Onomastics to measure cultural bias in medical research

Elian CARSENAT, NamSor Applied Onomastics, namsor.com

Dr. Evgeny Shokhenmayer, e-onomastics

Abstract

This project involves the analysis of about over ten million medical research articles from PubMed. We propose to evaluate the correlation between the onomastic class of the article authors and that of the citation authors. We will demonstrate that the cultural bias exists and also that it evolves in time. Between 2007 and 2008, the ratio of articles authored by Chinese scientists (or scientists with Chinese names) nearly tripled. We will evaluate how fast this surge in Chinese research material (or research material produced by scientists of Chinese origin) became cross-referenced by other authors with Chinese or non-Chinese names. We hope to find that the onomastics provide a good enough estimation of the cultural bias of a research community. The findings can improve the efficiency of a particular research community, for the benefit of Science and the whole humanity.

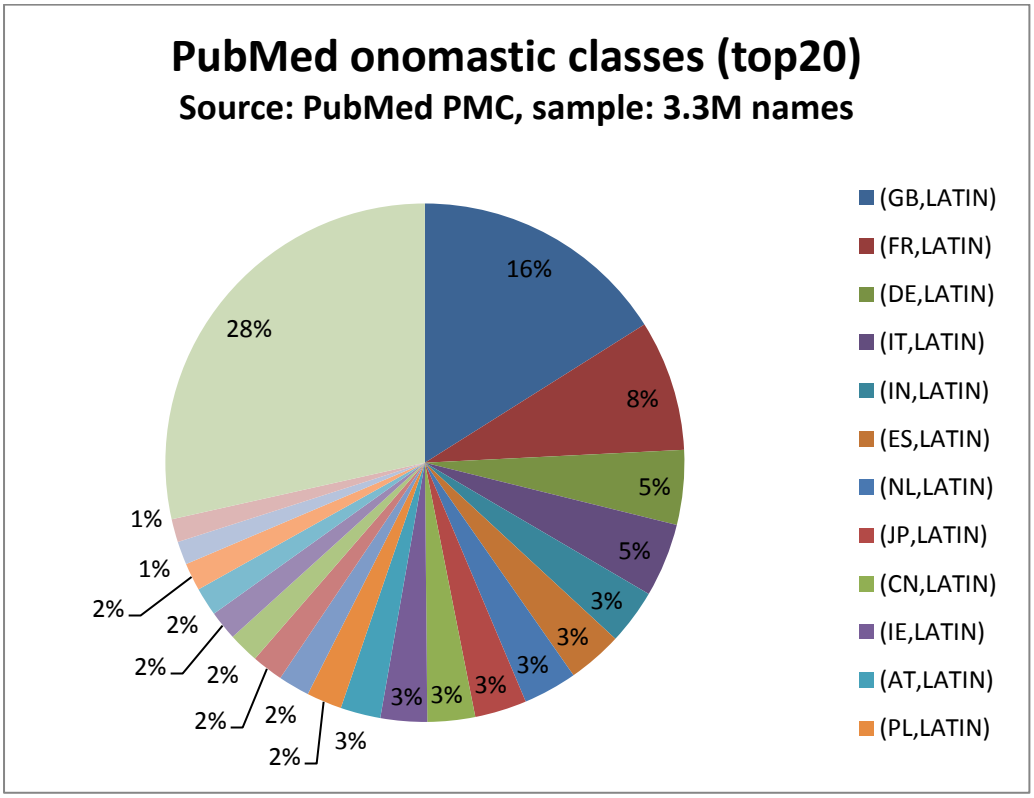
This paper was prepared for ICOS2014, the 25th International Congress of Onomastic Sciences, the premier conference in the field of name studies

Introduction

PubMed/PMC is a large collection of scientific publication in LifeSciences. We used the 2013 data dump for data mining, with 14 million articles and 3.3 million author names. Some of the names are duplicates due to different orthographies, inconsistent use of initials and other data quality issues.

We used NamSor software to allocate an onomastic class to each author name. NamSor software with initially designed to analyse the big data in the field of economic development[1], business and marketing. The method for anthroponomical classification can be summarized as follow: judging from the name only and the publicly available list of all ~150k Olympic athletes since 1896 (and other similar lists of names), for which national team would the person most likely run? Here, the United-States are typically considered as a melting pot of other 'cultural origins': Ireland, Germany, etc. and not as a onomastic class on its own.

The breakdown of author names by onomastic classes is represented below :



The largest groups of unique names in PubMed are British, French, German, Italian, Indian, Spanish, Dutch, etc.

An author with a French name might have a name from Brittany, Corsica or Limousin ... or he might have a Canadian French name, or a Belgium French name. Or he might be an American professor with a French ancestry.

[1] Onomastics and Big Data Mining, ParisTech Review 2013, arXiv:1310.6311 [cs.CY]

Scientists performance is often measured according to the number of publications, and the number of times a publication is cited by other publications (bibliometric rankings).

The table below shows the number of publications and the number of citations, by onomastic classes (top 20), as well as the ratio between the two metrics:

Onoma	A	C	Ratio (C/A)
(GB,LATIN)	557,177	1,664,415	3.0
(FR,LATIN)	272,150	743,471	2.7
(DE,LATIN)	192,778	448,103	2.3
(JP,LATIN)	172,866	361,682	2.1
(IT,LATIN)	187,564	323,771	1.7
(IE,LATIN)	86,161	422,103	4.9
(NL,LATIN)	102,982	321,787	3.1
(AT,LATIN)	78,199	339,819	4.3
(CN,LATIN)*	219,040	186,464	0.9
(IN,LATIN)	153,555	221,332	1.4
(ES,LATIN)	113,407	228,650	2.0
(PL,LATIN)	47,961	268,115	5.6
(SE,LATIN)	65,717	237,017	3.6
(FI,LATIN)	35,533	247,231	7.0
(KR,LATIN)	146,444	105,605	0.7
(TW,LATIN)*	88,822	162,132	1.8
(GR,LATIN)	51,564	196,056	3.8
(DK,LATIN)	42,403	181,199	4.3
(BE,LATIN)	44,647	162,146	3.6
(CH,LATIN)	32,295	162,495	5.0
***CN+TW**	**307,862**	**348,596**	**1.1**

This table tell us that scientists with British names have published 557 thousand articles in PubMed and have been cited 1.6 million times in other PubMed articles: the ratio is 3.

Articles written by authors with Italian names have been relatively less cited (with a ratio of 1.7) while the articles written by authors with Irish names or Finnish names have been more cited (with ratios respectively 4.9 and 7).

We cannot conclude on the overall performance of British, Italian or Finish scientists (many of them might be American scientists), but already we can observe interesting cultural biases emerging that cannot be explained by the imprecision of onomastic classification only. They raise interesting questions:

- can linguistic mastery of the English language explain why authors with British or Irish names have more citations?

- can features of a particular culture (ex. the Irish are excellent networkers and have great pubs) explain why scientific articles are more cited?

- do scientists with Italian names tend to cite more scientists with Foreign sounding names (English, Irish, etc.)?

- do scientists with Finish names tend to cite more scientists with Finish names?

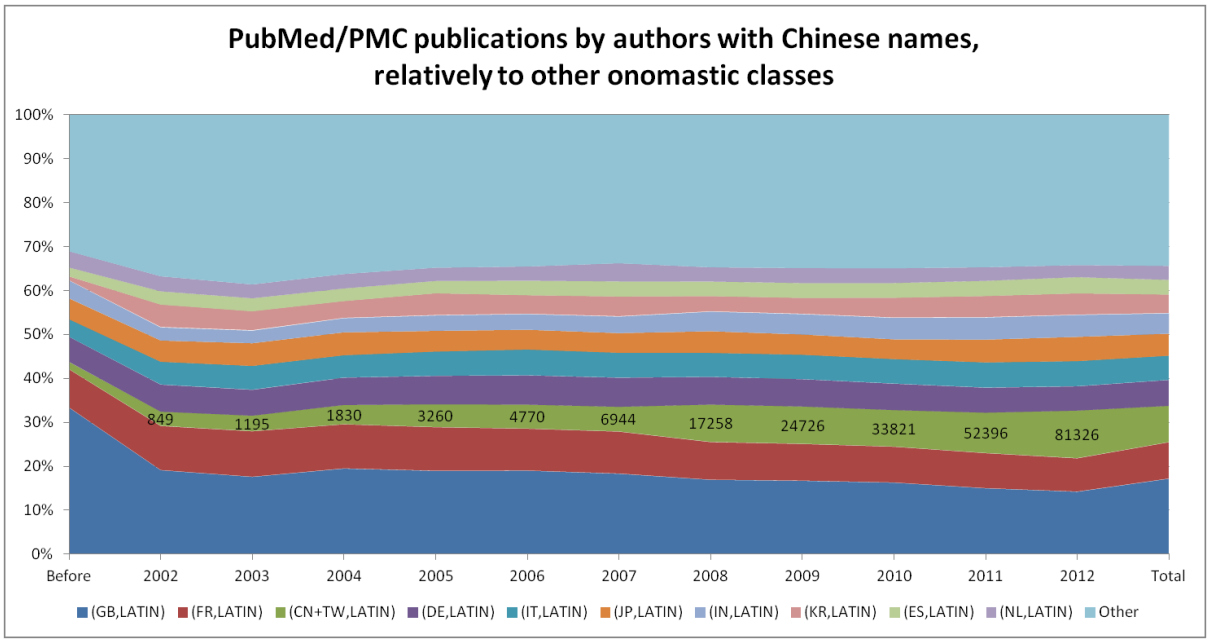
- are there additional cultural biases in the publication process itself (selection, curation, promotion of scientific publications)?

- is there a gender bias worth noting (ex. male scientists are more cited; a culture with less female scientists would get a higher ratio) ?

Altogether, scientists with Chinese names -with names from mainland China or Taiwan- have altogether produced 307 thousand articles and been cited 348 thousand times: a ratio of 1.1, in the low range. We will now focus the rest of this paper on Chinese names: publications authored by a scientist with a Chinese name, or citations of scientists with Chinese names.

Scientists with Chinese names in PubMed

Globally, the number of publications in life sciences has been growing exponentially. Many countries and institutions encourage scientists to publish and link performance to bibliometric rankings (ie. publications in reputable journals, number of citations, etc.)

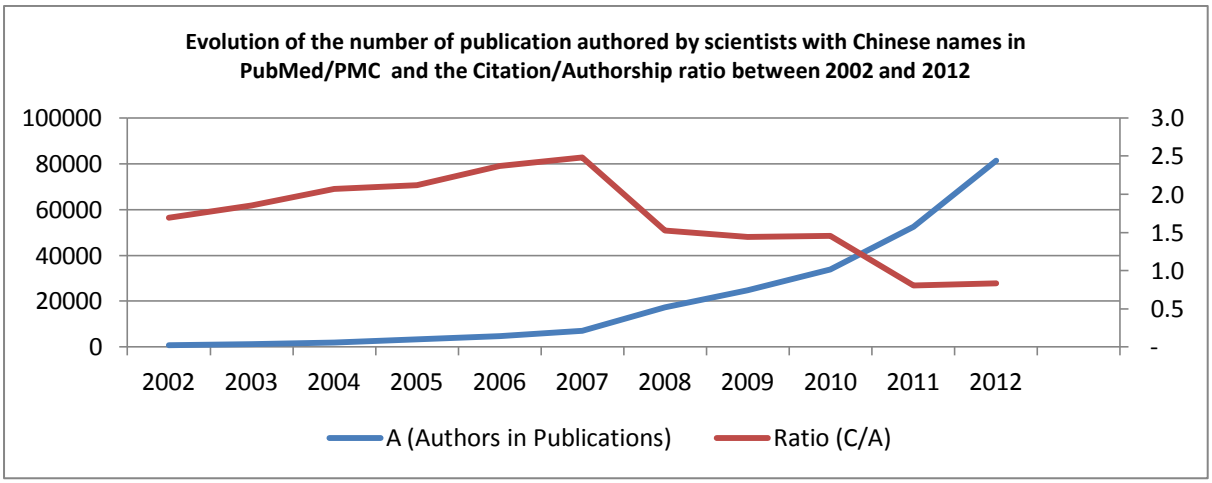


From this chart, we can observe,

- that the absolute number of publications authored by scientists with a Chinese name has nearly tripled between 2007 and 2008 (x2.5, from 7k to 17k);

- that the relative share of publications authored by scientists with a Chinese name (compared to other onomastic classes) is also growing steadily.

This growth in the number of publications by authors with Chinese names, in absolute and relative terms, is matched by a drop in the ratio of citation/authorship :



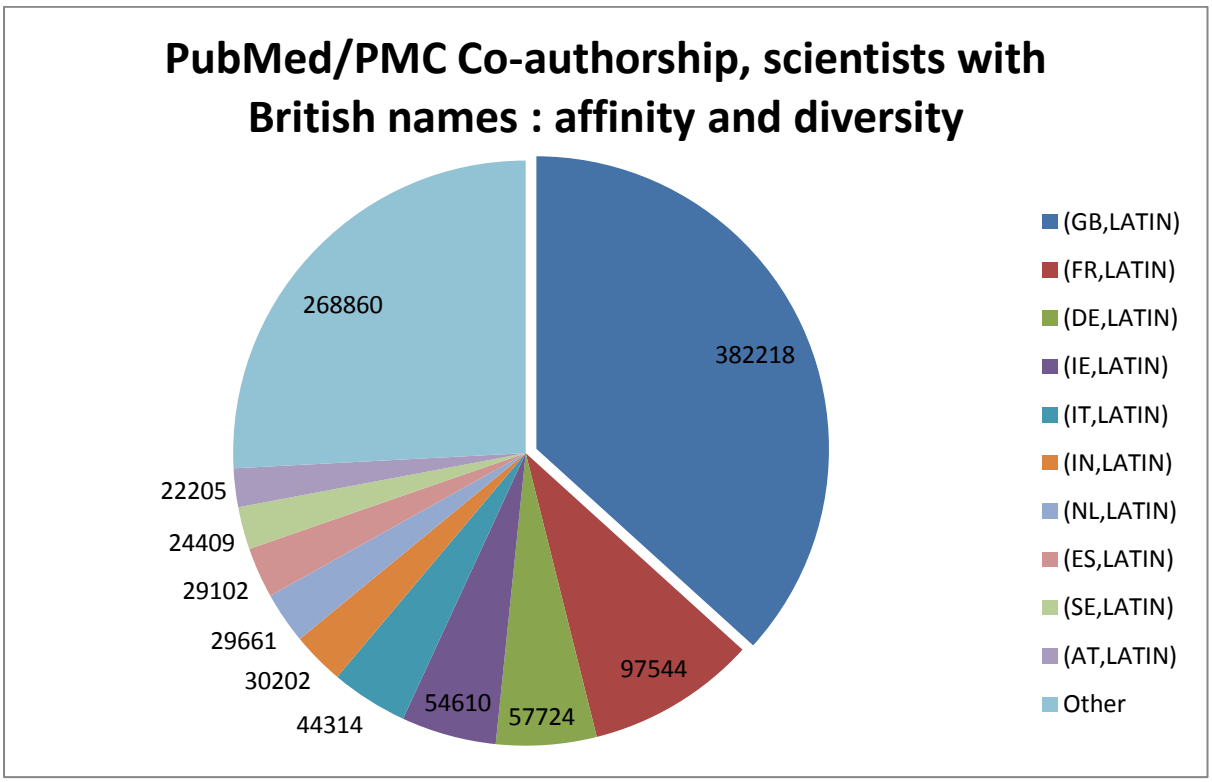
Year	A	C	Ratio (C/A)
2012	81326	68038	0.8
2011	52396	42371	0.8
2010	33821	49260	1.5
2009	24726	35715	1.4
2008	17258	26321	1.5
2007	6944	17234	2.5
2006	4770	11299	2.4
2005	3260	6910	2.1
2004	1830	3782	2.1
2003	1195	2211	1.9
2002	849	1436	1.7
Before	3477	3823	1.1

Next, we will look at co-authorships. We do expect co-authorships to be more frequent within a same onomastic class, because of the correlation with geography : scientists with an Italian name might live in Italy, work in the same University on a research project, publish together the result of their research. We also expect to find diversity: many publications are the result of an international cooperation ; scientists are internationally mobile; last but not least countries like the US, Switzerland attract talents from everywhere and as a result of this global 'brain drain' produce very international research teams.

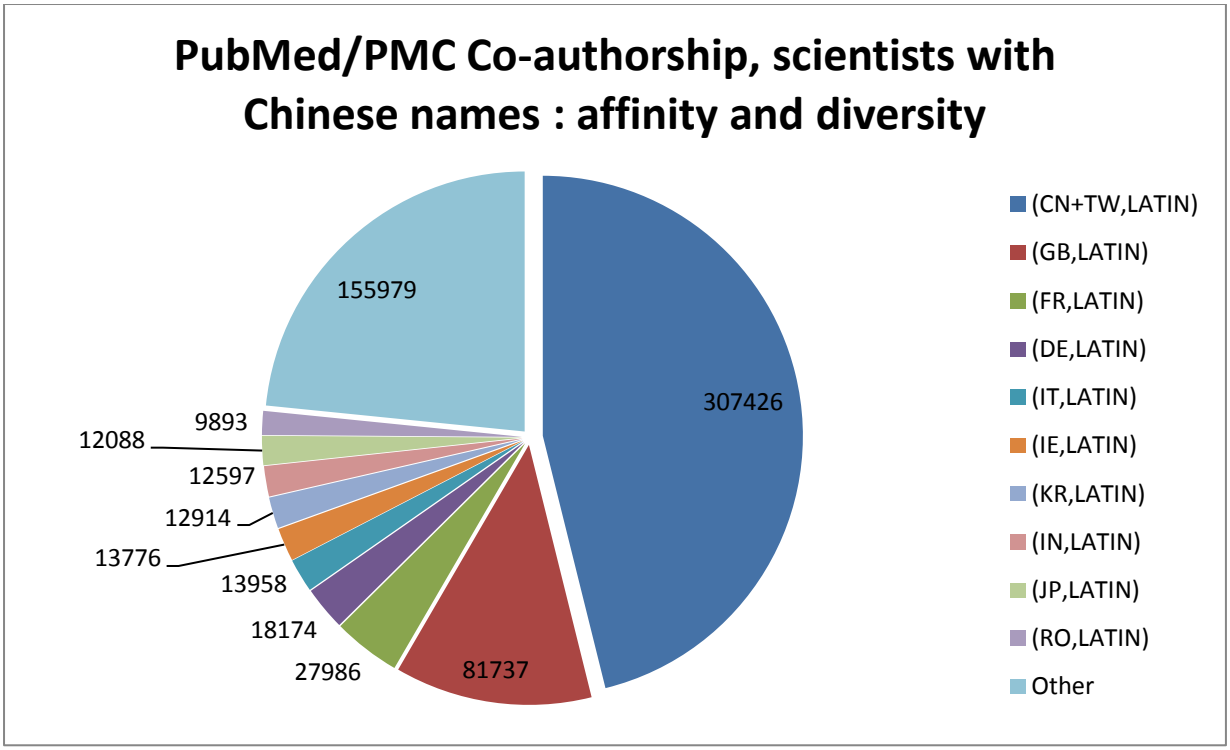
Both aspects, affinity and diversity, are reflected in the following matrix - displaying the number of co-authorships between onomastic classes:

Onoma	(GB,LATIN)	(FR,LATIN)	(CN,LATIN)	(DE,LATIN)	(JP,LATIN)	(IT,LATIN)	(IN,LATIN)	(NL,LATIN)	(KR,LATIN)	(ES,LATIN)	[…]
(GB,LATIN)	382218	94038	64678	76279	22348	38788	48313	44690	13184	24706	
(FR,LATIN)	97544	149321	22037	38419	11171	29738	16172	23120	4722	22213	
(IT,LATIN)	44314	33266	10990	19491	6610	174415	9683	10570	2101	17081	
(DE,LATIN)	57724	28115	14364	82552	6110	12564	9508	18716	3082	6844	
(JP,LATIN)	7813	3625	9647	2695	262912	2269	3046	1538	2533	1220	
(CN,LATIN)	12530	4030	180438	3552	5428	2165	3354	2311	4827	1163	
(ES,LATIN)	29102	22746	6094	10593	3683	16138	6307	6623	1350	79839	
(KR,LATIN)	7598	2613	8391	2081	4681	1173	2530	1095	168436	918	
(IN,LATIN)	30202	9681	9829	8762	4625	4481	64147	4148	2607	2653	
(NL,LATIN)	29661	15546	6415	14739	2326	6109	5680	43438	1349	3707	
[…]											

For example, the first column of the matrix (reflected in the pie chart below) shows that scientists with British names have a strong affinity to be co-author with scientists with British names, but also that they are likely to publish (in order) with scientists with French names, German names, Irish names, Italian names etc.



Scientists with Chinese names have an even stronger affinity to be co-authors with scientists with Chinese names; they are likely to publish (in order) with scientists with British names, French names, German names, Italian names, Irish names, Korean names etc.



Next, we will look at citations. In a perfect world, we expect citations to made based on the merits of scientific research only. We assume some 'invisible hand' will self-regulate the visibility of publications among research communities -so all relevant research is known by the experts of the field. If scientific excellence is equally distributed, we expect the number of publications citing authors of a particular onomastic class to be proportional to the number of authors of that particular onomastic class. However, the following table tells a different story.

Onomastic Class	Onoma Authored %	Onoma Self Citations %	Bias Factor
(GB,LATIN)	16.6%	17.0%	1.02
(FR,LATIN)	8.1%	7.6%	0.94
(IT,LATIN)	5.6%	3.8%	0.68
(DE,LATIN)	5.8%	6.1%	1.05
(CN+TW,LATIN)	9.2%	12.1%	**1.32**
(ES,LATIN)	3.4%	3.8%	1.13
(JP,LATIN)	5.2%	19.3%	3.73
(IE,LATIN)	2.6%	4.4%	1.73
(NL,LATIN)	3.1%	5.6%	1.83
(AT,LATIN)	2.3%	4.2%	1.79
(SE,LATIN)	2.0%	3.5%	1.76
(IN,LATIN)	4.6%	4.1%	0.89
(PT,LATIN)	1.9%	2.3%	1.17
(GR,LATIN)	1.5%	2.8%	1.82
(KR,LATIN)	4.4%	3.0%	0.68
(BE,LATIN)	1.3%	2.6%	1.98
(DK,LATIN)	1.3%	3.4%	2.65

In this table, we observe that authors with British names represent 16.6% of publications, but 17% of their citations : a bias factor of 1.02 (almost no bias). Conversely, we observe that authors with French names represent 8.1% of publications, but only 7.6% of their citations : a bias factor of 0.94 indicating that authors with French names tend to cite authors with foreign names more.

As for authors with Chinese names, they represent 9.2% of the publications, but 12.1% of their citations : a bias factor of 1.32 indicating that they tend to cite authors with Chinese names more.

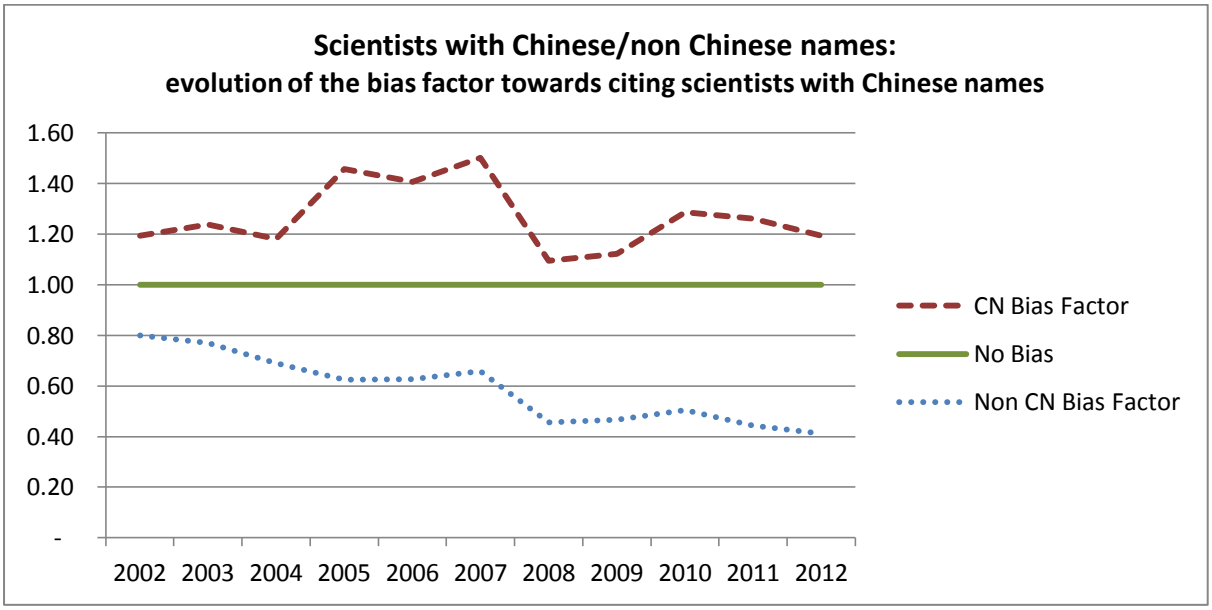
Authors with Chinese names have a positive bias in citing authors with Chinese names, however we can see other cases where the bias is even stronger: authors with Japanese names citing authors with Japanese names, authors with Danish names...

More interesting, the following table shows that -apart from authors with a Chinese name- every other onomastic class (British, French, Italian, German etc.) have a negative bias towards citing authors with a Chinese name.

Onomastic class	Chinese Onoma Citation Pct%	Bias Factor
(GB,LATIN)	3.9%	0.43
(FR,LATIN)	3.9%	0.42
(IT,LATIN)	3.9%	0.43
(DE,LATIN)	4.1%	0.44
(CN+TW,LATIN)	12.1%	1.32
(ES,LATIN)	4.0%	0.43
(JP,LATIN)	5.2%	0.56
(IE,LATIN)	4.0%	0.44
(NL,LATIN)	3.5%	0.38
(AT,LATIN)	4.1%	0.44
(SE,LATIN)	3.6%	0.40
(IN,LATIN)	5.9%	0.65
(PT,LATIN)	4.0%	0.43
(GR,LATIN)	3.9%	0.42
(KR,LATIN)	6.8%	0.74
(BE,LATIN)	3.8%	0.42
(DK,LATIN)	3.9%	0.42

Authors with a Chinese name tend to cite authors with a Chinese name more. Comparatively, scientists with non Chinese names (British, French, Italian, German etc.) have a bias factor of 0.46 and are 3 times less likely to cite publications authored by a scientist with a Chinese name.

We will now see of the biases factors evolve between 2002 and 2012:



According to this table, the positive bias factor of authors with Chinese names in citing other authors with Chinese names remains roughly stable. On the other hand, the negative bias factor of scientists with non-Chinese names in citing authors with Chinese names is generally increasing.

Manual controls

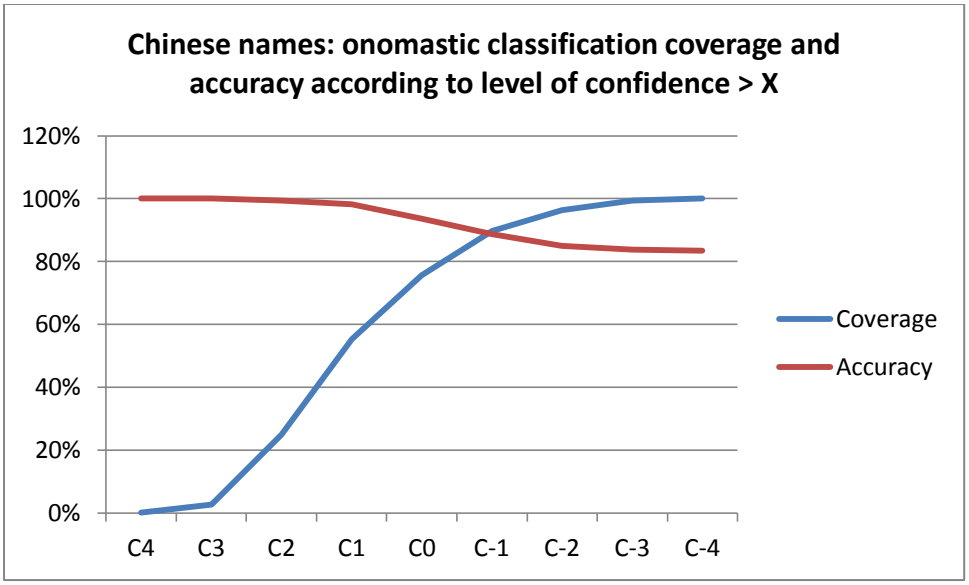
Given the large number of names automatically classified in a taxonomy based on geographic origin (China, etc.) we could not verify manually the entire database. We verified manually two randomly selected subsets:

- firstly, a list of 1280 names recognized by the software as Chinese names;

- secondly, a list of ~10000 names classified by the software into the full taxonomy (over 100 onomastic classes, corresponding to different countries of origin)

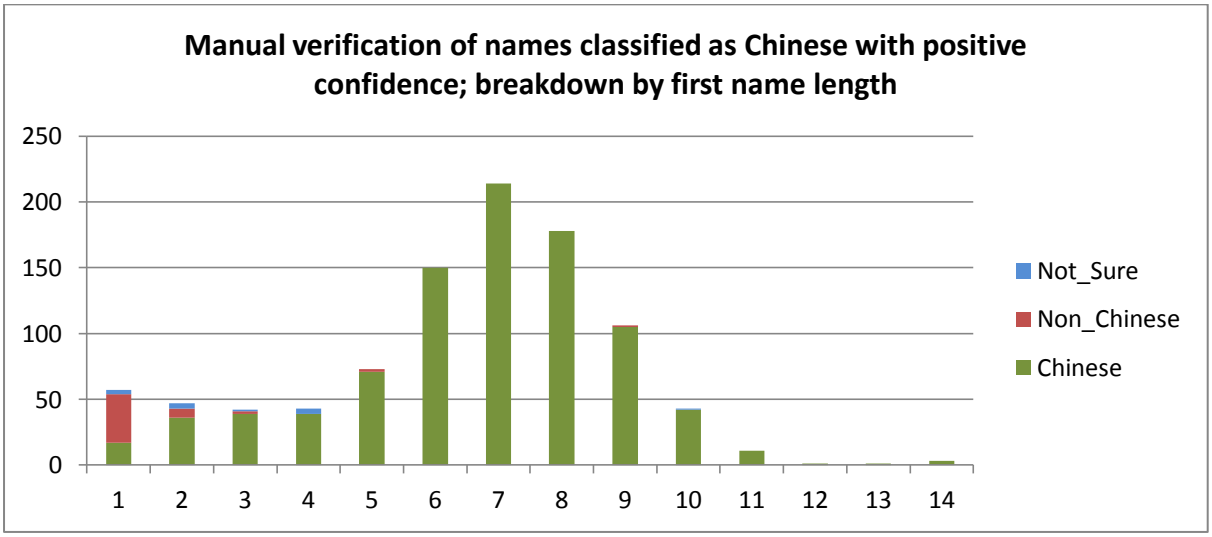
According to the first validation method, 83% of names the software recognized as Chinese were manually verified as Chinese; 2% unknown; 15% as non-Chinese (ie. mis-classifications).

The software outputs a confidence level. 76% of the names were classified with positive confidence. For the names recognized as Chinese with a positive confidence, 94% were manually verified as Chinese; 1% unknown; 4% as non-Chinese (ie. mis-classification).



In PubMed, many names do not have a full first name, only initials.

For names classified with positive confidence, we found that first names of just one or two character (ex. J or JH) accounted for 90% of mis-classifications. When the input includes a full name (as would generally be the case with other bibliometric sources such as Thomson WoS, Scopus or ORCID) the accuracy is 99%.



According to the second validation method, we can calculate the usual metrics used in classification : precision and recall.

10172 names were manually classified by a manual operator independently. In this method, errors could be made by the computer and also by the manual operator.

For the calculations below, we assume the assume the manual operator made no mistakes (this is not the case, error is human). The manual operator could classify 50% of names, left the rest as 'Not Sure'.

For Chinese, non Chinese names, the software precision was respectively 81% and 97% and the recall was 59% and 99%. For names classified by the software with positive confidence (52% of all names), the precision was 93% and the recall was 69%. Excluding the names with first name length < 2 (initials, such as J or JH) the precision was 97% and the recall was 72%.

If conversely, we assume that the computer made no mistakes, then we can compare the precision and recall of the operator with that of the computer:

	Chinese Names		Non Chinese Names	
All Names	Computer	Human	Computer	Human
Precision	81%	59%	97%	99%
Recall	59%	42%	99%	48%

	Chinese Names		Non Chinese Names	
Confidence>0	Computer	Human	Computer	Human
Precision	93%	69%	96%	99%
Recall	69%	49%	99%	48%

	Chinese Names		Non Chinese Names	
Confidence>0 && Len(firstName)>2	Computer	Human	Computer	Human
Precision	97%	72%	96%	100%
Recall	72%	51%	100%	48%

This method of cross validation between computer and human could be improved by having several manual checks by different operators to obtain a good validation sample.

Future work

For future work, we would data mine the large commercial bibliographic databases (Thomson WoS, Scopus and possibly ORCID) because they offer better data quality and useful additional information:

- firstly, they have the full name in addition to the short name cited with just initials; this significantly reduces the error rate of onomastic classification

- secondly, they link scientists to research institutions (affiliations) and geographies (country of affiliation) ; this allows additional analysis on the topic of Diasporas and brain drain, comparing -for example- the research output of Chinese / Chinese American scientists in the US with that of scientists of Mainland China;

- thirdly, those databases have a larger coverage in terms of scientific disciplines, allowing comparison between different fields of research.

Conclusions

Significant cultural biases exist, not only in the way scientists co-author publications together, but also in the way they make citations. Scientific publications authored by scientists with Chinese names are three times less cited by the international research community that they are cited by other scientists with Chinese names. We cannot conclude on the quality of Chinese research but we can challenge the commonly accepted idea that the volume of publications and citations alone indicate that China is becoming a superpower in Science and Technology.

Given the importance of bibliometric rankings in the way countries build and monitor public policies on Science and Education or international cooperation; in the way research institutions measure and reward scientific excellence of researchers and teams, those biases should be accounted for. Otherwise, international comparisons are not 'scientific', not fair and can lead to wrong decisions.