Inventor Data for Research on Migration and Innovation: A Survey and a Pilot (Stefano Breschi, Francesco Lissoni, and Gianluca Tarasconi)

Discussion by Bronwyn H. Hall[1]

July 2013

This paper is in two parts: the first part is an excellent survey of the literature on high-skilled migration and the second is an exploration of the use of data from a new dataset constructed by the authors. This new database merges their inventor dataset APE-INV with IBM's Global Name Recognition (IBM-GNR) database, which allows identification of the ethnicity of an inventor via surname analysis. The resulting database is entitled Ethic-Inv and it is used in the paper to provide a number of descriptive regressions that compare immigrant and "native" inventor productivity over time and across countries of origin and residence.

The substantive questions for which such data has been and will be useful are questions about the contribution of high-skilled immigrants to the economies of various countries, the potential loss of innovative activity when emigrants leave a country, and the innovative output of migrants who return to their home country with additional skills and/or knowledge. Policy questions in this area concern the design of immigration systems, and the mechanisms that might be adopted to encourage emigrants to return, or to attract innovators. In this discussion I offer a few suggestions that may advance the research agenda in this area, and then present a simulation model of invention and migration that clarifies the potential for bias in the use of patent data to track migration. One of the advantages of the present paper is that it does not rely solely on patent data to identify the migration of inventors.

The paper presents several tables that capture the productivity of foreign vs. local inventors at the country level, and over several cohorts. Looking at these tables suggests some ways to move forward on research in this area. For example, Table 2 shows that inventors migrating during the 1990s tended to have higher productivity than the residents of the countries they go to, whereas this is only true for the US, UK, and Germany during the 2001-2005 period. One explanation for variation over time is that macro-economic conditions in the destination or host countries affect the productivity of migrating workers. The 1990s may have been a period when productive inventors from former Soviet bloc countries found it easier to migrate than previously, whereas this impact would have diminished by the 2000s, leaving only the largest economies as those attractive to productive inventors. It is possible that booming economies in the US and UK prior to 2008 made them very attractive for new firm entry by inventors, as suggested by Wadha *et al*.'s research for the US, and that such entry was accompanied by greater patenting activity, simply because it was associated with new products and processes.

[1] University of California at Berkeley and University of Maastricht.

On the origin country side, it would be useful to model the relative costs of immigration, including language issues (some of which remain in spite of the authors' selection of countries that are not traditionally English, French, Spanish, Portuguese-speaking – e.g., India, Pakistan, Algeria, Morocco, Tunisia, plus countries with related languages such as Romania). Accounting for economic conditions in the home country as well as the difficulty of exit could be useful, especially for forecasting purposes.

General measurement issues

Studying the contribution of migrants to innovative activity requires data on migration and on innovative activity. Such data will be available in the aggregate separately, but rarely will one find an aggregate measure of immigrants engaged in innovative activity – the best one can hope for is some kind of measure of immigrants with high levels of education, and even then, the data may not be available by source country. The method pursued in this paper uses a micro approach to the problem, by assembling a dataset of inventors and tracking the location of their invention. Of necessity, the approach is restricted to the kind of invention that lends itself to patenting, and is therefore subject to some limitations: it will not capture innovative activity that is not patented, and it will not be able to assess even patentable inventive activity in countries with underdeveloped patent systems. The latter limitation is not likely to be important for the study of inventor migration to developed countries, but it could affect the measurement of spillovers from these inventors to their home countries, as the authors note in the final section of their paper (see Figure 1 in Breschi *et al.*).

The process of generating such a dataset presents the researcher with at least two difficulties. The first is to measure inventive activity, which is done here in the usual way by identifying inventors as those named on patent applications to various patent offices around the world. The difficulty with this measure is that inventive activity may be ongoing but may rarely turn up a patent application, so some inventors are not captured in a given time frame. The most serious consequence of this problem comes when one tries to identify changes in the location of an inventor. Here it is necessary to see at least two associated patents, and there will presumably be a fair amount of bias due to undercounting. In this paper, the problem is sidestepped to a certain extent by the use of surnames as a proxy for original location of the inventor. Later in this discussion, I present a stochastic model that illustrates the problem that would arise if patent data were used alone and I suggest a possible solution involving calibration using known distributions of patenting incidence per inventor in order to estimate the undercount.

The second problem that arises when using patent data for measuring the migration/invention nexus is identifying whether an inventor is a migrant. This is the problem addressed by the use of IBM-GNR database. This database is derived from a US immigration database generated in the early 1990s that associates names and surnames with the countries from which they come. The first obvious qualification is that the database does not contain US origin names, but even if it did, they would probably not be terribly useful, as they would overlap with a number of countries. The paper contains a good discussion of other possible biases in these data (e.g., weakness of Eastern European data

due to the vintage of the database, the importance of Mexico as an origin country in the US versus Spain in Europe, etc.).

However, it has to be said that it is not obvious that the discussion has exhausted the difficulties raised by using such a database to identify country of origin for inventors. Taking Europe as an example (this is the geographic area on which the paper focuses), we know from genetic evidence that Europeans are rather mixed across countries, and it would not be a surprise to learn that surnames are also mixed (although much less so, for linguistic reasons and because surnames are generally of recent vintage compared to genetics). One might expect that inventors are drawn from a set of people who are more likely to have migrated in the past, that is, they were more educated, or more enterprising, or even wealthier. If they retain their surnames, this fact will lead to an overestimate of innovation by recent migrants, some of whom will not be migrants at all. In the U.S., this is likely to be a fairly substantial problem, but it may affect Europe also. Countries such as France and the UK have experienced several waves of immigration throughout recent history.

Some issues with the use of patent data to track inventors

There are two main measurement concerns raised by the stream of research that infers inventor behavior from patent data. The first is that not all inventive activity is patented, and this fact can introduce biases that depend on the field of technology or the country of residence. There is not much that can be done about this absent a different source of inventive data, so it is better to simply be aware that this can be issue when evaluating the results. The second issue is internal to the data collected and may perhaps be mitigated by a suitable modeling strategy. This problem is the fact that inventors are only observed when they patent, and patenting is a relatively rare event, but one that is more common for more prolific inventors. This immediately suggests that migration is easier to observe when inventors apply for many patents than when they apply for few. One way to get an idea of the magnitude of the problem is to build a stochastic model of patenting and migration, and then simulate the model under various assumptions. I make a first start on this below.

It is important to emphasize that the present paper does not suffer terribly from such a bias because migration is measured by the presence of foreign surnames rather than by tracking inventors as they move. However, the last part of the paper does use patent inventor location to track returnees, and this part of the paper will suffer from the kind of bias identified here. It might be worthwhile to develop such a model in order to extract more precise estimates of the returnee distributions.

Assume that the probability that an inventor i applies for a patent is Poisson with λ_i and that the probability that the inventor moves is a small number that is the same for all inventors for simplicity. Inventors are assumed to be heterogeneous, so I draw the (permanent) propensity to patent λ_i for each inventor from a suitable distribution. Here I use the lognormal with mean 0.05, implying 20 years for each patent, and a Pareto with alpha parameter 2 and minimum 0.03, which also has a mean of 0.05. I observe the inventors for 25 years: over this period, some fraction (about one third for these parameters) will never

patent, and these are dropped from the simulation results, because they will never be observed in the data. To roughly calibrate the parameters of the distribution of inventor productivity, I used the numbers reported in Latham et al. (2006), which are for inventors named on US patents during the 1975-2006 period (about 30 years). The result of the calibration is in Figure 1, which shows the empirical distribution of numbers of patents per inventor together with the distribution generated by the two different simulation models (lognormal and Pareto). Given a more detailed distribution of patents per inventor, the calibration could doubtless be improved; it does appear that some kind of Pareto may be a better description of the distribution.

Table 1 shows the results of using the two distributions of inventor productivity together with two different probabilities of migration (0.1 per cent per year and 1 per cent per year) to simulate the probability that a move will be observed in patent data. The first set of columns (for the log normal distribution) show that if the move probability is tiny (0.1 per cent), 14 out of 641 inventors ever migrate during the period, but only 7 are actually observed to migrate because they patent on both sides of the move. For the Pareto, the number is even lower – only 4 inventors are observed to migrate, whereas 17 actually migrated. The second and fourth sets of columns show that if the annual probability of migration is 1 per cent, the situation is not much improved – only one quarter to one half of the inventors who migrate will be identified.

Figures 2 and 3 show the nature of the bias in more detail. These figures are based on the second (lognormal) and fourth (Pareto) sets of columns of Table 1. They show the distribution of productivity for all emigrants and for those emigrants that are observed in the patent data. Both distributions are shifted to the right, as expected, implying that more productive inventors are more likely to be observed. In the case of the lognormal, some productive inventors still fail to be observed, whereas the Pareto loses only those inventors who have fewer than 5 patents over the 25 year period. Of course, the precise nature of the bias will depend on a more detailed model of patenting productivity than is offered here.

The assumptions behind this simulation may be unrealistic, in that the migration probability is random across individuals and constant over time. If more productive individuals are more likely to migrate, we may be more likely to see them in the data. So it would be useful to calibrate the approach against some inventor data based on survey evidence. The advantage of a richer simulation model, one that also included covariates, is that developing such a model might enable researchers to infer the true migration data for inventors from the incomplete observations obtained from patent data.

Combining the economic factors that affect migration discussed earlier with a simulation model that recognizes the inherent partial observability in patent data seems to me a worthwhile endeavor. Breschi, Lissoni, and Tarasconi have made an excellent start towards increasing our understanding of high-skilled and inventive immigration and I look forward to seeing their future research develop along these lines.

Additional Reference

Latham, W., C. le Bas, and K. Touach (2006). The prolific inventor as a persistent inventor: an empirical study. In Latham and le Bas (eds.), *The Economics of Persistent Innovation*, Springer-Verlag.

Figure 1

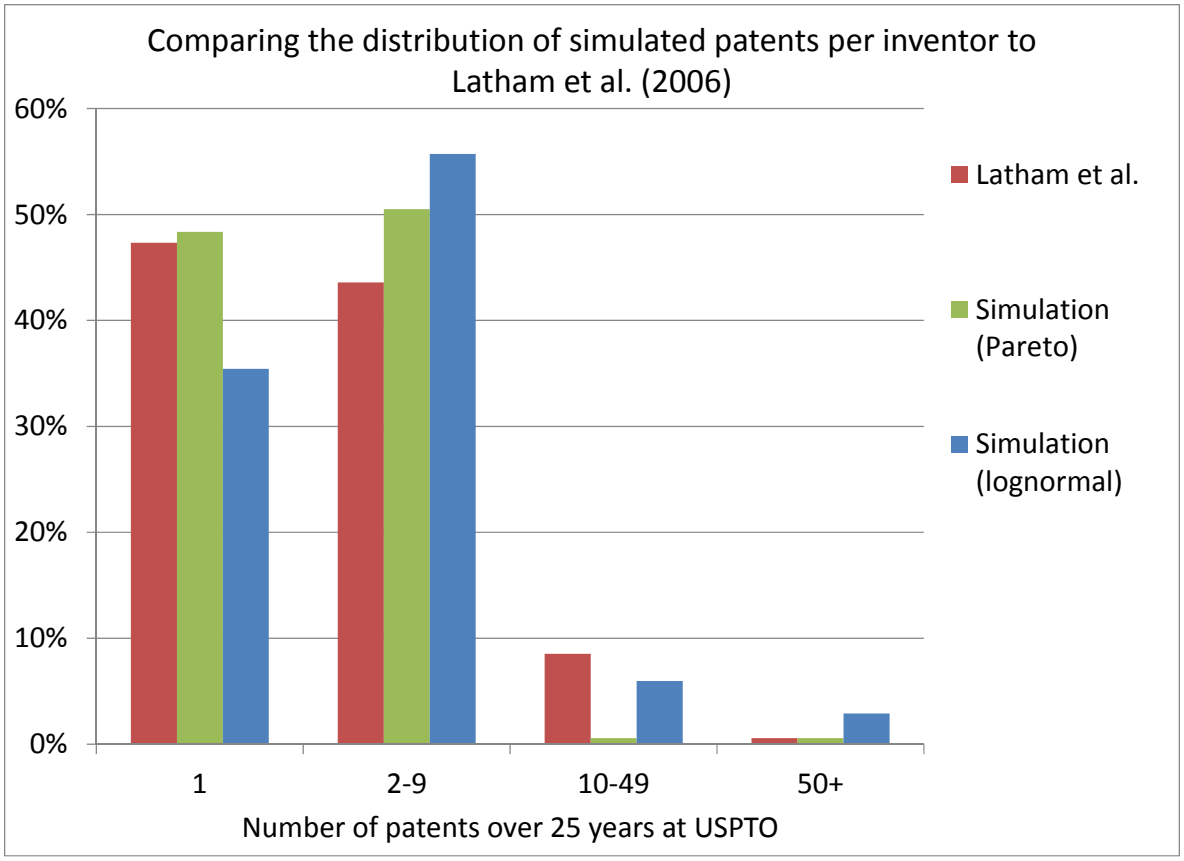

Table 1

Simulation of Migration of Heterogeneous Patenters over 25 years

	Prob(move) = 0.1%				*Prob(move) = 1%*				*Prob(move) = 0.1%*				*Prob(move) = 1%*			
Number of patents	*Stayers*	*Movers*	*Observed movers*	*Total*	*Stayers*	*Movers*	*Observed movers*	*Total*	*Stayers*	*Movers*	*Observed movers*	*Total*	*Stayers*	*Movers*	*Observed movers*	*Total*
1	226	6	0	232	181	51	0	232	330	9	0	339	264	75	0	339
2	123	0	0	123	104	19	9	123	179	4	1	183	147	36	11	183
3	79	3	2	82	65	17	12	82	82	0	0	82	70	12	4	82
4	44	1	1	45	37	8	5	45	44	4	3	48	36	12	8	48
5	35	2	2	37	29	8	7	37	18	0	0	18	15	3	3	18
6	25	2	2	27	22	5	5	27	8	0	0	8	6	2	2	8
7	25	0	0	25	18	7	5	25	8	0	0	8	5	3	3	8
8	12	0	0	12	11	1	0	12	4	0	0	4	2	2	2	4
9	14	0	0	14	12	2	2	14	4	0	0	4	4	0	0	4
10-19	39	0	0	39	28	11	9	39	4	0	0	4	3	1	1	4
20+	19	0	0	19	15	4	4	19	4	0	0	4	3	1	1	4
Total	**641**	**14**	**7**	**655**	**522**	**133**	**58**	**655**	**685**	**17**	**4**	**702**	**555**	**147**	**35**	**702**
Share	**97.9%**	**2.1%**	**1.1%**		**79.7%**	**20.3%**	**8.9%**		**97.6%**	**2.4%**	**0.6%**		**79.1%**	**20.9%**	**5.0%**	
	Each inventor has a patenting parameter drawn from a lognormal distribution with geometric mean = 0.05 (20 years per patent) and s.d. = 1. 345 inventors have no patents during the 25 year period and are dropped from the sample.								Each inventor has a patenting parameter drawn from a Pareto with minimum 0.03 and alpha==2. The resulting geometric mean was 0.05 298 inventors have no patents during the 25 year period and are dropped from the sample.							

Figure 2

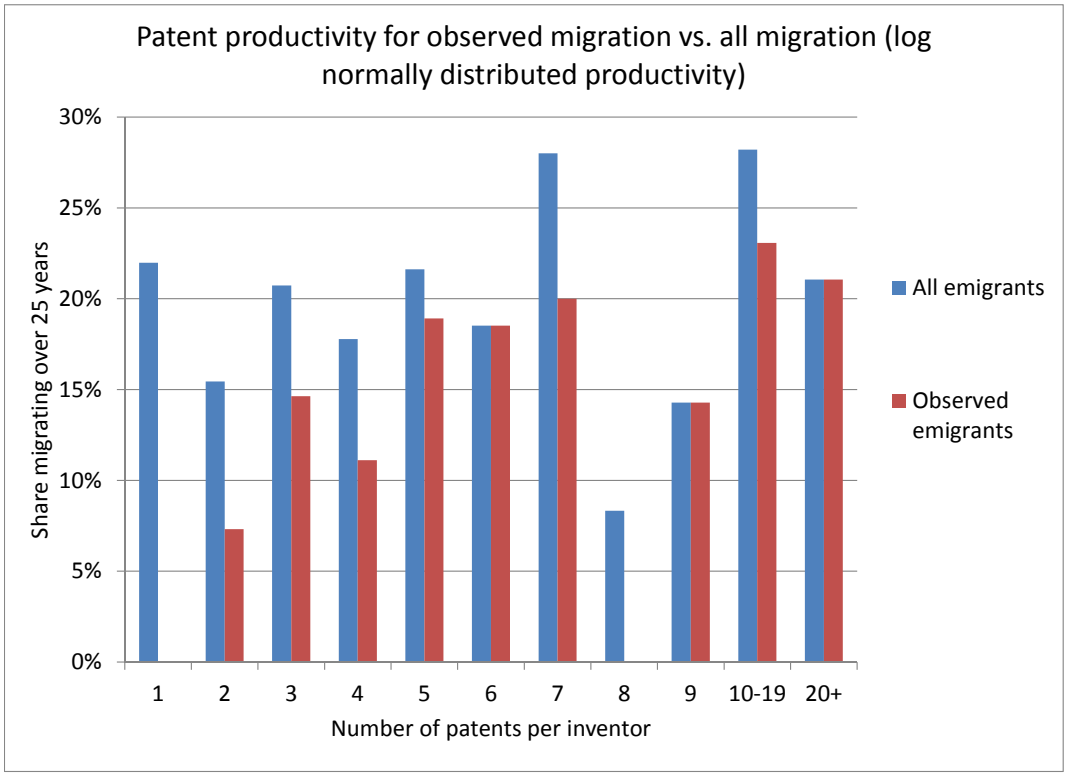

Figure 3