NamSor's performance in predicting the country of origin and ethnicity of 90,000 researchers based on their first and last names

Paul Sebo (paulsebo@hotmail.com)
University of Geneva

Research Article

Keywords: accuracy, classification, country, ethnicity, misclassification, NamSor, onomastics, performance

Posted Date: April 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1565759/v2

Abstract

Objective: We aimed to evaluate NamSor's performance in predicting the country of origin and ethnicity of researchers based on their first and last names.

Methods: We selected the 22 countries whose researchers authored at least 1000 medical publications in 2020 and whose percentage of migrants was <2.5% in 2020. Using PyMed, a Python library that gives access to PubMed, we extracted all publications in 2021 whose first authors were affiliated with universities or research institutes in the selected countries (N=89,906 articles). We calculated the proportion of misclassifications (=errorCodedWithoutNA) and the proportion of non-classifications (=naCoded) for three variables available in NamSor: “continent of origin” (Asia/Africa/Europe), “country of origin” and “ethnicity”. We created two other variables: “continent#2” ("Europe" replaced by "Europe/America/Oceania") and “country#2” ("Spain" replaced by “Spain/Hispanic American country” and "Portugal" replaced by "Portugal/Brazil"). We repeated the analyses by removing all results with accuracy<50%.

Results: For the full sample and the subsample, errorCodedWithoutNA was, respectively, 17.0% and 13.4% for “continent”, 6.7% and 3.6% for “continent#2”, 28.3% and 20.3% for “country”, 20.7% and 12.2% for “country#2”, and 21.2% and 15.7% for “ethnicity”, whereas naCoded was zero and 18.3% for all variables, except for “ethnicity” (zero and 10.9%).

Conclusions: NamSor is accurate in determining the continent of origin of individuals, especially when using the modified variable and/or restricting the analysis to names with inference accuracy ≥50%. The risk of misclassification is higher with country of origin or ethnicity, but decreases, as with continent of origin, when using the modified variable (country#2) and/or the subsample.

Introduction

Individuals are regularly discriminated against, for example because of their gender, their sexual orientation, their religion or their social or ethnic origin. The world of research is only a mirror of our society and does not escape these rejection behaviors. The study of discrimination in research mainly focused on gender inequalities, and numerous publications highlighted the major obstacles faced by women throughout their careers.[1–4] As a result, programs were launched in many countries to increase the representation of women in key academic positions and improve their career prospects.[5]

However, rejection behaviors can be related to other social categories in addition to gender. The origin of researchers seems to be a criterion of discrimination according to several recent publications. Researchers from low- and middle-income countries (LMICs), for example, were found to be underrepresented as authors of articles[6,7] or as members of editorial boards.[8]

To save time and resources in studies of inequalities by origin, researchers sometimes rely on NamSor, an online onomastic tool that infers origin from first and last names. However, because of its advantages, it is likely that more researchers will use it in the future. Indeed, NamSor combines three main advantages that are valuable to researchers: it is fast, cost-effective and can be applied retroactively to large datasets. The methodology used by the algorithm to determine the most likely origin of individuals is relatively opaque to non-specialists, but likely relies on large databases combining names with cultural, ethnic and linguistic backgrounds.

NamSor was used in several studies to estimate the origin of individuals. In a study comparing the number of citations, a proxy for scientific impact and relevance, for 13,000 articles published between 2015 and 2019 in fourteen high-impact general medical journals, it was found that articles by first/last authors with African names were cited less often than other articles.[9] In another study evaluating ethnic and gender disparities in 442 prize presentation sessions at two prestigious surgical conferences in the UK over a 21-year period, the authors showed that almost half of the presenters (48%) were white men, followed by Asian men (25%).[10] By contrast, there was only one black woman, one black man, and sixteen Asian women during these twenty-one years.

NamSor can help to determine both the gender and the origin of individuals. Its performance is high for gender inference, as demonstrated recently in a study comparing several gender detection tools,[11] but, to our knowledge, there is no published data on the accuracy of this tool for determining the origin of individuals.

Based on a database of scientific publications (PubMed) including authors' names and affiliations, the objective of this study was to evaluate the performance of NamSor for estimating the origin of researchers. Thanks to the progress made in data mining techniques, it is hypothesized that its performance is high for names of researchers from a large number of countries.

Methods

Selection of publications and their authors

We used data from SCImago Journal & Country Rank to retrieve all countries whose researchers authored at least 1000 scientific publications in 2020 in the field of medicine. SCImago Journal & Country Rank is a publicly available portal that includes scientific indicators for journals and countries developed from information in the Scopus® database.[12] Citation data are from over 34,000 titles and over 5,000 international publishers. Seventy-five countries met the inclusion criterion for the study, as shown in Table 1 (country #1: USA with 277,130 publications, country #75: Cuba with 1,059 publications).

We also used data from International Migrant Stock 2020, available on the United Nations Population Division portal, to obtain the percentage of migrants by country in 2020. Data on estimates of the number (or "stock") of international migrants are presented as a percentage of the total population, by age, sex, and country of destination, and are based on national statistics, in most cases obtained from population censuses.[13] We selected the 22 countries for which this proportion was below 2.5 percent (Table 1). We restricted the study to these countries only in order to obtain names of researchers that were as homogeneous as possible and representative of the selected countries. The proportion of migrants for these countries ranged from zero for Cuba to 2.2 percent for Japan and Poland.

Then, using PyMed,[14] a Python library that gives access to PubMed, we extracted all publications in 2021 with at least one author affiliated with a university or research institute located in the selected countries (N=120,104). We obtained a csv file in which the variable 'authors' had the following form (example for a publication authored by three researchers):

[{‘lastname’ : ‘x’, ‘firstname’ : ‘x’, ‘initials’ : ‘x’, ‘affiliation’ : ‘x’}, {‘lastname’ : ‘y’, ‘firstname’ : ‘y’, ‘initials’ : ‘y’, ‘affiliation’ : ‘y’}, {‘lastname’ : ‘z’, ‘firstname’ : ‘z’, ‘initials’ : ‘z’, ‘affiliation’ : ‘z’}]

Using Stata, we created the variable 'author1' (i.e., data for first authors only) and the variable 'country1' (i.e., country of affiliation of first authors). We removed the publications for which the affiliation to the selected countries did not concern the first author. The study database contained data for 89,906 publications.

NamSor Applied Onomastics

The authors' names were classified with NamSor Applied Onomastics, a name recognition software.[15] The software recognizes the linguistic or cultural origin of each name and assigns a gender (male or female) and/or an onomastic class (e.g., China, India). As the estimation is probabilistic, the software also provides a probability for the inference (‘probabilityCalibrated’) ranging from zero to one.

The names can be classified according to the continent of origin (three continents: Asia, Africa or Europe), the country of origin (e.g., China or India) and the ethnicity (e.g., Chinese or Indian). We created two other variables: continent#2 ("Europe" replaced by "Europe, America or Oceania") and country#2 ("Spain" replaced by “Spain or Hispanic American country” and "Portugal" replaced by "Portugal or Brazil"). We added these variables because a preliminary analysis of our data showed that a majority of researchers with Hispanic or Portuguese names who were affiliated with universities or research institutes in Brazil, Mexico or Cuba were considered to be from either Spain or Portugal.

Performance analysis

We evaluated NamSor’s performance by computing three efficiency metrics.[11,16] These metrics refer to the confusion matrix that contains three components, with 'c' corresponding to correct classifications, 'i' to misclassifications (i.e., a wrong continent, country or ethnicity assigned to a name) and 'u' to non-classifications (i.e., no continent, country or ethnicity assigned).

	Correct continent, country or ethnicity (predicted)	Incorrect continent, country or ethnicity (predicted)	Unknown (predicted)
Continent, country or ethnicity (actual)	c	i	u

errorCoded = (i + u) / (c + i + u)

errorCodedWithoutNA = (i) / (c + i)

naCoded = (u) / (c + i + u)

These performance metrics can be interpreted as follows: *errorCoded* estimates the proportion of misclassifications and non-classifications (this measure therefore penalizes both types of errors equally), *errorCodedWithoutNA* measures the proportion of misclassifications excluding non-classifications and *naCoded* measures the proportion of non-classifications.

We repeated the analyses by removing all results with inference accuracy <40%, <50%, <60% and <70%, respectively. All assignments made with an accuracy level below the selected threshold value were considered as non-classifications. We performed all analyses with STATA version 15.1 (College Station, TX, USA).

Ethical considerations

Since this study did not involve the collection of personal health-related data it did not require ethical review, according to current Swiss law.

Results

The main results of the study are presented in Tables 2, 3 and 4, for both the full sample and the four subsamples (sensitivity analyses). Table 2 shows for each of the 22 selected countries the number of researchers whose name origin was correctly classified by NamSor. These data are then summarized in Table 3 (confusion matrices) and Table 4 (performance metrics).

As shown in Table 2, the proportion of correct classifications varied widely by country, and was higher for “continent of origin”, compared to the other two variables tested. Most of the names were correctly identified for some countries, such as Polish, Pakistani and Vietnamese names. Other names were poorly recognized, for example Nepalese or Tanzanian names, and others were not recognized at all by NamSor, mainly Latin American names. No Brazilian, Mexican, Filipino or Cuban names were correctly identified. Brazilian names were considered as Portuguese, Mexican or Cuban names as Spanish.

The use of two modified variables (continent#2 and country#2) increased for all countries the proportion of correct classifications. In addition, by restricting the analyses to subsamples, NamSor's performance tended to increase gradually as the accuracy threshold value increased. For example, for "country of origin", the proportion of correct classifications for Japan was 84.8% for the full sample, 85.4% for a threshold value of 40%, 88.4% for a threshold value of 50%, 89.4% for a threshold value of 60%, and 90.5% for a threshold value of 70%. Similarly, the number of non-classifications also gradually increased as the accuracy threshold value increased. For example, for the same variable (country of origin) and the same country (Japan), the number of names classified by NamSor was 6431 for the full sample, 6373 with a cut-off value of 40%, 6080 with a cut-off value of 50%, 5949 with a cut-off value of 60%, and 5771 with a cut-off value of 70%.

As shown in the confusion matrices (Table 3), there was a decrease in the number of correct classifications as the threshold value for inference accuracy increased, due to a greater increase in the number of non-classifications relative to the decrease in the number of misclassifications. For example, for “country of origin”, the number of correct classifications was 64,499 for the full sample, 63,901 with a threshold value of 40%, 58,608 with a threshold value of 50%, 55,149 with a threshold value of 60%, and 50,679 with a threshold value of 70%.

Table 4 (accuracy metrics) confirms the results of the previous table. The proportion of misclassifications and non-classifications (i.e., errorCoded) was lowest for the full sample and increased gradually as the threshold value increased. With a cut-off value of 40%, errorCoded increased only slightly compared to the full sample because the number and proportion of non-classifications was low: 2334 (2.6%) for “continent of origin” and “country of origin”, and 6210 (6.9%) for “ethnicity”. Above 60%, errorCoded exceeded 25% for all variables tested. The use of the 50% cut-off value was probably the strategy that provided the best compromise between increasing the proportion of correct classifications on the one hand and increasing the proportion of non-classifications on the other.

For the full sample and the subsample with inference accuracy ≥50%, the proportion of misclassifications (=errorCodedWithoutNA) was, respectively, 17.0% and 13.4% for “continent of origin”, 6.7% and 3.6% for “continent#2”, 28.3% and 20.3% for “country of origin”, 20.7% and 12.2% for “country#2”, and 21.2% and 15.7% for “ethnicity”. Finally, the proportion of non-classifications (=naCoded) was zero and 18.3% for all variables, respectively, with the exception of “ethnicity” (zero and 10.9%).

Discussion

In this cross-sectional study, we examined the performance of NamSor in determining the origin of nearly 90'000 researchers affiliated with universities or research institutes in twenty-two different countries. We found NamSor to be accurate in determining the continent of origin, especially when using the modified variable (continent#2) and restricting the analysis to names with an inference accuracy ≥ 50%. For continent#2, the proportion of misclassifications (i.e., errorCodedWNA) was only 6.7% for the full sample and 3.6% for the subsample. However, we found that the risk of misclassification was higher with country of origin or ethnicity, but also decreased when using the modified variable (country#2) and the subsample.

Comparison with existing literature

Several authors used Namsor in the past to estimate the origin of individuals in their research, both in medicine[9,10] and in other disciplines,[17,18] but our study is the first to our knowledge to have evaluated its performance. We already evaluated NamSor's performance in determining the gender of individuals from their first and last names, and showed that the tool was accurate in the majority of cases (errorCodedWNA 2%).[11] However, we found that NamSor was much less efficient for some countries, for example for Chinese names.[19] We also found that the use of the accuracy parameter ('probabilityCalibrated') was not useful to improve the performance of NamSor for gender estimation.[20]

The results obtained in the current study were quite different. Asian names were in general relatively well recognized by NamSor. For example, 76% of the names of researchers affiliated with universities or research institutes in China were correctly classified for "country of origin" (and even 85% for "ethnicity"). These figures were 85% and 84%, respectively, for Japan. Furthermore, the use of the accuracy parameter greatly improved the performance of the tool for origin. The best compromise between improving NamSor's performance and increasing the number of non-classifications was obtained with a threshold value of 50%. With a threshold value of 40%, too few queries were considered as non-classifications (2.6% for "continent of origin" and "country of origin", and 6.9% for "ethnicity") to make a noticeable change in performance metrics. For example, for "continent of origin" and "country of origin", errorcodedWNA decreased only from 17.0% to 16.4% and from 28.3% to 27.0%, while these proportions decreased to 13.4% and 20.3%, respectively, for a threshold value of 50%.

As expected, using "continent of origin" yielded more accurate assignments than either "country of origin" or "ethnicity". This is indeed a logical finding since "continent of origin" consisted of only three categories, far fewer than the other two variables. For example, if authors with Chinese names were considered to be of Japanese origin, the continent of origin (i.e., Asia) would have been correctly estimated, unlike country of origin or ethnicity. However, if researchers using NamSor needed more precision for their study than simply assigning a continent of origin, the use of "ethnicity" would a priori allow more accurate queries than "country of origin". For example, for the total sample, errorCodedWNA was 21.2% for "ethnicity" and 28.3% for "country of origin". This difference persisted with the various subsamples.

As expected, it was the joint use of "continent#2" or "country#2", and the various subsamples with threshold values of 50% or more that really improved the performance of NamSor. For "continent#2" and a cut-off value of 50%, the proportion of misclassifications was only 3.6% in our study (vs. 17.0% for "continent of origin" and the total sample). For "country#2" and the same cut-off value of 50%, this proportion was 12.2% (vs. 28.3% for "country of origin" and the total sample). "Continent#2" led to more accurate assignments than "continent of origin", as many researchers with Spanish or Portuguese names were actually affiliated with universities or research institutes in Latin America. For the same reason, replacing “country of origin” by “country#2” (i.e., "Spain" by "Spain or Hispanic American country", and "Portugal" by "Portugal or Brazil") was also useful for improving NamSor's performance.

Anglo-Saxon countries (i.e., UK, USA, Canada, Australia and New Zealand) were not included in the study, as the proportion of migrants was too high in these countries. However, it is likely that if they were included we would observe misclassifications for the same reason as for names of Spanish or Portuguese origin. It would therefore make sense to use a third variable (country#3) that would add a modification to "country#2", replacing "UK", "USA", "Canada", "Australia" and "New Zealand" with "UK or USA or Canada or Australia or New Zealand". That said, to assess the relevance and usefulness of “country#3”, and more generally to confirm the results of our study, further studies would be needed in the future, also including names of Anglo-Saxon origin and determining the origin of individuals ideally by self-identification.

Limitations

Our study has a large sample size but has two main limitations. We restricted the study to twenty-two countries spread over four continents (Europe, Asia, Africa and America). As the performance of NamSor varies depending on the country examined, our results are not necessarily generalizable to other countries or regions (e.g., Oceania). In addition, the true origin of the researchers was based on their country of affiliation and was not determined by self-identification. Although we restricted the study to countries with less than 2.5% migrants to obtain the most homogeneous populations possible with names representative of the selected countries, there were inevitably foreign researchers in these countries. The results of our study are therefore probably an underestimate of the real performance of NamSor

Conclusion

NamSor is accurate in determining the continent of origin of individuals from their first and last names, especially when using the modified variable (i.e., continent#2) and restricting the analysis to names with inference accuracy ≥ 50%. The risk of misclassification is higher with country of origin or ethnicity, but decreases, as with continent of origin, when using the modified variable (i.e., country#2) and the subsample.

Declarations

Conflict of interest: None

Funding sources: None

Data availability statement: Data associated with this article are available in the Open Science Framework (Identifier: DOI 10.17605/OSF.IO/PVFCN).

References

1 Safdar B, Naveed S, Chaudhary AMD, Saboor S, Zeshan M, Khosa F. Gender Disparity in Grants and Awards at the National Institute of Health. *Cureus* 2021; **13**: e14644.

2 Richter KP, Clark L, Wick JA, *et al.* Women Physicians and Promotion in Academic Medicine. *N Engl J Med* 2020; **383**: 2148–57.

3 Sebo P, Clair C. Gender gap in authorship: a study of 44,000 articles published in 100 high-impact general medical journals. *Eur J Intern Med* 2021; S0953-6205(21)00313-7.

4 Sebo P. Gender disparity in publication associated with editor-in-chief gender: A cross-sectional study of fifty high-impact medical journals. *J Gen Intern Med* 2022; *In press*.

5 Gender equality in research and innovation. Eur. Comm. https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/democracy-and-rights/gender-equality-research-and-innovation_en (accessed April 1, 2022).

6 Merriman R, Galizia I, Tanaka S, Sheffel A, Buse K, Hawkes S. The gender and geography of publishing: a review of sex/gender reporting and author representation in leading general medical and global health journals. *BMJ Glob Health* 2021; **6**: e005672.

7 Busse CE, Anderson EW, Endale T, *et al.* Strengthening research capacity: a systematic review of manuscript writing and publishing interventions for researchers in low-income and middle-income countries. *BMJ Glob Health* 2022; **7**: e008059.

8 Nafade V, Sen P, Pai M. Global health journals need to address equity, diversity and inclusion. *BMJ Glob Health* 2019; **4**: e002018.

9 Sebo P. Fewer citations for authors with African names? A study of 13,000 articles published in high-impact general medical journals (*submitted)*.

10 Seehra JK, Lewis-Lloyd C, Koh A, *et al.* Publication Rates, Ethnic and Sex Disparities in UK and Ireland Surgical Research Prize Presentations: An Analysis of Data From the Moynihan and Patey Prizes From 2000 to 2020. *World J Surg* 2021; **45**: 3266–77.

11 Sebo P. Performance of gender detection tools: a comparative study of name-to-gender inference services. *J Med Libr Assoc JMLA* 2021; **109**: 414–21.

12 SJR - International Science Ranking. https://www.scimagojr.com/countryrank.php?year=2019 (accessed April 1, 2022).

13 International Migrant Stock | Population Division. https://www.un.org/development/desa/pd/content/international-migrant-stock (accessed April 1, 2022).

14 gijswobben/pymed. GitHub. https://github.com/gijswobben/pymed (accessed April 1, 2022).

15 Namsor: name checker for gender, origin and ethnicity classification. https://namsor.app/ (accessed April 1, 2022).

16 Santamaría L, Mihaljević H. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput Sci* 2018; **4**: e156.

17 Nagle F, Teodoridis F. Jack of all trades and master of knowledge: The role of diversification in new distant knowledge integration. *Strateg Manag J* 2020; **41**: 55–85.

18 de Rassenfosse G, Hosseini R. Discrimination against foreigners in the U.S. patent system. *J Int Bus Policy* 2020; **3**: 349–66.

19 Sebo P. How accurate are gender detection tools in predicting the gender for Chinese names? A study with 20,000 given names in Pinyin format. *J Med Libr Assoc JMLA* 2022; **110**: 205–11.

20 Sebo P. Are Accuracy Parameters Useful for Improving the Performance of Gender Detection Tools? A Comparative Study with Western and Chinese Names. *J Gen Intern Med* 2022; published online March 15. DOI:10.1007/s11606-022-07469-6.

Tables

Table 1. List of countries whose researchers authored at least 1000 scientific publications in 2020 in the field of medicine, and percentage of migrants per country in 2020

Rank	Country	Region	Number of documents published in 2020	Migrant stock in 2020 (as a percentage of the total population)	Country included in the study[1] (Y/N)
1	United States	Northern America	277130	15.3	N
2	China	Asiatic Region	172201	0.1	Y
3	United Kingdom	Western Europe	81178	13.8	N
4	Germany	Western Europe	62063	18.8	N
5	Italy	Western Europe	56413	10.6	N
6	Japan	Asiatic Region	48994	2.2	Y
7	Canada	Northern America	46214	21.3	N
8	India	Asiatic Region	44586	0.4	Y
9	Australia	Pacific Region	41640	30.1	N
10	France	Western Europe	41039	13.1	N
11	Spain	Western Europe	37726	14.6	N
12	Brazil	Latin America	30269	0.5	Y
13	Netherlands	Western Europe	29362	13.8	N
14	South Korea	Asiatic Region	28892	3.4	N
15	Turkey	Middle East	21840	7.2	N
16	Switzerland	Western Europe	21612	28.8	N
17	Iran	Middle East	21577	3.3	N
18	Russian Federation	Eastern Europe	17909	8.0	N
19	Sweden	Western Europe	17054	19.8	N
20	Belgium	Western Europe	14610	17.3	N
21	Poland	Eastern Europe	13993	2.2	Y
22	Denmark	Western Europe	12879	12.4	N
23	Taiwan	Asiatic Region	12421	N/A	N
24	Austria	Western	10245	19.3	N

		Europe				
25	Egypt	Africa/Middle East	9639	0.5	Y	
26	Mexico	Latin America	9347	0.9	Y	
27	Saudi Arabia	Middle East	9255	38.6	N	
28	Portugal	Western Europe	9145	9.8	N	
29	Israel	Middle East	8733	22.6	N	
30	South Africa	Africa	8432	4.8	N	
31	Greece	Western Europe	8384	12.9	N	
32	Norway	Western Europe	8292	15.7	N	
33	Pakistan	Asiatic Region	7620	1.5	Y	
34	Singapore	Asiatic Region	7458	43.1	N	
35	Ireland	Western Europe	7040	17.6	N	
36	Thailand	Asiatic Region	6610	5.2	N	
37	Malaysia	Asiatic Region	6511	10.7	N	
38	Finland	Western Europe	6452	7.0	N	
39	Hong Kong	Asiatic Region	6325	39.5	N	
40	New Zealand	Pacific Region	6201	28.7	N	
41	Czech Republic	Eastern Europe	6082	5.1	N	
42	Indonesia	Asiatic Region	5565	0.1	Y	
43	Argentina	Latin America	4901	5.0	N	
44	Chile	Latin America	4734	8.6	N	
45	Colombia	Latin America	4722	3.7	N	
46	Nigeria	Africa	4138	0.6	Y	
47	Hungary	Eastern Europe	3529	6.1	N	
48	Romania	Eastern Europe	3378	3.7	N	
49	Iraq	Middle East	3345	0.9	Y	
50	Ethiopia	Africa	2899	0.9	Y	
51	Bangladesh	Asiatic	2690	1.3	Y	

		Region			
52	Croatia	Eastern Europe	2455	12.9	N
53	Viet Nam	Asiatic Region	2378	0.1	Y
54	Serbia	Eastern Europe	2374	9.4	N
55	Ukraine	Eastern Europe	2330	11.4	N
56	United Arab Emirates	Middle East	2188	88.1	N
57	Lebanon	Middle East	2179	25.1	N
58	Tunisia	Africa	2023	0.5	Y
59	Slovenia	Eastern Europe	1905	13.4	N
60	Kenya	Africa	1880	2.0	Y
61	Slovakia	Eastern Europe	1872	3.6	N
62	Peru	Latin America	1819	3.7	N
63	Qatar	Middle East	1802	77.3	N
64	Morocco	Africa	1762	0.3	Y
65	Jordan	Middle East	1738	33.9	N
66	Nepal	Asiatic Region	1624	1.7	Y
67	Ghana	Africa	1512	1.5	Y
68	Bulgaria	Eastern Europe	1457	2.7	N
69	Philippines	Asiatic Region	1386	0.2	Y
70	Uganda	Africa	1346	3.8	N
71	Ecuador	Latin America	1162	4.4	N
72	Cyprus	Western Europe	1153	15.8	N
73	Tanzania	Africa	1123	0.7	Y
74	Lithuania	Eastern Europe	1119	5.3	N
75	Cuba	Latin America	1059	0	Y

[1] We selected for the study the 22 countries whose researchers authored at least 1000 medical publications in 2020 and whose percentage of migrants was <2.5% in 2020.

Table 2. Number of researchers whose name origin, sorted by continent, country and ethnicity, was correctly classified by NamSor (N=89,906 researchers from twenty-two countries). Data are presented for the full sample and for various accuracy thresholds

try of ation rchers	Continent, number of data	Continent, number (%) of correctly classified names	Country, number of data	Country, number (%) of correctly classified names	Ethnicity, number of data	Ethnicity, number (%) of correctly classified names
a		Asia		China		Chinese
ull le	7702	7462 (96.9)	7702	5837 (75.8)	7702	6506 (84.5)
racy %	7516	7290 (97.0)	7516	5772 (76.8)	7550	6461 (85.6)
racy %	6047	5882 (97.3)	6047	4862 (80.4)	7421	6383 (86.0)
racy %	5369	5226 (97.3)	5369	4312 (80.3)	7237	6253 (86.4)
racy %	4554	4434 (97.4)	4554	3646 (80.1)	6874	5953 (86.6)
n		Asia		Japan		Japanese
ull le	6431	6165 (95.9)	6431	5451 (84.8)	6431	5430 (84.4)
racy %	6373	6132 (96.2)	6373	5443 (85.4)	6321	5417 (85.7)
racy %	6080	5926 (97.5)	6080	5374 (88.4)	6266	5412 (86.4)
racy %	5949	5829 (98.0)	5949	5320 (89.4)	6193	5390 (87.0)
racy %	5771	5675 (98.3)	5771	5223 (90.5)	6087	5350 (87.9)
		Asia		India		Indian
ull le	5362	4698 (87.6)	5362	3406 (63.5)	5362	4307 (80.3)

racy %	5106	4537 (88.9)	5106	3325 (65.1)	5070	4213 (83.1)
racy %	3371	3177 (94.3)	3371	2530 (75.1)	4885	4139 (84.7)
racy %	2652	2542 (95.9)	2652	2066 (77.9)	4638	3993 (86.1)
racy %	1916	1857 (96.9)	1916	1521 (79.4)	4291	3748 (87.4)
l[1]		Europe		Portugal		Portuguese
ll le	2829	2666 (94.2)	2829	1635 (57.8)	2829	1790 (63.3)
racy %	2724	2584 (94.9)	2724	1610 (59.1)	2617	1737 (66.4)
racy %	2098	2032 (96.9)	2098	1429 (68.1)	2480	1685 (67.9)
racy %	1811	1766 (97.5)	1811	1317 (72.7)	2281	1615 (70.8)
racy %	1520	1491 (98.1)	1520	1162 (76.5)	2093	1537 (73.4)
d		Europe		Poland		Polish
ll le	18441	18106 (98.2)	18441	16816 (91.2)	18441	16466 (89.3)
racy %	18168	17862 (98.3)	18168	16731 (92.1)	17744	16245 (91.6)
racy %	16613	16401 (98.7)	16613	15814 (95.2)	17287	16026 (92.7)
	15744	15564 (98.9)	15744	15136 (96.1)	16676	15675

ıracy ⁄o						(94.0)
ıracy ⁄o	14761	14619 (99.0)	14761	14313 (97.0)	15885	15057 (94.8)
t		Africa		Egypt		Egyptian
ıll ›le	9476	8840 (93.3)	9476	8615 (90.9)	9476	7677 (81.0)
ıracy ⁄o	9280	8726 (94.0)	9280	8541 (92.0)	8783	7448 (84.8)
ıracy ⁄o	8145	7928 (97.3)	8145	7889 (96.9)	8372	7282 (87.0)
ıracy ⁄o	7466	7346 (98.4)	7466	7325 (98.1)	7842	6986 (89.1)
ıracy ⁄o	6631	6560 (98.9)	6631	6553 (98.8)	7184	6587 (91.7)
co[2]		Europe		Spain		Hispanic
ıll ›le	7005	6320 (90.2)	7005	4930 (70.4)	7005	4878 (69.6)
ıracy ⁄o	6785	6155 (90.7)	6785	4887 (72.0)	6134	4507 (73.5)
ıracy ⁄o	5304	4923 (92.8)	5304	4267 (80.5)	5535	4173 (75.4)
ıracy ⁄o	4595	4293 (93.4)	4595	3854 (83.9)	4693	3640 (77.6)
ıracy ⁄o	3866	3630 (93.9)	3866	3345 (86.5)	3588	2804 (78.2)
stan		Asia		Pakistan		Pakistanis
ıll	6810	6674 (98.0)	6810	6388 (93.8)	6810	5882 (86.4)

le						
iracy %	6744	6626 (98.3)	6744	6367 (94.4)	6507	5787 (88.9)
iracy %	6202	6160 (99.3)	6202	6035 (97.3)	6327	5690 (89.9)
iracy %	5872	5851 (99.6)	5872	5777 (98.4)	6132	5587 (91.1)
iracy %	5404	5387 (99.7)	5404	5340 (98.8)	5852	5381 (92.0)
nesia		Asia		Indonesia		Indonesian
ull le	3828	3403 (88.9)	3828	2980 (77.9)	3828	2820 (73.7)
iracy %	3692	3339 (90.4)	3692	2935 (79.5)	3397	2717 (80.0)
iracy %	3017	2883 (95.6)	3017	2644 (87.6)	3178	2634 (82.9)
iracy %	2732	2654 (97.1)	2732	2489 (91.1)	2948	2536 (86.0)
iracy %	2451	2411 (98.4)	2451	2291 (93.5)	2679	2366 (88.3)
ria		Africa		Nigeria		Nigerian
ull le	3370	3104 (92.1)	3370	2553 (75.8)	3370	2547 (75.6)
iracy %	3265	3018 (92.4)	3265	2522 (77.2)	3044	2481 (81.5)
iracy %	2695	2579 (95.7)	2695	2352 (87.3)	2899	2427 (83.7)

racy %	2493	2400 (96.3)	2493	2258 (90.6)	2721	2352 (86.4)
racy %	2272	2214 (97.5)	2272	2129 (93.7)	2505	2241 (89.5)
		Asia		Iraq		Iraqi
ll le	1006	829 (82.4)	1006	270 (26.8)	1006	247 (24.6)
racy %	903	742 (82.2)	903	249 (27.6)	771	212 (27.5)
racy %	507	436 (86.0)	507	171 (33.7)	661	194 (29.4)
racy %	335	286 (85.4)	335	129 (38.5)	513	154 (30.0)
racy %	225	195 (86.7)	225	95 (42.2)	391	119 (30.4)
pia		Africa		Ethiopia		Ethiopian
ll le	4030	3861 (95.8)	4030	3671 (91.1)	4030	3451 (85.6)
racy %	3960	3808 (96.2)	3960	3653 (92.3)	3795	3387 (89.3)
racy %	3685	3606 (97.9)	3685	3556 (96.5)	3671	3335 (90.9)
racy %	3589	3546 (98.8)	3589	3516 (98.0)	3513	3242 (92.3)
racy %	3489	3466 (99.3)	3489	3448 (98.8)	3359	3130 (93.2)
ladesh		Asia		Bangladesh		Bangladeshi

ll le	2491	2420 (97.2)	2491	1955 (78.5)	2491	1805 (72.5)
racy %	2445	2383 (97.5)	2445	1941 (79.4)	2328	1765 (75.8)
racy %	2054	2033 (99.0)	2054	1784 (86.9)	2232	1726 (77.3)
racy %	1866	1855 (99.4)	1866	1697 (90.9)	2096	1656 (79.0)
racy %	1667	1662 (99.7)	1667	1565 (93.9)	1934	1576 (81.5)
am		Asia		Vietnam		Vietnamese
ll le	1960	1895 (96.7)	1960	1842 (94.0)	1960	1809 (92.3)
racy %	1956	1895 (96.9)	1956	1842 (94.2)	1943	1804 (92.9)
racy %	1924	1886 (98.0)	1924	1837 (95.5)	1923	1793 (93.2)
racy %	1905	1876 (98.5)	1905	1833 (96.2)	1896	1779 (93.8)
racy %	1889	1864 (98.7)	1889	1828 (96.8)	1855	1752 (94.5)
ia		Africa		Tunisia		Tunisian
ll le	1632	1589 (97.4)	1632	1224 (75.0)	1632	1072 (65.7)
racy %	1547	1512 (97.7)	1547	1195 (77.3)	1452	1018 (70.1)
racy %	1103	1091 (98.9)	1103	999 (90.6)	1351	975 (72.2)

ıracy ⁄o	912	908 (99.6)	912	853 (93.5)	1181	896 (75.9)
ıracy ⁄o	720	716 (99.4)	720	684 (95.0)	995	798 (80.2)
a		Africa		Kenya		Kenyan
ıll le	1187	972 (81.9)	1187	665 (56.0)	1187	591 (49.8)
ıracy ⁄o	1153	953 (82.7)	1153	657 (57.0)	959	561 (58.5)
ıracy ⁄o	835	713 (85.4)	835	582 (69.7)	878	545 (62.1)
ıracy ⁄o	732	646 (88.3)	732	545 (74.5)	793	516 (65.1)
ıracy ⁄o	629	572 (90.9)	629	489 (77.7)	707	489 (69.2)
cco		Africa		Morocco		Moroccan
ıll le	1545	1469 (95.1)	1545	1091 (70.6)	1545	809 (52.4)
ıracy ⁄o	1466	1396 (95.2)	1466	1048 (71.5)	1234	706 (57.2)
ıracy ⁄o	934	914 (97.9)	934	789 (84.5)	1087	641 (59.0)
ıracy ⁄o	752	743 (98.8)	752	660 (87.8)	893	544 (60.9)
ıracy ⁄o	570	563 (98.8)	570	507 (89.0)	686	429 (62.5)
l		Asia		Nepal		Nepalese

ull le	1327	1196 (90.1)	1327	406 (30.6)	1327	900 (67.8)
uracy %	1209	1102 (91.2)	1209	383 (31.7)	1239	854 (68.9)
uracy %	655	613 (93.6)	655	233 (35.6)	1168	813 (69.6)
uracy %	436	409 (93.8)	436	151 (34.6)	1054	734 (69.6)
uracy %	271	254 (93.7)	271	87 (32.1)	914	625 (68.4)
a		Africa		Ghana		Ghanaian
ull le	1383	1251 (90.5)	1383	1036 (74.9)	1383	947 (68.5)
uracy %	1349	1225 (90.8)	1349	1025 (76.0)	1205	909 (75.4)
uracy %	1098	1043 (95.0)	1098	945 (86.1)	1115	888 (79.6)
uracy %	1009	977 (96.8)	1009	905 (89.7)	1011	857 (84.8)
uracy %	917	893 (97.4)	917	839 (91.5)	941	824 (87.6)
ppines		Asia		Philippines		Hispanic
ull le	1113	141 (12.7)	1113	0	1113	421 (37.8)
uracy %	1018	129 (12.7)	1018	0	795	342 (43.0)
uracy %	510	78 (15.3)	510	0	655	304 (46.4)

racy %	357	64 (17.9)	357	0	495	261 (52.7)
racy %	226	45 (19.9)	226	0	366	196 (53.6)
ania		Africa		Tanzania		Tanzanian
ll le	673	544 (80.8)	673	293 (43.5)	673	291 (43.2)
racy %	617	503 (81.5)	617	272 (44.1)	529	257 (48.6)
racy %	387	318 (82.2)	387	212 (54.8)	467	238 (51.0)
racy %	307	248 (80.8)	307	177 (57.7)	402	222 (55.2)
racy %	211	170 (80.6)	211	121 (57.4)	342	192 (56.1)
2		Europe		Spain		Hispanic
ll le	305	296 (97.1)	305	261 (85.6)	305	243 (79.7)
racy %	296	288 (97.3)	296	256 (86.5)	279	226 (81.0)
racy %	237	234 (98.7)	237	225 (94.9)	258	216 (83.7)
racy %	220	218 (99.1)	220	212 (96.4)	219	191 (87.2)
racy %	188	187 (99.5)	188	184 (97.9)	177	157 (88.7)

[1] The table shows the number of names correctly classified for this country, after replacing for the variable "continent" the category "Europe" by the category "Europe, America or Oceania", and for the variable "country" the category "Portugal" by the category "Portugal or Brazil".

[2] The table shows the number of names correctly classified for this country, after replacing for the variable "continent" the category "Europe" by the category "Europe, America or Oceania", and for the variable "country" the category "Spain" by the category "Spain or Hispanic American country".

Table 3. Confusion matrices for the origin of the names of 89,906 researchers using various accuracy threshold values

ıble	Number (%) of correctly classified names	Number (%) of misclassified names	Number (%) of unclassified names
sample (i.e., no hold value ted)			
ntinent of origin , Africa or Europe)	74619 (83.0)	15287 (17.0)	0
ntinent#2[1]	83901 (93.3)	6005 (6.7)	0
untry of origin	64499 (71.7)	25407 (28.3)	0
untry#2[2]	71325 (79.3)	18581 (20.7)	0
nicity	70889 (78.9)	19017 (21.1)	0
racy of the inference %			
ntinent of origin , Africa or Europe)	73178 (81.4)	14394 (16.0)	2334 (2.6)
ontinent#2[1]	82205 (91.4)	5367 (6.0)	2334 (2.6)
untry of origin	63901 (71.1)	23671 (26.3)	2334 (2.6)
untry#2[2]	70654 (78.6)	16918 (18.8)	2334 (2.6)
thnicity	69054 (76.8)	14642 (16.3)	6210 (6.9)
racy of the inference %			
ntinent of origin , Africa or Europe)	63667 (70.8)	9834 (10.9)	16405 (18.3)
ntinent#2[1]	70856 (78.8)	2645 (2.9)	16405 (18.3)
untry of origin	58608 (65.2)	14893 (16.5)	16405 (18.3)
untry#2[2]	64529 (71.7)	8972 (10.0)	16405 (18.3)
nicity	67519 (75.1)	12597 (14.0)	9790 (10.9)
racy of the inference %			
ntinent of origin	58970 (65.5)	8133 (9.1)	22803 (25.4)

, Africa or Europe)			
ntinent#2[1]	65247 (72.6)	1856 (2.0)	22803 (25.4)
untry of origin	55149 (61.3)	11954 (13.3)	22803 (25.4)
untry#2[2]	60532 (67.3)	6571 (7.3)	22803 (25.4)
nicity	65079 (72.4)	10348 (11.5)	14479 (16.1)
racy of the inference %			
ntinent of origin , Africa or Europe)	53557 (59.6)	6591 (7.3)	29758 (33.1)
ntinent#2[1]	58865 (65.5)	1283 (1.4)	29758 (33.1)
untry of origin	50679 (56.4)	9469 (10.5)	29758 (33.1)
untry#2[2]	55370 (61.6)	4778 (5.3)	29758 (33.1)
nicity	61311 (68.2)	8394 (9.3)	20201 (22.5)

[1] “Europe” replaced by “Europe, America or Oceania”

[2] "Spain" replaced by “Spain or Hispanic American country” and "Portugal" replaced by "Portugal or Brazil”

Table 4. Performance metrics (i.e., errorCoded, errorCodedWithoutNA and naCoded) for the origin of the names of 89,906 researchers using various accuracy threshold values

riable	errorCoded[1]	errorCodedWithoutNA[2]	naCoded[3]
ll sample (i.e., no threshold value ected)			
Continent of origin (Asia, Africa or rope)	0.1700	0.1700	0
Continent#2[4]	0.0668	0.0668	0
Country of origin	0.2826	0.2826	0
Country#2[5]	0.2067	0.2067	0
Ethnicity	0.2115	0.2115	0
curacy of the inference ≥40%			
Continent of origin (Asia, Africa or rope)	0.1861	0.1644	0.0260
Continent#2[4]	0.0857	0.0613	0.0260
Country of origin	0.2893	0.2703	0.0260
Country#2[5]	0.2141	0.1932	0.0260
Ethnicity	0.2319	0.1749	0.0691
curacy of the inference ≥50%			
Continent of origin (Asia, Africa or rope)	0.2919	0.1338	0.1825
Continent#2[4]	0.2119	0.0360	0.1825
Country of origin	0.3481	0.2026	0.1825
Country#2[5]	0.2823	0.1221	0.1825
Ethnicity	0.2490	0.1572	0.1089
curacy of the inference ≥60%			
Continent of origin (Asia, Africa or rope)	0.3441	0.1212	0.2536
Continent#2[4]	0.2743	0.0277	0.2536
Country of origin	0.3866	0.1781	0.2536

Country#2[5]	0.3267	0.0979	0.2536
Ethnicity	0.2761	0.1372	0.1611
curacy of the inference ≥70%			
Continent of origin (Asia, Africa or rope)	0.4043	0.1096	0.3310
Continent#2[4]	0.3453	0.0213	0.3310
Country of origin	0.4363	0.1574	0.3310
Country#2[5]	0.3841	0.0794	0.3310
Ethnicity	0.3181	0.1204	0.2247

[1] errorCoded = proportion of misclassifications (i.e., wrong continent, country or ethnicity assigned to a name) and non-classifications (i.e., no continent, country or ethnicity assigned)

[2] errorCodedWithoutNA = proportion of misclassifications excluding non-classifications

[3] naCoded = proportion of non-classifications

[4] “Europe” replaced by “Europe, America or Oceania”

[5] "Spain" replaced by “Spain or Hispanic American country” and "Portugal" replaced by "Portugal or Brazil”