The Ethnic Composition of US Inventors: Evidence Building from Ethnic Names in US Patents

William R. Kerr*
Harvard Business School
Boston MA

7 November 2007

Abstract

This study applies an ethnic-name database to individual patent records granted by the United States Patent and Trademark Office. This process characterizes the ethnic composition of US inventors with greater detail than previously available, even extending to within-firm analyses. There exists a dramatic increase in the contributions of Chinese and Indian scientists and engineers to US technology formation in the 1990s. The Chinese contribution noticeably levels off after 2000, however, and the Indian contribution declines.

JEL Classification: F15, F22, J44, J61, O31.

Key Words: Innovation, Research and Development, Patents, Scientists, Engineers, Inventors, Ethnicity, Immigration.

*Comments are appreciated and can be sent to wkerr@hbs.edu. This paper summarizes Kerr (2007b). I am grateful to William Lincoln and Debbie Strumsky for data assistance. This research is supported by the National Science Foundation, HBS Research, the Innovation Policy and the Economy Group, and the MIT George Schultz Fund.

1 Introduction

The contributions of immigrants to US technology formation are staggering: while foreign-born account for just over 10% of the US working population, they represent 25% of the US science and engineering (SE) workforce and nearly 50% of those with doctorates. Even looking within the Ph.D. level, ethnic researchers make an exceptional contribution to science as measured by Nobel Prizes, election to the National Academy of Sciences, patent citation counts, and so on.[1] Moreover, ethnic entrepreneurs are very active in commercializing new technologies, especially in the high-tech sectors (e.g., Saxenian 2002a). The magnitude of these ethnic contributions raises many research and policy questions: debates regarding the appropriate quota for H1-B temporary visas, the possible crowding out of native students from SE fields, the brain-drain or brain-circulation effect on sending countries, and the future prospects for US technology leadership are just four examples.[2]

Econometric studies quantifying the role of ethnic scientists and engineers for technology formation and diffusion are often hampered, however, by data constraints. It is very difficult to assemble sufficient cross-sectional and longitudinal variation for large-scale panel exercises.[3] This paper describes a new approach for quantifying the ethnic composition of US inventors with previously unavailable detail. The technique exploits the inventor names contained on the micro-records for all patents granted by the United States Patent and Trademark Office (USPTO) from January 1975 to April 2007.[4] Each patent record lists one or more inventors, with 7.5 million inventor names associated with the 4.3 million patents. The USPTO grants patents to inventors living within and outside of the US, with each group accounting for about half of patents over the 1975-2007 period.

This study maps into these inventor names an ethnic-name database typically used for commercial applications. This approach exploits the idea that inventors with the surnames Chang or Wang are likely of Chinese ethnicity, those with surnames Rodriguez or Martinez of Hispanic ethnicity, and so on. The match rates range from 92%-98% for US domestic inventor records, depending upon the procedure employed, and the process affords the distinction of nine ethnicities: Chinese, English, European, Hispanic/Filipino, Indian/Hindi, Japanese, Korean, Russian, and Vietnamese. Moreover, because the matching is done at the micro-level, greater detail on the ethnic composition of inventors is available annually on multiple dimensions: technologies, cities, companies, and so on.

[1] For example, Stephan and Levin (2001), Burton and Wang (1999), Johnson (1998, 2001), and Streeter (1997).

[2] Representative papers are Lowell (2000), Borjas (2004), Saxenian (2002b), and Freeman (2005) respectively.

[3] While the decennial Census provides detailed cross-sectional descriptions, its longitudinal variation is necessarily limited. On the other hand, the annual Current Population Survey provides weaker cross-sectional detail and does not ask immigrant status until 1994. The SESTAT data offer a better trade-off between the two dimensions but suffer important sampling biases with respect to immigrants (Kannankutty and Wilkinson 1999).

[4] The project initially employed the NBER Patent Data File, compiled by Hall et al. (2001), that includes patents granted by the USPTO from January 1975 to December 1999. The current version now employs an extended version developed by HBS Research that includes patents granted through early 2007. Some of the descriptive calculations have not been updated from their 1975-1999 values (noted in text).

Section 2 describes the ethnic-name matching strategy and the commercial database employed. The section also discusses the advantages and disadvantages for empirical estimations of the resulting dataset. Section 3 concludes with some aggregate descriptive statistics for the ethnic composition of US inventors.

2 Ethnic-Name Matching Technique

2.1 Melissa Ethnic-Name Database and Name-Matching Technique

Ethnic-name databases suffer from two inherent limitations — not all ethnicities are covered, and included ethnicities usually receive unequal treatment. The strength of the ethnic-name database obtained from the Melissa Data Corporation is the identification of Asian ethnicities, especially Chinese, Indian/Hindi, Japanese, Korean, Russian, and Vietnamese names. The database is comparatively weaker for looking within continental Europe. For example, Dutch surnames are collected without first names, while the opposite is true for French names. The Asian comparative advantage and overall cost effectiveness led to the selection of the Melissa database, as well as the European amalgamation employed in the matching technique. In total, nine ethnicities are distinguished: Chinese, English, European, Hispanic/Filipino, Indian/Hindi, Japanese, Korean, Russian, and Vietnamese.[5]

The second limitation is that commercial databases vary in the number of names they contain for each ethnicity. These differences reflect both uneven coverage and that some ethnicities are more homogeneous in their naming conventions. For example, the 1975 to 1999 Herfindahl indices of foreign inventor surnames for Korean (470) and Vietnamese (1121) are significantly higher than Japanese (132) and English (164) due to frequent Korean surnames like Kim (16%) and Park (12%) and Vietnamese surnames like Nguyen (29%) and Tran (12%).

Two polar matching strategies are employed to ensure coverage differences do not overly influence ethnicity assignments.

> *Full Matching:* This procedure utilizes all of the name assignments in the Melissa database and manually codes any unmatched surname or first name associated with 100 or more inventor records. This technique further exploits the international distribution of inventor names within the patent database to provide superior results. The match rate for this restricted procedure is 97% (98% US, 97% foreign). This rate should be less than 100% with the Melissa database as not all ethnicities are included.

[5]The largest ethnicity in the US SE workforce absent from the ethnic-name database is Iranian, which accounted for 0.7% of bachelor-level SEs in the 1990 Census. Kerr (2007b) further describes the Melissa ethnic-name database, the name-matching process, the international name distribution technique, and the apportionment of matches using MSA characteristics. The Melissa database is typically employed for direct-mail advertisements. I am grateful to the MIT George Schultz Fund for financial assistance in its purchase.

Restricted Matching: A second strategy employs a uniform name database using only the 3000 and 200 most common surnames and first names, respectively, for each ethnicity. These numerical bars are the lowest common denominators across the major ethnicities studied. The match rate for this restricted procedure is 88% (92% US, 86% foreign).

For matching, names in both the patent and ethnic-name databases are capitalized and truncated to ten characters. Approximately 88% of the patent name records have a unique surname, first name, or middle name match in the Full Matching procedure (77% in the Restricted Matching), affording a single ethnicity determination with priority given to surname matches. For inventors residing in the US, representative probabilities are assigned to non-unique matches using the masters-level SE communities in Metropolitan Statistical Areas (MSAs). Ethnic probabilities for the remaining 3% of records (mostly foreign) are calculated as equal shares.

2.2 Inventors Residing in Foreign Countries and Regions

Visual confirmation of the top 1000 surnames and first names in the USPTO records confirms the name-matching technique works well. Kerr (2007b) also documents the fifty most common surnames of US-based inventors for each ethnicity, along with their relative contributions. While some inventors are certainly misclassified, the measurement error in aggregate trends building from the micro-data is minor. The Full Matching procedure is the preferred technique and underlies the trends presented in the next section, but most applications find negligible differences when the Restricted Matching dataset is employed instead.

The application of the ethnic-name database to the inventors residing outside of the US provides a natural quality-assurance exercise for the technique. Inventions originating outside the US account for just under half of USPTO patents, with applications from Japan comprising about 48% of this foreign total. Kerr (2007b) documents the results of applying the ethnic-matching procedures for countries and regions grouped to the ethnicities identifiable with the database. The results are very encouraging. First, the Full Matching procedure assigns ethnicities to a large percentage of foreign records, with the match rates greater than 93% for all countries but India (84%). In the Restricted Matching procedure, a matching rate of greater than 73% holds for all regions.

Second, the estimated inventor compositions are reasonable. The weighted average is 86% in the Full Matching procedure, and own-ethnicity contributions are greater than 80% in the UK, China, India, Japan, Korea, and Russia regardless of the matching procedure employed. Like the US, own-ethnicity contributions should be less than 100% due to foreign researchers. The high success rate using the Restricted Matching procedure indicates that the ethnic-name database performs well without exploiting the international distribution of names, although power is lost with Europe. Likewise, uneven coverage in the Melissa database is not driving the ethnic composition trends.

2.3 Advantages and Disadvantages of Name-Matching Technique

The matched records describe the ethnic composition of US scientists and engineers with previously unavailable detail: incorporating the major ethnicities working in the US SE community; separating out detailed technologies and manufacturing industries; providing MSA and state statistics; and providing annual metrics. Moreover, the assignment of patents to corporations and institutions affords firm-level and university-level characterizations (e.g., the ethnic composition of IBM's inventors filing computer patents from San Francisco in 1985). The next section provides graphical descriptions along these various dimensions, and Kerr (2007a) is a sample application.

The ethnic-name procedure does, however, have two potential limitations for empirical work that should be highlighted. First, the approach does not distinguish foreign-born ethnic researchers in the US from later generations working as SEs. The procedure can only estimate total ethnic SE populations, and these levels are to some extent measured with time-invariant error due to the name-matching approach. The resulting data are very powerful, however, for panel econometrics that employ changes in these ethnic SE populations for identification. Moreover, Census and INS records confirm Asian changes are primarily due to new SE immigration for this period, substantially weakening this concern when examining these groups.

The name-matching technique also does not distinguish finer divisions within the nine major ethnic groupings. For ethnic network analyses, it would be advantageous to separate Mexican from Chilean scientists within the Hispanic ethnicity, to distinguish Chinese engineers with ethnic ties to Taipei versus Beijing versus Shanghai, and so on. These distinctions are not possible with the Melissa database, and researchers should understand that measurement error from the broader ethnic divisions may bias their estimated coefficients downward depending upon the application.[6] Nevertheless, Section 3 demonstrates how the deep variation available with the ethnic patenting data provides a rich description of US ethnic invention.

3 Ethnic Composition of US Inventors

Table 1 describes the ethnic composition of US inventors for 1975-2004.[7] The trends demonstrate a growing ethnic contribution to US technology development, especially among Chinese and Indian scientists. Ethnic inventors are more concentrated in high-tech industries like computers and pharmaceuticals and in gateway cities relatively closer to their home countries (e.g., Chinese in San Francisco, European in New York, and Hispanic in Miami). The final three rows demonstrate a close correspondence of the estimated ethnic composition to the country-of-birth composition of the US SE workforce in the 1990 Census.

[6] When mapping the ethnic patenting data to country-level data for international diffusion estimations, researchers will also need to cluster their standard errors to reflect the multiple country-to-ethnicity mappings.

[7] Granted patents are grouped by application year. Kerr (2007b) presents additional descriptive statistics and discusses rationales for looking at percentages of patents granted versus raw patent counts.

Figure 1 illustrates the evolving ethnic composition of US inventors from 1975-2004. The omitted English share declines from 83% to 72% during this period. Looking across all technology categories, the European ethnicity is initially the largest foreign contributor to US technology development. Like the English ethnicity, however, the European share of US domestic inventors declines steadily from 8% in 1975 to 6% in 2004. This declining share is partly due to the exceptional growth over the thirty years of the Chinese and Indian ethnicities, which increase from under 2% to 8% and 4%, respectively. The Indian ethnic contribution declines somewhat after 2000, mostly due to changes within the computer technology sector.

Figure 2 documents the total ethnic contribution by the six broad technology groups into which patents are often classified: Chemicals, Computers and Communications, Drugs and Medical, Electrical and Electronic, Mechanical, and Others. The miscellaneous group includes patents for agriculture, textiles, furniture, and the like. Growth in ethnic patenting is clearly stronger in high-tech sectors than in more traditional industries. Figures 3 and 4 provide more detailed glimpses within the Chinese and Indian ethnicities. These two ethnic groups are clearly important contributors to the stronger growth in ethnic contributions among high-tech sectors, where Chinese inventors supplant European researchers as the largest ethnic contributor to US technology formation.

These aggregate trends are a small sample of the descriptions that can be developed through the resulting data. While most applications thus far have focused on ethnicity-industry-year variation useful for macroeconomic exercises, current work is exploiting the firm-level assignment of US patents. This research is considering the role of US ethnic researchers in foreign direct investment by US multinationals; a second project is considering the impact of recent H1B visa reforms for firms that are heavily dependant on ethnic scientists and engineers. It is hoped that similar research projects will realize the full potential of this empirical strategy.

References

[1] Borjas, George, "Do Foreign Students Crowd Out Native Students from Graduate Programs?", NBER Working Paper 10349 (2004).

[2] Burton, Lawrence, and Jack Wang, "How Much Does the U.S. Rely on Immigrant Engineers?", NSF SRS Issue Brief (1999).

[3] Freeman, Richard, "Does Globalization of the Scientific/Engineering Workforce Threaten U.S. Economic Leadership?", NBER Working Paper 11457 (2005).

[4] Griliches, Zvi, "Patent Statistics as Economic Indicators: A Survey", *Journal of Economic Literature* 28:4 (1990), 1661-1707.

[5] Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg, "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools", NBER Working Paper 8498 (2001).

[6] Johnson, Jean, "Statistical Profiles of Foreign Doctoral Recipients in Science and Engineering: Plans to Stay in the United States", NSF SRS Report (1998).

[7] Johnson, Jean, "Human Resource Contribution to U.S. Science and Engineering From China", NSF SRS Issue Brief (2001).

[8] Kannankutty, Nirmala, and R. Keith Wilkinson, "SESTAT: A Tool for Studying Scientists and Engineers in the United States", NSF SRS Report (1999).

[9] Kerr, William, "Ethnic Scientific Communities and International Technology Diffusion", *Review of Economics and Statistics* forthcoming (2007a).

[10] Kerr, William, "The Ethnic Composition of US Inventors", Harvard Business School Working Paper 08-006 (2007b).

[11] Lowell, B. Lindsay, "H1-B Temporary Workers: Estimating the Population", The Center for Comparative Immigration Studies Working Paper 12 (2000).

[12] Saxenian, AnnaLee, with Yasuyuki Motoyama and Xiaohong Quan, *Local and Global Networks of Immigrant Professionals in Silicon Valley* (San Francisco, CA: Public Policy Institute of California, 2002a).

[13] Saxenian, AnnaLee, "Silicon Valley's New Immigrant High-Growth Entrepreneurs", *Economic Development Quarterly* 16:1 (2002b), 20-31.

[14] Stephan, Paula, and Sharon Levin, "Exceptional Contributions to US Science by the Foreign-Born and Foreign-Educated", *Population Research and Policy Review* 20:1 (2001), 59-79.

[15] Streeter, Joanne, "Major Declines in Admissions of Immigrant Scientists and Engineers in Fiscal Year 1994", NSF SRS Issue Brief (1997).

Fig. 1: Ethnic Share of US Domestic Patents

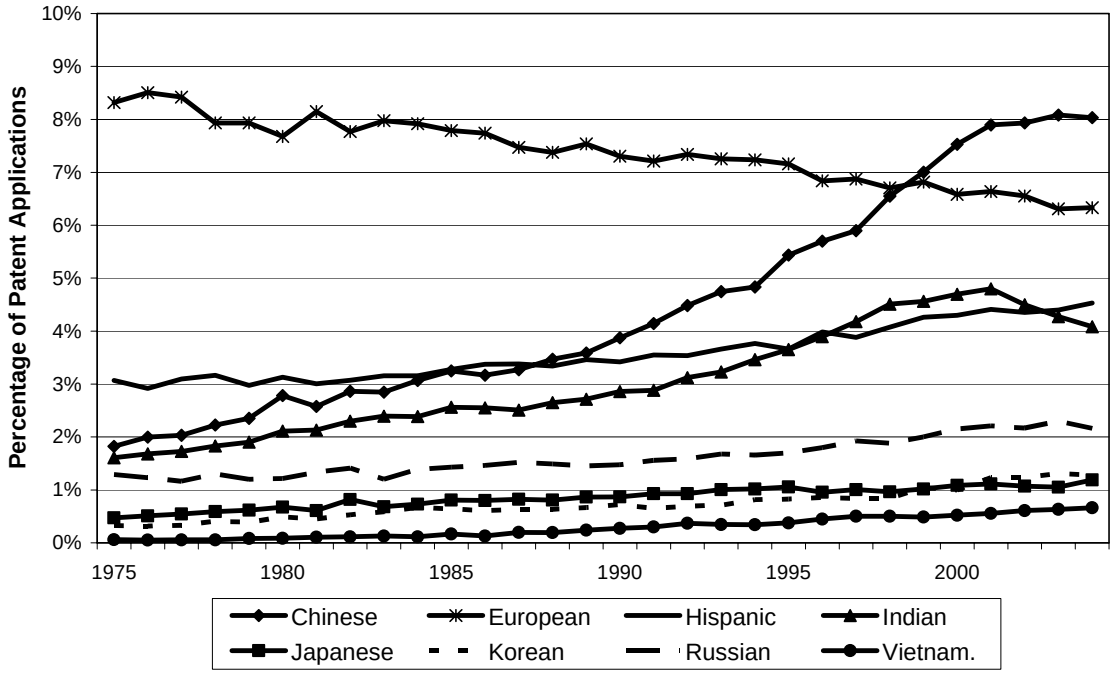

Fig. 2: Total US Ethnic Share by Technology

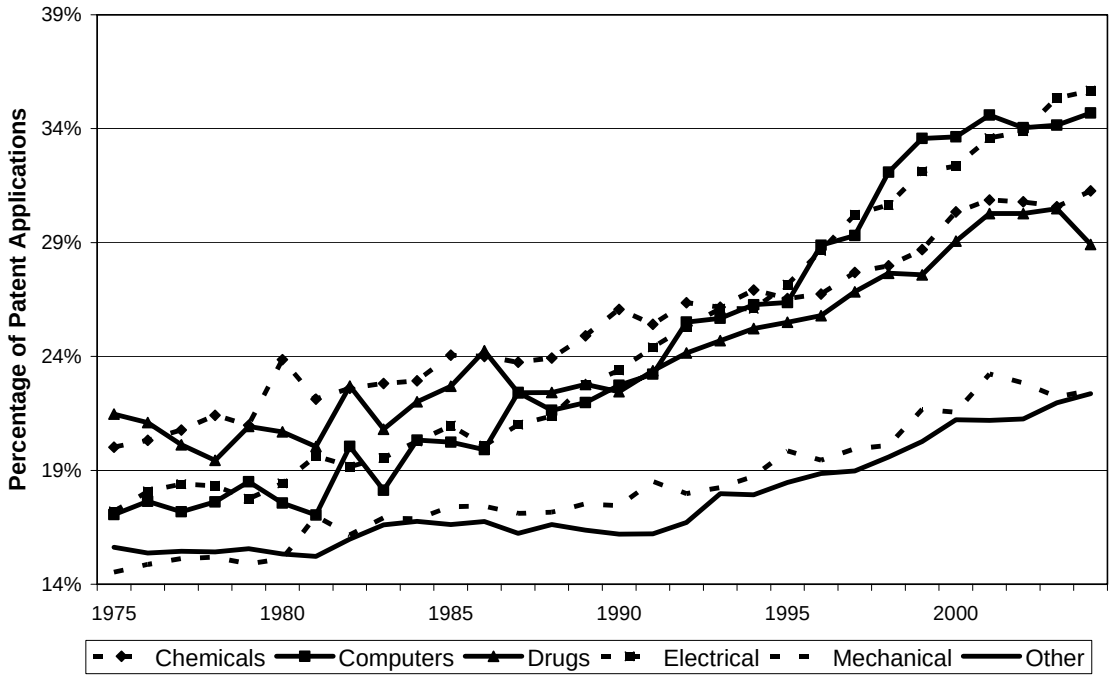

Fig. 3: Chinese Contribution by Technology

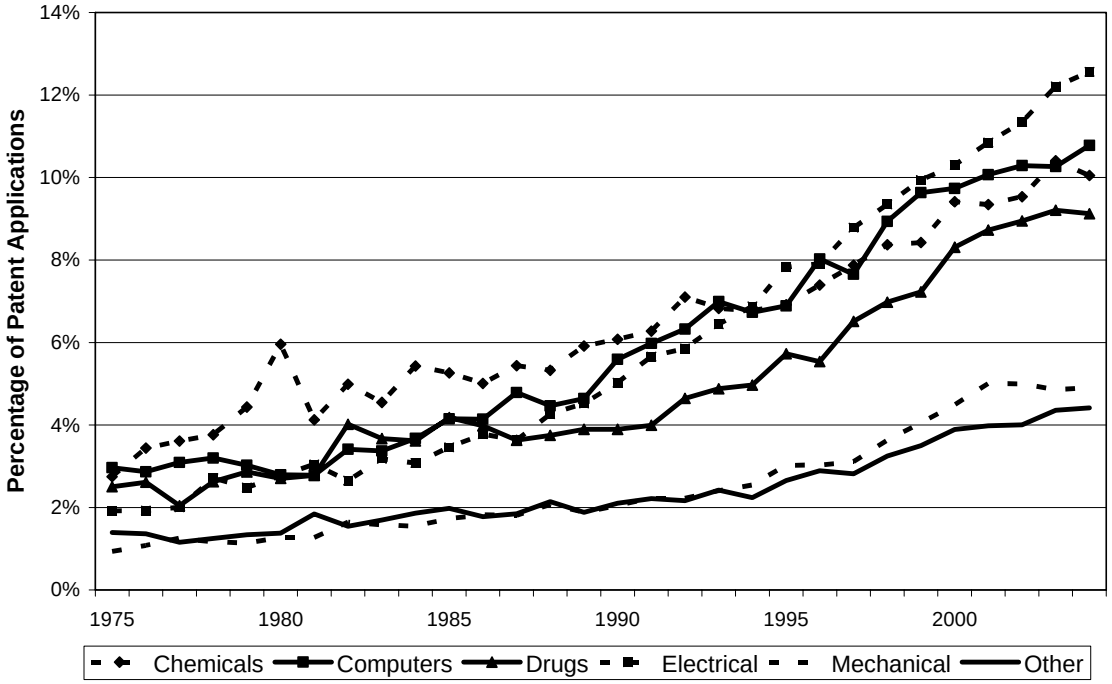

Fig. 4: Indian Contribution by Technology

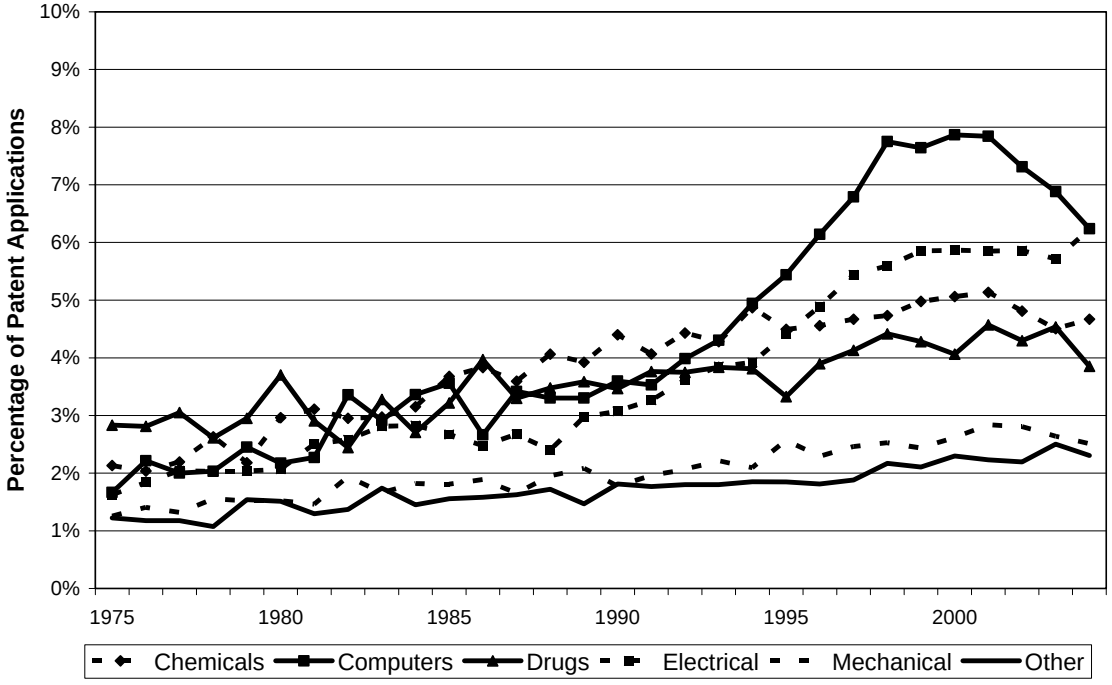

Table 1: Descriptive Statistics for Inventors Residing in US

	Ethnicity of Inventor								
	English	Chinese	European	Hispanic	Indian	Japanese	Korean	Russian	Vietnam.
	A. Ethnic Inventor Shares Estimated from US Inventor Records								
1975-1979	82.5%	2.2%	8.2%	3.0%	1.9%	0.6%	0.4%	1.2%	0.1%
1980-1984	81.1%	2.9%	7.9%	3.1%	2.4%	0.7%	0.6%	1.3%	0.1%
1985-1989	79.8%	3.6%	7.5%	3.3%	2.8%	0.8%	0.7%	1.4%	0.2%
1990-1994	77.6%	4.7%	7.2%	3.5%	3.4%	0.9%	0.8%	1.5%	0.4%
1995-1999	74.0%	6.6%	6.8%	3.9%	4.5%	0.9%	0.9%	1.8%	0.5%
2000-2004	71.0%	8.5%	6.4%	4.2%	4.8%	1.0%	1.2%	2.2%	0.6%
Chemicals	73.7%	7.1%	7.6%	3.6%	4.2%	0.9%	0.9%	1.7%	0.3%
Computers	71.3%	7.9%	6.3%	3.7%	6.1%	1.1%	1.0%	2.0%	0.7%
Pharmaceuticals	73.3%	6.9%	7.4%	4.3%	3.9%	1.1%	1.0%	1.8%	0.3%
Electrical	72.0%	8.0%	6.8%	3.7%	4.6%	1.1%	1.2%	2.0%	0.7%
Mechanical	80.6%	3.2%	7.2%	3.4%	2.4%	0.7%	0.6%	1.6%	0.2%
Miscellaneous	81.5%	2.9%	7.0%	3.8%	2.1%	0.6%	0.6%	1.4%	0.2%
Top MSAs as a	KC (89)	SF (14)	NOR (12)	MIA (16)	AUS (6)	SF (2)	BAL (2)	BOS (3)	AUS (2)
Percentage of MSA's	WS (88)	LA (8)	STL (11)	SA (9)	SF (6)	SD (2)	LA (2)	NYC (3)	SF (1)
Patents	NAS (88)	AUS (6)	NYC (11)	WPB (7)	BUF (5)	LA (2)	SF (2)	SF (3)	PRT (1)
	B. Ethnic Scientist and Engineer Shares Estimated from 1990 US Census Records								
Bachelors Share	87.6%	2.7%	2.3%	2.4%	2.3%	0.6%	0.5%	0.4%	1.2%
Masters Share	78.9%	6.7%	3.4%	2.2%	5.4%	0.9%	0.7%	0.8%	1.0%
Doctorate Share	71.2%	13.2%	4.0%	1.7%	6.5%	0.9%	1.5%	0.5%	0.4%

Notes: MSAs - AUS (Austin), BAL (Baltimore), BOS (Boston), BUF (Buffalo), KC (Kansas City), LA (Los Angeles), MIA (Miami), NAS (Nashville), NOR (New Orleans), NYC (New York City), PRT (Portland), SA (San Antonio), SD (San Diego), SF (San Francisco), STL (St. Louis), WPB (West Palm Beach), and WS (Winston-Salem). MSAs are identified from inventors' city names using city lists collected from the Office of Social and Economic Data Analysis at the University of Missouri, with a matching rate of 99%. Manual recoding further ensures all patents with more than 100 citations and all city names with more than 100 patents are identified. 1990 Census statistics are calculated by country-of-birth using the groupings listed in Kerr (2007b); English provides a residual in the Census statistics.