How Much Does Immigration Boost Innovation? Data Appendix

By Jennifer Hunt and Marjolaine Gauthier–Loiselle*

We measure the extent to which skilled immigrants increase innovation in the United States. The 2003 National Survey of College Graduates shows that immigrants patent at double the native rate, due to their disproportionately holding science and engineering degrees. Using a 1940–2000 state panel, we show that a one percentage point increase in immigrant college graduates' population share increases patents per capita by 9–18 percent. Our instrument for the change in the skilled immigrant share is based on the 1940 distribution across states of immigrants from various source regions and the subsequent national increase in skilled immigration from these regions.
JEL: J61, O31
Keywords: Immigration, innovation, patenting, productivity, growth.

Data Appendix

National Survey of College Graduates

The data were collected between October 2003 and August 2004 by the U.S. Bureau of the Census, on behalf of the National Science Foundation. The data consist of a stratified random sample of people reporting having a bachelor's degree or higher on the long form of the (April) 2000 census, who were under age 76 and living in the United States or its territories including Puerto Rico in the reference week of October 1, 2003. We drop respondents living outside the United States and define as immigrants those born outside the United States. Missing information is imputed with a hot deck procedure, and imputed values are not flagged. More information on the data is provided at www.nsf.gov/statistics/showsrvy.cfm?srvy_CatID=3&srvy_Seri=7#fn1. The data are available at www.nsf.gov/statistics/sestat/.

* Hunt: Department of Economics, McGill University, 855 Sherbrooke Street West, Montreal, QC, H2A 2T7, Canada, and National Bureau of Economic Research (NBER), jennifer.hunt@mcgill.ca. Gauthier–Loiselle: Department of Economics, Princeton University, Fisher Hall, Princeton, N.J. 08544, mgauthie@princeton.edu. We are grateful to David Munroe for excellent research assistance, and for helpful comments to Francisco Alvarez–Cuadrado, Leah Brooks, David Card, Lee Fleming, Rachel Friedberg, David Green, Francisco Gonzales, Judy Hellerstein, Chad Jones, Bill Kerr, Daniel Parent, Giovanni Peri, Steve Pischke, Regina Riphahn, Eric Stuen and Dee Suttiphisal, seminar participants at the London City University, London School of Economics, NBER (Productivity and Labor Studies), Nürnberg, Simon Fraser, University College London and the SoLE/EALE Transatlantic Conference, and several friends holding patents. We thank Bill Kerr, Nicole Fortin, and Jim Hirabayashi of the USPTO for data and Deven Parmar for obtaining and formatting the USPTO data. Hunt is also affiliated with the CEPR, IZA and DIW-Berlin, and acknowledges the Social Science and Humanities Research Council of Canada grant number 410-2006-0257 for financial support.

Patents

We combine two patent series from the U.S. Patent and Trademark Office (USPTO). The first series was compiled for us by the USPTO based on their electronic records which begin in 1963. This series is utility patents by state and year of application. Year of application is preferred to year of grant as it is a more accurate match to the time of invention. The second series (U.S. Department of Commerce 1977) is from paper–based USPTO records of patents by state and grant year 1883–1976 (application year is not available pre–1963). Grants lag applications by a median of three years between 1950 and 1963 (according to our US–wide calculations based on Lexis–Nexis), so we lead this series three years. Patent grants are also more volatile than patent applications (Bronwyn H. Hall 2004), so we smooth the series with a three year moving average. Finally, because for 1930–1960 plants and designs cannot be separated from utility patents, we leave them in for the whole series, calculate by state the average percent gap in the overlap years of the two series (18 percent on average), and reduce the old series by this percent. We then merge the series, using the adjusted paper series values only for pre–1963. The USPTO attributes a patent to a state according to the home address of the first–listed inventor.

We have also used an extract from the Harvard Business School patent data file, which contains information on utility patents granted from 1975 to 2007, arranged by year of application and patent class. We have aggregated the patent classes to six categories using the classification in Hall, Adam B. Jaffe and Manuel Trajtenberg (2001) and our own classification of patent classes created since 1999. In particular, we attribute classes 506 and 977 to chemical patents; classes 398, 701–720, 725 and 726 to computers and communication patents; and classes 901 and 903 to mechanical patents. We have not been able to find definitions for some patent classes created in 2006 or later (which affects some patents applied for in earlier years), and a small number of patents have a missing patent class. For the application years we used, 0.04 percent of patents are not allocated to one of our six categories. To examine patents by category, we have simply attributed 1974 values (most patents granted in 1975 were applied for in 1974 or earlier) to 1971, then used 1971, 1981, 1991 and 2001 patent values, and 1970, 1980, 1990 and 2000 values for the dependent variables. Some small states do not have patents in every category in every year, and in the analysis of log patents these observations are missing.

The extract also contains the number of citations made to patents in each patent class, state and application year. These may be viewed as a proxy for the quality of the patent. We calculate citations per patent from 1974 onwards for each state. We then run a regression of this ratio on a trend for each state from 1974–1980, and use the resulting coefficient to predict the 1971 value of citations per patent for each state. We then return to our original, longer patent series obtained directly from the USPTO, and multiply the patents by the ratio for 1971 onwards. We can then study citations, or quality–adjusted patents, for 1971, 1981, 1991 and 2001.

Immigration, education, age, occupation, labor force status

We use extracts from the Integrated Public Use Microdata Series for the United States Census, available at usa.ipums.org/usa/, and aggregate to the state level using the weights provided. Variables computed as shares (other than the excluded instruments) are computed as shares of the population or workers aged 18–65, and average population age is the average age of people aged 18–65. Immigrants are people born outside the United States. We use the census–provided edurec variable to identify college graduates (16 years of education or more in the 1940–1980 censuses, and a college or higher degree in the 1990 and 2000 censuses) and high–school dropouts (11 or fewer years of education). People with post–college education are people with 17 or more years of education in the 1940–1980 censuses, and a post–college degree in 1990 and 2000. This is the highest level of education that can be distinguished for the whole 1940–2000 period. Alaska and Hawaii are not in the 1940 and 1950 IPUMS and we drop them from the analysis. The SIC codes we count as electrical engineering are 321, 322, 342, 350, 371, 372.

Other data

We use Bureau of Economic Analysis data for total state population (used to weight the regressions) and for state personal income per capita (available from 1929 onwards, unlike gross state product which is not available for our whole period). The data are available at www.bea.gov/regional/spi/.

Department of Defense procurement contracts by state are available on paper for the early years in U.S. Department of Defense (1979). The later years are available online at www.fpds.gov. Some measurement error in the attribution to states is involved, as recipient firms may subcontract the work to firms in other states. Also, in the electronic records for 1978–1983, 1986 and 1989 (of which only 1980 is relevant for the paper), the California numbers seem to be too small by a factor of 1000, so we have multiplied them by 1000. (We have obtained scanned versions of the paper documents for these years: the values for the non–problematic states and years are only approximately the same as those online, but the problematic California years are indeed about 1000 times higher than the online version.) We attribute the 1951 value to 1950, and set changes in the values involving 1940, for which data are not available, to zero.

We obtain the land area of each state from the US. Census Bureau at www.census.gov/population/censusdata/90den_stco.txt.

REFERENCES

Hall, Bronwyn H. "Exploring the Patent Explosion," *Journal of Technology Transfer*, December 2004, *30* (1), 35–48.

—, Adam B. Jaffe, and Manuel Trajtenberg, "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools," Working Paper 8498, National Bureau of Economics Research October 2001.

U.S. Department of Commerce, Patent and Trademark Office, *Technology Assessment and Forecast, Seventh Report*, Washington, D.C.: U.S. G.P.O, 1977.

U.S. Department of Defense, *Prime Contract Awards by State, Fiscal Years 1951–1978*, Washington, D.C.: OASD (Comptroller), Directorate for Information Operations and Control, 1979.