and almost entirely focussed on the US as a destination country, and China and India as origin countries. This largely ignores the fact that highly skilled migration to the US originates not only from developing countries, but also from Western Europe, South Korea, and Japan (Docquier and Marfouk, 2006; Widmaier and Dumont, 2011; see also Freeman, 2010). [6]

A series of papers by William Kerr and co-authors) has exploited two sources of information:

- the NBER Patent Data File, by Hall et al. (2001), which includes information on name, surnames, and addresses of inventors

[5] In addition, the demand of highly skilled migrants by destination countries provides youth in origin ones with an incentive to get higher education. Such migration-induced demand for higher education allows origin countries to keep their university system going, even in fields for which local demand of graduates is lacking, and, with it, some absorptive capacity of foreign science and technology.

[6] Besides, advanced European countries, while not as attractive as the US, host a non-negligible amount of highly skilled migrants from the East and South of the Old Continent, as well as from several former colonies

- the Melissa ethnic-name database, a commercial repository of names and surnames of US residents, classified by ethnicity, mainly used for direct-mail advertisements.

Names and surnames from the two sources are matched, so to assign each inventor to one out of nine broad ethnic groups, the most distinctive being the Asian ones. As for knowledge flows, Kerr (2008) focusses on citations running from patents filed at the USPTO by inventors from outside the US to patents filed by US residents over the 10 preceding years (company self-citations excluded). 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. It is found that co-ethnic citation groups are on average 50% more numerous than mixed ethnic ones.

Kerr (2008) further uses patent data as regressors in a first-difference panel data econometric exercise concerning origin countries of immigration into the US, with economic growth as the dependent variable. Migrants' patents in the US are found to affect substantially their home countries' growth rates. 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. Notice that this positive results is in contrast with the weak evidence provided, for India, by Agrawal et al. (2011).[7]

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. They find that US multinationals with a high share or quantity of ethnic patents invest and innovate more in their ethnic inventors' origin countries, and at the same time they rely 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.

As for returnee inventors, Agrawal et al. (2011) manage to identify 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 as India is concerned, 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 (2014) exploits the information on inventors' nationality contained in PCT patent applications filed at the USPTO (more in the following section). The author estimates the impact of foreign inventors on the extent of international technological collaborations between origin and

[7] Notice also that regressions do not consider Western Europe and Japan as countries of origin, in order to avoid reverse causality problems: 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.

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.

2.3 Methodological issues

The importance assumed by inventor data in the geography of innovation literature has pushed several scholars to improve the quality and transparency of their data mining efforts, and to discuss how this may affect research results (Li et al., 2014; Martínez et al., 2013; Marx et al., 2009; Pezzoni et al., 2012; Raffo and Lhuillery, 2009). We sum up here some previously unexplored implications for studies on the localization and "ethnicity" of citations (for more details, see Appendix 1). [8]

Ideally, a good disambiguation algorithm would minimize both "false negatives" (maximise "recall") and "false positive" (maximise "precision")[9]. Unfortunately, a trade-off exists between the two objectives, which requires making choices based on the consequences of each type of error for the subsequent analysis.

This has two consequences for the analysis of ethnic citations:

1) High precision/Low recall algorithms lead to underestimating the number of personal self-citations and overestimating that of co-ethnic citations (the opposite holds for low precision/high recall).

2) When applied to inventor sets from different countries of origin, the same matching rules may return different results in terms of precision and recall.

So far, patent-based studies on migration and innovation have not tackled these issues. Kerr (2007) and extensions make use of non-disambiguated inventor data. Agrawal et al. (2008, 2011) and Almeida et al. (2014) do not provide details on the disambiguation techniques they have used, while Alnuaimi et al. (2012) apply a "perfect matching" techniques (only inventors with exactly the same name and surname are considered as the same person), which works as an extreme case of high precision / low recall algorithm.

Precision and recall issues also appear when assigning inventors to a country of origin, based on their names/surnames. We discuss this matter in Appendix 3. Differently from name disambiguation, most of the studies reviewed above discuss openly this methodological issue, and decide to go for maximizing precision. For example, Agrawal et al. (2008) identify Indian inventors based on a very narrow list of Indian surnames, which are highly frequent in both India and validated by experts as indicative of recent migration status. This implies a tendency to limit the definition of ethnic inventor to first-generation migrants, which in turn hides the assumption that the strength of ethnic ties weakens with time. While making sense, the assumption is not very precise about the generational timing of the decay (at which generation do ethnic ties dissolve?) and does not consider the possibility of "ethnic revival" phenomena, such as those induced by home countries' policies aimed at reviving contacts with their diasporas; or by second- and third-generation migrants being affected by identitarian politics (Kuznetsov, 2006, 2010).

---

[8] The wave of interest for disambiguated inventor data has produced several open access inventor datasets. Two of them are: (i) 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 (http://www.esf-ape-inv.eu/index.php?page=3#EP-INV); and (ii) the US Patent Inventor Database, developed by Lee Fleming and associates, which contains USPTO data (http://dvn.iq.harvard.edu/dvn/dv/patent)

[9] Precision and recall rates are measured as follows: $Precision = \frac{tp}{tp+fp}$; $Recall = \frac{tp}{tp+fn}$
where: $tp\ (fp) = number\ of\ true\ (false) positives$; $tn\ (fn) = number\ of\ true\ (false) negatives$

Similar concerns arise when dealing with information on patent applicants. All the studies we reviewed in sections 2.1 and 2.2 state that they exclude company self-citations from the analysis. Yet, they are silent on the methodology followed in order to uniquely identify companies, and do not mention the issues of business groups. This is in contrast with recent concerted efforts to harmonize company names as found on patent data (Du Plessis et al., 2009; Peeters et al., 2010; Thoma et al., 2010).

Using patent applicant data in the absence of name harmonization and information on business groups can be equated to a high precision/low recall disambiguation technique (as discussed above for the case of inventors). When applied to localization studies, it leads to underestimating self-citations and overestimating the co-location of knowledge flows. When applied to the analysis of international knowledge flows, it underplays the role of multinationals as carriers of knowledge towards migrants' home countries, which in itself is an issue of substantive interest.

3. Propositions and data

In this section we first synthetize our research questions by means of a set of empirical propositions. We then describe our dataset, including relevant information on methodology.

3.1 Research questions: diaspora and brain gain effect

We are interested in exploring the role of "ethnic ties" (social ties involving migrant inventors on the basis of their common country of origin) in the diffusion of knowledge, both at the national and at the international level.

Ethnic ties between expatriates in the same destination country are interesting insofar they represent an instance of social bonds that may exist independently from those derived from professional experiences and/or physical proximity, though they may interact with both. Ethnic ties may have been formed in the destination country (as a result of homophily in the choice of acquaintances and friends Currarini et al., 2009) or inherited from the home country (as with chain migration). In both cases, they represent an instance of vitality and relevance of a community of expatriates, to which we will refer as a diaspora. We state a "diaspora effect" to exist when foreign inventors of the same ethnic group and active in the same country of destination have a higher propensity to cite one another's patents, as opposed to patents by other inventors, other things being equal and excluding self-citations at the company level. We test for its existence by adapting the JTH methodology, as described in section 2, and build a sample of cited-citing & cited-control patent pairs. Cited patents are all signed by at least a foreign inventor in a given destination country (in our case, the US), while citing and control patents are signed by inventors (foreign and local) from within the same country. Pairs by the same company or business group are excluded. We then estimate the simple model:

$$Probability\ of\ citation = f(co-ethnicity;\ spatial\ distance; social\ distance;\ controls) \qquad (1)$$

where the observations are patent pairs and the dependent variable is a binary one, which takes value one if the two patents in the pair are linked by a citation. The main variable of interest, *co-ethnicity*, is a dummy taking value one when both patents in the pair have been invented by one or several inventors from the same country of origin. As for spatial distance, this is derived from the addresses of inventors of the two patents in the pairs, which allow computing various measures, such as co-location at the metropolitan area

level, distance between inventors' addresses (ZIP codes' centroids), and/or co-location at the state level (in case one or both patents have multiple inventors with different addresses, we consider the minimum distance). Social distance refers to geodesic distances on the network of inventors (minimum social distance between inventors on the two patents of the pair). Interactions between co-ethnicity and both forms of distance are considered, too, so to test whether co-ethnicity and proximity in the physical or social space are substitute or complements. As for controls, they mostly refer to the characteristics of patents in the pair (especially the citing/control patents), based on the large literature on the determinants of patent citations (Hall et al., 2005; Harhoff et al., 2003). We provide full details of our sampling scheme and specification in the next two subsections (3.2 and 3.3).

Ethnic ties may also play a role at the international level. Most importantly, they may induce a "brain gain" effect, by which inventors in the countries of origin of migrant inventors cite disproportionately the latter's patents. We are interested in considering them separately from other brain gain sources, such as returnee inventors' self-citations. We are also interested in weighing their importance against multinational companies' self-citations. We test for a "brain gain" to exist and being induced by ethnic ties also by adapting the JTH methodology and building a sample of cited-citing & cited-control patent pairs. Cited patents are all signed by at least a foreign inventor in the US, while citing and control patents are signed by inventors (foreign and local) from outside the US. Pairs by the same inventors (returnee inventors) are excluded, while pairs from the same company or business group are included. We then proceed to regression analysis, by adapting equation (1), as follows:

$$Probability\ of\ citation = f(home\ country;\ same\ company; controls) \quad (2)$$

where the dependent variable is the same as in (1), and the main variable of interest is now *home country*, a dummy variable that takes value one if the inventor of the citing (control) patent in the pair resides in the country of origin of the foreign inventor in the cited one. *Same company* is also a dummy that takes value one if both patents in the pair have been filed by the same company or business group. Controls are as in (1), plus social distance and spatial distance[10].

Finally, we consider that international knowledge flows generated by migrant inventors may reach different destinations than their countries of origin, and still be mediated by ethnic ties, as when an *international diaspora* exists (in analogy with evidence fro trade, as reported by Rauch, 2001). We test for this by modifying equation (2) as follows

$$Probability\ of\ citation = f(co-ethnicity;\ same\ company; controls) \quad (3)$$

that is by replacing the home country effect with co-ethnicity (we also allow for specifications including both), while keeping the sampling scheme unaltered.

The "home country" effect and the "co-ethnicity" effect, at an international scale, may not coincide. In particular, an "international diaspora" may exist which is not associated to any brain gain for the CoO (and vice versa).

[10] Networks of inventors may span across countries, so it is necessary to include social distance in (2). As for spatial distance, this cannot be measured with co-location dummies, since by construction inventors of cited patents do not reside in the same country (a fortiori, the same city) of inventors of citing(control) ones. Still, the distance between migrant inventors' destination country and their home country (and between cities therein) may vary and affect the probability of citation.

Migrant inventors in the US from a specific CoO may be well connected to their home country, but not with migrant inventors from the same CoO to other destination countries. This may occur because a sizeable international community does not exist (migrant inventors from that specific CoO all go to the same destination country) or because migration to different destination countries originate from different cities/regions of the same CoO or occurred at different points in time (migrants to different destination countries belong to different migration cohorts). It may also be that migrant inventors are connected to the home country in their role of employees of one or a few multinationals, with sites in both their destination and home country; or because of specific diaspora policies run by their home country's government.

Conversely, migrants from the same CoO in different destination countries may be well connected among themselves (through social ties or, possibly, as fellow employees of the same multinationals), but not to their home countries. They may have fled the latter for political or religious reasons, or simply their home country does not have any absorptive capacity of the knowledge they generate (for lack of R&D investments or relevant policy). It may also be that an international diaspora has existed for so long that ties with the CoO have become less relevant than those between important communities abroad.

3.2 Data

3.2.1 Patent and inventor data

Our data results from matching names and surnames of inventors in the EP-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).

The EP-INV inventor database contains information on uniquely identified inventors listed on patent applications filed at the European Patent Office (EPO). For short, we will often refer to patent applications simply as "patents", whether granted or not.

EP-INV contains information dating back to the opening of EPO (1978) and it is continuously updated with raw data coming from PatStat, the Worldwide Patent Statistical Database published regularly by the European Patent Office[11]. For this paper, EP-INV was updated to the October 2013 release of PatStat, which is reliable for patent applications with priority dates up until 2010. Information on inventors includes their home address (sometimes replaced by the patent applicant's address), as harmonized and linked to Eurostat and/or OECD territorial units (NUTS3 and TL3, respectively) by RegPat, a OECD product also derived by PatStat.

Name disambiguation for inventors in EP-INV is performed by making use of Massacrator 2.0, a 3-step algorithm described at length by Pezzoni et al. (2012). Massacrator 2.0's matches inventors on the basis of edit distances between all tokens comprised in the inventors' name-and-surname text strings, and then filters the matches by exploiting information on both the inventors and their patents[12]. Massacrator 2.0

[11] 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.

[12] As an example, consider "Dmitriy Yavid", a Russian inventor with a 2-token name-and-surname text string, and his fellow countryman "Sergei Vladimirovich Ivanov", with a 3-token name-and-surname string. As all of their tokens are pretty different, the two inventors will not be matched. Instead, "Dmitriy Yavid" and "Dimitriy Victorovich Yavid" will be matched, as, of the former's two tokens, one is identical to a token in the latter's, and another differs for just one character. The "Dmitriy Yavid" - "Dimitriy Victorovich Yavid" match will be then retained as valid if the two inventors' patents are either similar in contents, citation patterns, priority year, location in space, or property regime (same applicant); or if the two inventors have common co-inventors, or co-inventors who worked together. Otherwise they will be discarded as false matches.

does not produce a unique dataset, but several ones, each of which is calibrated against a benchmark dataset in order to return a different combination of precision and recall. For this paper we started from the "balanced" calibration (which returns a precision rate of 88%, and a recall of 68%, when tested against a benchmark of French inventors) and slightly modified it. The modification consists in considering as positive cases (that is, the same person) all matched inventors whose patents are linked by at least one citation, irrespective of other filter criteria. This presumably allows for higher recall, and directly address the problem of over-estimation of ethnic citations discussed in section 2.3.1[13]. For all citing patents entering our final sample, we also checked manually the inventors whose names and surnames are at edit (Levensthein) distances lower than 4.

The *IBM-GNR system* is a commercial product that performs various tasks, including the association of names and surnames to one or (more often) several "countries of association" (from now on: CoAs). 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. As the original dataset included only non-US citizens, the US itself is never listed among the possible CoAs.

When fed with either a name or a surname or both, IBM-GNR returns a list of CoAs and two scores of interest:

- "frequency", which indicates to which percentile of the frequency distribution in the CoA the name or surname belongs to, for each CoA;
- "significance", which approximates the frequency distribution of the name or surname across all CoA. [14]

We treat this information by means of an original algorithm (named Ethnic-INV) that we describe in Appendix 2, along with some descriptive statistics. Here it suffices to sum it up as follows.

We consider all inventors i ($i=1\dots N$) who reside in the US and whose name-surname combination is associated to a vector of CoA that includes at least one of the 10 Countries of Origin (CoO; $c=1\dots 10$) of our interest, namely: China, India, Iran, Japan, and South Korea (for Asia); and France, Germany, Italy, Poland, and Russia (for Europe). These countries figure among the top 20 countries of origin of high skilled migrants to the US according to OECD/DIOC data (Widmaier and Dumont, 2011). None of them has English or Spanish, which are the most widely spoken languages in the US, as official languages.

For each inventor-CoO couple (i,c) , we calculate three indicators:

1. The frequency of *i's* first name in English- and Spanish-speaking CoA [15]

[13] To the extent that this modification induces higher recall at the price of lowering precision, it may lead to over-estimating the phenomenon of returnee inventorship (when the same inventor is first found to be active away from her country of origin, and then back to it).

[14] For example, an extremely common Italian surname such as Rossi will be associated both to Italy and to France, which hosts a significant Italian minority; but in Italy 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. When it comes to significance, the highest percentage of inventors named Rossi will be found in Italy (say 80), followed by France (which has been historically the most important destination countries of Italian migrants besides the US), and several South-American countries (where to many Italians emigrated in the past), with much smaller values.

[15] The intuition is as follows. An inventor with a typical Indian surname, such as Laroia, but named John or Luis is unlikely to be a recent Indian migrant into the US; this is because John and Luis are high-frequency names, respectively, in English-speaking and Spanish-speaking countries. More likely, he will be born in the US, possibly from mixed parents. On the contrary, Rajiv Laroia is more likely to be a first- or second-generation Indian immigrant, as Rajiv is high-frequency name in India, a zero-frequency name

2. The product of the significances attached to *i*'s name(s) and surname, for CoO *c*.

3. The significance attached to *i*'s surname, , for CoO *c*. [16]

We then set several possible threshold values for the three indicators (ceiling values for indicator n.1; floor values for the other two). For each set of values, inventor *i* is considered either "ethnic" from country of origin c or simply non-ethnic (with respect to *c*), depending on whether all indicators cross the thresholds or not.

Finally, we proceed to calibration against a benchmark dataset of PCT patent applications, which contains information on the inventors' nationality, which is the best proxy we have at hand for the country of origin (Miguelez and Fink, 2013). Calibration consists in altering the threshold values in order to obtain different combinations of precision and recall. In what follows we use a "high recall" calibration, one that minimizes false negatives (it tries to avoid mistaking ethnic inventors for non-ethnic), but allows for many false positives (it mistakes many non-ethnic inventors for ethnic ones). This may introduce a negative bias in our estimates of the co-ethnicity and "home country" parameters in equations (1) - (3) to the extent that, conditional on co-ethnicity to positively affect the probability of citations, a loose measurement of the latter may lead us to treat as co-ethnic two inventors who are not so, and do not cite each other's patents [17].

Figures A3.1-2 in Appendix 3 report, respectively, the share of foreign inventors and of foreign inventors' patent applications in the US, from 1980 to 2010, for the 10 CoO of interest (as calculated with the "high recall" calibration; patents assigned to one or another CoO as long as they include at least one of inventor from such CoO). The observed trends are close to those reported by (Kerr, 2007), with the only exception of Indian inventors' patents in the 2000s, for which Kerr observes a decline and we do not. As for values, they are in the same order of magnitude, with our estimates for Chinese inventors' patent share being around 1 point smaller than Kerr's, and those for Indians' share 1 point higher.

As for the geographic distribution of ethnic inventors, figures A3.13-12 in Appendix 3 report the "Location Quotient" (LQ) of inventors from each CoO of interest, across all states in the US. The LQ is modelled upon the classic "relative specialization index" in international trade, varies from -1 to 1, and is defined as follows:

$$LQ_{CoO,State} = \frac{Share_{CoO,State} - 1}{Share_{CoO,State} + 1}$$

$$\text{where:} \quad Share_{CoO,State} = \frac{Nr\ inventors_{CoO,State}/Nr\ inventors_{CoO,US}}{Nr\ inventors_{State}/Nr\ inventors_{US}}$$

High (close to 1) values of *standardized* $LQ_{CoO,State}$ indicate that the observed State host a higher-than-average share of foreign inventors from the CoO of interest. Visual inspection of maps confirms that our

in Spanish-speaking countries, and a low-frequency name in English-speaking countries (some of which host sizeable communities of Indian descent).

[16] The intuition is as follows: the indicator n.2 may have a high value due exclusively to a very high value of the significance for the name, with a moderate value for the significance of the surname. We wish the latter not to be too low.

[17] Going for a "high precision" calibration would avoid this problem, but it would also impose a considerable loss in terms of number of observations, which could affect negatively the significance of our estimates. In future versions of the paper, however, we will check the robustness of our results by running our regression exercise also on the "high precision" calibration.

algorithms work well on Asian CoO, whose inventors are concentrated in US technological powerhouses (such as California, for all CoO; and Texas, for Indians). On the contrary, it may bring in too many late descendants of immigrants from European CoO (as with Polish and German inventors, whose LQ is high for several states in the Midwestern and the Great Lakes' area).[18] We deal with the disparities in the precision of our Ethnic-Inv algorithm by running some robustness check of our results based on the information on the nationality of inventors, for the subset of inventors who also have patents in the WIPO-PCT database.

3.2.2 Sampling for the JTH exercise and regressions

Following Agrawal et al. (2008) we focus on patents by ethnic inventors in the US, but do not follow their inventor-based sampling scheme. Rather, we stick more closely to the original patent-based JTH sampling scheme, as adapted by Breschi and Lissoni (2009) to EPO patents and to the necessity to measure social distances (distances on the network of inventors). Compared to the original JTH matched sampling scheme, we also control more accurately for the technological classification of patents (as suggested by Thompson and Fox-Kean, 2005) .[19]

Subject to the "high recall" calibration of the Ethnic-Inv algorithm, we select all EPO patent applications from the EP-INV database, with priority years comprised between 1990 and 2010, and at least one inventor with residence in the US, but a CoO included among the ten of our interest. Our starting sample includes 88,522 inventors and 174,160 patents. We then retain only the applications that have received at least (either directly, or indirectly, via their patent family) one forward citation from another EPO patent application (either directly, or indirectly, via its patent family).[20] Overall, 127,664 of the 174,160 patents received at least one forward citation from 498,090 citing patents, for a total number of citing-cited patent pairs equal to 1,050,821.

Notice that the same cited patent enters our sample as many times as the number of citations it receives. The same applies to each citing patent that cites more than one cited patent. This will require correcting for non-independence of errors when conducting our econometric exercises.

We then proceed to collect information on:

- the citing and cited patents' applicants (as harmonized by the EEE-PPAT project and published with the October 2013 release of PatStat, supplemented by manual checks) [21]
- the inventors' identity, as from the EP-INV database
- the inventors' addresses (country and, for US-residents, the MSAMSA), also from EP-INV
- the citing patents' priority year and technological field (IPC group) [22]

[18] As for Italians and Russians, they exhibit a high LQ in New York state, which host both recent migrants and the descendants of early ones), while Louisiana is among the states with LQ>1 for French inventors, which also sounds suspect.

[19] Agrawal et al.'s (2008) sampling scheme also differs from JTH in that, once identified the cited-citing patent pairs, it proceed to build the control sample by matching on cited patents, instead of citing ones.

[20] Several patents from the same or, more commonly, different patent offices, form a « family » whenever they share one or several priority filings (or "priorities"). Roughly speaking, the family includes all patent documents that protect the same invention, so it is good practice to measure the citation links between two patents from the same office (in our case, the EPO) by counting all citations running between the families the two patents belong to (Harhoff et al., 2003). Several definitions of family exist, of which we adopt here the simplest one: all patents in the family must share exactly the same priorities (for more definition and a discussion, see Martinez, 2011).

[21] On the EEE-PPAT harmonization of applicants' names, see: https://www.ecoom.be/nl/EEE-PPAT and http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-RA-11-008 (last visited, May 2014)

On this basis, we build two different samples, a "local" and an "international" one.

For the local sample we retain all cited-citing pairs in which the citing patent comprises among its inventors at least one US-resident, and we exclude all self-citations at the applicant level, as well as all self-citations at the inventor level, where the self-citing inventor belongs to one of the 10 CoO of interest. For each citing patent, we randomly select a control patent that satisfies the following conditions:

- it does not cite the cited patent in cited-citing pair
- it has the same priority year and is classified under the same IPC groups of the citing patent [23]
- it comprises among its inventor at least one US-resident

This leaves us with 1,044,888 observations – excluding all the cited-citing pairs (and their respective controls) for which controls cannot be computed, half of which are cited-citing pairs, the other half cited-control pairs. Both pairs are generated by the combination of 90,020 cited patents, 195,863 citing ones and 279,997 controls. Table 1 (part 1) reports details by CoO. As expected, more than half the observations come from the two largest CoO, China and India. The only European country in the same order of magnitude is Germany.

Intuitively, the local sample includes all the citations addressed to "ethnic" patents by US-resident inventors. Were we to find that ethnicity affecting the observed citation patterns, this would point at the existence of a diaspora effect (social ties among migrants in the same destination country, from the same country of origin).

As for the "international" sample we retain all cited-citing pairs in which the citing patent has no US-resident inventors. For each citing patent, we randomly select a control patent that satisfies the following conditions:

- it does not cite the cited patent in cited-citing pair
- it has the same priority year and is classified under the same IPC groups of the citing patent
- it does not comprises among its inventor any US-resident

This leaves us with 1,050,236 observations – excluding all the cited-citing pairs (and their respective controls) for which controls cannot be computed, half of which are cited-citing pairs, the other half cited-control pairs. Both pairs are generated by the combination of 105,059 cited patents, 266,629 citing ones, and 390519controls. Table 1 (part 2) shows that the distribution of observations by CoO of cited patents' inventors is very much the same as that for the local sample.

When running a test of proportions, the cited-citing pairs and the cited-control pairs in both the local and the international samples will be treated as different subsamples (in a "cases vs controls" setting). In the regression setting, they will be "stacked" and flagged as different by means of the binary variable *Citation* (=1 for cited-citing pairs, =0 for cited-control pairs), which will then be our dependent variable in the econometric exercises. In addition, in the regressions we further control for the characteristics of patents

[22] IPC is the International Patent Classification, which is maintained and regularly update by the World Intellectual Property Organization (WIPO). We use IPC version 8. IPC groups are the second finest level of aggregation. For details: http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf (last visited, May 2014)

[23] Notice that the same patent may be assigned to several IPC groups. Therefore our matching criteria require the citing patent and its control to be classified under the same number of IPC groups, and to share them all

that we did not consider when performing the matched sampling, including finer levels of technological classification (as suggested by Singh and Marx, 2013).

Table 1. Local and international samples: nr of patents, pairs, and observations; by country of origin of cited patents' inventors

	cited patents		citing patents		cited-citing pairs		obs (3)
	Nr	%	nr	%	nr	%	nr
1. Local sample (citations from within the US)							
China	27,502	25.35%	73,838	20.80%	124,841	23.90%	249,682
Germany	17,549	16.18%	63,101	17.78%	87,939	16.83%	175,878
France	6,915	6.37%	26,674	7.52%	33,128	6.34%	66,256
India	33,181	30.58%	97,562	27.49%	162,228	31.05%	324,456
Iran	2,987	2.75%	12,445	3.51%	14,551	2.79%	29,102
Italy	4,259	3.93%	18,873	5.32%	23,368	4.47%	46,736
Japan	4,934	4.55%	19,988	5.63%	24,140	4.62%	48,280
Korea	5,221	4.81%	20,475	5.77%	25,934	4.96%	51,868
Poland	1,758	1.62%	7,001	1.97%	8,042	1.54%	16,084
Russia	4,185	3.86%	14,963	4.22%	18,273	3.50%	36,546
Total (1)	108,491	100.00%	354,920	100.00%	522,444	100.00%	1,044,888
Total (2)	90,020		195,863		438,345		876,690
2. International sample (citations from outside the US)							
China	31,336	24.75%	88,802	21.86%	128,293	24.43%	256,586
Germany	21,522	17.00%	72,860	17.94%	88,988	16.95%	177,976
France	8,255	6.52%	29,365	7.23%	34,125	6.50%	68,250
India	38,016	30.02%	115,071	28.33%	158,513	30.19%	317,026
Iran	3,311	2.61%	12,042	2.96%	13,358	2.54%	26,716
Italy	5,021	3.97%	19,879	4.89%	23,171	4.41%	46,342
Japan	6,193	4.89%	23,321	5.74%	26,619	5.07%	53,238
Korea	5,963	4.71%	21,126	5.20%	24,571	4.68%	49,142
Poland	2,089	1.65%	7,391	1.82%	8,348	1.59%	16,696
Russia	4,915	3.88%	16,369	4.03%	19,132	3.64%	38,264
Total (1)	126,621	100.00%	406,226	100.00%	525,118	100.00%	1,050,236
Total (2)	105,059		266,629		432,681		865,362

(1) Total = sum of observations by country of origin (same patent may be recorded under >1 country)
(2) Total = sum of distinct observations
*(3) Nr observations per country = Nr cited-citing pairs * 2*

For all patent pairs in the two samples, we produce the following dummy variables, which will enter as independent variables in the regressions:

1. *Co-ethnicity* : =1 if at least one inventor in the cited patent and one inventor in the citing (control) one are from the same CoO

2. *Social distance S* (with S=0,1,2,>3,+∞) : =1 if the minimum geodesic distance between cited patent and the citing (control) is equal to *S* . Formally: $S = min \{Sij\}$ with S_{ij}=geodesic distance between inventor *i* (*i*=1…I) on the cited patent and inventor *j* (*i*=1…*J*) on the citing (control) one, as calculated on the entire network of inventors, for all inventors on the cited and the citing (control) patents. Notice that: for *i*=*j* →S=0; if all *i* and all *j* belong to disconnected network components: S=+∞. For each year *t* we calculate a different network of inventors, based on co-inventorship patterns of all patents with priority years (t-1, t-5). [24]

[24] This amounts to assuming that social ties generated by co-inventorship decay after 5 years, unless renewed vy furthe collaborations. For more details, see Breschi and Lissoni (2009).

3. Miles : shortest distance (in miles) between the two patents, based on their inventors' addresses ; in some specification we consider also the quadratic term, in others we replace it with a set of 10 dummies, ranging from *0-25 miles* (reference case) to *over 2500 miles* [25].

4. Characteristics of the citing (control) patent, such as: its technological field (OST-30 classification, as from Tarasconi and Coffano, 2014), the number of claims (*claims*), the number of backward citations to prior art (*backward citations*) and to non-patent literature (*NPL citations*), as well as its technological proximity to the cited patent (nr of overlapping IPC-7 codes – *overlap IPCs 7* – and nr of overlapping full IPC codes, out all codes assigned to the patents).

For the patent pairs in the local sample we also calculate:

5. *Same MSA* and *Same State*: =1 if at least one inventor in the cited patent and one inventor in the citing (control) are located in the same metropolitan statistical area (MSA) and same US State, respectively

For patent pairs in the international sample we also calculate:

6. *Home country* : =1 if at least one inventor in citing (control) patent is located in the CoO of one of the inventors of the cited patent

7. *Same country* : 1 if at least one inventor in the cited patent and one inventor in the citing (control) are located in the same country, outside the US[26]

8. Other measures of country proximity, such as: border-sharing (*Contiguous countries*), *Former colonial relationship*, and language-sharing (*English*, =1 if at least one inventor of the citing (control) patent is located in an English-speaking country; and *Similarity to English*, a language similarity index ranging from 0 to 1, adapted from Miguelez, 2014)

9. *Same company* : =1 if applicants of the cited and the citing (control) patents are the same

10. *Returnee* : =1 if the inventor of the cited and the citing (control) patents are the same (notice that this implies *Social distance 0* = 1)

Table 2 reports the descriptive statistics for all variables in both samples; for details by country, see tables A3.13-22 in the appendix.

[25] For each combination of an inventor *i* from the cited patent and inventor j from the citing (control) patent we calculate the great-circle distance between the centroid of the respective ZIP codes; we then retain the minimum distance. In case of missing values at the ZIP code level, the centroid of the city was used (or the county, if the city's was missing, too)

[26] Notice that in the international sample, no inventor of the citing (control) patent can be located in the US, where instead it is located at least one ethnic inventor. This leaves open the possibility that one or several inventors in both the cited and the citing (control) patents are both located outside the US and in the same country, which is not necessarily the CoO of the inventor(s) of the cited patent.

Table 2. Local and international samples: descriptive statistics

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	1044888	0.500	0.500	0	1
Co-ethnicity	1044888	0.123	0.329	0	1
Same MSA	1044888	0.138	0.345	0	1
Same State	1044888	0.218	0.413	0	1
Miles	1044888	933.973	877.760	0	5085.412
Soc. Dist. 0	1044888	0.009	0.092	0	1
Soc. Dist. 1	1044888	0.008	0.091	0	1
Soc. Dist. 2	1044888	0.006	0.080	0	1
Soc. Dist. 3	1044888	0.008	0.088	0	1
Soc. Dist. >3	1044888	0.238	0.426	0	1
Soc. Dist. ∞	1044888	0.731	0.444	0	1
#claims	1044888	8.502	12.799	0	259
backward citations	1044888	4.582	3.148	0	87
NPL citations	1044888	1.325	2.451	0	57
overlap IPCs 7 digits	1044888	1.132	1.467	0	27
overlap IPCs 7 digits / all IPCs	1044888	0.284	0.285	0	1
overlap IPCs	1044888	0.827	1.566	0	53
2. International sample (citations from outside the US)					
Citation	1050236	0.500	0.500	0	1
Co-ethnicity	1050236	0.081	0.273	0	1
Home country	1050236	0.086	0.281	0	1
Same company	1050236	0.028	0.164	0	1
Contiguous countries	1050236	0.035	0.183	0	1
Former colonial relationship	1050236	0.200	0.400	0	1
Same country	1050236	0.037	0.190	0	1
English	1050236	0.174	0.380	0	1
Similarity to English	1050236	0.248	0.259	0	1
Miles	1050236	4461.819	1931.623	0	11498.1
Soc. Dist. 1	1050236	0.006	0.074	0	1
Soc. Dist. 2	1050236	0.004	0.067	0	1
Soc. Dist. 3	1050236	0.005	0.068	0	1
Soc. Dist. >3	1050236	0.201	0.401	0	1
Soc. Dist. ∞	1050236	0.780	0.414	0	1
#claims	1050236	9.917	11.832	0	442
backward citations	1050236	4.003	3.197	0	98
backward NPL citations	1050236	0.996	2.064	0	76
overlap IPCs 7 digits	1050236	1.087	1.268	0	32
overlap IPCs 7 digits / all IPCs	1050236	0.783	1.368	0	54
overlap IPCs	1050236	0.315	0.297	0	1
Returnee					

4. Results and discussion

4.1 Within-US knowledge flows

We test the existence of a diaspora effect on data from the local sample (citations within the US), first by running some simple tests of proportions, then by means of a logit regression. We run both the tests separately by CoO of the cited patents' inventors.

Table 3 reports the results of two simple tests of proportions, the former comparing the co-location rate of cited-citing patent pairs to that of control pairs, the latter doing the same for the co-ethnicity. We interpret any positive evidence on the co-ethnicity of citations as indicative of the existence of a diaspora effect, for the CoO considered. As for co-location, we are interested in checking whether our "ethnic" data reproduce the JTH basic results (which were obtained for all US-resident inventors), and in assessing to what extent they owe to social and ethnic ties.

Table 3. Co-location and co-ethnicity of citations: test of proportions, by CoO

	Co-location %			Co-ethnicity %			
CoO	**citing**	**controls**	**z**	**citing**	**controls**	**z**	***obs***
China	17.54%	11.02%	46.57*	24.96%	18.83%	37.06*	124841
Germany	16.36%	9.81%	40.70*	7.56%	7.13%	3.49*	87939
France	17.90%	11.33%	23.92*	3.74%	3.64%	0.7	33128
India	16.73%	10.60%	50.78*	17.56%	15.23%	17.91*	162228
Iran	20.01%	11.74%	19.30*	1.79%	1.28%	3.53*	14551
Italy	15.89%	9.55%	20.59*	1.99%	1.93%	0.5	23368
Japan	16.54%	10.96%	17.80*	2.92%	2.52%	2.71*	24140
Korea	17.02%	10.69%	20.87*	3.24%	2.69%	3.70*	25934
Poland	14.04%	9.13%	9.73*	0.68%	0.87%	-1.35	8042
Russia	17.32%	10.41%	19.10*	3.39%	2.50%	5.08*	18273

* significance at 99%

The co-location test replicates, to a large extent, JTH's classic results, with percentages that do not vary sensibly across CoO. As for the co-ethnicity test, this is positive and significant for all CoO, with the only exception of France and Poland, but it hides important cross-country differences. In particular, it seems to suggest that diaspora effects may be the strongest for China, India, Russia, and South Korea. The former two record the highest percentages of co-ethnic citing-cited and control-cited pairs, which is explained by the large number of Chinese and Indian inventors in the sample. They also report the largest difference between the co-ethnicity percentages of cited-citing and cited-control pairs (6.13 and 2.33, respectively). As for Russia, it records the third largest difference in absolute value (0.89, while Iran the largest ratio (1.4, with Russia being the second, 1.36).

Table 4 reports the results of six different specifications of equation (1), which we estimate by means of a logit regression, without distinguishing by CoO of ethnic inventors. The first specification reproduces Agrawal et al.'s (2008) basic exercise for Indian inventors in the US, which focusses on co-ethnicity and MSA co-location; the second introduces social distances between inventors; the third one also controls for patent characteristics; and the last three control for further measures of spatial distance, besides MSA co-location.

Table 4 – Probability of citation from within the US, as a function of co-ethnicity, spatial & social distance, and controls -- Logit regression

	(1)	(2)	(3)	(4)	(5)	(6)
Same MSA	0.549***	0.369***	0.345***	0.0312**	0.0638***	0.0676***
	(0.00638)	(0.00664)	(0.00694)	(0.0126)	(0.0146)	(0.0145)
Co-ethnic	0.210***	0.174***	0.168***	0.165***	0.167***	0.167***
	(0.00622)	(0.00630)	(0.00660)	(0.00660)	(0.00660)	(0.00660)
Co-ethnic * MSA	-0.0678***	-0.0633***	-0.0779***	-0.0796***	-0.0931***	-0.0820***
	(0.0163)	(0.0168)	(0.0175)	(0.0175)	(0.0176)	(0.0175)
Same State				0.0916***	0.0955***	0.109***
				(0.00772)	(0.00772)	(0.00821)
ln(Miles)				-0.0269***	-0.158***	
				(0.00195)	(0.00690)	
ln(miles)^2					0.0122***	
					(0.000602)	
Dist. 25-50miles						-0.232***
						(0.0160)
Dist. 50-100miles						-0.257***
						(0.0178)
Dist. 100-150miles						-0.245***
						(0.0193)
Dist. 150-250miles						-0.258***
						(0.0173)
Dist. 250-500miles						-0.226***
						(0.0158)
Dist. 500-1000miles						-0.188***
						(0.0163)
Dist. 1000-1500miles						-0.227***
						(0.0169)
Dist. 1500-2500miles						-0.264***
						(0.0161)
Dist. over 2500 miles						-0.201***
						(0.0175)
Soc. Dist. 1		-1.185***	-1.201***	-1.175***	-1.086***	-1.189***
		(0.0804)	(0.0819)	(0.0819)	(0.0822)	(0.0819)
Soc. Dist. 2		-2.042***	-2.070***	-2.039***	-1.932***	-2.055***
		(0.0775)	(0.0792)	(0.0792)	(0.0795)	(0.0791)
Soc. Dist. 3		-2.543***	-2.580***	-2.543***	-2.427***	-2.554***
		(0.0750)	(0.0765)	(0.0765)	(0.0768)	(0.0765)
Soc. Dist. >3		-3.276***	-3.266***	-3.219***	-3.087***	-3.225***
		(0.0704)	(0.0713)	(0.0714)	(0.0717)	(0.0714)
Soc. Dist. ∞		-3.421***	-3.356***	-3.305***	-3.172***	-3.313***
		(0.0703)	(0.0713)	(0.0713)	(0.0716)	(0.0713)
ln(#claims)			0.00479***	0.00473***	0.00503***	0.00495***
			(0.00147)	(0.00147)	(0.00147)	(0.00147)
ln(1 + backward citations)			0.362***	0.363***	0.363***	0.363***
			(0.00403)	(0.00403)	(0.00403)	(0.00403)
ln(1 + NPL citations)			-0.0263***	-0.0259***	-0.0272***	-0.0263***
			(0.00333)	(0.00334)	(0.00334)	(0.00334)
ln(1 + overlap IPCs 7 digits)			0.921***	0.919***	0.919***	0.920***
			(0.00350)	(0.00351)	(0.00351)	(0.00351)
OST-30 F.E.	no	no	yes	yes	yes	yes
Constant	-0.0989***	3.275***	2.303***	2.419***	2.611***	2.477***
	(0.00111)	(0.0703)	(0.0717)	(0.0724)	(0.0734)	(0.0731)
Observations	1,044,888	1,044,888	1,044,888	1,044,888	1,044,888	1,044,888
chi2	9766	16657	93062	93370	93368	93642
ll	-719188	-710428	-679307	-679011	-678782	-678848
r2_p	0.00700	0.0191	0.0621	0.0625	0.0628	0.0627

The table reports estimated parameters (s) ; Clustered standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

Estimated coefficients in column (1) have the same sign and are of the same order of magnitude as those in Agrawal et al.'s (2008; table 5): co-ethnicity affects positively the probability to observe a citation link between two patents, but its marginal effect is smaller than that of MSA co-location. The interaction term between co-ethnicity and co-location is negative, which suggests a substitution effect. When controlling for social distance (column 2) the estimated coefficients for both co-ethnicity and co-location drops

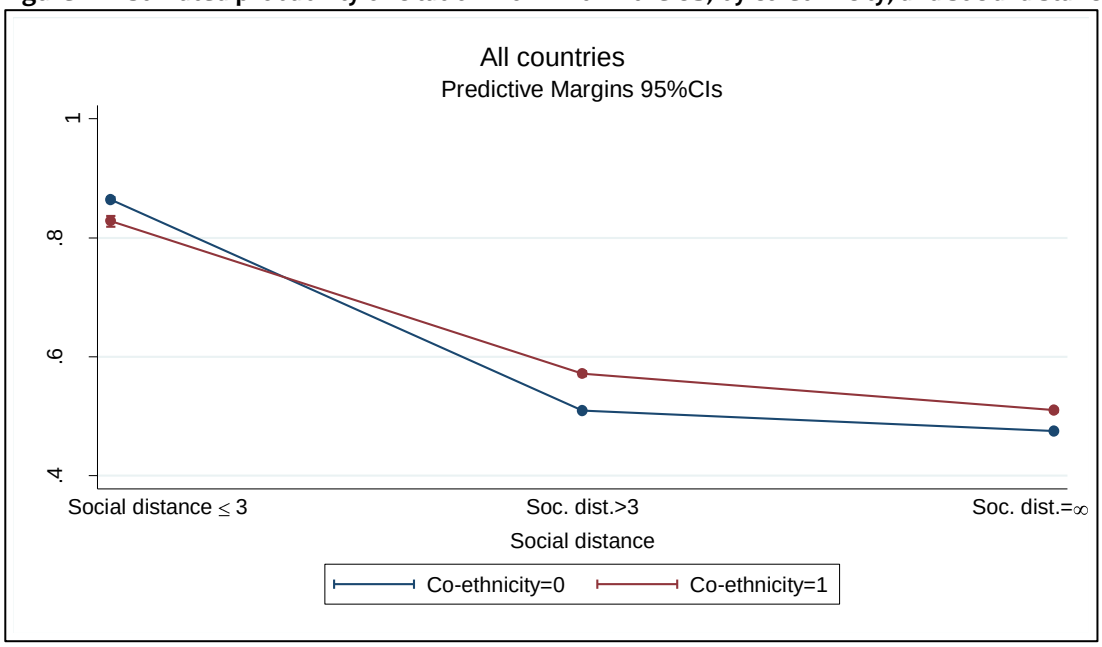
sharply, as social distance affects negatively the probability of citation and it is positively correlated with spatial distance, as suggested by Breschi and Lissoni (2009). We also notice that the marginal effect of social distance reduces sharply when the latter increases (the absolute value of coefficients first increases sharply, then less and less). This is also in accordance with previous findings.

Controlling for the characteristics of patents (column 3) does not alter much the coefficients of interest, which we take as a sign that the original sampling scheme was valid. Adding controls for spatial distance alters the estimated co-efficient of co-location (columns 4, 5 and 6), but neither those for social distance and co-ethnicity.

When interacting social distance dummies and co-ethnicity we get positive and significant coefficients only for China, India and, to less extent, France (results available on request).

Figure 1 reports the estimated probability of citation, as derived from the estimation of model (6) in table 4 plus interactions between co-ethnicity and social distance. We notice that the marginal effect of co-ethnicity (which can be guessed by comparing the vertical position of the two lines) is slightly negative for low social distances (less than 3 degrees of separation between the inventor teams) and positive thereafter. This suggests that, as with spatial distance, co-ethnicity is a substitute social connection to those induced by co-location in space or proximity in a professional network. We also notice that even at its maximum (for finite social distances higher than 3), its marginal effect is much lower than that of social distance, at low social distances (compare the vertical distance between the lines in the graph with the inclination of both lines between the first two points).

Figure 1 – Estimated probability of citation from within the US, by co-ethnicity, and social distance



In table 5 we allow for the estimated coefficient of co-ethnicity to vary across CoO, first without interaction with MSA co-location (column 1), then with interaction (column 2). The importance of co-ethnicity for the probability of citation clearly varies by CoO. Its estimated coefficient is clearly positive and significant only for Asian countries (slightly less for Japan in column 1 and for Iran in column 2) and Russia. It is even negative, albeit slightly or not significant, for France and Poland. As for the interaction term, this is negative and significant only for China and India, positive and significant for Russia, and not significant for

all the other CoO. The coefficients for social distance and other controls (unreported) do not differ much from those in table 4.

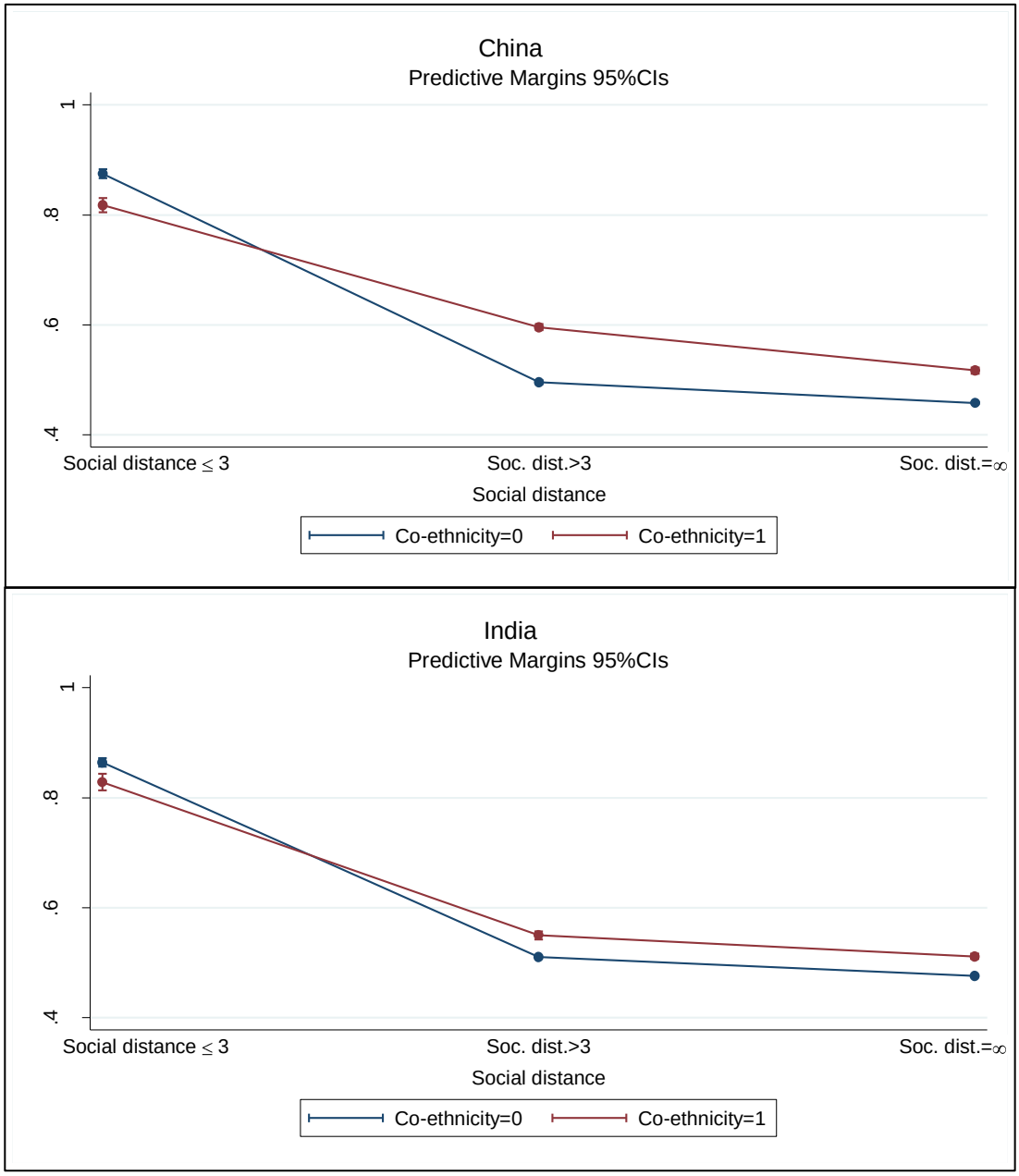
Figure 2 reports the estimated probability of citation, as a function of co-ethnicity and social distance, for China and India. Overall, the position and shape of the lines remind closely to those in figure 1, but we can guess that the marginal effect of co-ethnicity is much higher for China than for India (and for all other CoO, unreported, with the exception of Iran).

Table 5 – Probability of citation from within the US, as a function of co-ethnicity by Country of Origin, spatial & social distance, and controls, -- Logit regression

	(1)	(2)	(2-cont.)	(2-cont.)
Same MSA	0.140***	0.151***		
	(0.0107)	(0.0110)		
Same State	0.0916***	0.0912***		
	(0.00773)	(0.00773)		
ln(Miles)	-0.0269***	-0.0270***		
	(0.00195)	(0.00195)	*Co-ethnicity* Same MSA*	
China co-ethnic	0.229***	0.250***	China * Same MSA	-0.135***
	(0.00858)	(0.00943)		(0.0254)
Germany co-ethnic	0.0232	0.0165	Germany * Same MSA	0.0458
	(0.0178)	(0.0193)		(0.0535)
France co-ethnic	-0.0709*	-0.0392	France * Same MSA	-0.179
	(0.0421)	(0.0466)		(0.110)
India co-ethnic	0.124***	0.133***	India * Same MSA	-0.0584**
	(0.00867)	(0.00944)		(0.0256)
Iran co-ethnic	0.229**	0.169	Iran * Same MSA	0.244
	(0.0982)	(0.114)		(0.234)
Italy co-ethnic	0.0194	0.0507	Italy * Same MSA	-0.199
	(0.0692)	(0.0758)		(0.185)
Japan co-ethnic	0.0961*	0.126**	Japan * Same MSA	-0.190
	(0.0572)	(0.0623)		(0.159)
Korea co-ethnic	0.129**	0.153***	Korea * Same MSA	-0.129
	(0.0529)	(0.0584)		(0.139)
Poland co-ethnic	-0.240	-0.258	Poland * Same MSA	0.105
	(0.179)	(0.204)		(0.399)
Russia co-ethnic	0.287***	0.221***	Russia * Same MSA	0.452**
	(0.0615)	(0.0666)		(0.189)
Co-ethnicity Same MSA*	*No*	*Yes (see right)*		
Constant	2.423***	2.420***		
	(0.0724)	(0.0724)		
Social distance dummies	yes	yes		
Citing patent characteristics	yes	yes		
OST-30 FE	yes	yes		
Observations	1,044,888	1,044,888		
chi2	93382	93502		
ll	-678936	-678914		
r2_p	0.0626	0.0626		

Clustered standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

Figure 2 – Estimated probability of citation from within the US, by co-ethnicity, and social distance (from column 2 in table 5) , for China and India as CoO



Cross-CoO differences in the size and significance of the co-ethnicity coefficients may depend either from the demographic composition of social groups from the same CoO (share of first vs second-generation migrants and/or long established ethnic minorities) or from their social structure (cohesiveness of the social group, depending on its internal ties: familiar, linguistic, cultural, and economic). These characteristics depend, in turn, on how we define the boundaries of the group, that is on how well our Ethnic-INV algorithm works for each specific CoO: the lower its precision, the more likely it will be that we mix first generation migrants with established communities (think of Italian young PhDs and Italian-American communities in NY State or New Jersey) or migrants from different CoO, possibly linked by the same language (think of French vs Quebecois migrants into the US; or of German vs Austrian and Swiss ones).

Table 6 – Probability of citation from within the US, as a function of co-ethnicity or co-nationality -– Logit regression

	CO-ETHNICITY		CO-NATIONALITY	
	(1)	(2)	(3)	(4)
Same MSA	0.139***	0.138***	0.137***	0.137***
	(0.0215)	(0.0215)	(0.0215)	(0.0215)
Same State	0.155***	0.156***	0.155***	0.156***
	(0.0147)	(0.0148)	(0.0147)	(0.0148)
ln(Miles)	-0.0106***	-0.0111***	-0.0108***	-0.0111***
	(0.00378)	(0.00378)	(0.00378)	(0.00378)
Co-ethnicity/ Co-nationality §	0.221***		0.282***	
	(0.0109)		(0.0128)	
China §		0.289***		0.334***
		(0.0129)		(0.0154)
Germany §		0.0451		0.135***
		(0.0406)		(0.0496)
France §		-0.115		0.0278
		(0.0787)		(0.0878)
India §		0.148***		0.225***
		(0.0174)		(0.0220)
Iran §		0.700*		1.236
		(0.366)		(0.756)
Italy §		0.203		0.250
		(0.140)		(0.167)
Japan §		0.206*		0.156
		(0.105)		(0.124)
Korea §		0.106		0.243**
		(0.0970)		(0.118)
Poland §		-1.189		-1.515*
		(0.734)		(0.890)
Russia §		0.408***		0.444***
		(0.121)		(0.136)
Social distance dummies	yes	yes	yes	yes
Citing patent characteristics	yes	yes	yes	yes
OST-30	yes	yes	yes	yes
Constant	2.963***	2.970***	2.959***	2.964***
	(0.161)	(0.161)	(0.161)	(0.161)
Observations	237,866	237,866	237,866	237,866
chi2	20195	20296	20217	20279
ll	-154744	-154700	-154702	-154681
r2_p	0.0615	0.0617	0.0617	0.0618

§ Co-ethnicity in columns 1 and 2 ; co-nationality in columns 3 and 4
Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

One way to assess the relative weight of substantive factors vs measurement errors is to make use of a different definition of "ethnic" inventor. In table 6 we exploit information on inventors' nationality from PCT patents, for all inventors in our original sample who had at least one patent in the WIPO-PCT database by Miguelez and Fink (2013). We then retain only the cited patents (and the related citing and control ones) in which the inventors' countries of origin and of nationality coincide. This reduces the sample to around one fifth of the initial one[27]. We then run two sets of regressions very similar to those in tables 4 and 5: in the first set we maintain co-ethnicity as our explanatory variable of interest – using the reduced sample

[27] We proceeded as follows. Based on information on patent families provided by PatStat, we first identified all patents in the WIPO-PCT database that are equivalents of EP-INV patents in our sample. Within each pair of equivalent patents we name-matched inventors on the EPO patent to inventors on the WIPO-PCT one: around 90% of positive matches result from perfect name string matching, the remaining from a combination of Soundex matching of surname and first given name (around 9%), 2-gram string matching or manual checking (less than 125). This allowed us to assign a nationality to all inventors in the EP-INV database with at least one patent in the WIPO-PCT database. On the notion of patent family and equivalents, see Martinez (2011).

(one fifth of the original); in the second, we replace it with "co-nationality", which is more stringent (=1 if at least one inventor in the cited patent and one inventor in the citing (control) one have the same nationality, and not just the same CoO). When comparing the estimated coefficients for co-ethnicity and co-nationality across similar specifications (column 1 to 3, and column 2 to 4) we notice that, in general, the co-nationality one is larger. This suggests that "ethnic" inventors identified by the Ethnic-Inv may be comprise late-generation migrants or ethnic communities whose mutual bonds are not as strong as those between first-generation migrants. The most striking result concerns Germany, whose co-ethnicity coefficient is never (neither tables 4 and 5, nor in column 2 of table 6), but whose co-nationality coefficient is both positive and significant. This is not the case, however, for the other European countries: for France, Italy, and Poland neither co-ethnicity nor co-nationality are never significant; for Russia, both co-ethnicity and co-nationality are positive and significant, and, in table 6, they do not differ much. As for Asian countries, co-nationality is clearly larger than co-ethnicity for China, India, and Korea, while the opposite hold for Japan (and we have no good explanation for that; as for Iran, co-nationality is larger than co-ethnicity but not significant, possibly due to the very small number of Iranian nationals in the WIPO-PCT database). Overall, this suggests that, with the exception of Germany, no European countries among those we considered exhibit a diaspora effect, and this is not just a statistical artefact due to a measurement error problem.

One further way to assess the extent at which our main results may depend on the reliability of the name-based definition of ethnic inventors consists in running separate regressions for different technological classes (of the cited patents). In table 7 we report the results for such exercise, by considering seven macro-classes (OST-7 classification, as from Tarasconi and Coffano, 2014), the first four being relatively more science-based than the last two (and with the fifth in between).[28] We expect the ratio between first-generation migrants and late-generation ones (or members of established ethnic minorities) to be higher among "ethnic" inventors of science-based patents, than in non-science-based ones. This is because inventors of science-based patents are more likely to hold a PhD and/or to work in universities; and we now from the literature that first-generation migrants are over-represented among PhD holders and academic faculty. Inventors of non-science-based patents, on the contrary, are less educated, which suggests that "ethnic" individuals among them are less likely to be foreign-born.

From table 7, we can see that Pharmaceuticals & Biotech is only technological class with the most instances of a positive and significant (at 95%) coefficient for co-ethnicity (six CoO out of ten), followed by and Chemicals & Materials and Electrical engineering & Electronics (four CoO), Industrial Processes (three CoO) and Instruments (two). Mechanical engineering & Transport has just one case and Consumer goods none. This is in line with our expectations. At the same time, these results are not much in contrast with our main ones, as European CoO (with the usual exception of Russia) never exhibit more than one positive and significant co-ethnicity coefficient: in Chemicals & Materials for Germany, in Electrical engineering & Electronics for France, in Pharma & Biotech for Italy, and in Industrial processes for Poland. On the contrary, the co-ethnicity coefficient for China is positive in all science-based classed plus Industrial

[28] The relative closeness to science of different patents' technological classes can be assessed by comparing various indicators, such as the average ratio between backward citations to prior art (pre-existing patents) or non-patent literature (which is largely made of scientific publications; Callaert et al., 2006); and the share of academic patents, that is patents whose inventors were, at the time of filing, working for a university (Lissoni. 2011). In either cases, Pharmaceuticals & Biotech invariably emerges as the most science-based, followed by Chemicals & Materials, Instruments, and Electrical engineering & Electronics. The remaining classes follow at considerable distance, with the possible exception of Industrial processes, which may include a non-negligible number of academic patents, depending on the country.

processes, and almost the same applies to India (with the coefficient for Mechanical engineering & Transport being significant, instead of that for Instruments). As for Japan and Korea, they have two positive and significant coefficients, while Russia has three (all in science-based classes). Iran has just one, due to small number problems.

Table 7 – Probability of citation from within the US, as a function of co-ethnicity, by technological class of cited patents -– Logit regression

	Electrical eng.; Electronics	Instruments	Chemicals; Materials	Pharma & Biotech.	Industrial processes	Mechanical eng.; Transport	Consumer goods; Civil eng.
Same MSA	0.176***	0.187***	0.0640***	0.0702***	0.144***	0.147***	-0.0299
	(0.0189)	(0.0200)	(0.0203)	(0.0180)	(0.0346)	(0.0547)	(0.0678)
Same State	0.00834	0.0529***	0.161***	0.243***	0.0164	-0.0680	-0.0640
	(0.0133)	(0.0142)	(0.0153)	(0.0129)	(0.0256)	(0.0420)	(0.0542)
ln(Miles)	-0.0284***	-0.0336***	-0.0195***	-0.00290	-0.0570***	-0.0839***	-0.0808***
	(0.00366)	(0.00361)	(0.00350)	(0.00313)	(0.00604)	(0.00964)	(0.0134)
China	0.185***	0.0866***	0.279***	0.238***	0.115***	0.0599	-0.167
	(0.0175)	(0.0207)	(0.0125)	(0.0112)	(0.0343)	(0.0723)	(0.122)
Germany	-0.0273	0.0228	0.110***	0.0570*	-0.0126	-0.0216	-0.253**
	(0.0388)	(0.0296)	(0.0319)	(0.0299)	(0.0467)	(0.0745)	(0.102)
France	0.157*	-0.195**	-0.146**	-0.161***	-0.390**	-0.287	-0.235
	(0.0848)	(0.0833)	(0.0701)	(0.0608)	(0.158)	(0.243)	(0.219)
India	0.126***	0.00237	0.188***	0.131***	-0.000351	0.0881*	-0.155*
	(0.0126)	(0.0219)	(0.0163)	(0.0156)	(0.0317)	(0.0517)	(0.0912)
Iran	0.181	0.291*	0.188	0.671**	0.738*	-0.221	
	(0.127)	(0.175)	(0.281)	(0.298)	(0.402)	(0.333)	
Italy	-0.115	-0.123	0.0393	0.195**	-0.321	0.391	-0.696
	(0.142)	(0.138)	(0.129)	(0.0928)	(0.233)	(0.432)	(0.439)
Japan	-0.150	0.0997	0.171*	0.203**	0.187	-0.414	0.499
	(0.102)	(0.117)	(0.0912)	(0.0820)	(0.181)	(0.378)	(0.589)
Korea	0.0261	0.314***	0.0787	0.0437	0.354**	0.0157	-0.0444
	(0.0947)	(0.110)	(0.0822)	(0.0803)	(0.168)	(0.315)	(0.553)
Poland	-0.139	-0.171	0.0775	-0.618**	1.476**	-1.666*	0.926
	(0.443)	(0.364)	(0.272)	(0.298)	(0.746)	(1.009)	(1.150)
Russia	0.188**	0.220*	0.520***	0.446***	0.209	0.158	0.353
	(0.0946)	(0.115)	(0.116)	(0.112)	(0.213)	(0.405)	(0.487)
Interaction co-ethnicity * MSA	no	no	no	no	no	no	no
Constant	2.333***	2.435***	2.386***	3.058***	1.894***	1.819***	0.997***
	(0.153)	(0.140)	(0.133)	(0.179)	(0.168)	(0.245)	(0.326)
Observations	339,141	315,500	300,606	364,534	118,760	44,868	23,275
chi2	37072	29493	28964	34434	14784	6365	2995
ll	-220908	-202990	-191013	-232835	-73960	-28390	-14967
r2_p	0.0603	0.0718	0.0833	0.0785	0.102	0.0871	0.0725

Clustered standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

In Appendix 4 we run a few more robustness checks. First, we test whether our results depend exclusively from the most important high-tech clusters within the US, which attract a disproportionate number of highly skilled migrants. We focus on the top six MSAs by number of patent applications in our sample (S.Francisco, S.José, NY, Dallas, Boston, and S.Diego) and on the top ten MSA pairs with the highest number of citations running in one or another direction (that is, the ten most important city corridors for

citation flows; see table A4.1). We then run a series of regressions, controlling for the fixed effects of either the top six MSAs or the top ten city corridors (table A4.2). Our main results remain unaltered.

We also consider the possibility of cohort effects, with different generations of migrant inventors (from the same CoO) having different propensities to share knowledge with members of their communities. In order to control for that, we run two regressions, with year fixed effects (where the year corresponds to the priority date of the cited patents; table A4.3). Again, our main results remain unchanged.

Finally, we consider the possibility that the high significance of several coefficients in tables 4 and 5 may depend on the very large number of observations in our sample – which may decrease the variance of the estimators. We run again the regressions in table 5 with samples of reduced size, by applying the bootstrap technique described by Greene (2008, p.596) and Wooldridge (2002, p.378). As reported in table A4.4, the coefficients are maintained, but the standard errors increase as the size of the subsamples diminishes. Despite this, significance is always maintained for India and China, as well as for Russia with the exception of the last case (smallest sample). In regressions 4 and 8, with many dummies, not all subsamples lead to convergence, so results are based on a smaller set of replications. Estimates based of 1% subsample do not include the last column, since any of the subsample was able to converge.

4.2 *International knowledge flows*

Coming to international knowledge diffusion, we test for the existence of a "brain gain" effect and of an "international diaspora" one by running two parallel sets of logit regressions, corresponding to equations (2) and (3), on the international sample (citations from outside the US). In the first set, the variable of interest is the *Home country* dummy, in the second it is, once again, *Co-ethnicity*. Otherwise, all other regressors are the same.

Contrary to the previous exercise, we do not drop self-company citations, but we control for them. In this way the two variables of interest so that the two variables of interest (*Home country* or *Co-ethnicity*) will capture the weight of home country or ethnic ties relative to intra-organizational ties (knowledge transfer mediated by multinational companies). Finally, we had to drop *Returnee* from our regressor list because, for several CoO, it predicted perfectly the value of the dependent variable (which implies that returnee inventors, albeit very few, bring with them the knowledge they produced abroad).

Table 8 reports the results of the basic specifications, which do not distinguish by CoO of ethnic inventors. From column (1) we can see that the estimated coefficient for the *Home country* dummy is positive and is significant. However, it is much smaller than that for the *Same company* dummy, which implies a much smaller marginal effect; the same comparison applies, in absolute value, to *Social distance* dummies and for all dummies concerning cross-country proximity effects. From column (2) we observe that the interaction term between *Home country* and *Same company* is not significant, which suggest that the two are neither complementary nor substitute (home country ties are not especially important when occurring within the same company). More interestingly, the estimated co-efficient for *Co-ethnicity* (column 3) appears to be much larger, with a positive and significant interaction with *Same company* (column 4). These results are suggestive of the existence of an international diaspora effect, much stronger than the brain gain, possibly enhanced by knowledge circulation within multinationals.

As with the diaspora effect, however, this evidence does not apply to all CoO. Table 9 reports the results of two regressions in which we allow for the estimated coefficients of *Home country* (column 1) and *Co-ethnicity* (column 2) to vary across CoO. Coefficients for *Home country* are positive, significant and

comparable to those of *Co-ethnicity* only for China, Korea, and Russia. As for other countries that exhibited a strong diaspora effect within the US, India seems to exhibit an international diaspora effect, too, but no brain gain, as the *Home country* dummy is not significant; while Iran exhibits none (notice that Iranian inventors outside the US, and especially in Iran itself, are very few). As for the remaining countries we observe that co-ethnicity seems at work for both France (where it is stronger than the home country effect) and Germany (where there is no home country effect); Poland exhibits a home country effect, Italy none. One possible explanation for the findings concerning France and Germany is that our name-based CoO attribution mixes inventors from different countries or regions (France and French Canada; Germany, Switzerland and Austria), which introduces a measurement error affecting the *Home country* coefficient. For this reasons we replicate the robustness check based on the inventors' nationality we already performed when looking for the diaspora effect.

Table 8– Probability of citation from outside the US, as a function of "home country" effect or co-ethnicity, spatial & social distance, and controls -– Logit regression

	HOME COUNTRY		CO-ETHNICITY	
	(1)	(2)	(3)	(4)
Home country / Co-ethnicity §	0.0305***	0.0301***	0.108***	0.104***
	(0.00703)	(0.00714)	(0.00739)	(0.00750)
Same company	1.137***	1.135***	1.135***	1.113***
	(0.0181)	(0.0203)	(0.0181)	(0.0198)
Home country / Co-ethnicity * Same company §		0.0120		0.127***
		(0.0378)		(0.0432)
ln(Miles)	0.0194***	0.0195***	0.0201***	0.0208***
	(0.00284)	(0.00284)	(0.00284)	(0.00284)
Contiguous countries	0.187***	0.187***	0.184***	0.186***
	(0.0151)	(0.0151)	(0.0151)	(0.0151)
Former colonial relationship	0.0607***	0.0607***	0.0622***	0.0625***
	(0.00653)	(0.00653)	(0.00653)	(0.00653)
Same country	0.115***	0.115***	0.0945***	0.0948***
	(0.0194)	(0.0195)	(0.0193)	(0.0193)
English	0.380***	0.380***	0.379***	0.379***
	(0.00791)	(0.00791)	(0.00791)	(0.00791)
Similarity to English	0.226***	0.226***	0.220***	0.221***
	(0.0106)	(0.0106)	(0.0106)	(0.0106)
Soc. Dist. 1	-2.045***	-2.047***	-2.065***	-2.080***
	(0.168)	(0.168)	(0.168)	(0.168)
Soc. Dist. 2	-2.871***	-2.872***	-2.888***	-2.898***
	(0.166)	(0.166)	(0.166)	(0.166)
Soc. Dist. 3	-3.201***	-3.201***	-3.220***	-3.227***
	(0.164)	(0.164)	(0.164)	(0.164)
Soc. Dist. >3	-3.554***	-3.554***	-3.573***	-3.582***
	(0.160)	(0.160)	(0.160)	(0.160)
Soc. Dist. ∞	-3.755***	-3.755***	-3.774***	-3.783***
	(0.160)	(0.160)	(0.160)	(0.160)
Citing patent characteristics	yes	yes	yes	yes
OST F.E.	yes	yes	yes	yes
Constant	1.893***	1.894***	1.909***	1.912***
	(0.160)	(0.160)	(0.160)	(0.160)
Observations	1,050,236	1,050,236	1,050,236	1,050,236
chi2	124935	124950	125147	125115
ll	-657730	-657730	-657644	-657639
r2_p	0.0965	0.0965	0.0966	0.0966

§ « Home country » effect in columns 1 and 2 ; co-nationality in columns 3 and 4
Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table 9– Probability of citation from outside the US, as a function of "home country" effect, co-ethnicity (by Country of Origin), spatial & social distance, and controls -– Logit regression

	HOME COUNTRY	CO-ETHNICITY
	(1)	(2)
Same company	1.136***	1.135***
	(0.0181)	(0.0181)
Home country / Co-ethnicity – China §	0.179***	0.179***
	(0.0257)	(0.0212)
Home country / Co-ethnicity – Germany §	-0.0118	0.0896***
	(0.00868)	(0.00997)
Home country / Co-ethnicity – France §	0.0600***	0.182***
	(0.0233)	(0.0246)
Home country / Co-ethnicity – India §	0.0297	0.0919***
	(0.0404)	(0.0282)
Home country / Co-ethnicity – Iran §	0.330	0.0691
	(1.078)	(0.242)
Home country / Co-ethnicity – Italy §	-0.0493	0.0256
	(0.0468)	(0.0462)
Home country / Co-ethnicity – Japan §	0.0189	0.0207
	(0.0152)	(0.0154)
Home country / Co-ethnicity – Korea §	0.415***	0.399***
	(0.0428)	(0.0431)
Home country / Co-ethnicity – Poland §	1.225**	0.192
	(0.599)	(0.272)
Home country / Co-ethnicity – Russia §	0.630***	0.501***
	(0.157)	(0.101)
Country proximity controls	yes	yes
Social distance dummies	yes	yes
Citing patent characteristics	yes	yes
OST F.E.	yes	yes
Constant	1.880***	1.897***
	(0.160)	(0.160)
Observations	1,050,236	1,050,236
chi2	125051	125209
ll	-657656	-657590
r2_p	0.0966	0.0967

§ « Home country » effect in column 1 ; co-nationality in column 2
Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

We consider only the cited patents with inventors whose nationality, as reported by PCT patents, coincides with the CoO we attribute them. This reduces our sample to about one tenth of the original one. We then run again our logit regressions, with either Home country, Co-ethnicity or Co-nationality as main regressor. Results are reported in table 10. The estimated coefficient for *Home country* in column (1) is double than the corresponding one in table 8, while that for *Co-ethnicity* is pretty much the same, and it remains larger. When replacing it with Co-nationality the result does not change much.

When introducing separate coefficients for different CoO, we confirm our main results for China, France, Germany, and Korea (the relative values or estimated coefficients for Home country and *Co-ethnicity* do

not change, and those for *Co-ethnicity* and *Co-nationality* do not differ much). For other countries we get more mixed results. For India, we confirm the absence of any home country effect, but we also lose any co-ethnicity effect (same for co-nationality). For Japan, that in the main regressions exhibited no effect whatsoever, we now find positive effects with all three alternative regressors, with similar coefficients; almost the same applies to Italy, with the exception of the *Co-nationality* dummy, which remains non-significant. As for Russia, only *Co-ethnicity* remains significant. Results for Poland and Iran are not reliable, due to small number problems (in the case of Iran we often have perfectly predicted results).

Table 10 – Probability of citation from outside the US, as a function of "home-country" effect, co-ethnicity or co-nationality (also by Country of Origin) -– Logit regression

	HOME COUNTRY		CO-ETHNICITY		CO-NATIONALITY	
	(1)	(2)	(3)	(4)	(5)	(6)
Home country / Co-ethnicity / Co-Nationality	0.0725***		0.103***		0.0929***	
	(0.0198)		(0.0185)		(0.0185)	
Same company	1.030***	1.021***	0.995***	1.022***	1.028***	1.020***
	(0.0478)	(0.0397)	(0.0465)	(0.0397)	(0.0481)	(0.0397)
Home / Co-ethn. / Co-Nat. * Same company	-0.0229		0.0953		-0.0200	
	(0.0753)		(0.0790)		(0.0749)	
Home / Co-ethn. / Co-Nat. - China [§]		0.171***		0.155***		0.183***
		(0.0445)		(0.0374)		(0.0399)
Home / Co-ethn. / Co-Nat. - Germany [§]		-0.00449		0.0672**		0.0171
		(0.0264)		(0.0262)		(0.0250)
Home / Co-ethn. / Co-Nat. - France [§]		0.0842*		0.102**		0.0932**
		(0.0498)		(0.0487)		(0.0469)
Home / Co-ethn. / Co-Nat. - India [§]		-0.153*		-0.00685		-0.0251
		(0.0824)		(0.0580)		(0.0646)
Home / Co-ethn. / Co-Nat. - Iran [§]		-		-0.308		-
				(0.829)		
Home / Co-ethn. / Co-Nat. - Italy [§]		0.266**		0.160		0.240**
		(0.132)		(0.114)		(0.114)
Home / Co-ethn. / Co-Nat. - Japan [§]		0.103**		0.116***		0.121***
		(0.0411)		(0.0413)		(0.0408)
Home / Co-ethn. / Co-Nat. - Korea [§]		0.512***		0.494***		0.499***
		(0.0985)		(0.101)		(0.0983)
Home / Co-ethn. / Co-Nat. - Poland [§]		2.002		0.782		0.411
		(1.231)		(0.686)		(0.856)
Home / Co-ethn. / Co-Nat. - Russia [§]		0.380		0.481**		0.389
		(0.395)		(0.227)		(0.288)
Country proximity controls	yes	yes	yes	yes	yes	yes
Social distance dummies	yes	yes	yes	yes	yes	yes
Citing patent characteristics	yes	yes	yes	yes	yes	yes
OST F.E.	yes	yes	yes	yes	yes	yes
Constant	2.389***	2.364***	2.396***	2.380***	2.389***	2.369***
	(0.324)	(0.325)	(0.325)	(0.325)	(0.324)	(0.325)
Observations	166,672	166,672	166,672	166,672	166,672	166,671
chi2	17144	17156	17168	17174	17157	17163
ll	-106681	-106660	-106670	-106659	-106676	-106658
r2_p	0.0766	0.0768	0.0767	0.0768	0.0766	0.0768

[§] « Home country » effect in columns 1 and 2 ; co-ethnicity in columns 3 and 4 ; co-nationality in columns 5 and 6
Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

5. Discussion and conclusions

By means of patent and inventor data, we have investigated whether social ties binding migrants from the same country of origin help diffusing technical knowledge. We distinguish between diffusion in the migrant community within the same destination country ("diaspora effect") and "knowledge remittances" towards the country of origin ("brain gain effect"). We also consider the possibility of an "international diaspora" effect, by which knowledge can reach migrants from the same country of origin in different destination country. We focus on the US as main destination country and on five Asian and five European countries of origin, which we selected among the most important sources of highly skilled migration in the US.

Our empirical exercise has made use of a large and entirely novel sample of patents filed by "ethnic" inventors in the US, from 1980 to 2010. Ethnic inventors are defined as inventors whose Country of Origin (CoO) falls in a list of 10 countries that the OECD rank among the most important sources of highly skilled migration towards the US. Inventor are assigned to one or another CoO based on linguistic analysis of names and surnames. Robustness checks are conducted on subsamples of inventors whose nationality is known and overlapping with the CoO.

We find evidence of a diaspora effect to exist for all Asian countries in our sample (China, India, and, to a decreasing extent, Korea, Japan and Iran) and for one European country (Russia). However, the marginal effect of co-ethnicity is secondary to the effect of proximity in the physical space (co-location at the city or State level) and of other inventors' chief form of social proximity, namely close positioning on the network of inventors. In addition, co-ethnicity ties appear to act as substitute of both spatial and social proximity, that is to kick-in between spatially or socially distant inventors. Overall, these results appear in line with those obtained by Agrawal et al. (2008, 2011) for India, and to extend them to China and other Asian countries, but not European ones, with the exception of Russia.

As for the brain gain effect, we find that diasporic ties do not necessarily imply a knowledge transfer to the home country. In particular, we see none of this for one of the two most important inventor diasporas in the US, namely the Indian one. This may have to do more with the absorptive capacities of the country of origin, than with the international dimension of the diffusion process under consideration. In fact, for some countries that do not exhibit a brain gain effect, we find an "international diaspora" effect, which presents some analogy with Rauch's and Trindade's (2002) findings for the Chinese ethnic diaspora in trade. This is again the case for India, but also for some European countries, such as France and Germany, that did not even exhibit a diaspora effect within the US. This latter result is however not very robust and will require further investigation. In particular, we ought to test to what extent results on the brain gain and international diaspora effects are sensitive to the quality of inventor names' disambiguation and company names' harmonization.

Besides strengthening the results of this paper, our future research plans include investigating the role of ethnic ties in the formation of network of inventors, so to reconsider their role in determining the form of social proximity that we know to dominate knowledge transfer between individuals. Besides, we plan to , extend the analysis conducted in this paper to Europe, instead of the US, as the focal destination region.

This extension will contribute, among other things, to casting light on a policy–sensitive topic such as the comparative attractiveness of Europe and the US as destinations for migrant scientists and engineers (Cerna and Chou, 2014; Guild, 2007).

Appendix 1 – Inventor names' disambiguation

A key element of name disambiguation algorithms consists in measuring the edit or phonetic distance between similar names/surnames, and setting some thresholds under which different names/surnames are considered the same ("matching"). Further information contained in the patent documents, as well as benchmarking is then used to validate the matches ("filtering"). Ideally, a good algorithm would minimize both "false negatives" (maximise "recall") and "false positive" (maximise "precision"). False negatives occur whenever two inventors, whose names or surnames have been spelled or abbreviated differently on different patents, are treated as different persons. False positives occur when homonyms and quasi-homonyms are treated as the same person. Unfortunately, a trade-off exists between the two objectives, which requires making choices based on the consequences of each type of error for the subsequent analysis.

This has two consequences for the analysis of ethnic citations, based upon the linguistic analysis of inventors' names/surnames:

1. High precision/Low recall algorithms lead to underestimating the number of personal self-citations and overestimating that of co-ethnic citations. This is because all variants of the same inventor's name and surname will be, most likely, classified as belonging to the same ethnic group (for example, "Vafaie Mehrnaz" and "Vafaie Mehranz" will be both classified as Iranian, but a low recall algorithms may end up treating them as different persons, when instead they are one) When considering the two most important countries of origin of migrant inventors in the US, China and India, and before disambiguating inventors, we calculate a co-ethnic citation rate of respectively 20.5 and 15.2, which drop at 18.8 and 13.3 if we recalculate after disambiguation. . When applying the JTH methodology, this problem can be magnified by the presence of very prolific inventors, who are responsible for a large number of both cited and citing patents., and thus have the potential to generate a large number of false co-ethnic citations.[29]
2. When applied to inventor sets from different countries of origin, the same matching rules return different results in terms of pre-filtering precision and recall, due to cross-country differences in the average length of text strings containing names and surnames, and in the relative frequency of common names and surnames [30].

Three complementary strategies may help tackling these problems. The first one consists in making the best possible use of the contextual information contained in patents (that is, to correct for matching errors at the filtering stage). The second consists in using different algorithms to produce more than one datasets, each of which with different combinations of precision and recall, and using them to test the

[29] High precision/Low recall algorithms may also lead to underestimating the number of returnee inventors. If our Russian inventor patent as "Yavid Dimitriy" and as "Yavid Dimitriy" in Russia, he will not be counted as as a returnee (but his self-citations will be counted as a knowledge flow mediated by ethnicity). However, we suspect this to be a relatively minor problem, as figures of returnee inventors appear too low for their order of magnitude to change with a change in algorithms.

[30] Chinese and Korean names and 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 or Kim). The opposite holds for Russian surnames.

robustness of results. The third one consists in calibrating the disambiguation algorithm by collecting information on linguistic specificities of each country of origin, and exploit them at the matching stage. The information retrieval and computational costs increase when moving from the first to the third strategy. For this reason, our disambiguation algorithm (Massacrator 2.0) does not follow the third one.

Appendix 2 – Ethnic classification of inventors

When fed with a name and/or a surname, the *IBM-GNR system* returns a list of CoAs and two scores: [31]

- "frequency", which indicates to which percentile of the frequency distribution of names or surnames the name or surname belongs to, for each CoA;
- "significance", which approximates the frequency distribution of the name or surname across all CoA. [32]

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 A2.1 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 the algorithm is as follows:

I. We consider only inventors in the EP-INV database with at least one patent filed as US residents, or who cite at least one patent filed by US residents, and we assign them to either one of the 10 CoO of our interest, or leave her "unassigned" (which means she may be either a US "native" – whatever it means - or a migrant from other countries)

II. The 10 CoO of our interest are China, India, Iran, Japan, and South Korea (for Asia) and France, Germany, Italy, Poland, and Russia (for Europe). They share two characteristics: they belong to the top 20 CoO of highly skilled migrants in the US, according to OECD/DIOC stock figures for 2005/06 (Widmaier and Dumont, 2011); and their official language is neither English nor Spanish or Portuguese, which is a prerequisite for our algorithm to make sense when applied to migration into the US.[33]

III. For each inventor, we calculate three indicators of her likely CoO:
   a. The frequency of her name(s) in English- and Spanish-speaking CoA [34]

[31] Information on IBM-GNR reported here comes from IBM online documentation (http://www-01.ibm.com/support/knowledgecenter/SSEV5M/SSEV5M_welcome.html?lang=en; last visit: 19/1/2015) 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 them by deduction. For an application to a research topic close to ours, see Jeppesen and Lakhani (2010).

[32] For example, 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. When it comes to significance, the highest percentage of inventors names Nguyen will be found in Vietnam (say 80), followed by France (which has been historically the most important destination countries of Vietnamese migrants besides the US), and several Asian countries, with much smaller values.

[33] Language is an issue to the extent that our tools cannot distinguish English-speaking migrant inventors from US ones, nor Spanish-speaking migrants from one country of origin or another. This is why we cannot include in our analysis important origin countries such as the UK, Canada, Mexico and Cuba. We also have not yet included Ukraine and Taiwan, as this will require merging them with Russia and China, respectively. Two other countries in the top 20 list we have not included are Vietnam (too few observations among inventors) and Egypt (whose migrants into the US we cannot tell apart from those from other Arab-speaking countries).

[34] The intuition is as follows. An inventor in the USwith a typical Indian surname, such as Laroia, but named John is unlikely to be a recent Indian migrant; this is because John. More likely, he will be born in the US, possibly from mixed parents. On the contrary, Rajiv Laroia is more likely to be a first- or second-generation Indian immigrant, as Rajiv is high-frequency name in India, and a low-frequency name in English-speaking countries (some of which host sizeable communities of Indian descent).

b. The product of the significances attached to her name and to the surname, for each CoA coinciding with one of the 10 CoO of our interest. Notice that, in principle, we could find that an inventor is associated to more than one of the 10 countries of interest, either via her name or her surname (for example, a French inventor of Italian descent may have a French name and an Italian surname). However, these cases are not frequent.
c. The significance attached to the surname in the CoA associated to indicator n.2. [35]

As a result, we will have, for each inventor, one (or very few) candidate CoO and three indicators of potential success of this "candidacy".

IV. We set six possible threshold values for indicator n.1 (from 10 to 100, with steps of 20), eleven threshold values for indicator n.2 (from 0 to 10000, with steps of 1000), and six threshold values for indicator n.3 (from 50 to 100, with steps of 10). We consider 102 combinations of such threshold values ("calibrations"), and for each combination we assign each inventor to one or another CoO (or to no CoO at all). Each inventor is therefore associated to one vector of 102 dummies (one for each calibration) and a specific CoO, with dummy=1 indicating that the inventor comes for that CoO, and dummy=0 that she does not (no CoO assigned). [36]
V. We apply steps I. to IV. also to inventors in the WIPO-PCT database by Fink and Miguelez (2013), which report the inventors' nationality, which we use as benchmark to evaluate the precision and recall rates obtained by each calibration, for each CoO. We then identify Pareto-optimal calibration, namely the calibrations whose precision rate cannot be improved upon without losing out on the recall rate, and viceversa (blue dots in figures A2.2, which report the calibration results for China and Italy). Notice that the Pareto-optimal calibrations are not necessarily the same for all CoO; again from figure A2.2, one can the Pareto-Optimal calibrations for China are more convex than those for Italy. In other words, they imply a much less sharp trade-off between precision and recall: while for Italy we can attain a 70% precision rate only at the cost of reducing the recall rate to 10%, for China we reduce the latter only to 60%. The precision-recall trade-off can be considered a measure of the quality of our algorithm, per country. In general, quality is higher for Asian countries (with the exception of Iran) than for European.
VI. Finally, we retain for our analysis two calibrations per CoO: a "high recall" calibration (one that ensures the highest recall value, conditional on precision being at least 30%); and a "high precision" calibration, one that requires precision to be no less than 70% . High recall values may include a large number of false positives (inventors wrongly assigned to one or another of the 10 CoO of interest), but also accommodate for a looser definition of migrant inventors, one that includes late-generation migrants. The latter's validity depends on the strength of ties binding such migrants to other US residents of the same descent and/or to their countries of origin (on which we have no *a priori* information).

[35] The intuition is as follows: the indicator n.2 may have a high value due exclusively to a very high value of the significance for the name, with a moderate value for the significance of the surname. We wish the latter not to be too low.

[36] Keeping with the example from the previous footnotes, Rajiv Laroia will be associated to CoO=India, with a vector containing n<102 zeroes and 102-n ones. The ones are all associated with "high recall" combinations of high threshold values for indicator n.1 and low threshold values for nr.2 and nr.3 (such as, respectively, 70-5000-60; see figure 1), while the zeroes will be associated with "high precision" combinations (low threshold values for indicator n.1 and high threshold values for nr.2 and nr.3; such as, respectively, 30-8000-80). Rajiv Laroia will be confirmed having CoO=India only in the high recall case, but not in the high precision case (for which indicator nr.1 is too high). In practice, the high precision combination leaves the door open to Rajiv Laroia's CoO being the UK, and to Rajiv Laroia being possibly of Indian descent, but with no ties to India or to Indian migrants in the US.

In the present version of the paper, we make use only of "high recall" calibration results. To further compare data quality across CoO, we inspect the frequency distribution of values taken by indicator n.2 (figure A2.3). The more right (left) skewed the distribution, the better(worse) the quality: the most striking comparison here is between India and Italy, with the former clearly exhibiting higher quality. According to this measure, too, quality is generally higher for Asian countries (with the exception of Iran) than for European ones.

Figure A2.1 From inventor data to the Ethnic-INV database

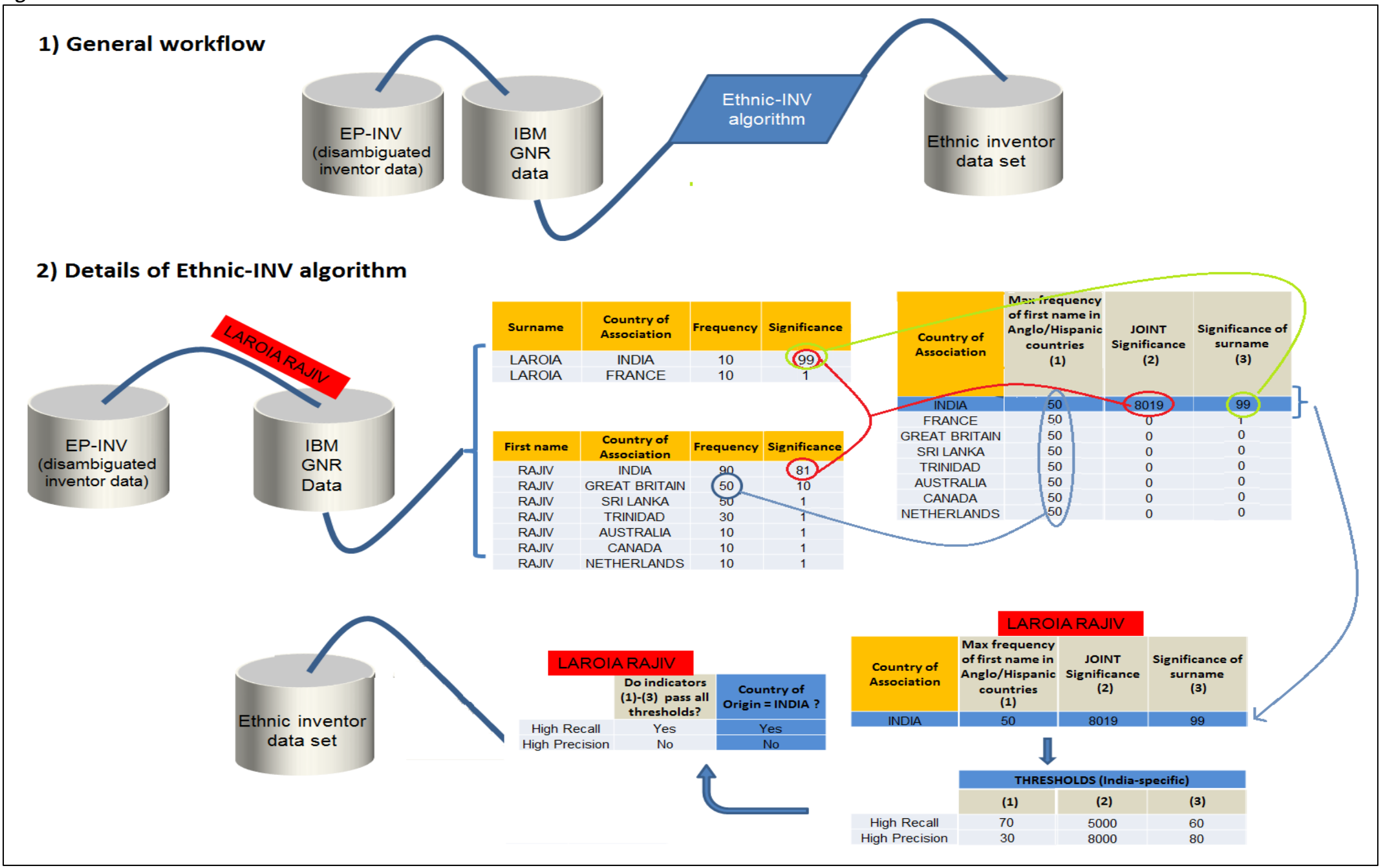


Figure A2.2a - Ethnic-INV algorithm calibration results: China and Italy

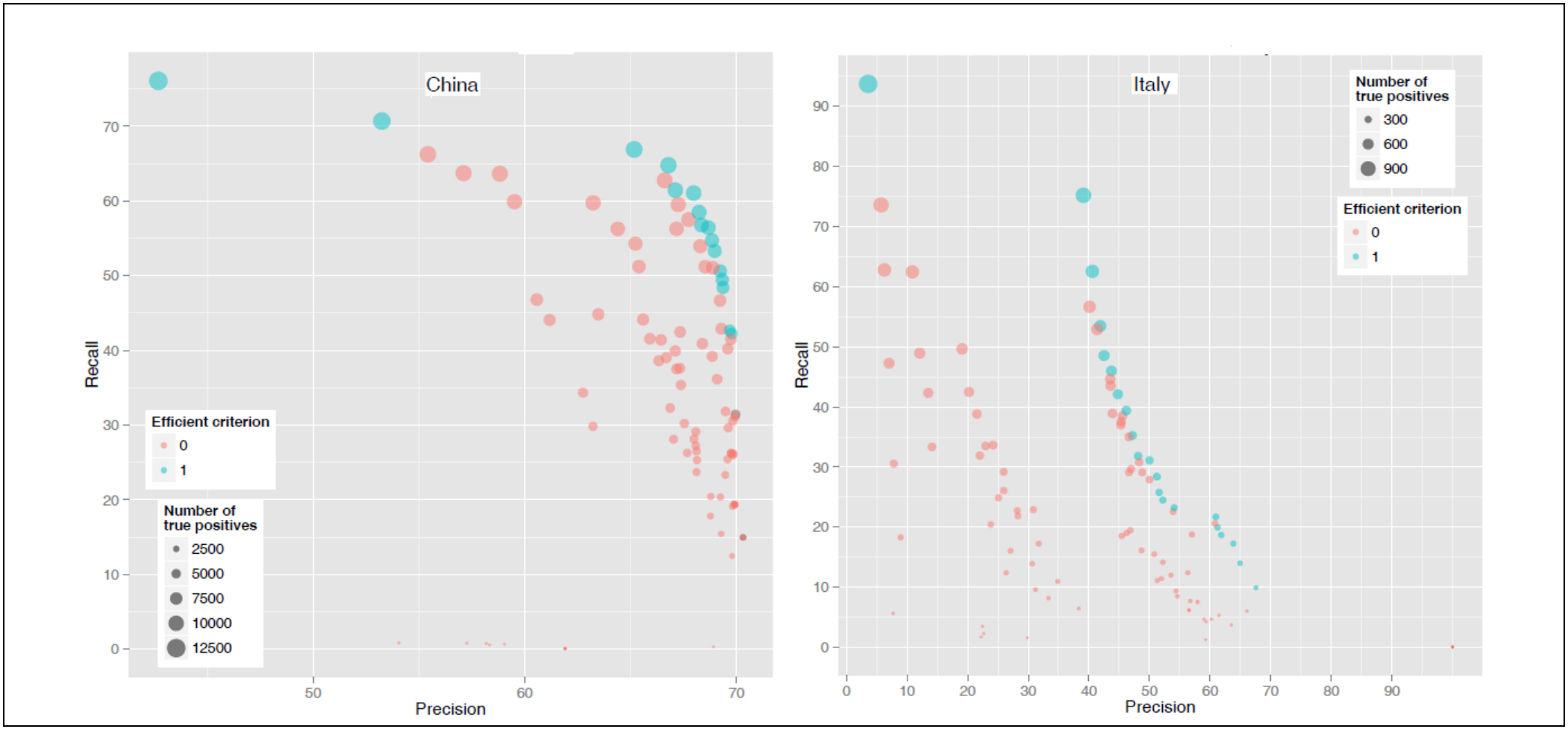
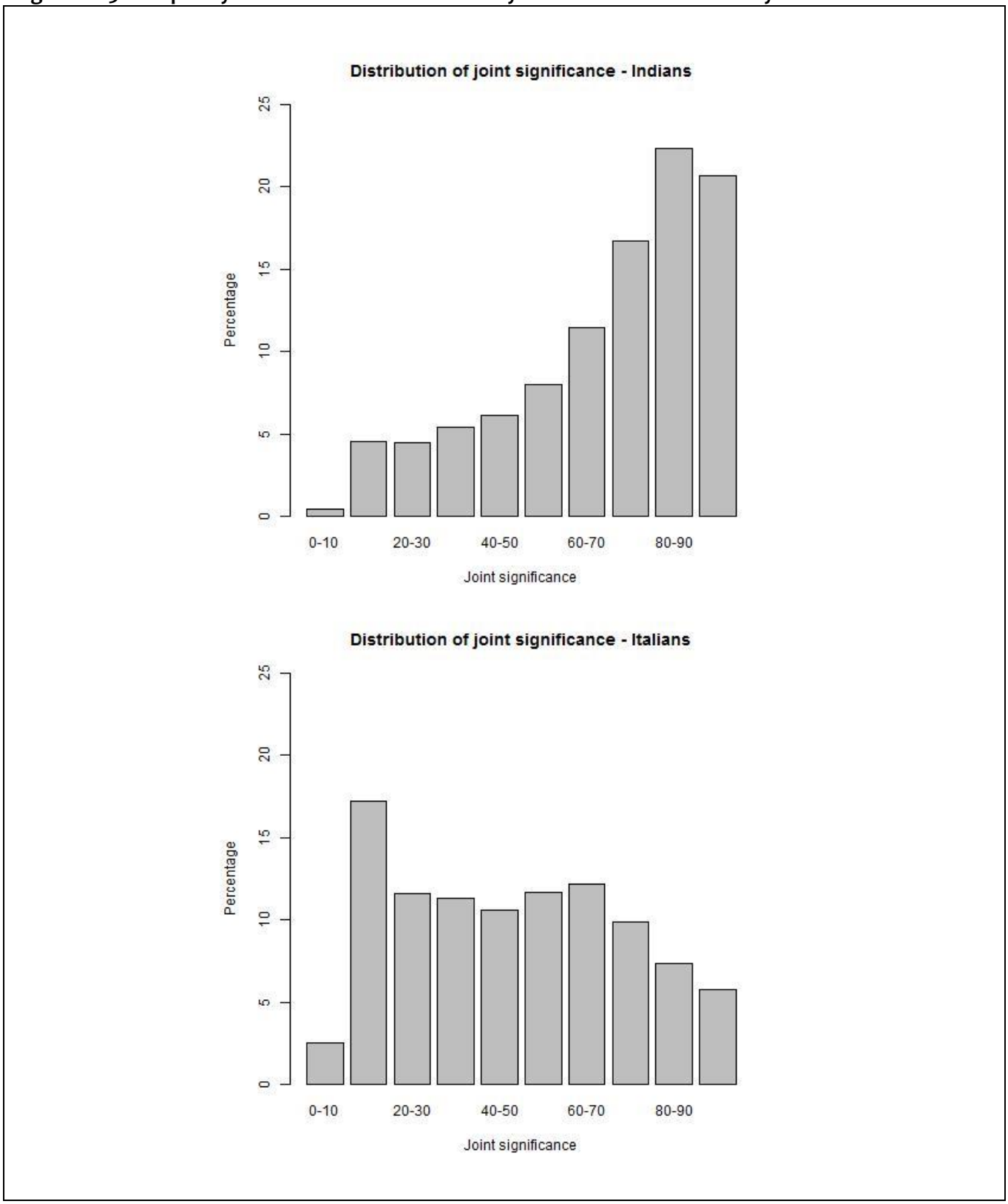


Figure A2.3 - Frequency distribution of values taken by indicator n.2: India vs. Italy



Appendix 3 – Additional tables

Figure A3.1 – Share of ethnic inventors of EPO patent applications by US residents; by CoO

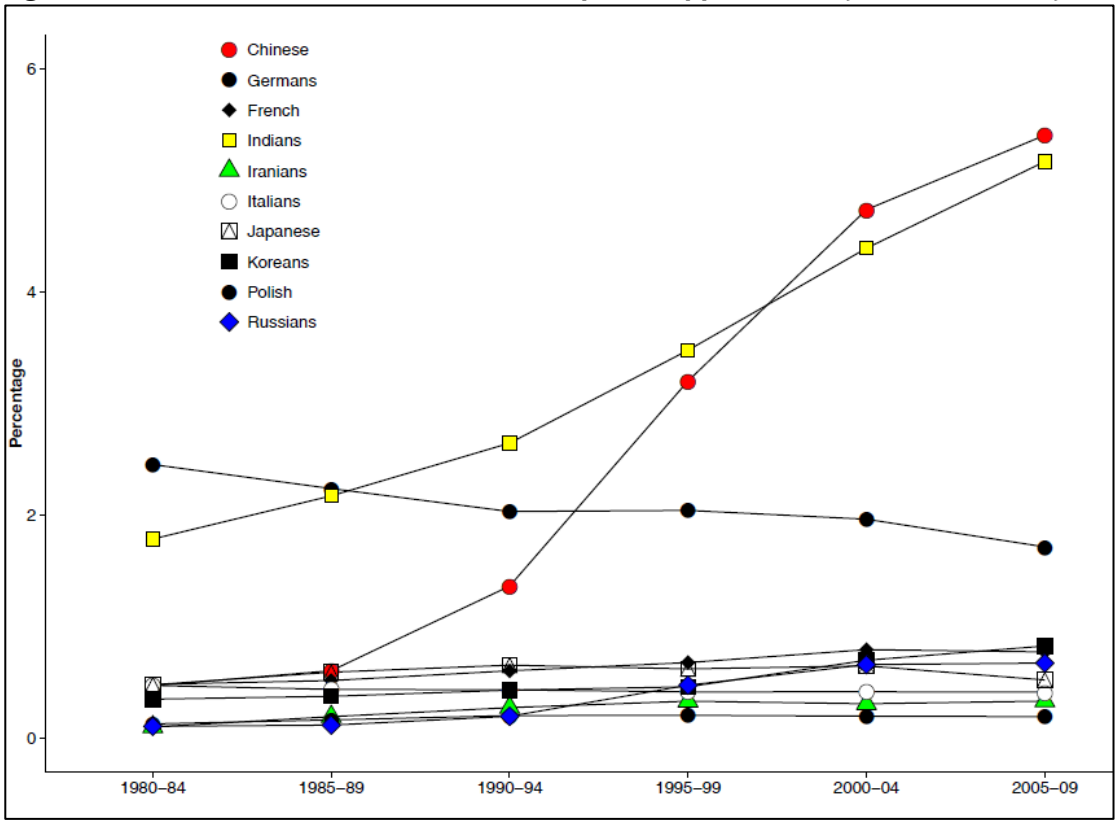


Figure A3.2 – Ethnic inventors' share of EPO patent applications by US residents; by CoO

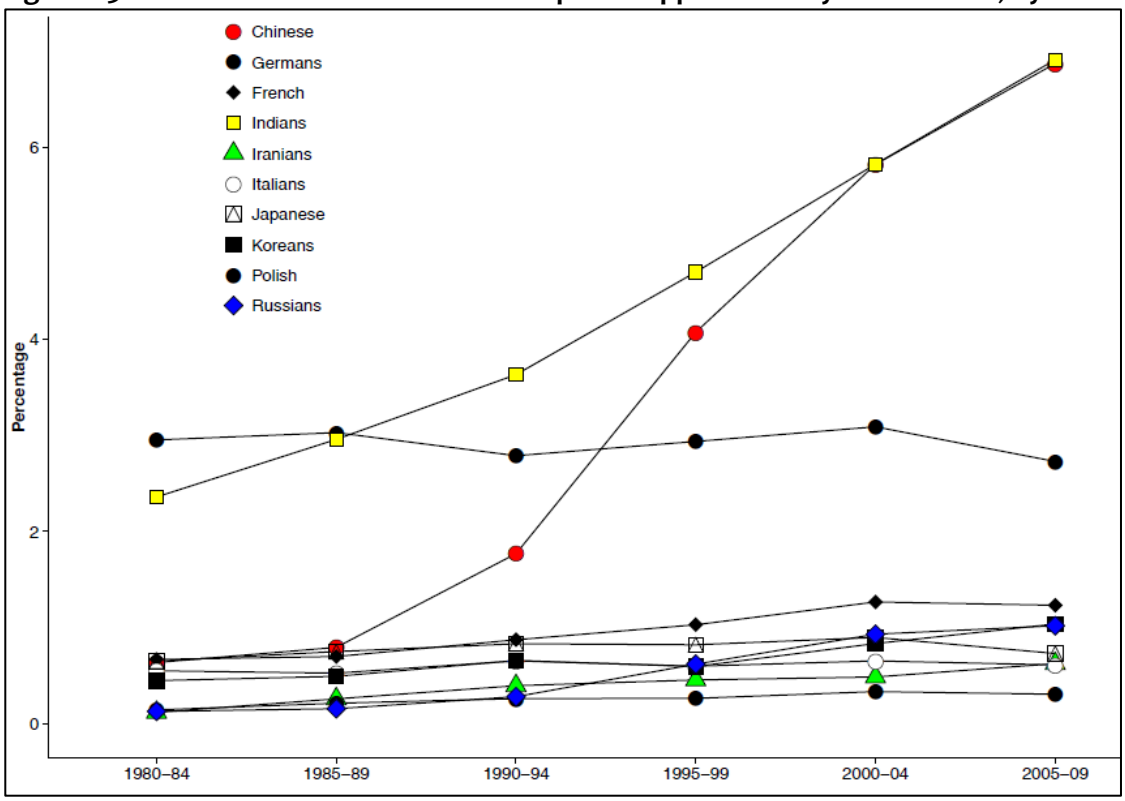


Figure A3.3

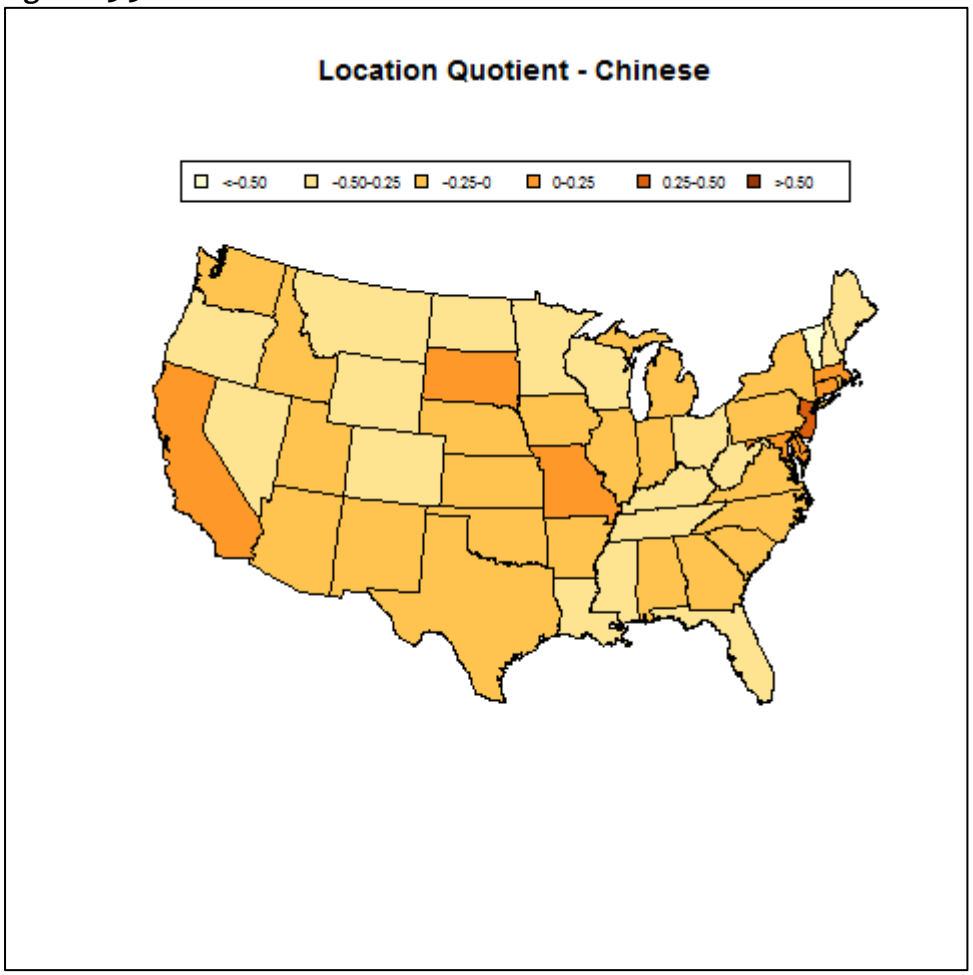


Figure A3.4

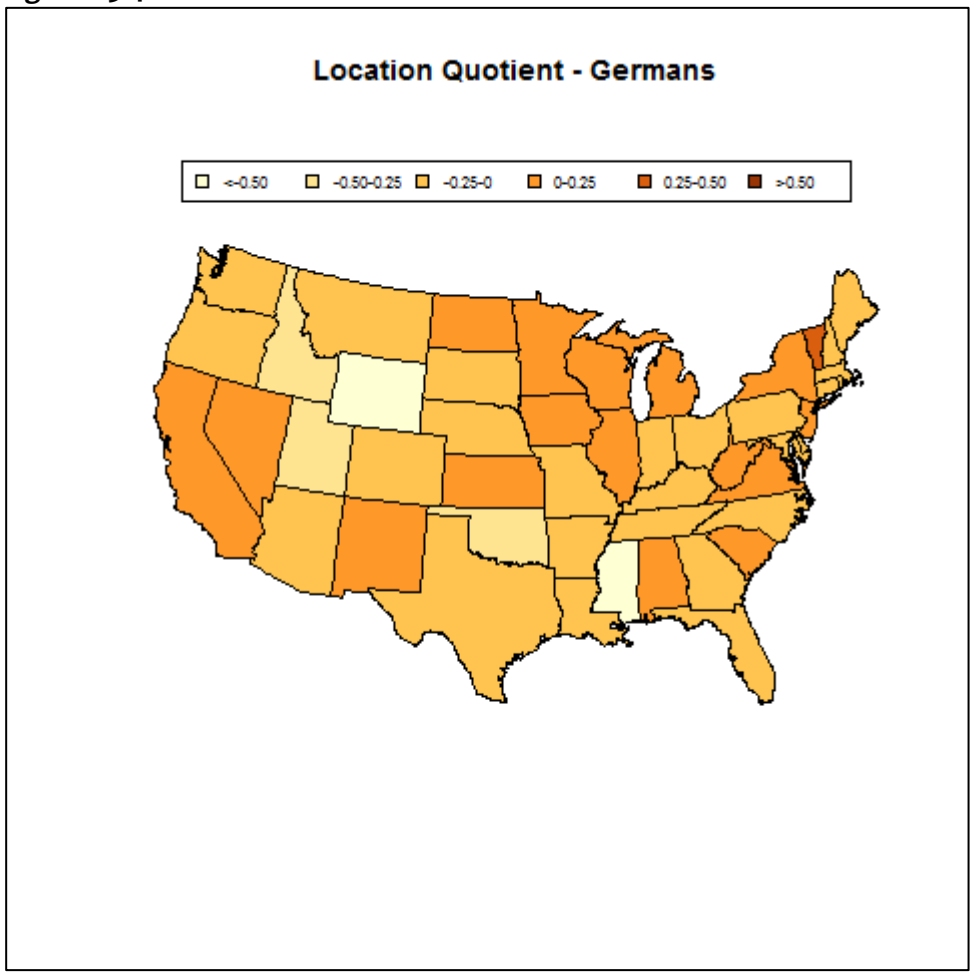


Figure A3.5

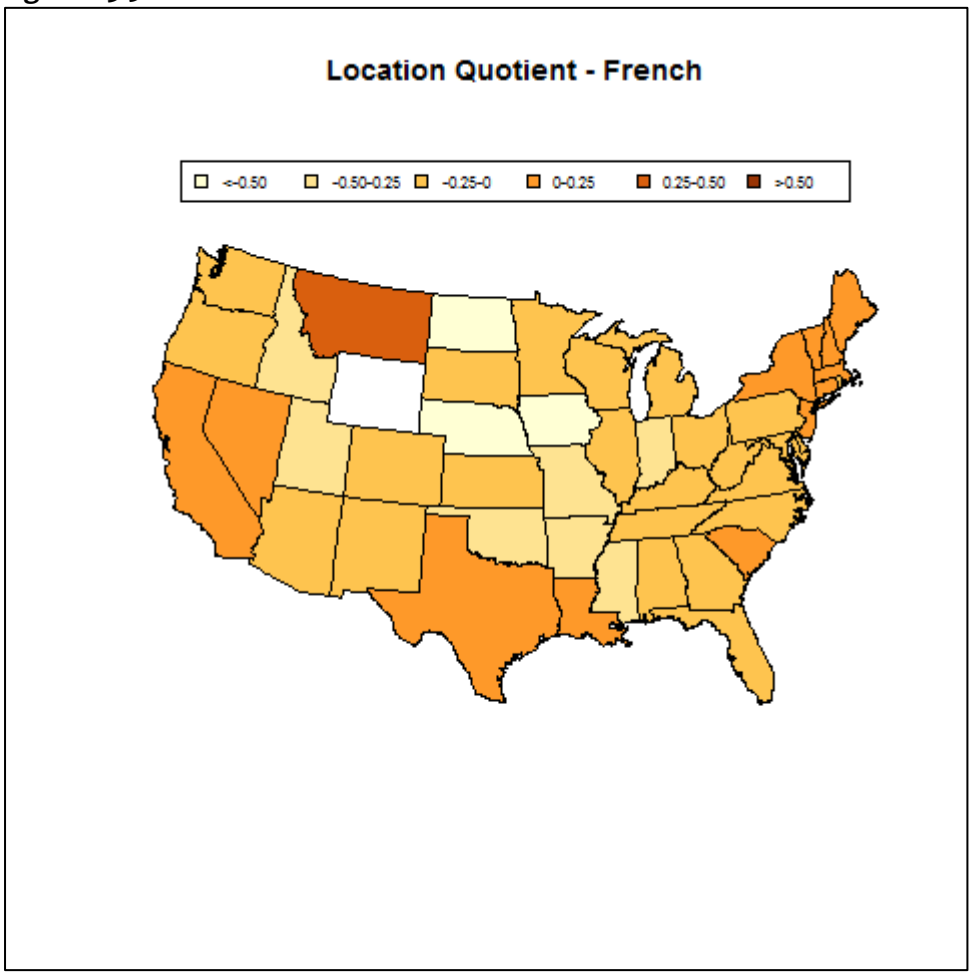


Figure A3.6

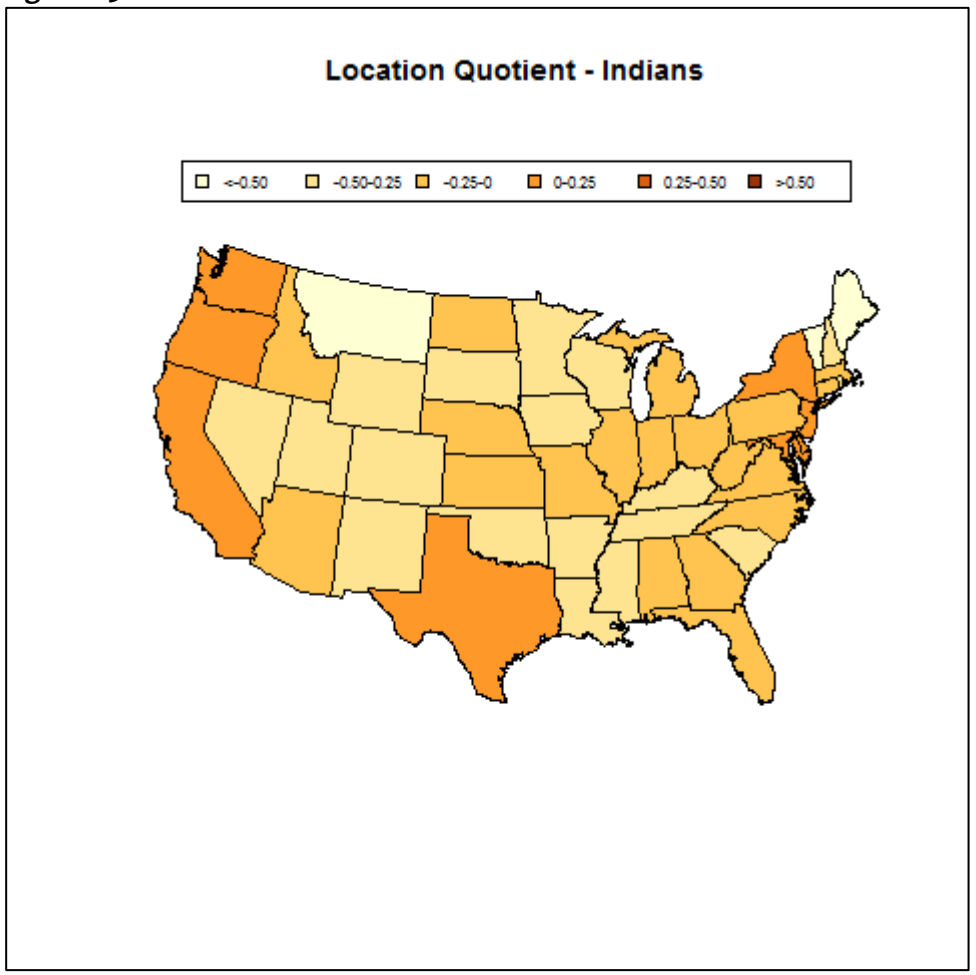


Figure A3.7

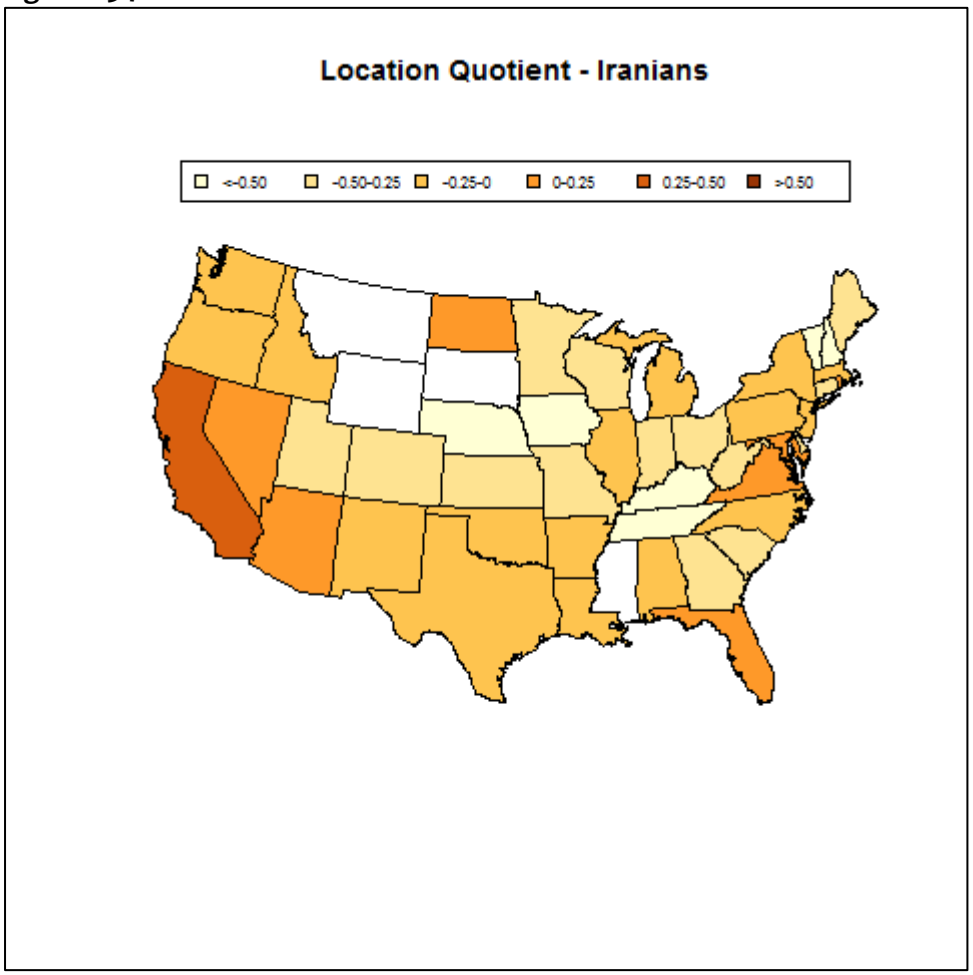


Figure A3.8

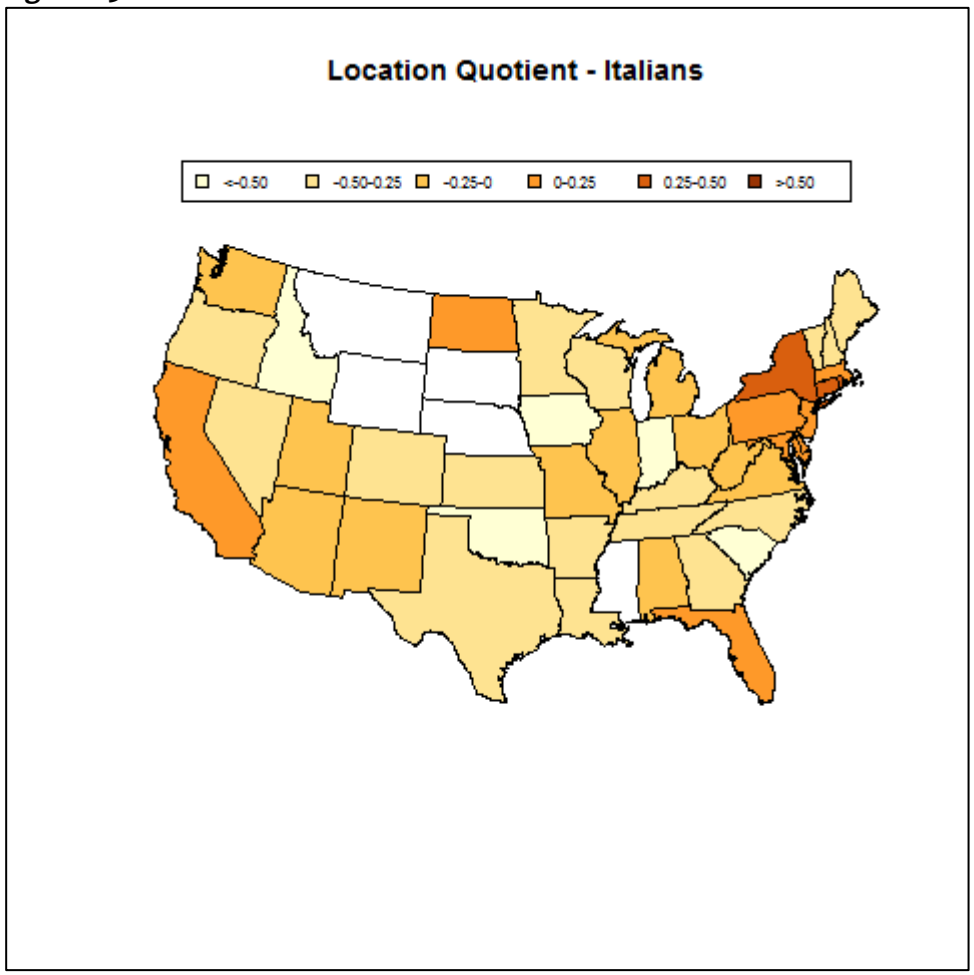


Figure A3.9

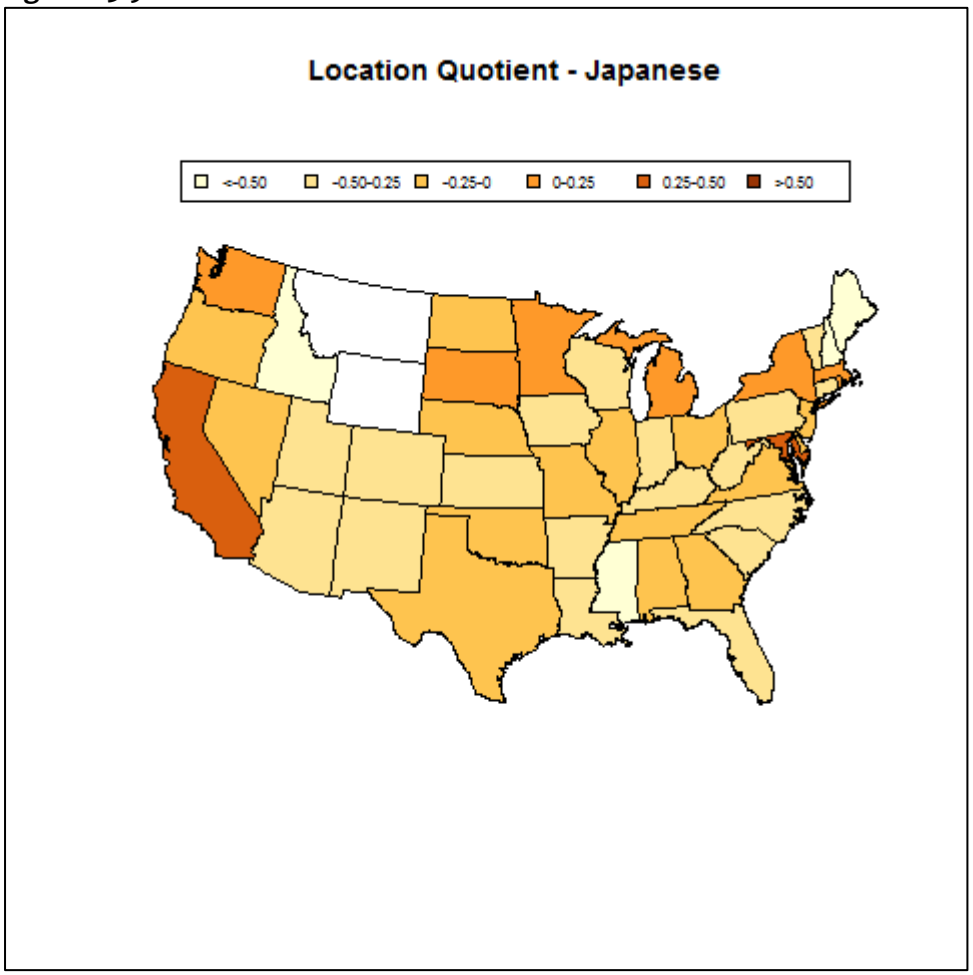


Figure A3.10

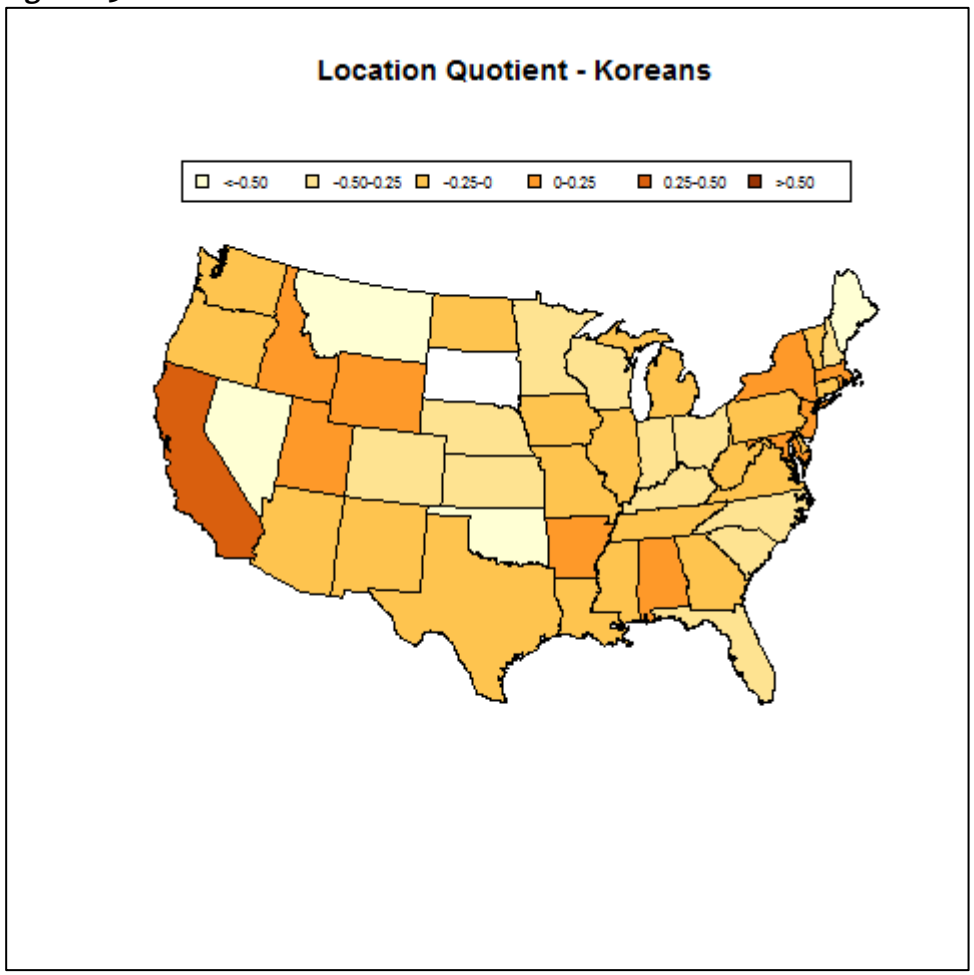


Figure A3.11

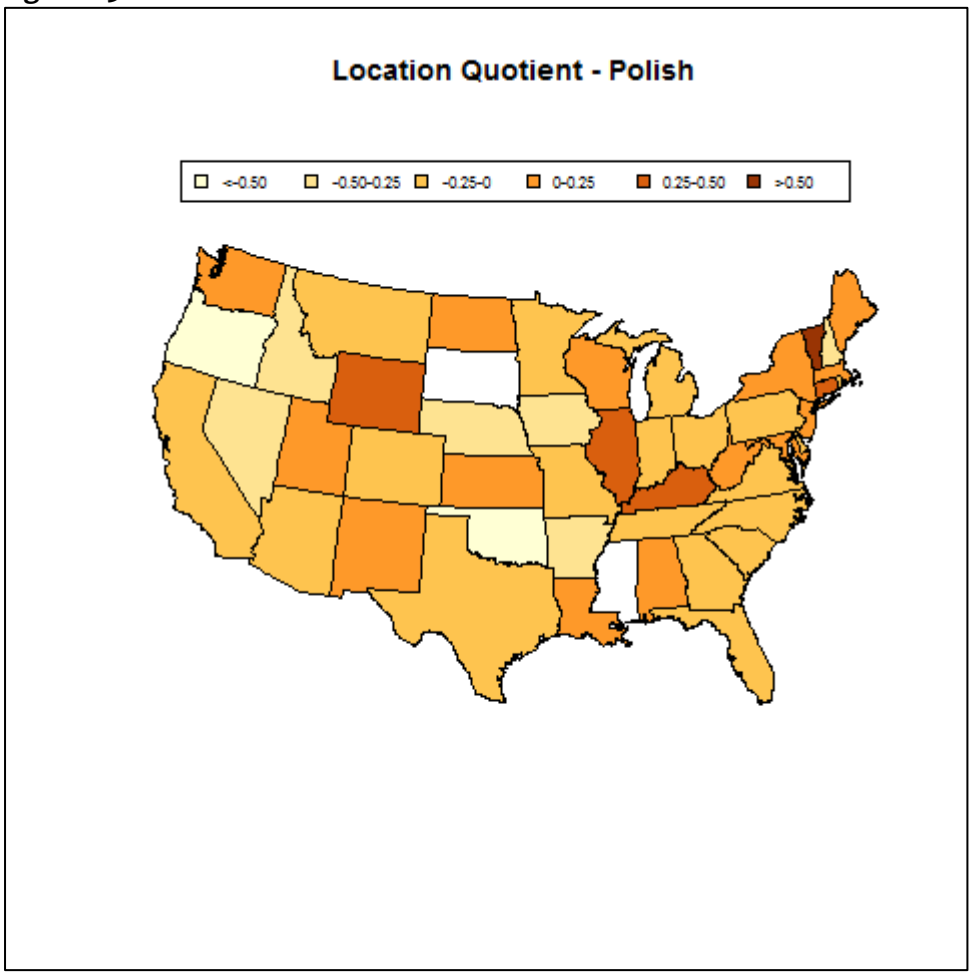


Figure A3.12

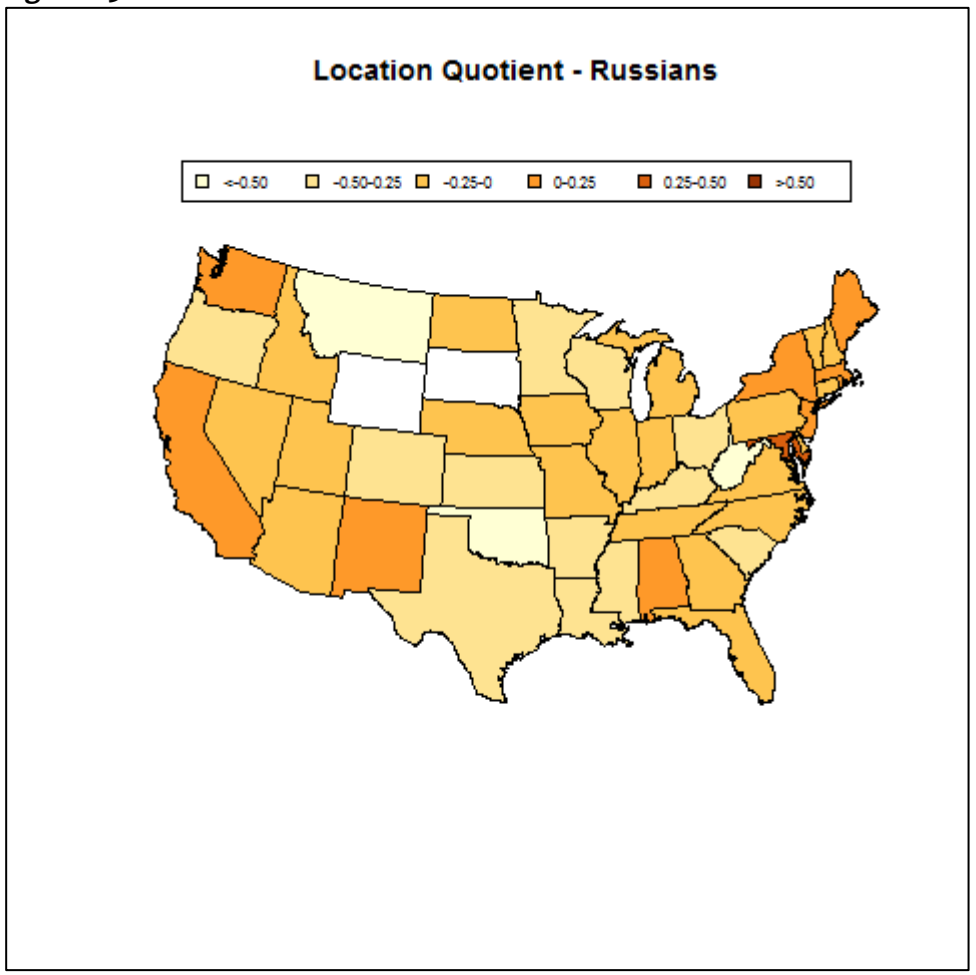


Table A3.13 Local and international samples: descriptive statistics. China

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	249682	0.500	0.500	0	1
Co-ethnicity	249682	0.219	0.414	0	1
Same MSA	249682	0.143	0.350	0	1
Same State	249682	0.226	0.418	0	1
Miles	249682	937.849	890.481	0	5081.5
Soc. Dist. 0	249682	0.010	0.099	0	1
Soc. Dist. 1	249682	0.010	0.099	0	1
Soc. Dist. 2	249682	0.007	0.086	0	1
Soc. Dist. 3	249682	0.010	0.100	0	1
Soc. Dist. >3	249682	0.297	0.457	0	1
Soc. Dist. ∞	249682	0.666	0.472	0	1
#claims	249682	7.810	12.638	0	235
backward citations	249682	4.516	3.183	0	87
NPL citations	249682	1.557	2.632	0	57
overlap IPCs 7 digits	249682	1.250	1.680	0	27
overlap IPCs 7 digits / all IPCs	249682	0.269	0.270	0	1
overlap IPCs	249682	0.893	1.773	0	53
2. International sample (citations from outside the US)					
Citation	256586	0.500	0.500	0	1
Co-ethnicity	256586	0.040	0.196	0	1
Home country	256586	0.025	0.157	0	1
Same company	256586	0.025	0.155	0	1
Contiguous countries	256586	0.036	0.186	0	1
Former colonial relationship	256586	0.207	0.405	0	1
Same country	256586	0.022	0.146	0	1
English	256586	0.181	0.385	0	1
Similarity to English	256586	0.242	0.259	0	1
Miles	256586	4606.738	1822.956	0	11498.1
Soc. Dist. 1	256586	0.003	0.056	0	1
Soc. Dist. 2	256586	0.005	0.069	0	1
Soc. Dist. 3	256586	0.004	0.062	0	1
Soc. Dist. >3	256586	0.005	0.071	0	1
Soc. Dist. ∞	256586	0.247	0.432	0	1
#claims	256586	0.736	0.441	0	1
backward citations	256586	9.713	12.006	0	383
backward NPL citations	256586	3.967	3.260	0	98
overlap IPCs 7 digits	256586	1.187	2.253	0	76
overlap IPCs 7 digits / all IPCs	256586	1.168	1.401	0	31
overlap IPCs	256586	0.303	0.288	0	1

Table A3.14 Local and international samples: descriptive statistics. Germany

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	175878	0.500	0.500	0	1
Co-ethnicity	175878	0.073	0.261	0	1
Same MSA	175878	0.131	0.337	0	1
Same State	175878	0.212	0.409	0	1
Miles	175878	910.112	850.556	0	5085.412
Soc. Dist. 0	175878	0.008	0.091	0	1
Soc. Dist. 1	175878	0.009	0.095	0	1
Soc. Dist. 2	175878	0.006	0.077	0	1
Soc. Dist. 3	175878	0.007	0.082	0	1
Soc. Dist. >3	175878	0.210	0.407	0	1
Soc. Dist. ∞	175878	0.760	0.427	0	1
#claims	175878	8.957	12.856	0	259
backward citations	175878	4.726	3.134	0	68
NPL citations	175878	1.152	2.320	0	49
overlap IPCs 7 digits	175878	1.096	1.384	0	23
overlap IPCs 7 digits / all IPCs	175878	0.298	0.296	0	1
overlap IPCs	175878	0.817	1.473	0	28
2. International sample (citations from outside the US)					
Citation	177976	0.500	0.500	0	1
Co-ethnicity	177976	0.237	0.425	0	1
Home country	177976	0.301	0.459	0	1
Same company	177976	0.038	0.190	0	1
Contiguous countries	177976	0.029	0.168	0	1
Former colonial relationship	177976	0.191	0.393	0	1
Same country	177976	0.072	0.259	0	1
English	177976	0.163	0.369	0	1
Similarity to English	177976	0.272	0.261	0	1
Miles	177976	4161.850	2128.470	0	11083.11
Soc. Dist. 1	177976	0.004	0.064	0	1
Soc. Dist. 2	177976	0.010	0.098	0	1
Soc. Dist. 3	177976	0.007	0.083	0	1
Soc. Dist. >3	177976	0.006	0.079	0	1
Soc. Dist. ∞	177976	0.166	0.373	0	1
#claims	177976	0.807	0.395	0	1
backward citations	177976	9.791	11.533	0	442
backward NPL citations	177976	4.115	3.195	0	98
overlap IPCs 7 digits	177976	0.803	1.913	0	76
overlap IPCs 7 digits / all IPCs	177976	1.050	1.187	0	19
overlap IPCs	177976	0.329	0.307	0	1

Table A3.15 Local and international samples: descriptive statistics. France

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	66256	0.500	0.500	0	1
Co-ethnicity	66256	0.037	0.188	0	1
Same MSA	66256	0.146	0.353	0	1
Same State	66256	0.230	0.421	0	1
Miles	66256	923.499	878.817	0	5024.276
Soc. Dist. 0	66256	0.010	0.098	0	1
Soc. Dist. 1	66256	0.008	0.088	0	1
Soc. Dist. 2	66256	0.006	0.079	0	1
Soc. Dist. 3	66256	0.008	0.089	0	1
Soc. Dist. >3	66256	0.234	0.423	0	1
Soc. Dist. ∞	66256	0.734	0.442	0	1
#claims	66256	8.367	12.562	0	197
backward citations	66256	4.634	3.170	0	64
NPL citations	66256	1.263	2.393	0	50
overlap IPCs 7 digits	66256	1.190	1.579	0	24
overlap IPCs 7 digits / all IPCs	66256	0.294	0.290	0	1
overlap IPCs	66256	0.884	1.730	0	40
2. International sample (citations from outside the US)					
Citation	68250	0.500	0.500	0	1
Co-ethnicity	68250	0.107	0.310	0	1
Home country	68250	0.115	0.319	0	1
Same company	68250	0.036	0.185	0	1
Contiguous countries	68250	0.037	0.189	0	1
Former colonial relationship	68250	0.220	0.414	0	1
Same country	68250	0.055	0.228	0	1
English	68250	0.184	0.387	0	1
Similarity to English	68250	0.248	0.257	0	0.6666667
Miles	68250	4113.260	2182.504	0	11045.67
Soc. Dist. 1	68250	0.006	0.078	0	1
Soc. Dist. 2	68250	0.007	0.084	0	1
Soc. Dist. 3	68250	0.006	0.075	0	1
Soc. Dist. >3	68250	0.006	0.075	0	1
Soc. Dist. ∞	68250	0.185	0.388	0	1
#claims	68250	0.791	0.407	0	1
backward citations	68250	9.805	11.698	0	292
backward NPL citations	68250	4.012	3.141	0	55
overlap IPCs 7 digits	68250	0.990	2.077	0	33
overlap IPCs 7 digits / all IPCs	68250	1.144	1.391	0	22
overlap IPCs	68250	0.328	0.302	0	1

Table A3.16 Local and international samples: descriptive statistics. India

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	324456	0.500	0.500	0	1
Co-ethnicity	324456	0.164	0.370	0	1
Same MSA	324456	0.137	0.343	0	1
Same State	324456	0.210	0.408	0	1
Miles	324456	928.169	870.342	0	5082.868
Soc. Dist. 0	324456	0.007	0.085	0	1
Soc. Dist. 1	324456	0.007	0.086	0	1
Soc. Dist. 2	324456	0.006	0.078	0	1
Soc. Dist. 3	324456	0.008	0.086	0	1
Soc. Dist. >3	324456	0.227	0.419	0	1
Soc. Dist. ∞	324456	0.745	0.436	0	1
#claims	324456	8.729	12.964	0	235
backward citations	324456	4.534	3.136	0	87
NPL citations	324456	1.252	2.342	0	53
overlap IPCs 7 digits	324456	1.071	1.356	0	26
overlap IPCs 7 digits / all IPCs	324456	0.281	0.284	0	1
overlap IPCs	324456	0.789	1.475	0	47
2. International sample (citations from outside the US)					
Citation	68250	0.500	0.500	0	1
Co-ethnicity	68250	0.107	0.310	0	1
Same country	68250	0.115	0.319	0	1
Home company	68250	0.036	0.185	0	1
Contiguous countries	68250	0.037	0.189	0	1
Former colonial relationship	68250	0.220	0.414	0	1
Same country	68250	0.055	0.228	0	1
English	68250	0.184	0.387	0	1
Similarity to English	68250	0.248	0.257	0	0.6666667
Miles	68250	4113.260	2182.504	0	11045.67
Soc. Dist. 1	68250	0.006	0.078	0	1
Soc. Dist. 2	68250	0.007	0.084	0	1
Soc. Dist. 3	68250	0.006	0.075	0	1
Soc. Dist. >3	68250	0.006	0.075	0	1
Soc. Dist. ∞	68250	0.185	0.388	0	1
#claims	68250	0.791	0.407	0	1
backward citations	68250	9.805	11.698	0	292
backward NPL citations	68250	4.012	3.141	0	55
overlap IPCs 7 digits	68250	0.990	2.077	0	33
overlap IPCs 7 digits / all IPCs	68250	1.144	1.391	0	22
overlap IPCs	68250	0.328	0.302	0	1

Table A3.17 Local and international samples: descriptive statistics. Iran

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	29102	0.500	0.500	0	1
Co-ethnicity	29102	0.015	0.123	0	1
Same MSA	29102	0.159	0.365	0	1
Same State	29102	0.269	0.444	0	1
Miles	29102	1002.144	909.576	0	5073.808
Soc. Dist. 0	29102	0.008	0.089	0	1
Soc. Dist. 1	29102	0.008	0.091	0	1
Soc. Dist. 2	29102	0.007	0.082	0	1
Soc. Dist. 3	29102	0.007	0.081	0	1
Soc. Dist. >3	29102	0.193	0.395	0	1
Soc. Dist. ∞	29102	0.777	0.416	0	1
#claims	29102	8.570	12.272	0	227
backward citations	29102	4.618	2.997	0	50
NPL citations	29102	0.930	1.919	0	26
overlap IPCs 7 digits	29102	0.940	1.097	0	26
overlap IPCs 7 digits / all IPCs	29102	0.308	0.304	0	1
overlap IPCs	29102	0.716	1.271	0	39
2. International sample (citations from outside the US)					
Citation	26716	0.500	0.500	0	1
Co-ethnicity	26716	0.003	0.054	0	1
Home country	26716	0.000	0.012	0	1
Same company	26716	0.023	0.149	0	1
Contiguous countries	26716	0.032	0.175	0	1
Former colonial relationship	26716	0.187	0.390	0	1
Same country	26716	0.028	0.166	0	1
English	26716	0.176	0.381	0	1
Similarity to English	26716	0.239	0.257	0	0.6666667
Miles	26716	4733.390	1790.990	0	11026.53
Soc. Dist. 1	26716	0.004	0.064	0	1
Soc. Dist. 2	26716	0.005	0.070	0	1
Soc. Dist. 3	26716	0.006	0.078	0	1
Soc. Dist. >3	26716	0.004	0.061	0	1
Soc. Dist. ∞	26716	0.154	0.361	0	1
#claims	26716	0.827	0.378	0	1
backward citations	26716	10.390	12.109	0	265
backward NPL citations	26716	3.999	3.075	0	37
overlap IPCs 7 digits	26716	0.847	1.847	0	33
overlap IPCs 7 digits / all IPCs	26716	0.945	1.016	0	13
overlap IPCs	26716	0.326	0.309	0	1

Table A3.18 Local and international samples: descriptive statistics. Italy

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	46736	0.500	0.500	0	1
Co-ethnicity	46736	0.020	0.139	0	1
Same MSA	46736	0.127	0.333	0	1
Same State	46736	0.207	0.405	0	1
Miles	46736	947.778	882.275	0	4929.127
Soc. Dist. 0	46736	0.008	0.089	0	1
Soc. Dist. 1	46736	0.008	0.087	0	1
Soc. Dist. 2	46736	0.008	0.088	0	1
Soc. Dist. 3	46736	0.007	0.084	0	1
Soc. Dist. >3	46736	0.206	0.405	0	1
Soc. Dist. ∞	46736	0.763	0.425	0	1
#claims	46736	9.112	12.926	0	235
backward citations	46736	4.570	3.089	0	44
NPL citations	46736	1.354	2.508	0	57
overlap IPCs 7 digits	46736	1.183	1.518	0	27
overlap IPCs 7 digits / all IPCs	46736	0.301	0.289	0	1
overlap IPCs	46736	0.876	1.628	0	43
2. International sample (citations from outside the US)					
Citation	46342	0.500	0.500	0	1
Co-ethnicity	46342	0.043	0.204	0	1
Home country	46342	0.043	0.203	0	1
Same company	46342	0.026	0.158	0	1
Contiguous countries	46342	0.035	0.183	0	1
Former colonial relationship	46342	0.208	0.406	0	1
Same country	46342	0.027	0.161	0	1
English	46342	0.178	0.382	0	1
Similarity to English	46342	0.254	0.257	0	1
Miles	46342	4388.072	1969.418	0	11270.66
Soc. Dist. 1	46342	0.004	0.064	0	1
Soc. Dist. 2	46342	0.006	0.077	0	1
Soc. Dist. 3	46342	0.005	0.070	0	1
Soc. Dist. >3	46342	0.004	0.063	0	1
Soc. Dist. ∞	46342	0.187	0.390	0	1
#claims	46342	0.794	0.404	0	1
backward citations	46342	10.138	11.766	0	383
backward NPL citations	46342	3.932	3.059	0	69
overlap IPCs 7 digits	46342	0.997	2.094	0	39
overlap IPCs 7 digits / all IPCs	46342	1.089	1.228	0	18
overlap IPCs	46342	0.325	0.301	0	1

Table A3.19 Local and international samples: descriptive statistics. Japan

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	48280	0.500	0.500	0	1
Co-ethnicity	48280	0.027	0.163	0	1
Same MSA	48280	0.137	0.344	0	1
Same State	48280	0.229	0.421	0	1
Miles	48280	995.816	912.131	0	5085.159
Soc. Dist. 0	48280	0.006	0.080	0	1
Soc. Dist. 1	48280	0.006	0.080	0	1
Soc. Dist. 2	48280	0.004	0.066	0	1
Soc. Dist. 3	48280	0.005	0.069	0	1
Soc. Dist. >3	48280	0.213	0.409	0	1
Soc. Dist. ∞	48280	0.765	0.424	0	1
#claims	48280	8.991	13.192	0	247
backward citations	48280	4.504	3.206	0	64
NPL citations	48280	1.626	2.814	0	45
overlap IPCs 7 digits	48280	1.191	1.462	0	27
overlap IPCs 7 digits / all IPCs	48280	0.287	0.278	0	1
overlap IPCs	48280	0.872	1.523	0	32
2. International sample (citations from outside the US)					
Citation	53238	0.500	0.500	0	1
Co-ethnicity	53238	0.276	0.447	0	1
Home country	53238	0.284	0.451	0	1
Same company	53238	0.047	0.212	0	1
Contiguous countries	53238	0.032	0.176	0	1
Former colonial relationship	53238	0.189	0.392	0	1
Same country	53238	0.124	0.330	0	1
English	53238	0.163	0.370	0	1
Similarity to English	53238	0.231	0.259	0	1
Miles	53238	4032.253	2150.099	0	11046.71
Soc. Dist. 1	53238	0.004	0.063	0	1
Soc. Dist. 2	53238	0.008	0.091	0	1
Soc. Dist. 3	53238	0.005	0.073	0	1
Soc. Dist. >3	53238	0.004	0.067	0	1
Soc. Dist. ∞	53238	0.171	0.376	0	1
#claims	53238	0.807	0.394	0	1
backward citations	53238	10.232	12.029	0	442
backward NPL citations	53238	4.003	3.214	0	79
overlap IPCs 7 digits	53238	1.072	2.176	0	41
overlap IPCs 7 digits / all IPCs	53238	1.139	1.283	0	19
overlap IPCs	53238	0.314	0.290	0	1

Table A3.20 Local and international samples: descriptive statistics. Korea

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	51868	0.500	0.500	0	1
Co-ethnicity	51868	0.030	0.170	0	1
Same MSA	51868	0.139	0.345	0	1
Same State	51868	0.224	0.417	0	1
Miles	51868	930.996	894.078	0	4841.666
Soc. Dist. 0	51868	0.009	0.095	0	1
Soc. Dist. 1	51868	0.008	0.088	0	1
Soc. Dist. 2	51868	0.007	0.082	0	1
Soc. Dist. 3	51868	0.008	0.091	0	1
Soc. Dist. >3	51868	0.236	0.425	0	1
Soc. Dist. ∞	51868	0.732	0.443	0	1
#claims	51868	8.501	12.768	0	197
backward citations	51868	4.590	3.150	0	58
NPL citations	51868	1.313	2.435	0	50
overlap IPCs 7 digits	51868	1.117	1.407	0	22
overlap IPCs 7 digits / all IPCs	51868	0.273	0.282	0	1
overlap IPCs	51868	0.787	1.449	0	25
2. International sample (citations from outside the US)					
Citation	49142	0.500	0.500	0	1
Co-ethnicity	49142	0.046	0.210	0	1
Home country	49142	0.047	0.211	0	1
Same company	49142	0.022	0.147	0	1
Contiguous countries	49142	0.031	0.172	0	1
Former colonial relationship	49142	0.203	0.403	0	1
Same country	49142	0.026	0.158	0	1
English	49142	0.175	0.380	0	1
Similarity to English	49142	0.241	0.259	0	0.6666667
Miles	49142	4595.749	1834.877	0	11043.31
Soc. Dist. 1	49142	0.003	0.056	0	1
Soc. Dist. 2	49142	0.005	0.071	0	1
Soc. Dist. 3	49142	0.004	0.060	0	1
Soc. Dist. >3	49142	0.004	0.065	0	1
Soc. Dist. ∞	49142	0.196	0.397	0	1
#claims	49142	0.788	0.409	0	1
backward citations	49142	9.974	11.764	0	240
backward NPL citations	49142	4.073	3.318	0	98
overlap IPCs 7 digits	49142	0.993	2.050	0	58
overlap IPCs 7 digits / all IPCs	49142	1.078	1.257	0	32
overlap IPCs	49142	0.301	0.291	0	1

Table A3.21 Local and international samples: descriptive statistics. Poland

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	16084	0.500	0.500	0	1
Co-ethnicity	16084	0.008	0.088	0	1
Same MSA	16084	0.116	0.320	0	1
Same State	16084	0.166	0.372	0	1
Miles	16084	923.325	842.636	0	4849.524
Soc. Dist. 0	16084	0.013	0.111	0	1
Soc. Dist. 1	16084	0.010	0.099	0	1
Soc. Dist. 2	16084	0.005	0.068	0	1
Soc. Dist. 3	16084	0.006	0.079	0	1
Soc. Dist. >3	16084	0.213	0.409	0	1
Soc. Dist. ∞	16084	0.754	0.431	0	1
#claims	16084	8.531	12.631	0	209
backward citations	16084	4.652	3.135	0	64
NPL citations	16084	1.281	2.515	0	49
overlap IPCs 7 digits	16084	1.171	1.586	0	18
overlap IPCs 7 digits / all IPCs	16084	0.304	0.299	0	1
overlap IPCs	16084	0.835	1.543	0	19
2. International sample (citations from outside the US)					
Citation	16696	0.500	0.500	0	1
Co-ethnicity	16696	0.005	0.068	0	1
Home country	16696	0.001	0.033	0	1
Same company	16696	0.022	0.147	0	1
Contiguous countries	16696	0.032	0.177	0	1
Former colonial relationship	16696	0.198	0.398	0	1
Same country	16696	0.021	0.143	0	1
English	16696	0.167	0.373	0	1
Similarity to English	16696	0.259	0.260	0	1
Miles	16696	4518.551	1765.418	0	11026.99
Soc. Dist. 1	16696	0.003	0.059	0	1
Soc. Dist. 2	16696	0.005	0.074	0	1
Soc. Dist. 3	16696	0.004	0.060	0	1
Soc. Dist. >3	16696	0.002	0.045	0	1
Soc. Dist. ∞	16696	0.196	0.397	0	1
#claims	16696	0.789	0.408	0	1
backward citations	16696	10.101	11.914	0	212
backward NPL citations	16696	4.160	3.312	0	79
overlap IPCs 7 digits	16696	0.966	2.130	0	34
overlap IPCs 7 digits / all IPCs	16696	1.060	1.171	0	16
overlap IPCs	16696	0.327	0.306	0	1

Table A3.22 Local and international samples: descriptive statistics. Russia

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	36546	0.500	0.500	0	1
Co-ethnicity	36546	0.029	0.169	0	1
Same MSA	36546	0.139	0.346	0	1
Same State	36546	0.223	0.416	0	1
Miles	36546	948.115	889.507	0	5080.685
Soc. Dist. 0	36546	0.012	0.107	0	1
Soc. Dist. 1	36546	0.008	0.088	0	1
Soc. Dist. 2	36546	0.005	0.071	0	1
Soc. Dist. 3	36546	0.005	0.073	0	1
Soc. Dist. >3	36546	0.206	0.405	0	1
Soc. Dist. ∞	36546	0.764	0.425	0	1
#claims	36546	7.761	12.165	0	195
backward citations	36546	4.724	3.139	0	64
NPL citations	36546	1.254	2.458	0	45
overlap IPCs 7 digits	36546	0.947	1.148	0	21
overlap IPCs 7 digits / all IPCs	36546	0.293	0.302	0	1
overlap IPCs	36546	0.686	1.263	0	27
2. International sample (citations from outside the US)					
Citation	38264	0.500	0.500	0	1
Co-ethnicity	38264	0.012	0.110	0	1
Home country	38264	0.005	0.071	0	1
Same company	38264	0.020	0.141	0	1
Contiguous countries	38264	0.034	0.182	0	1
Former colonial relationship	38264	0.181	0.385	0	1
Same country	38264	0.026	0.160	0	1
English	38264	0.164	0.370	0	1
Similarity to English	38264	0.255	0.261	0	0.6666667
Miles	38264	4508.039	1839.596	0	11053.67
Soc. Dist. 1	38264	0.004	0.067	0	1
Soc. Dist. 2	38264	0.004	0.063	0	1
Soc. Dist. 3	38264	0.003	0.057	0	1
Soc. Dist. >3	38264	0.003	0.057	0	1
Soc. Dist. ∞	38264	0.185	0.389	0	1
#claims	38264	0.800	0.400	0	1
backward citations	38264	9.564	12.032	0	383
backward NPL citations	38264	4.036	3.149	0	79
overlap IPCs 7 digits	38264	0.920	1.975	0	24
overlap IPCs 7 digits / all IPCs	38264	0.961	1.069	0	20
overlap IPCs	38264	0.321	0.306	0	1

Appendix 4 – Additional tables

Table A4.1. Top-10 cross-MSA citation corridors

MSA name	MSA name	Citations (both directions)
San Jose-Sunnyvale-Santa Clara, CA	San Francisco-Oakland-Fremont, CA	8931.80
New York-Northern New Jersey-Long Island, NY-NJ-PA	San Francisco-Oakland-Fremont, CA	7194.53
New York-Northern New Jersey-Long Island, NY-NJ-PA	Boston-Cambridge-Quincy, MA-NH	6846.82
New York-Northern New Jersey-Long Island, NY-NJ-PA	San Diego-Carlsbad-San Marcos, CA	6834.77
San Francisco-Oakland-Fremont, CA	Boston-Cambridge-Quincy, MA-NH	6702.78
New York-Northern New Jersey-Long Island, NY-NJ-PA	San Jose-Sunnyvale-Santa Clara, CA	5909.32
San Francisco-Oakland-Fremont, CA	San Diego-Carlsbad-San Marcos, CA	5059.78
New York-Northern New Jersey-Long Island, NY-NJ-PA	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	4866.75
San Jose-Sunnyvale-Santa Clara, CA	Boston-Cambridge-Quincy, MA-NH	4496.28
New York-Northern New Jersey-Long Island, NY-NJ-PA	Chicago-Joliet-Naperville, IL-IN-WI	3638.95

Table A4.2. Baseline regressions with additional FE

	(1)	(2)	(5)	(6)
Same MSA	0.162***	0.162***	0.132***	0.132***
	(0.0108)	(0.0108)	(0.0110)	(0.0110)
Same State	0.113***	0.113***	0.0691***	0.0685***
	(0.00822)	(0.00822)	(0.00859)	(0.00859)
ln(Miles)	-0.0245***	-0.0245***	-0.0348***	-0.0349***
	(0.00199)	(0.00199)	(0.00210)	(0.00210)
Co-ethnic	0.153***		0.153***	
	(0.00608)		(0.00609)	
Co-ethnic*MSA				
China		0.229***		0.228***
		(0.00859)		(0.00859)
Germany		0.0209		0.0203
		(0.0178)		(0.0178)
France		-0.0712*		-0.0718*
		(0.0422)		(0.0422)
India		0.123***		0.124***
		(0.00868)		(0.00868)
Iran		0.234**		0.228**
		(0.0981)		(0.0985)
Italy		0.0122		0.0156
		(0.0692)		(0.0693)
Japan		0.0943*		0.0969*
		(0.0572)		(0.0572)
Korea		0.130**		0.129**
		(0.0530)		(0.0529)
Poland		-0.256		-0.241
		(0.179)		(0.179)
Russia		0.284***		0.288***
		(0.0614)		(0.0616)
Productive MSA FE	yes	yes	no	no
Top-10 FE	no	no	yes	yes
Soc.dist dummies	yes	yes	yes	yes
Patent characteristics	yes	yes	yes	yes
OST FE	yes	yes	yes	yes
Constant	2.420***	2.421***	2.460***	2.462***
	(0.0725)	(0.0725)	(0.0725)	(0.0725)
Observations	1,044,888	1,044,888	1,044,888	1,044,888
chi2	93252	93335	93644	93727
ll	-678767	-678680	-678845	-678760
r2_p	0.0628	0.0629	0.0627	0.0628

Table A4.3. Baseline regressions with year FE

	(1)	(2)	(3)
Same MSA	0.141***	0.141***	0.152***
	(0.0107)	(0.0107)	(0.0110)
Same State	0.0836***	0.0836***	0.0833***
	(0.00774)	(0.00775)	(0.00774)
ln(Miles)	-0.0279***	-0.0280***	-0.0280***
	(0.00195)	(0.00195)	(0.00195)
Co-ethnic	0.145***		0.157***
	(0.00611)		(0.00663)
Co-ethnic * MSA			-0.0775***
			(0.0175)
China		0.214***	
		(0.00865)	
Germany		0.0240	
		(0.0179)	
France		-0.0706*	
		(0.0422)	
India		0.120***	
		(0.00869)	
Iran		0.231**	
		(0.0979)	
Italy		0.0192	
		(0.0694)	
Japan		0.0970*	
		(0.0573)	
Korea		0.101*	
		(0.0531)	
Poland		-0.245	
		(0.178)	
Russia		0.290***	
		(0.0617)	
Social distance dummies	yes	yes	yes
Citing patent characteristics	yes	yes	yes
OST FE	yes	yes	yes
Year FE	yes	yes	yes
Constant	2.452***	2.459***	2.449***
	(0.0775)	(0.0776)	(0.0775)
Observations	1,044,888	1,044,888	1,044,888
ll	-678128	-678055	-678119
r2_p	0.0637	0.0638	0.0637

Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table A4.4. Baseline regressions with year FE

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	sample of 10% size, 50 reps.				sample of 5% size, 50 reps.				sample of 1% size, 50 reps.		
Same MSA	0.141***	0.152***	0.140***	0.152***	0.141***	0.152***	0.140***	0.152***	0.141**	0.152**	0.140**
	(0.0221)	(0.0209)	(0.0221)	(0.0211)	(0.0279)	(0.0279)	(0.0277)	(0.0273)	(0.0677)	(0.0673)	(0.0638)
Same State	0.0919***	0.0915***	0.0912***	0.0909***	0.0919***	0.0915***	0.0912***	0.0909***	0.0919*	0.0915	0.0912
	(0.0137)	(0.0137)	(0.0138)	(0.0138)	(0.0203)	(0.0203)	(0.0202)	(0.0206)	(0.0559)	(0.0559)	(0.0626)
ln(Miles)	-0.0266***	-0.0268***	-0.0269***	-0.0270***	-0.0266***	-0.0268***	-0.0269***	-0.0270***	-0.0266**	-0.0268**	-0.0269**
	(0.00381)	(0.00382)	(0.00380)	(0.00336)	(0.00580)	(0.00579)	(0.00579)	(0.00641)	(0.0120)	(0.0120)	(0.0133)
Co-ethnic	0.153***	0.165***			0.153***	0.165***			0.153***	0.165***	
	(0.0137)	(0.0160)			(0.0196)	(0.0216)			(0.0375)	(0.0400)	
Co-ethnic * MSA		-0.0791*				-0.0791				-0.0791	
		(0.0404)				(0.0502)				(0.101)	
China			0.229***	0.250***			0.229***	0.250***			0.229***
			(0.0162)	(0.0195)			(0.0229)	(0.0252)			(0.0444)
Germany			0.0213	0.0143			0.0213	0.0143			0.0213
			(0.0346)	(0.0375)			(0.0467)	(0.0559)			(0.106)
France			-0.0741	-0.0433			-0.0741	-0.0433			-0.0741
			(0.0877)	(0.102)			(0.130)	(0.132)			(0.205)
India			0.124***	0.133***			0.124***	0.133***			0.124*
			(0.0163)	(0.0199)			(0.0270)	(0.0262)			(0.0665)
Iran			0.228	0.169			0.228	0.169			0.228
			(0.173)	(0.225)			(0.271)	(0.259)			(0.668)
Italy			0.0185	0.0495			0.0185	0.0495			0.0185
			(0.124)	(0.140)			(0.207)	(0.237)			(0.434)
Japan			0.0899	0.118			0.0899	0.118			0.0899
			(0.116)	(0.130)			(0.141)	(0.162)			(0.410)
Korea			0.128	0.152			0.128	0.152			0.128
			(0.120)	(0.110)			(0.168)	(0.173)			(0.334)
Poland			-0.240	-0.258			-0.240	-0.258			-0.240
			(0.439)	(0.490)			(0.693)	(0.688)			(0.721)
Russia			0.286**	0.221*			0.286*	0.221			0.286
			(0.122)	(0.117)			(0.148)	(0.158)			(0.369)
China * MSA				-0.135**				-0.135**			
				(0.0559)				(0.0685)			
Germany * MSA				0.0479				0.0479			
				(0.121)				(0.174)			
France * MSA				-0.174				-0.174			
				(0.264)				(0.315)			
India * MSA				-0.0584				-0.0584			
				(0.0633)				(0.0811)			
Iran * MSA				0.245				0.245			
				(0.531)				(0.766)			

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	sample of 10% size, 50 reps.				sample of 5% size, 50 reps.				sample of 1% size, 50 reps.		
Italy * MSA				-0.198				-0.198			
				(0.392)				(0.640)			
Japan * MSA				-0.184				-0.184			
				(0.292)				(0.415)			
Korea * MSA				-0.128				-0.128			
				(0.327)				(0.463)			
Poland * MSA				0.105				0.105			
				(0.810)				(0.897)			
Russia * MSA				0.453				0.453			
				(0.432)				(0.534)			
Soc.dist. dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Patent characteristics	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
OST FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Constant	2.420***	2.417***	2.422***	2.419***	2.420***	2.417***	2.422***	2.419***	2.420***	2.417***	2.422***
	(0.177)	(0.177)	(0.177)	(0.174)	(0.260)	(0.261)	(0.260)	(0.201)	(0.651)	(0.651)	(0.659)
Observations	1,044,888	1,044,889	1,044,890	1,044,891	1,044,892	1,044,893	1,044,894	1,044,895	1,044,896	1,044,897	1,044,898
chi2	26668	26655	48699	171796	17663	17859	26982	.	2943	2961	.
ll	-679100	-679090	-679014	-678991	-679100	-679090	-679014	-678991	-679100	-679090	-679014
r2_p	0.0625	0.0625	0.0626	0.0626	0.0625	0.0625	0.0626	0.0626	0.0625	0.0625	0.0626

Appendix 5 – "Brain gain" regression analysis: additional tables

Table a5.1

	HOME COUNTRY	CO-ETHNICITY		HOME COUNTRY	CO-ETHNICITY
	(1)	(2)		(1)	(2)
Same company	1.130***	1.110***			
	(0.0203)	(0.0198)			
China §	0.177***	0.176***	China * Same company §	0.148	0.240
	(0.0259)	(0.0213)		(0.229)	(0.193)
Germany §	-0.00445	0.0882***	Germany * Same company §	-0.151***	0.0535
	(0.00886)	(0.0102)		(0.0454)	(0.0571)
France §	0.0332	0.170***	France * Same company §	0.380***	0.252**
	(0.0243)	(0.0252)		(0.101)	(0.123)
India §	0.0233	0.0866***	India * Same company §	0.211	0.257
	(0.0411)	(0.0286)		(0.240)	(0.201)
Iran §	0.330	0.0504	Iran * Same company §		
	(1.078)	(0.245)			
Italy §	-0.0840*	0.00544	Italy * Same company §	0.750***	0.495*
	(0.0483)	(0.0474)		(0.268)	(0.260)
Japan §	0.00262	0.00690	Japan * Same company §	0.282***	0.276***
	(0.0157)	(0.0159)		(0.0806)	(0.0818)
Korea §	0.442***	0.432***	Korea * Same company §	-0.647***	-0.706***
	(0.0435)	(0.0439)		(0.188)	(0.181)
Poland §	1.225**	0.152	Poland * Same company §		
	(0.599)	(0.279)			
Russia §	0.652***	0.487***	Russia * Same company §	-0.927	
	(0.159)	(0.102)		(1.091)	
Country proximity controls	yes	yes			
Social distance dummies	yes	yes			
Citing patent characteristics	yes	yes			
OST F.E.	yes	yes			
Constant	1.883***	1.901***			
	(0.160)	(0.160)			
Observations	1,050,236	1,050,214			
chi2	125033	125174			
ll	-657625	-657570			
r2_p	0.0966	0.0967			

§ « Home country » effect in column 1 ; co-nationality in column 2. Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table a5.2

	HOME COUNTRY	CO-ETHNICITY	NATIONALITY
Same company	1.017***	0.988***	1.015***
	(0.0479)	(0.0466)	(0.0483)
China §	0.184***	0.153***	0.186***
	(0.0446)	(0.0376)	(0.0401)
Germany §	0.0197	0.0723***	0.0384
	(0.0275)	(0.0272)	(0.0261)
France §	0.0451	0.0725	0.0618
	(0.0532)	(0.0514)	(0.0501)
India §	-0.166**	-0.00910	-0.0249
	(0.0847)	(0.0590)	(0.0658)
Iran §		-0.309	
		(0.829)	
Italy §	0.126	0.145	0.163
	(0.146)	(0.120)	(0.123)
Japan §	0.0559	0.0786*	0.0756*
	(0.0430)	(0.0432)	(0.0427)
Korea §	0.546***	0.534***	0.537***
	(0.1000)	(0.102)	(0.0995)
Poland §	2.003	0.782	0.411
	(1.231)	(0.686)	(0.856)
Russia §	0.379	0.466**	0.389
	(0.396)	(0.229)	(0.288)
China * Same company §	-0.845**	0.104	-0.172
	(0.377)	(0.347)	(0.364)
Germany * Same company §	-0.198**	-0.0174	-0.182**
	(0.0864)	(0.0940)	(0.0864)
France * Same company §	0.311*	0.336*	0.269
	(0.171)	(0.189)	(0.164)
India * Same company §	0.391	0.0961	-0.00983
	(0.441)	(0.360)	(0.368)
Iran * Same company §			
Italy * Same company §	0.926**	0.184	0.638
	(0.434)	(0.410)	(0.402)
Japan * Same company §	0.619***	0.540***	0.603***
	(0.196)	(0.194)	(0.197)
Korea * Same company §	-0.827*	-0.916**	-0.959**
	(0.457)	(0.426)	(0.426)
Poland * Same company §			
Russia * Same company §			
Country proximity controls	yes	yes	yes
Social distance dummies	yes	yes	yes
Citing patent characteristics	yes	yes	yes
OST F.E.	yes	yes	yes
Constant	2.386***	2.391***	2.382***
	(0.325)	(0.325)	(0.325)
Observations	166,671	166,668	166,670
chi2	17141	17166	17142
ll	-106642	-106650	-106644
r2_p	0.0769	0.0768	0.0769

§ « Home country » effect in column 1 ; co-ethnicity in column 2; Co-nationality in column 3. Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

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(5), 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(2), 258-269.

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(1), 43-55.

Alcacer, J., Gittelman, M., 2006. Patent citations as a measure of knowledge flows: The influence of examiner citations. The Review of Economics and Statistics, 88(4), 774-779.

Almeida, P., Phene, A., Li, S., 2014. The Influence of Ethnic Community Knowledge on Indian Inventor Innovativeness. Organization Science.

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(9), 1534-1543.

Blomström, M., Kokko, A., 1998. Multinational corporations and spillovers. Journal of Economic surveys, 12(3), 247-277.

Borgatti, S.P., Everett, M.G., 1997. Network analysis of 2-mode data. Social networks, 19(3), 243-269.

Boschma, R., Frenken, K., 2011. The emerging empirics of evolutionary economic geography. Journal of Economic Geography, 11(2), 295-307.

Breschi, S., 2011, The geography of knowledge flows. In: P. Cooke, B.T. Asheim, R. Boschma, R. Martin, D. Schwartz, F. Tödtling (Eds.). Handbook of Regional Innovation and Growth. Edward Elgar Publishing.

Breschi, S., Lissoni, F., 2001. Localised knowledge spillovers vs. innovative milieux: Knowledge "tacitness" reconsidered. Papers in Regional Science, 80(3), 255-273.

Breschi, S., Lissoni, F., 2005, Knowledge networks from patent data. In: H.F. Moed, W. Glänzel, U. Schmoch (Eds.). Handbook of quantitative science and technology research. Springer Science+Business Media, Berlin, pp. 613-643.

Breschi, S., Lissoni, F., 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. Journal of Economic Geography, 9(4), 439-468.

Cerna, L., Chou, M.-H., 2014. The regional dimension in the global competition for talent: Lessons from framing the European Scientific Visa and Blue Card. Journal of European Public Policy, 21(1), 76-95.

Currarini, S., Jackson, M.O., Pin, P., 2009. An economic model of friendship: Homophily, minorities, and segregation. Econometrica, 77(4), 1003-1045.

Docquier, F., Marfouk, A., 2006, International migration by educational attainment (1990-2000). In: Ç. Özden, M. Schiff (Eds.). International migration, remittances and the brain drain. The World Bank - Palgrave Macmillan, New York, pp. 151-199.

Du Plessis, M., Van Looy, B., Song, X., Magerman, T., 2009. Data production methods for harmonized patent indicators: Assignee sector allocation, Luxembourg.

Ellison, G., Glaeser, E.L., Kerr, W., 2007. What causes industry agglomeration? Evidence from coagglomeration patterns, National Bureau of Economic Research.

Foley, C.F., Kerr, W.R., 2011. Ethnic Innovation and US Multinational Firm Activity, National Bureau of Economic Research, Inc.

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(5), 393-406.

Guild, E., 2007. EU Policy on Labour Migration: A First Look at the Commission's Blue Card Initiative. CEPS Policy brief(145).

Hall, B.H., Jaffe, A., Trajtenberg, M., 2005. Market value and patent citations. RAND Journal of economics, 16-38.

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.

Harhoff, D., Scherer, F.M., Vopel, K., 2003. Citations, family size, opposition and the value of patent rights. Research Policy, 32(8), 1343-1363.

Henderson, V., 1997. Externalities and industrial development. Journal of urban economics, 42(3), 449-470.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. The Quarterly Journal of Economics, 108(3), 577-598.

Jeppesen, L.B., Lakhani, K.R., 2010. Marginality and problem-solving effectiveness in broadcast search. Organization science, 21(5), 1016-1033.

Kapur, D., 2001. Diasporas and technology transfer. Journal of Human Development, 2(2), 265-286.

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(2), 391-407.

Kerr, W.R., 2007. The Ethnic Composition of US Inventors. 08-006.

Kerr, W.R., 2008. Ethnic Scientific Communities and International Technology Diffusion. Review of Economics and Statistics, 90(3), 518-537.

Kerr, W.R., Lincoln, W.F., 2010. The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention. Journal of Labor Economics, 28(3), 473-508.

Krugman, P., 2011. The new economic geography, now middle-aged. Regional Studies, 45(1), 1-7.

Krugman, P.R., 1991, Geography and trade. MIT press.

Kuznetsov, Y. (Ed.), 2006. Diaspora networks and the international migration of skills: how countries can draw on their talent abroad. World Bank Publications, Washington, DC.

Kuznetsov, Y. (Ed.), 2010. Talent Abroad Promoting Growth and Institutional Development at Home: Skilled Diaspora as Part of the Country. World Bank, Washington, DC (http://hdl.handle.net/10986/10117).

Li, G.-C., Lai, R., D'Amour, A., Doolin, D.M., Sun, Y., Torvik, V.I., Yu, A.Z., Fleming, L., 2014. Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). Research Policy, 43(6), 941-955.

Lissoni, F., 2010. Academic inventors as brokers. Research Policy, 39(7), 843-857.

Lissoni, F., 2012. Academic patenting in Europe: An overview of recent research and new perspectives. World Patent Information, 34(3), 197-205.

Lissoni, F., Llerena, P., Sanditov, B., 2013. Small Worlds in Networks of Inventors and the Role of Academics: An Analysis of France. Industry and Innovation, 20(3), 195-220.

Martínez, C., 2011. Patent families: When do different definitions really matter? Scientometrics, 86(1), 39-63.

Martínez, C., Azagra-Caro, J.M., Maraut, S., 2013. Academic Inventors, Scientific Impact and the Institutionalisation of Pasteur's Quadrant in Spain. Industry and Innovation, 20(5), 438-455.

Marx, M., Strumsky, D., Fleming, L., 2009. Mobility, skills, and the Michigan non-compete experiment. Management Science, 55(6), 875-889.

Meyer, J.-B., 2001. Network Approach versus Brain Drain: Lessons from the Diaspora. International Migration, 39(5), 91-110.

Meyer, J.-B., Brown, M., 1999, Scientific diasporas: A new approach to the brain drain. In: UNESCO (Ed.).

Miguelez, E., 2014, **Inventor diasporas and internationalization of technology** Cahiers du GREthA, Université de Bordeaux.

Miguelez, E., Fink, C., 2013. Measuring the International Mobility of Inventors: A New Database, World Intellectual Property Organization-Economics and Statistics Division.

Nanda, R., Khanna, T., 2010. Diasporas and domestic entrepreneurs: Evidence from the Indian software industry. Journal of Economics & Management Strategy, 19(4), 991-1012.

Pandey, A., Aggarwal, A., Devane, R., Kuznetsov, Y., 2006, The Indian Diaspora: a unique case? In: Y. Kuznetsov (Ed.). Diaspora Networks and the International Migration of Skills. World Bank Publications, Washington, DC pp. 71-97.

Peeters, B., Song, X., Callaert, J., Grouwels, J., Van Looy, B., 2010. Harmonizing harmonized patentee names: an exploratory assessment of top patentees, Luxembourg.

Pezzoni, M., Lissoni, F., Tarasconi, G., 2012. How To Kill Inventors: Testing The Massacrator© Algorithm For Inventor Disambiguation, Groupe de Recherche en Economie Théorique et Appliquée - Université Bordeaux IV.

Raffo, J., Lhuillery, S., 2009. How to play the "Names Game": Patent retrieval comparing different heuristics. Research Policy, 38(10), 1617-1627.

Rauch, J.E., 2001. Business and social networks in international trade. Journal of economic literature, 1177-1203.

Rauch, J.E., Trindade, V., 2002. Ethnic Chinese networks in international trade. Review of Economics and Statistics, 84(1), 116-130.

Saxenian, A., 2006, The new argonauts: Regional advantage in a global economy. Harvard University Press.

Saxenian, A., Motoyama, Y., Quan, X., 2002, Local and global networks of immigrant professionals in Silicon Valley. Public Policy Instit. of CA.

Singh, J., Marx, M., 2013. Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity. Management Science, 59(9), 2056-2078.

Tarasconi, G., Coffano, M., 2014, Crios-Patstat Database: Sources, Contents and Access Rules. Center for Research on Innovation, Organization and Strategy, CRIOS. http://ssrn.com/abstract=2404344 or http://dx.doi.org/10.2139/ssrn.2404344.

Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B.H., Harhoff, D., 2010. Harmonizing and combining large datasets–An application to firm-level patent and accounting data, National Bureau of Economic Research.

Thompson, P., 2006. Patent citations and the geography of knowledge spillovers: evidence from inventor- and examiner-added citations. The Review of Economics and Statistics, 88(2), 383-388.

Thompson, P., Fox-Kean, M., 2005. Patent citations and the geography of knowledge spillovers: A reassessment. American Economic Review, 450-460.

Veugelers, R., Cassiman, B., 2004. Foreign subsidiaries as a channel of international technology diffusion: Some direct firm level evidence from Belgium. European Economic Review, 48(2), 455-476.

Wadhwa, V., Rissing, B., Saxenian, A., Gereffi, G., 2007a. Education, Entrepreneurship and Immigration: America's New Immigrant Entrepreneurs, Part II. Part II (June 11, 2007).

Wadhwa, V., Saxenian, A., Rissing, B., Gereffi, G., 2007b. America's new Immigrant entrepreneurs: Part I. Duke Science, Technology & Innovation Paper(23).

Widmaier, S., Dumont, J.-C., 2011, Are recent immigrants different? A new profile of immigrants in the OECD based on DIOC 2005/06. OECD Publishing, Paris.