THE INTERNAL MIGRATION OF INDIAN SCIENTISTS, 1981-2003, FROM AN ANALYSIS OF SURNAMES

RAMESH KUNDRA[1]

GRANT LEWISON[2]

1   National Institute of Science, Technology and Development Studies, Dr KS Krishnan Marg, Pusa, New Delhi 110012, India

 2. Evaluametrics Ltd, 8 Nylands Avenue, Kew, Richmond, Surrey, TW9 4HH, England

Abstract

Although many Indian surnames are common across the whole country, some are specifically associated with just one of the 35 states and union territories that comprise India today. For example, Reddy comes from Andhra Pradesh and Das, Ghosh and Roy from West Bengal. We investigated the extent to which researchers with names associated with some of the larger states were writing scientific papers in those states, and in other ones, and to see how these relative concentrations (relative to the whole of India) had changed since the early 1980s. We found that West Bengalis, for example, were now significantly less concentrated in their home state than formerly, and that their concentrations elsewhere were strongly influenced by the state's geographical distance from West Bengal and, to a lesser extent, by the correlation between the scientific profile of their host state and their own preferences (which favored physics and engineering over biology and mathematics). Thus they were strongly represented in nearby Bihar, Assam and Orissa, and much less so in Tamilnadu and Kerala.

1.   Introduction: the use of names to identify ethnic, national, regional and family origins

There is now a significant amount of literature on the use of names, particularly surnames or paternal family names, as identifiers of the origins of individuals. One of the first studies was in 1875 by Darwin[1] in which he determined the rate of consanguineous marriages or inbreeding (between cousins) by an analysis of those couples who had the same surname. More recently, names have been used as a proxy for race or ethnicity in order to compare the morbidity and mortality of different groups, particularly immigrants to the USA[2], to Canada[3, 4, 5, 6, 7] and to the UK[8, 9, 10, 11]. In France, surnames have been used to assist in the location of potential bone marrow donors [12] and in Sweden, to check if non-Swedish children were obtaining appropriate health care[13].
Another application is to study the structure and movements of populations. One of the first studies was of Korean immigrants to the USA, using the name "Kim" which is very common in Korea[14]. Piazza and his colleagues[15] did this in Italy and found that the surname method turned out to be surprisingly robust, and gave results comparable with

census data and the names in telephone directories. In South America, names have been used as an indicator of the admixture of Spanish and Aymara Indian genes[16]. In Russia, Startseva *et al*[17] applied mathematical techniques to describe the distributions of surnames in two regions, and relate these to the incidence of genetic disorders. Degioanni *et al*[18] examined the migration of Italians into France, and were able to identify the provinces most subject to emigration, and the French *departments* most likely to have immigrants. Lewison and Igic[19] looked at the possible "ethnic cleansing" of scientists in the former Yugoslavia – by Serbs of Croats, or *vice versa* – by categorizing names as "typically Serb" or "typically Croat", and comparing their presence in the respective countries before and after the civil war from 1991-93.

In this study, we used Indian surnames, many of which can be associated with particular parts of the country, to categorize Indian scientists as originating in those regions, and we then examined to what extent they were now more evenly spread across the country, having left their region of origin to take up scientific posts elsewhere. We examined the situation in three time periods: 1981-83, 1991-93 and 2001-03. It would be expected that in the first period there would be a high concentration of scientists in their region of origin, and that this would gradually diminish over the next 20 years as they migrated to other parts of India. As we shall see, this is largely the case.

2. *Introduction: science in India*

There have been several studies designed to map demographically India's scientific and technological developments. For example, the Council of Scientific Industrial Research (CSIR) attempted in 1966 to indicate on the map of India the location of various government, public and private research institutes and other centres of research[20]. There have also been surveys[21, 22] of the publication activities of Indian researchers and scientists engaged in specific fields. Sharma *et al*.[23] studied the geographical distribution of scientific activity across India and related it to regional demographical and economic factors. The purpose was to understand the geographical concentration of scientists and the factors responsible for it.

For administrative purposes, India is divided up into 28 states and 7 "union territories" (UT), notably including Delhi, Chandigarh and Pondicherry, see map in Figure 1.
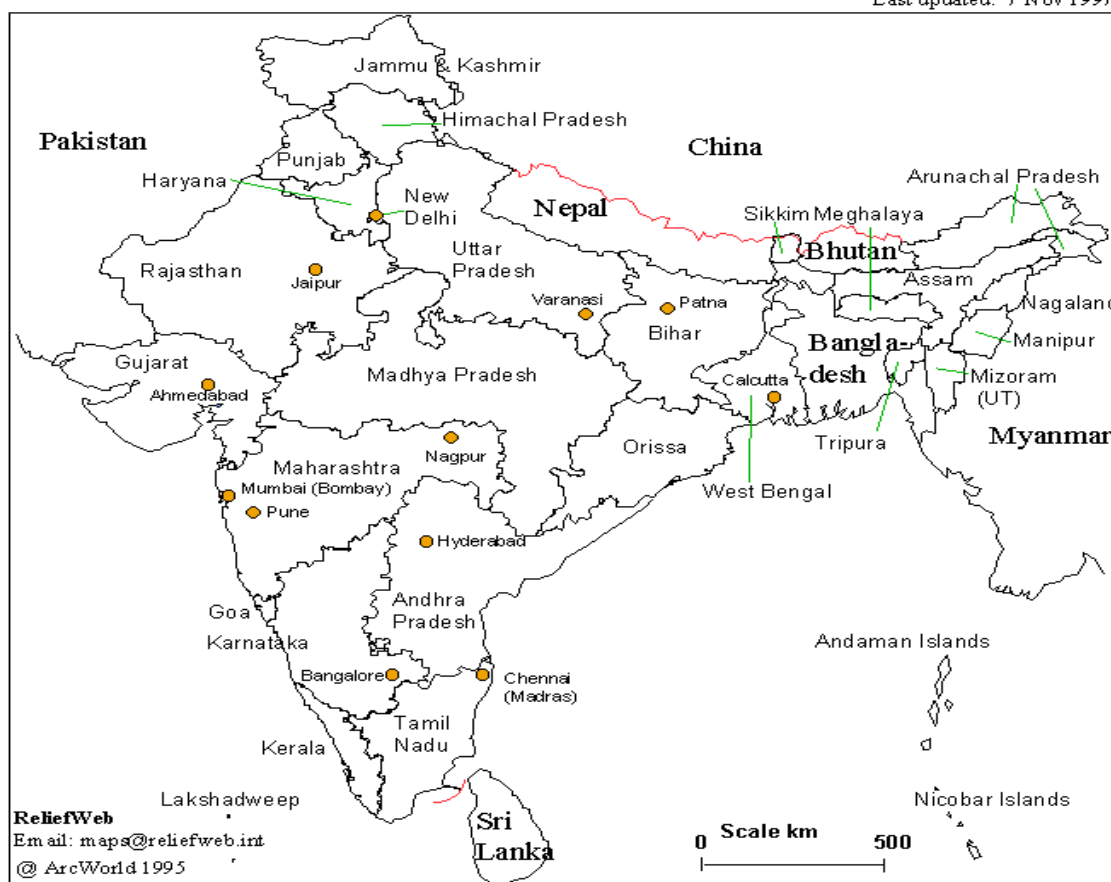
Figure 1. Map of India, showing states and principal cities.

Within Indian scientific society certain castes have played a major role in the development of science and technology. The present study not only identifies the demographic sources of these castes but also their movements from their place of origin. This was based on the last, or family, names (surnames) given by authors on scientific papers. Normally in India most of the castes tend to use their given names first, followed by a family, caste or clan name. There are exceptions when a title that indicates the person's caste or clan is used. In southern India, where the society is more matriarchal, the given name may be preceded by the mother's family name. In our study we have used researchers' surnames to characterize their demographic origins.

Because of its size and its varied historical and cultural traditions, many Indians have family or surnames that indicate their provenance, although of course some families have long resided in other parts of the country. Nevertheless, names such as Das and Ghosh can be particularly associated with West Bengal in the east, and Nair, Thomas, Joseph and George with Kerala in the far south. Some names may derive from two or more states in one of the five regions which we used for the purposes of our analysis. These were made up as shown in Table 1, which lists the states and UTs with digraph codes used in the figures.

Table 1. List of states and UTs that have been grouped into five Indian regions (C = centre, E = east, N = north, S = south, W = west) for analysis purposes. *Note: some small states and UTs have been omitted as their scientific output is too small to affect the analysis.*

Central states/UT of India	Eastern states/UT of India	Northern states/UT of India	Southern states/UT of India	Western states/UT of India
BH Bihar	AR Arunachal Pradesh	HP Himachal Pradesh	AN Andaman & Nicobar Isl.	DN Dadar & Nagar Haveli
JH Jharkhand	AS Assam	HY Haryana	AP Andhra Pradesh	GJ Gujarat
MP Madhya Pradesh	OR Orissa	JK Jammu & Kashmir	KA Karnataka	GO Goa
UP Uttar Pradesh	SK Sikkim	ND New Delhi	KE Kerala	MS Maharashtra
	TP Tripura	PJ Punjab	PC Pondicherry	
	WB West Bengal	RJ Rajasthan	TN Tamilnadu	
		UT Uttaranchal Pradesh		

We were interested to see to what extent Indians with names associated with a particular part of the country were still living and carrying out research in that state or region, and whether there had been changes over the 22 years covered by our study. We would have expected that initially there would have been a high concentration of Bengali names in West Bengal (WB) and nearby states, for example, but that over time the relative concentrations elsewhere would have increased and that in WB would have declined. We were also interested to see in which directions Indians from particular states and regions had traveled, and whether their participation in the common enterprise of science had made the country more homogeneous.

3.   Methodology

The data for this study were all taken from the Science Citation Index © Thomson Scientific, CD-ROM version. Three sets of bibliographic data on Indian papers (articles only) were downloaded to file, for the years 1981-83, 1991-93 and 2001-03. The data included the authors' names, their addresses and the full source of the article. The addresses were analysed and any internationally-coauthored papers were removed from the three files.

The authors' surnames for the 2001-03 papers were listed in descending order of frequency of occurrence. There were 30,116 all-India papers, 28,647 different authors but only 10,308 different surnames. Of these, 4356 occurred just once but Singh and Kumar both appeared more than 2400 times. The top 450 surnames, accounting for just over 60% of the total, were classified (by RK) as being particularly associated with one or more states. Of these names, 288 were associated with a single state (representing 31% of all names), and 378 could be associated with one of the five regions listed in Table 1 (42% of all names). The leading names for the four states with the most particular surnames are listed in Table2.

Table 2. Leading surnames for Indians from four states with the most individual names: West Bengal, Tamil Nadu, Maharashtra and Uttar Pradesh.

West Bengal	Tamil Nadu	Maharashtra	Uttar Pradesh
Das	Srinivasan	Kulkarni	Mathur
Ghosh	Srinivas	Patil	Ahmad
Roy	Ramesh	Nayak	Ali
Banerjee	Sarma	Naik	Goel
Pal	Natarajan	Deshpande	Goyal
Mukherjee	Subramanian	Mahajan	Bhatnagar
Bhattacharya	Radhakrishnan	Naidu	Dwivedi
	Balasubramania		
Ray	n	Kale	Pathak

The Science Citation Index lists all the authors' names and all their addresses, but the individual names are not associated with a particular address, therefore it is impossible to ascribe a researcher to a state or UT unless all the addresses are from the same entity. We therefore coded every address on all the papers for the 3 x 3 years with its digraph state or UT code. This was a major task, because although for many addresses the state name is given, or the city is large and well-known, there were also many papers from small cities which had to be checked for their geographical location. Fortunately, almost all of them featured on the Web and could be found with their state – often because they were attempting to attract visitors. Once each address had been coded with its digraph, it was possible to check whether these were all the same or involved inter-state collaboration. As with international collaboration, inter-state partnerships have increased within India: they comprised 3.8% in 1981-83; 6.0% in 1991-93 and 10.7% in 2001-03 of the all-India papers – but still represented only a small fraction of the total output. The multi-state papers were discarded from the analysis: this left 24,236 for 1981-83; 21,281 for 1991-93 and 26,894 for 2001-03.

Table 3. Distribution of Indian single-state papers by major field, 1981-83, 1991-93 and 2001-03.

	1981-	%	1991-	%	2001-	%

	83		93		03	
Biology	3909	*16.1*	1863	*8.8*	1735	*6.5*
Biomedical research	2700	*11.1*	2256	*10.6*	2994	*11.1*
Chemistry	5706	*23.5*	5486	*25.8*	7557	*28.1*
Clinical medicine	3421	*14.1*	3106	*14.6*	4392	*16.3*
Earth & space	1069	*4.4*	1278	*6.0*	1236	*4.6*
Engineering & tech.	2095	*8.6*	2791	*13.1*	3208	*11.9*
Mathematics	880	*3.6*	247	*1.2*	394	*1.5*
Physics	3908	*16.1*	3910	*18.4*	4493	*16.7*
Others	548	*2.3*	344	*1.6*	883	*3.3*
Total	24236	*100.0*	21281	*100.0*	26892	*100.0*

The journals in which the papers were published were assigned to one of eight major fields using the system developed by CHI Research Inc. in the USA[i] and used for the US Science and Engineering Indicators. Table 3 shows the distribution of papers between these fields in the three 3-year periods.

Biology and mathematics have clearly suffered, whereas chemistry has increased, as has clinical medicine, though to a lesser extent. However, the main purpose of this analysis was to serve as a possible explanation of any anomalies that might be found in the migration patterns of the scientists.

For the analysis of the locations of the scientists, sets of names, either from a single state (Tamil Nadu or West Bengal) or from one of the five regions, were applied as a filter to the spreadsheet column with the authors' names so as to reveal those papers with one or more of the given set of surnames. The state/UT codes for the set of papers could then be listed and compared with the list of address codes for all the single state papers for the given years so as to show the ratio of percentage presences. For example, in 1981-83, West Bengal published 11.1% of all single-state papers, and Tamil Nadu, 7.2%. But for papers with one or more WB names, WB accounted for 39.1% of the total but Tamil Nadu for only 2.2%. So the relative concentration of West Bengalis was 39.1/11.1 = 3.53 in WB and 2.2/7.2 = 0.30 in TN.

Finally, the major fields of authors from these two states and the five regions were tabulated so as to compare them with those of all-India in order to see if their interests and training were significantly different from the average as this might have influenced their migration patterns. This was done in the same way as for the analysis of addresses.

4. Results

Figure 2 shows the distribution of single-state Indian papers by region for the three 3-year periods. Clearly there has been a shift towards the south at the expense of the centre and,

to a lesser extent, of the north. This is the background against which migration patterns must be viewed.
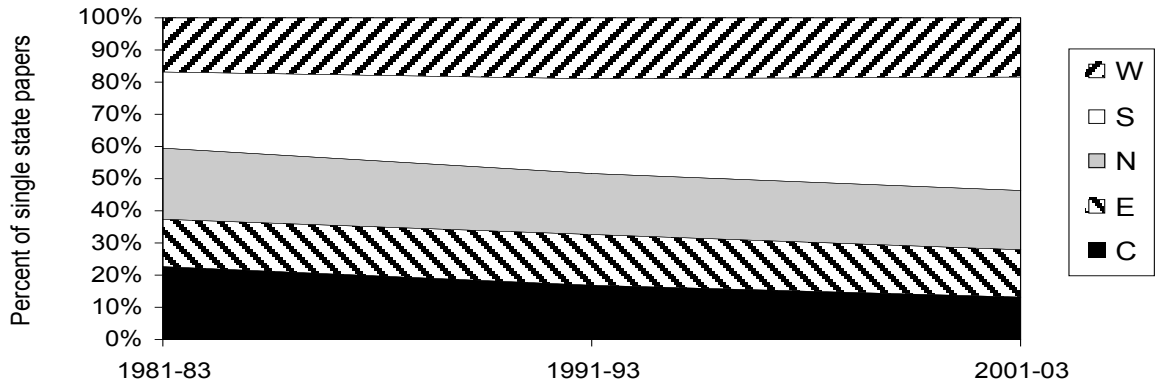

Figure 2.: Distribution of single-state Indian papers in the SCI in three 3-year periods by region (for states and UTs included, see Table 1).
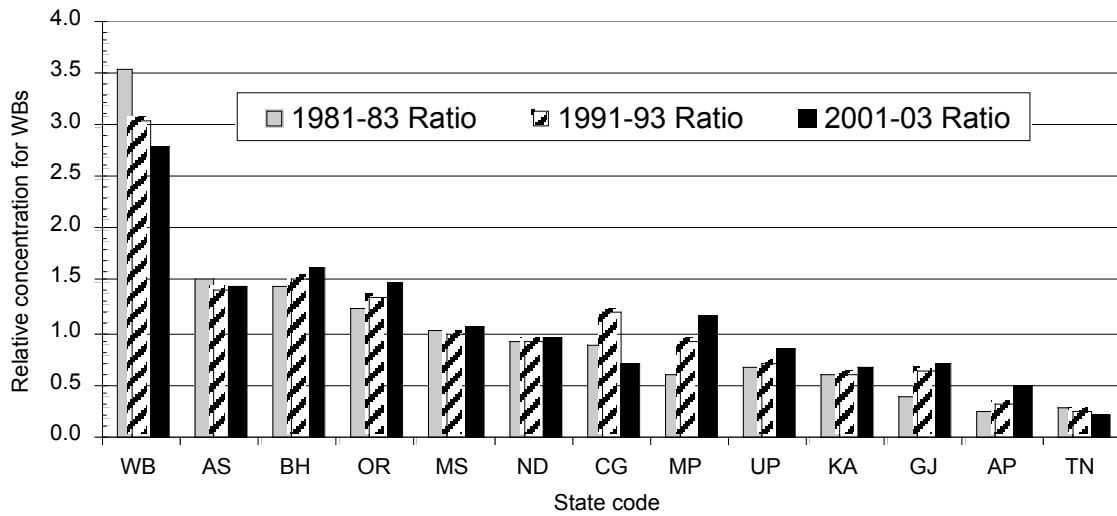

Figure 3. Relative concentration of West Bengalis in leading Indian states and UTs in three 3-year periods (SCI articles). For codes, see Table 1.
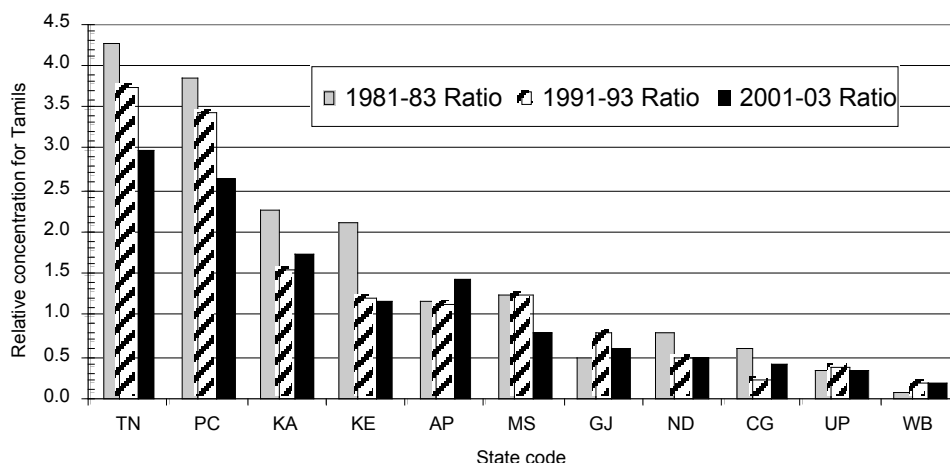
Figure 4. Relative concentration of Tamils in leading Indian states and UTs in three 3-year periods (SCI articles). For codes, see Table 1.

Figures 3 and 4 show the relative concentrations of West Bengalis and Tamils, respectively, in the largest states and UTs (in terms of their scientific output) in the three 3-year periods. They are ordered by their mean relative concentrations of WBs and Tamils over the whole nine years.

The pattern is rather clear. Both West Bengalis and Tamils show a decreasing relative concentration with time in their own states, and this relative concentration falls off quite rapidly with distance from the state, except for Maharashtra (MS). But whereas Tamils are now less frequently present in the neighbouring states and UTs of Pondicherry (PC), Karnataka (KA) and Kerala (KE), West Bengalis are now *more* frequently present in their nearby states of Bihar (BH) and Orissa (OR). They are also more likely to be in the central capital, Delhi (ND), and the new city Chandigarh (CG, capital of two states, Haryana HY and Punjab PJ, since partition). They are also more likely to be found in Tamil Nadu (TN) than are Tamils in West Bengal. So overall, West Bengalis are more likely to migrate internally than are Tamils. This may, of course, reflect the greater scientific opportunities in the south in recent years seen in Figure 2.

Although the differences are not marked, it appears that Tamils are relatively strong in chemistry and engineering, whereas Bengalis favour physics. Both groups appear relatively weak in biology (compared to all Indians) and in mathematics. Could these preferences explain some of the apparent anomalies in the migration patterns?

Table 4. Major fields researched by Tamils and West Bengalis, 1981-83, 1991-93 and 2001-03 and by all Indians, and ratios of relative concentration.

	TN	WB	All IN	%TNs	%WBs	% IN	TN/IN	WB/IN
Biology	548	1520	7507	6.7	7.6	10.4	0.64	0.73
Biomedical res	929	2376	7950	11.3	11.8	11.0	1.03	1.08
Chemistry	2415	5094	18749	29.4	25.3	25.9	1.14	0.98

Clinical medicine	1066	2919	10919	13.0	14.5	15.1	0.86	0.96
Earth and space	405	826	3583	4.9	4.1	4.9	1.00	0.83
Engineering & tech	1100	2348	8094	13.4	11.7	11.2	1.20	1.05
Mathematics	117	313	1521	1.4	1.6	2.1	0.68	0.74
Physics	1388	4259	12311	16.9	21.2	17.0	0.99	1.25
Others	241	442	1775	2.9	2.2	2.5	1.20	0.90
Total	8209	20097	72409					

The overall distribution of Indian papers by major field was given in Table 3. However, there are some differences between the states and Table 5 shows the pattern for the ten leading states in terms of overall output. This table suggests that Maharashtra (MS), which is the leading state in terms of output, may be attracting Bengalis, despite the distance between Mumbai and Kolkata, because of its above-average commitment to physics and below-average commitment to biology, which accords with

Table 5. Distribution (percent) of papers in major fields from all India (single state only) and from the 10 leading states in output: 1981-83, 1991-93 and 2001-03. States ordered by their output in the three 3-year periods overall.

Field\state	All IN	MS	UP	WB	ND	KA	TN	AP	GJ	KE	CG
Biology	10.4	4.2	13.1	7.3	8.7	10.2	7.8	11.1	8.0	11.7	5.5
Biomedical res	11.0	8.8	9.5	12.6	13.8	14.3	10.0	12.9	6.9	10.0	17.4
Chemistry	25.9	30.8	24.1	25.4	15.0	25.1	26.8	35.4	33.8	27.3	10.0
Clinical med	15.1	12.8	16.4	7.1	27.6	9.1	15.9	10.2	11.0	13.7	53.4
Earth and space	4.9	5.9	6.7	3.6	2.9	5.8	2.1	7.1	8.5	6.5	0.9
Engin'g & tech	11.2	11.7	11.9	12.3	11.6	12.7	16.2	8.9	7.6	15.7	1.2
Mathematics	2.1	2.6	2.6	2.4	1.4	1.8	2.4	1.6	1.3	1.0	2.3
Physics	17.0	20.9	13.0	28.1	16.6	16.5	16.3	10.2	19.9	12.0	8.3
Others	2.5	2.3	2.7	1.3	2.3	4.4	2.4	2.5	2.9	2.1	1.1

their preferences. Kerala (KE) and Andra Pradesh (AP) are the reverse, with a low interest in physics but an above-average interest in biology. The relative presence of Bengalis is much lower here than in Karnataka (KA) whose field distribution is similar to that of India overall. Delhi (ND) and Chandigarh (CG) are both strong in clinical medicine and fairly strong in biomedical research, but weak in chemistry; this may partly explain the far lower relative presence of Tamils there compared with that of Bengalis.

We turn now to the distribution of papers with authors not from a single state but from each of the five regions. There were, as mentioned above, 378 surnames associated with a single region: 132 from the south, 93 from the east, 59 from the north, 51 from the centre and 43 from the west. When the bibliographic files were filtered to reveal the sets of papers with an author from each group in turn, the relative concentrations of authors with given surnames (listed as column headers in Table 6) could be calculated for each of the states (listed as row labels).

Table 6. Relative concentration of authors from five state groups in the papers of these same state groups in the SCI, 1981-83, 1991-93 and 2001-03.

State\name	Centre	East	North	South	West
Centre	2.26	0.74	1.23	0.39	0.44
East	0.62	2.41	0.31	0.24	0.69
North	1.22	0.73	2.47	0.52	0.74
South	0.35	0.51	0.42	2.11	0.71
West	0.93	1.17	0.70	0.94	2.57

The ratios shown on the diagonal of the matrix are all greater than two, and most of the other ratios are less than unity, showing that internal migration is only partial. There is clearly a good deal of mixing between the centre and north, and the west is attracting researchers from all over the country. However there are relatively few researchers from the north and the south working in the east, nor from the centre and north working in the south despite its expansion in output.

5. Discussion and conclusions

The study selected and analyzed the names based on the prevalent castes in different states of India. Hindus, who form the dominant caste, are considered to be an Indo-European group of people with languages descended from Sanskrit – Hindi, Gujarati, Punjabi, Bengali, Marathi – which are spoken by about 70% of the population in the northern and central parts of the country. Sanskrit names tend to be compounds due to the popularity of the names of gods or their titles but are often written as separate words (Ram Gopal rather than Ramgopal). Muslims are the next largest caste who mostly speak Urdu and their names are mostly of Arabic origin with some Persian (Farsee) borrowings. In the south, the Dravidian languages - Tamil, Telugu, Kannada, Malayalam – are spoken by 20-25% of the population, but many names are of Sanskrit origin due to Hindu influence though a few Dravidian names are widely used. By tradition three names: the name of the village, the father's name and the personal name (in that order) are used. Western influence is now seen as families adopt just personal and family names; the tactic of converting their traditional names has created much confusion. Trade names are also used as surnames - some of these have developed quite recently from English, such as Merchant, Engineer and Contractor. Anglo-Indians (people of mixed English and

Indian ancestry) were traditionally railway workers and used names that reflected this. One also finds some regional variations such as:

> North : final '-a' usually left out (Raja becomes Raj);
>
> South : '-a' retained, often with '-n' or '-m' added (Raja becomes Rajan or

Rajam);

> East (esp. Bengal) : Sanskrit 'a' often becomes 'o', 'v' becomes 'b', 's' becomes 'sh', 'j' becomes 'y' and 'ks' becomes 'x'. The Sanskrit ending '-ra' becomes '-er' in many parts of the country (Chandra => Chander).
>
> Within the north, east, south and west regions there are variations among states located in these regions and through these indicators one can identify the region from which the person originates.

Science has been actively pursued in northern and central India in 1981-83 but the activity shifted to southern India in 1991-1993 and still further in 2001-2003 while in other states scientific activity remained static in the entire twenty year period. It was the contributions of Tamils and Bengalis in their respective states that enhanced the scientific productivity in southern India and maintained it in eastern India. There are several possible reasons for this. Some new institutions may have been created; there may be a greater tendency for scientists from southern India to publish papers in international journals; and there may simply be more scientists in these regions.

The study reveals that the states of Maharashtra and Gujarat lead in terms of scientific output. Chemistry is the favoured major field due to the presence of many chemical and pharmaceutical industries and research institutes, especially in Mumbai and Baroda, and the National Chemical Laboratory in Pune. West Bengal has strength in physics. Kerala, Tamilnadu and Andhra Pradesh show less interest in physics but do much work in biology. In northern India clinical medicine plays a dominating role especially in Delhi (notably from the All India Institute of Medical Sciences and the Indian Council of Medical Research) and in Chandigarh (where there is the. Postgraduate Institute of Medical Education & Research).

1. G.F. Darwin: Marriages between first cousins in England and their effects. Journal Royal Statistical Society, 1875; **38**, 153-184.

2. B.C. Choi, A.J. Hanley, E.J. Holowaty and D. Daley: Use of surname to identify individuals of Chinese ancestry. American Journal of Epidemiology 1993; **138**: 723-734.

3. E.Y. Tjam How to find Chinese research participants: use of a phonologically-based surname search method. Canadian Journal of Public Health, 2001; **92**(2), 138-142

4. H. Quan, W.A. Ghali, S. Dean, C.M. Norris, PD. Faris, P.D. Galbraith, M.M. Graham and M.L. Knudtson: Validity of using surnames to define Chinese ethnicity (meeting abstract) Journal of General Internal Medicine, 2003; **18** (s1), 186

5. H. Quan, W.A. Ghali, S. Dean, C.M. Norris, P.D. Faris, P.D. Galbraith, M.M. Graham and M.L. Knudtson: Validity of using names to define south Asian ethnicity (meeting abstract) Journal of General Internal Medicine, 2003; **18** (s1), 186

6. H Quan, W.A. Ghali, S. Dean, C.M. Norris, P.D. Galbraith, P.D. Faris, M.M. Graham and M.L. Knudtson: Validity of using surname to define Chinese ethnicity. Canadian Journal of Public Health, 2004; **95**(4), 314

7. A.J. Coldman, T. Braun and R.P. Gallagher: The classification of ethnic status using name information. Journal of Epidemiology Community Health, 1988; **42**, 390-395.

8. R. Balarajan, L. Bulusu, A.M. Adelstein and V. Shukla: Patterns of mortality among migrants to England and Wales from the Indian sub-continent. British Medical Journal, 1985; **289**, 1185-1187

9. A. Nicoll, K. Bassett and S.J, Ulijaszek: What's in a name? Accuracy of using surnames and forenames in ascribing Asian ethnic identity in English populations. Journal of Epidemiology Community Health, 1986;**40**, 364-368

10. S. Harding, H. Dews and S.L. Simpson: The potential to identify South Asians using a computerized algorithm to classify names. Population Trends, 1999; **97**, 46-49

11. C. Cummins, H. Winter, K.K. Cheng, R. Maric, P. Silcocks and C. Varghese: An assessment of the Nam Pehchan computer program for the identification of names of south Asian ethnic origin. Journal of Public Health Medicine, 1999; **21**(4): 401-406.

12. A. Degioanni, P. Darlu and C. Raffoux: Analysis of the French National Registry of unrelated bone marrow donors, using surnames as a tool for improving geographical localization of HLA haplotypes. European Journal of Human Genetics, 2003; **11**: 794-801.

13. D. Lagerberg, M. Magnusson and C. Sendelinp: Surname as a marker of ethnicity. A study from child health services shows that immigrant respective Swedish families seem to be isolated in different ways. Lakartidningen, 2005; **102**(30-31): 2145-2148.

14. E.H. Shin and E.Y. Yu EY: Use of surnames in ethnic research: the case of Kims in the Korean-American population. Demography, 1984; **21**(3): 347-360.

15. A. Piazza A, S. Rendine, G. Zei, A. Moroni and L.L. Cavelli-Sforza: Migration rates of human populations from surname distributions Nature, 1987; **329**, 714-716

16. R. Chakraborty R, Barton SA, Ferrell RE and Schull WJ. Ethnicity determination by names among the Aymara of Chile and Bolivia. Human Biology, 1989; **61**(2): 159-177.

17. E.A. Startseva, G.I. Elchinova, R.A. Mamedova and E.K. Ginter: The use of the migration index, the parameter of surname diversity, and the values of entropy and redundancy of surname distribution for description of population-structure. Genetika, 1994; **30** (7), 978-981

18. A. Degioanni, A. Lisa, G. Zei and P.Darlu: Italian surnames and Italian migration to France 1891-1940 Population (Paris), 1996; **51**(6), 1153-1180

19. G. Lewison and R. Igic: Yugoslav politics, "ethnic cleansing" and co-authorship in science Scientometrics, 1999; **44**, 183-192

20. M. Bhatia: Science in India (Research, Survey & Planning Organization, New Delhi) 1966

21. P.S. Nagpaul, J.R. Bhatia, & A. Rahman: Trends of research in electronics. Research Survey & Planning Organization, 1966, New Delhi.

22. A. Rahman & S. Malik: Trends in research in Botany, Research Survey & Planning Organization, 1966, New Delhi.

23. K.D. Sharma, B.M. Gupta and Kundra, Ramesh: Geographical Distribution of Indian Publishing Scientists. Journal of Scientific Industrial. Research, *1979, 38, 471-477.*