Can first or last name uniqueness help to identify diaspora researchers from any country?

Mike Thelwall†
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton, UK

Abstract

Purpose: Diaspora researchers work in one country but have ancestral origins in another, either through moves during a research career (mobile diaspora researchers) or by starting research in the target country (embedded diaspora researchers). Whilst mobile researchers might be tracked through affiliation changes in bibliometric databases, embedded researchers cannot. This article reports an evidence-based discussion of which countries’ diaspora researchers can be partially tracked using first or last names, addressing this limitation.

Design/methodology/approach: A frequency analysis of first and last names of authors of all Scopus journal articles 2001-2021 for 200 countries or regions.

Findings: There are great variations in the extent to which first or last names are uniquely national, from Monserrat (no unique first names) to Thailand (81% unique last names). Nevertheless, most countries have a subset of first or last names that are relatively unique. For the 50 countries with the most researchers, authors with relatively national names are always more likely to research their name-associated country, suggesting a continued national association. Lists of researchers’ first and last name frequencies and proportions are provided for 200 countries/regions.

Research limitations: Only one period is tracked (2001-2021) and no attempt was made to validate the ancestral origins of any researcher.

Practical implications: Simple name heuristics can be used to identify the international spread of a sample of most countries’ diaspora researchers, but some manual checks of individual names are needed to weed out false matches. This can supplement mobile researcher data from bibliometric databases.

Originality/value: This is the first attempt to list name associations for the authors of all countries and large regions, and to identify the countries for which diaspora researchers could be tracked by name.

† Corresponding author: Mike Thelwall (Email: m.thelwall@wlv.ac.uk; ORCID: 0000-0001-6065-205X)
Keywords: First name; Given name; Last name; Family name; Author nationality; Researcher mobility; Diaspora researchers.

1 Introduction

Every country has diaspora communities in the sense of people with its historical roots that live elsewhere, including as first, second or subsequent generation migrants. These are recognized as important for economics, trade, and tourism (Gelb, et al., 2021; Li et al., 2020; Liu et al., 2020). For example, India, China, and Mexico receive substantial international personal remittances, presumably from nationals living overseas or subsequent generations still connected to local families (McAuliffe & Triandafyllidou, 2021). In academia, diaspora communities can include first generation migrants that move overseas before, during and after education, as well as subsequent generations. People that change country after starting research are often studied as an aspect of “academic mobility”, but other diaspora researchers may share some of their advantages, such as language skills, cultural awareness, and personal connections to their ancestral land. Identifying this community can be helpful to monitor their spread and locations as well as for specific outreach activities, such as seeking culturally aware overseas expertise to help with local tasks.

Academic mobility, in the sense of institutional moves after the start of a research career, is often encouraged by research organizations and funders (e.g., Baumert et al., 2022) to support collaboration, cross-fertilization of ideas and to discourage parochialism (Flanagan, 2015; Siekierski et al., 2018; Teichler, 2015), or to address local skills shortages (Tremblay, 2005). From a national perspective, it can also be useful to assess the value of diaspora researchers to their home country and to international scientific knowledge (Bassecoulard et al., 2003; Marmolejo-Leyva et al., 2015; Xie & Freeman, 2020), to discover why researchers emigrate (Kasa, 2022) or to investigate whether scientific immigration is beneficial (Mendoza et al., 2020). For many of these purposes, it may also be helpful to identify all diaspora researchers, including those that moved before starting research or who are second or subsequent generation immigrants. These are called “embedded” diaspora researchers here, in contrast to “mobile” diaspora researchers that moved after starting research (equated with publishing here).

There have been periodic attempts to assess researcher mobility through surveys, such as through the EU’s MORE 1 to MORE 4 (www.more-4.eu; Bojica et al., 2022; Boring et al., 2015) and samples of foreign-born scientists in selected countries (e.g., Franzoni et al., 2012). These are all limited in scope and scope and suffer from sample selection bias and non-response bias. For example, MORE 4 (https://www.more-4.eu/surveys) extracted academic email addresses from university websites in
Europe and selected other countries for its survey. On a much larger scale and without the problem of non-response bias, researchers can be tracked over time from their publications, identifying those that move from one country to another (e.g., Sanliturk et al., 2023). In particular, author IDs, including ORCID profiles, can be used to track researcher national affiliation changes (Zhao et al., 2020), but this only identifies internationally mobile researchers and excludes all embedded diaspora researchers, such as people that move abroad for their Ph.D. (e.g., Schäfer & El Dali, 2021).

A third strategy is to identify the national origin or ethnicity of researchers from their first and last names. Unlike author ID/affiliation analysis, name analysis may be able to identify embedded diaspora researchers alongside their internationally mobile counterparts. Unlike a survey, it is relatively quick, simple, and scalable to the world. It has several disadvantages, however, including misidentification through marital name changes and an absence of information about whether the researcher emigrated or one of their ancestors did. If this approach works, it could be used to identify embedded diaspora researchers to complement lists of mobile diaspora researchers that can be identified through affiliation changes in bibliometric databases. Name-based heuristics have already been used for individual countries, including for researchers with Russian heritage (Karaulova et al., 2019), and for internal migrants within India (Lewison, & Kundra, 2008). Names have also been used to classify researchers into seven geographic regions (Roe et al., 2014). One preprint has attempted to exploit bibliometric affiliations to create lists of last names associated with countries. It compared the lists of names generated against 5 countries and 7 country groups for which there were existing lists of last names, finding high precision for both methods (Robinson-Garcia et al., 2015). This paper did not analyze first names or identify the countries for which the approach would be most successful, however.

The most comprehensive attempt so far to identify all (embedded and mobile) researchers associated with a country from their names used extensive manual classifications of first and last names to identify publishing researchers with Greek heritage (Ioannidis et al., 2021). This amount of work may be impractical for other countries, and it is not clear whether Greece has uniquely distinctive names, such as due to their conversion from the Greek alphabet. Name-based heuristics are also routinely used to infer ethnicity within single countries for health and population studies, but these typically focus on a limited set of ethnicities and not national origins (Shah et al., 2010). One much larger experiment used Wikipedia national origin information to classify broad national groups (e.g., Hispanic, East Asian) through names with a high degree of accuracy for many groups, but this did not
focus on individual countries (Ye et al., 2017).

The attempted comprehensive manual strategy used for Greece (Ioannidis et al., 2021) relies on first or last names, as recorded in Scopus, often having recognizably Greek origins. This strategy could also be applied to other countries with relatively distinctive names to identify their associated researchers. A less labor-intensive variant would be to focus on the most distinctive names for a country, if any, and use them to identify a sample of researchers to track internationally through their names. For example, if 40% of Malian first or last names were unique to Mali, then the spread of Malian (mobile and embedded) academics internationally could be estimated by tracking the 40% with uniquely Malian names. This article assesses the extent to which either approach is possible for 200 countries or regions by investigating how often first or last names are distinctively national and discussing social and historical factors that influence the extent to which name matching can be used to track the national origins of researchers. The following research questions underpin this goal. For convenience here, the term “country” is sometimes used inclusively of regions that are recorded separately in the Scopus country field (e.g., Hong Kong, China).

- RQ1: How nationally distinctive are researcher first and last names in the countries/regions recorded in Scopus?
- RQ2: Are researcher first or last names most unique for countries/regions?
- RQ3: Which factors affect the uniqueness of researcher first and last names in a country/region? This is addressed based on a review rather than new empirical evidence but is included for completeness.
- RQ4: Do researcher first and last names that occur disproportionately often in a country/region tend to be associated with that country, even for researchers in other countries?
- RQ5: From which countries/regions is it easiest to track diaspora researchers through first or last names?

2 Methods

The research design was to extract all first and last names of researchers from Scopus for a recent period and then to compare the frequency of the names between countries/regions.

2.1 Data

The data used was a copy of all journal articles in Scopus with a publication date from 2001 to 2021 as downloaded in January-February 2022 using the Scopus
Applications Programming Interface (API). The 21 years of the current century 2001-2021 at the time of data collection were chosen to give recent data with a large enough period for robust name frequency information. Scopus was used in preference to the Web of Science for its greater coverage. Dimensions probably has even greater coverage but is less controlled in content. Google Scholar has far greater coverage but no API (Martín-Martín et al., 2021). The Scopus API was used to download records for all journal articles for 2001 to 2021 in January-February 2022.

Author first and last names were extracted from the downloaded Scopus journal article records, when present, as well as Scopus author identifiers and the first country affiliation listed, assuming this to be the main one. All names were in the Latin alphabet, so non-Latin names (e.g., Arabic, Chinese, Greek, Thai) were recorded in Latin versions. No attempt was made to correct any name errors and it is likely that given/first and family/last names were occasionally reversed in countries, such as China and Malaysia, where it is common to write the family name first. Only the first part of each last name was extracted in cases where there was a last name containing a space, unless it was one of a few common prepositions (including nobiliary particles): al, de, del, von, von/van der, von/van den, de la. The initial name list was found to be incomplete for The Netherlands and Belgium, so the data for these two countries was recalculate with additional prepositions (op de, van ‘t, ten). Last names in some countries, including Spain are often doubled (and not hyphenated), but only the first part of any last name was used because it seems likely that researchers moving to countries where this is unusual would often shorten their last names. Article titles, abstracts, and keywords were also recorded from the API at the same time (for RQ4).

2.2 Name frequency calculations

Unique authors were identified by Scopus author ID, ignoring authors without an ID. Each author’s national affiliation was recorded from the first journal article in the data set. This was repeated for last names, and again for both considering only first authors. For the year range chosen, there were 249 country or region affiliations in Scopus, but data was gathered only for the 200 with the most researchers due to software array size limitations. The largest omitted territory was the British Overseas Territory of Gibraltar (167 journal articles) and the smallest was Saint Martin (French Part) (4 journal articles).

Four datasets were produced, each of which was a frequency table with a list of 200 countries/regions (as recorded in Scopus) as columns and a large set of rows, one for each name (first or last), recording the number of authors with that name in each country/region. Two tables analyzed first authors only (for RQ4) and two tables...
analyzed all authors (for all other RQs). From these tables, the proportion of each name occurring in each country/region could be calculated. The results report this information in various different ways, as described in the figure and table captions.

3 Results

3.1 RQ1: How nationally distinctive are researchers’ first and last names by country?

The results refer only to the declared affiliations of authors of Scopus-indexed journal articles 2001-2021. There are huge variations in the extent to which the first and last names of researchers in a country/region are never found in other countries/regions (Table 1, Table 2: columns 3 and 7) during the period 2001-2021.

Table 1. First and last name national uniqueness statistics for the 25 countries or regions with the highest average percentage of last name in the country (2001-2021). A “national” first or last name is used only by researchers in one country. A complete list of countries/regions is in the supplementary materials (https://doi.org/10.6084/m9.figshare.21954467).

<table>
<thead>
<tr>
<th>Country/region</th>
<th>Authors with a first name</th>
<th>% of national first names</th>
<th>Average % of first name in country</th>
<th>First name rank</th>
<th>Authors with a last name</th>
<th>% of national last names</th>
<th>Average % of last name in country</th>
<th>Last name rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thailand</td>
<td>85,162</td>
<td>42%</td>
<td>81%</td>
<td>3</td>
<td>96,790</td>
<td>81%</td>
<td>88%</td>
<td>1</td>
</tr>
<tr>
<td>Japan</td>
<td>1,059,927</td>
<td>5%</td>
<td>87%</td>
<td>1</td>
<td>1,217,638</td>
<td>8%</td>
<td>87%</td>
<td>2</td>
</tr>
<tr>
<td>Lithuania</td>
<td>15,245</td>
<td>8%</td>
<td>55%</td>
<td>15</td>
<td>19,240</td>
<td>73%</td>
<td>86%</td>
<td>3</td>
</tr>
<tr>
<td>China</td>
<td>4,488,356</td>
<td>3%</td>
<td>85%</td>
<td>2</td>
<td>4,988,411</td>
<td>1%</td>
<td>83%</td>
<td>4</td>
</tr>
<tr>
<td>Turkey</td>
<td>217,560</td>
<td>6%</td>
<td>75%</td>
<td>4</td>
<td>255,292</td>
<td>26%</td>
<td>80%</td>
<td>5</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>137,269</td>
<td>4%</td>
<td>41%</td>
<td>31</td>
<td>452,961</td>
<td>29%</td>
<td>77%</td>
<td>6</td>
</tr>
<tr>
<td>Iran</td>
<td>277,009</td>
<td>4%</td>
<td>66%</td>
<td>7</td>
<td>350,691</td>
<td>23%</td>
<td>76%</td>
<td>7</td>
</tr>
<tr>
<td>Laos</td>
<td>1,179</td>
<td>60%</td>
<td>72%</td>
<td>5</td>
<td>1,370</td>
<td>65%</td>
<td>75%</td>
<td>8</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>18,780</td>
<td>24%</td>
<td>59%</td>
<td>14</td>
<td>27,596</td>
<td>57%</td>
<td>74%</td>
<td>9</td>
</tr>
<tr>
<td>Indonesia</td>
<td>98,259</td>
<td>20%</td>
<td>53%</td>
<td>17</td>
<td>121,959</td>
<td>35%</td>
<td>72%</td>
<td>10</td>
</tr>
<tr>
<td>Greece</td>
<td>71,132</td>
<td>5%</td>
<td>50%</td>
<td>19</td>
<td>101,776</td>
<td>38%</td>
<td>72%</td>
<td>11</td>
</tr>
<tr>
<td>Madagascar</td>
<td>2,084</td>
<td>25%</td>
<td>37%</td>
<td>38</td>
<td>2,950</td>
<td>48%</td>
<td>71%</td>
<td>12</td>
</tr>
<tr>
<td>Poland</td>
<td>163,794</td>
<td>3%</td>
<td>63%</td>
<td>10</td>
<td>199,191</td>
<td>33%</td>
<td>71%</td>
<td>13</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>61,820</td>
<td>3%</td>
<td>39%</td>
<td>34</td>
<td>89,460</td>
<td>43%</td>
<td>70%</td>
<td>14</td>
</tr>
<tr>
<td>India</td>
<td>629,418</td>
<td>10%</td>
<td>61%</td>
<td>12</td>
<td>982,934</td>
<td>12%</td>
<td>69%</td>
<td>15</td>
</tr>
<tr>
<td>Georgia</td>
<td>3,615</td>
<td>5%</td>
<td>31%</td>
<td>56</td>
<td>6,628</td>
<td>36%</td>
<td>69%</td>
<td>16</td>
</tr>
<tr>
<td>Latvia</td>
<td>5,775</td>
<td>9%</td>
<td>40%</td>
<td>32</td>
<td>8,441</td>
<td>54%</td>
<td>69%</td>
<td>17</td>
</tr>
<tr>
<td>Finland</td>
<td>65,723</td>
<td>3%</td>
<td>47%</td>
<td>21</td>
<td>81,530</td>
<td>26%</td>
<td>67%</td>
<td>18</td>
</tr>
<tr>
<td>Hungary</td>
<td>47,033</td>
<td>8%</td>
<td>62%</td>
<td>11</td>
<td>59,104</td>
<td>24%</td>
<td>65%</td>
<td>19</td>
</tr>
<tr>
<td>Slovakia</td>
<td>24,042</td>
<td>3%</td>
<td>24%</td>
<td>81</td>
<td>33,917</td>
<td>46%</td>
<td>65%</td>
<td>20</td>
</tr>
<tr>
<td>Mongolia</td>
<td>2,335</td>
<td>38%</td>
<td>65%</td>
<td>9</td>
<td>3,101</td>
<td>37%</td>
<td>65%</td>
<td>21</td>
</tr>
<tr>
<td>Uganda</td>
<td>8,543</td>
<td>9%</td>
<td>14%</td>
<td>120</td>
<td>10,160</td>
<td>37%</td>
<td>64%</td>
<td>22</td>
</tr>
<tr>
<td>Romania</td>
<td>51,161</td>
<td>6%</td>
<td>37%</td>
<td>41</td>
<td>63,560</td>
<td>28%</td>
<td>64%</td>
<td>23</td>
</tr>
<tr>
<td>Nigeria</td>
<td>47,983</td>
<td>15%</td>
<td>44%</td>
<td>26</td>
<td>78,489</td>
<td>25%</td>
<td>63%</td>
<td>24</td>
</tr>
<tr>
<td>South Korea</td>
<td>514,026</td>
<td>5%</td>
<td>69%</td>
<td>6</td>
<td>556,771</td>
<td>1%</td>
<td>63%</td>
<td>25</td>
</tr>
</tbody>
</table>
Can first or last name uniqueness help to identify diaspora researchers from any country?

Table 2. First and last name national uniqueness statistics for the 25 countries or regions with the lowest average percentage of last name in the country/region. A “national” first or last name is used only by researchers in one country.

<table>
<thead>
<tr>
<th>Country/region</th>
<th>Authors with a first name</th>
<th>% of national first names</th>
<th>Average % of first name in country</th>
<th>First name rank</th>
<th>Authors with a last name</th>
<th>% of national last names</th>
<th>Average % of last name in country</th>
<th>Last name rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virgin Islands (UK)</td>
<td>26</td>
<td>0%</td>
<td>2%</td>
<td>191</td>
<td>36</td>
<td>8%</td>
<td>11%</td>
<td>176</td>
</tr>
<tr>
<td>Paraguay</td>
<td>1,683</td>
<td>2%</td>
<td>3%</td>
<td>186</td>
<td>1,990</td>
<td>5%</td>
<td>10%</td>
<td>177</td>
</tr>
<tr>
<td>El Salvador</td>
<td>814</td>
<td>1%</td>
<td>2%</td>
<td>192</td>
<td>947</td>
<td>5%</td>
<td>10%</td>
<td>178</td>
</tr>
<tr>
<td>Andorra</td>
<td>106</td>
<td>3%</td>
<td>6%</td>
<td>168</td>
<td>130</td>
<td>4%</td>
<td>10%</td>
<td>179</td>
</tr>
<tr>
<td>Barbados</td>
<td>357</td>
<td>3%</td>
<td>6%</td>
<td>170</td>
<td>436</td>
<td>6%</td>
<td>10%</td>
<td>180</td>
</tr>
<tr>
<td>New Zealand</td>
<td>50,477</td>
<td>2%</td>
<td>5%</td>
<td>175</td>
<td>63,946</td>
<td>4%</td>
<td>10%</td>
<td>181</td>
</tr>
<tr>
<td>Guatemala</td>
<td>1,968</td>
<td>2%</td>
<td>4%</td>
<td>182</td>
<td>2,199</td>
<td>6%</td>
<td>10%</td>
<td>182</td>
</tr>
<tr>
<td>Hong Kong, China</td>
<td>53,289</td>
<td>1%</td>
<td>8%</td>
<td>157</td>
<td>66,369</td>
<td>1%</td>
<td>10%</td>
<td>183</td>
</tr>
<tr>
<td>Montserrat</td>
<td>20</td>
<td>0%</td>
<td>5%</td>
<td>176</td>
<td>39</td>
<td>5%</td>
<td>10%</td>
<td>184</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>6,122</td>
<td>2%</td>
<td>5%</td>
<td>178</td>
<td>6,996</td>
<td>2%</td>
<td>9%</td>
<td>185</td>
</tr>
<tr>
<td>Falkland Islands</td>
<td>69</td>
<td>0%</td>
<td>0%</td>
<td>199</td>
<td>88</td>
<td>6%</td>
<td>9%</td>
<td>186</td>
</tr>
<tr>
<td>Grenada</td>
<td>551</td>
<td>6%</td>
<td>8%</td>
<td>152</td>
<td>640</td>
<td>4%</td>
<td>9%</td>
<td>187</td>
</tr>
<tr>
<td>Panama</td>
<td>1,864</td>
<td>6%</td>
<td>9%</td>
<td>146</td>
<td>2,135</td>
<td>5%</td>
<td>9%</td>
<td>188</td>
</tr>
<tr>
<td>Virgin Islands (USA)</td>
<td>157</td>
<td>6%</td>
<td>8%</td>
<td>149</td>
<td>202</td>
<td>3%</td>
<td>9%</td>
<td>189</td>
</tr>
<tr>
<td>Bahamas</td>
<td>196</td>
<td>5%</td>
<td>7%</td>
<td>159</td>
<td>237</td>
<td>5%</td>
<td>8%</td>
<td>190</td>
</tr>
<tr>
<td>Cayman Islands</td>
<td>96</td>
<td>0%</td>
<td>1%</td>
<td>197</td>
<td>120</td>
<td>3%</td>
<td>8%</td>
<td>191</td>
</tr>
<tr>
<td>Jamaica</td>
<td>2,217</td>
<td>9%</td>
<td>13%</td>
<td>122</td>
<td>3,113</td>
<td>4%</td>
<td>8%</td>
<td>192</td>
</tr>
<tr>
<td>Bermuda</td>
<td>144</td>
<td>1%</td>
<td>2%</td>
<td>194</td>
<td>187</td>
<td>4%</td>
<td>8%</td>
<td>193</td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>8,310</td>
<td>6%</td>
<td>10%</td>
<td>139</td>
<td>9,347</td>
<td>2%</td>
<td>8%</td>
<td>194</td>
</tr>
<tr>
<td>Dominican Republic</td>
<td>1,198</td>
<td>6%</td>
<td>9%</td>
<td>141</td>
<td>1,355</td>
<td>3%</td>
<td>7%</td>
<td>195</td>
</tr>
<tr>
<td>Honduras</td>
<td>1,175</td>
<td>3%</td>
<td>5%</td>
<td>173</td>
<td>1,345</td>
<td>3%</td>
<td>6%</td>
<td>196</td>
</tr>
<tr>
<td>Nicaragua</td>
<td>961</td>
<td>3%</td>
<td>5%</td>
<td>174</td>
<td>1,133</td>
<td>3%</td>
<td>6%</td>
<td>197</td>
</tr>
<tr>
<td>Cape Verde</td>
<td>247</td>
<td>8%</td>
<td>11%</td>
<td>133</td>
<td>279</td>
<td>3%</td>
<td>6%</td>
<td>198</td>
</tr>
<tr>
<td>Dominica</td>
<td>82</td>
<td>6%</td>
<td>10%</td>
<td>136</td>
<td>101</td>
<td>2%</td>
<td>5%</td>
<td>199</td>
</tr>
<tr>
<td>Macao, China</td>
<td>3,933</td>
<td>2%</td>
<td>3%</td>
<td>184</td>
<td>4,352</td>
<td>1%</td>
<td>2%</td>
<td>200</td>
</tr>
</tbody>
</table>

In terms of (purely) “national” names (defined here to be names only used by the researchers of one country), whilst 81% of Thai-affiliated authors of Scopus journal articles 2001-2021 have (purely) national last names, the same is true for only 1% of authors in Macao, China. First names have smaller differences: whilst 60% of authors from Laos have national first names, none from the Cayman Islands do. Overall, however, it is uncommon for a last name to be national, and very uncommon for a first name to be national.

Some names may occur predominantly in one country, even if they occasionally occur in other countries due to migration or other reasons. These might be termed partly national names. The “percentage national” or “percentage [country]” of a name is defined here to be the percentage of authors with that name who are affiliated with a given country (not reported in the tables). Thus, Babatunde is 57% national...
for Nigeria (or 57% Nigerian) since 150 of the 263 Babatundes in Scopus have a Nigerian affiliation.

There are substantial international differences in the extent to which first and last names tend to occur predominantly in one country/region (Table 1, Table 2: columns 4 and 8). On average, the last name of a researcher in Thailand will be used 88% of the time in Thailand, but the last name of a researcher in Macao, China will be used only 2% of the time in Macao, China (i.e., 98% of researchers sharing a last name with a Macao researcher are outside Macao, China). Similarly, the first name of a researcher in Japan will be used 87% of the time in Japan, but the first name of a researcher in the Falkland Islands (Malvinas) will be used only 0% (to 0 decimal places) in the Falkland Islands (i.e., virtually all researchers sharing a first name with a Falkland Islands researcher are outside of the Falkland Islands).

3.2 RQ2: Are first or last names the most unique for countries?

At the country level, there is a trend for countries that have a higher percentage of (purely) national first names also to have a larger percentage of (purely) national last names. There are exceptions, however, and the relationship is not close to a straight line (Figure 1). For example, a high proportion of researchers’ last names in Lithuania are national whereas few first names are. In most cases, last names are more likely to be national than first names, but a few countries (e.g., Cambodia) break this trend. On average (averaging across countries rather than names), 9% of first names are national and 19% of last names are national.

![Figure 1. The percentage of national last names against the percentage of national first names by country (all authors of Scopus journal articles 2002-2021).](image)
There is a stronger trend for countries with a higher average first name percentage in the country to also have a higher average last name percentage in that country (Figure 2) and the pattern is closer to a straight line. Again, there are outliers (e.g., last names are far more national than first names in Uganda) and last names are generally more national than first names. On average (averaging across countries rather than names), the first name percentage national is 23% and for last names it is 38%.

Figure 2. Last name average percentage in country/region against first name average percentage in country/region (all authors of Scopus journal articles 2002-2021).

3.3 RQ3: Which factors affect the uniqueness of researcher first and last names in a country?

This section reviews factors that may lead to international differences in first and last names, extending a previous list (Mateos, 2007). Tables 1 and 2 suggest a size effect on name distinctiveness and there is indeed a trend for countries/regions with greater numbers of authors to have a higher percentage of national first and last name shares (Figure 3, 4), although the pattern is not linear. The trend is a natural side effect of size. If names were distributed randomly around the world, then larger countries/regions would naturally have higher shares of each name so Figures 3 and 4 would be monotonically increasing lines. The substantial deviations from monotonically increasing lines together with the differing distributions shown above confirm that other factors are at work. These include the following.
Figure 3. Average percentage of first names in a country against the number of authors for 200 countries/regions in Scopus (the x axis scale is logarithmic).

Figure 4. Average percentage of last names in a country/region against the number of authors for 200 countries/regions in Scopus (the x axis scale is logarithmic).
Can first or last name uniqueness help to identify diaspora researchers from any country?

• **Historical territory changes**: Overlapping or similar cultures in different countries/regions can be the result of historical border changes, leaving a legacy of partly shared names and languages. This is the case between Belgium and the Netherlands, for example (partly shared versions of Dutch/Flemish; e.g., Piet is relatively common in both countries) and perhaps also for constituent countries of the Austro-Hungarian Empire, and the Middle East (where most countries are recent). These would tend to decrease the national uniqueness of names.

• **Immigration**: Immigration into a country (whether for academic jobs or for other reasons) would tend to increase the proportion of names that had a low national percentage by introducing names that had been previously rare or absent (e.g., Priya from India to the UK). This may apply less to first names if the immigrants choose to be known by host country’s first names (Edwards, 2006).

• **Emigration and colonization**: Emigration from a country would tend to reduce the proportion of national names and lower the national percentage of other names by spreading a sample of those names to other countries (e.g., Priya from India to the UK and USA). Systematic mass emigration and colonization (e.g., from Europe to the Americas) would have this effect on a large scale. Historic colonization is presumably the main cause of the low proportion of national UK names.

• **International religions**: International religions associated with naming conventions would tend to reduce the uniqueness of names in countries with substantial numbers of followers. In the UK, for example, indigenous first names have been partly replaced by biblical names (e.g., John), and Islamic names (e.g., Mohammed) are common in many countries.

• **Recent national language or cultural shifts**: The increasing adoption of previously minority languages within a country can increase the prevalence of relatively unique names. For example, emphasizing Welsh in Wales, Hebrew in Israel, or Irish in Ireland may increase the update of traditional first names associated with those languages. This would presumably increase the prevalence of national names within the host country. In the USA, the widespread coining of new African-American first names since the 1960s as part of the Black Power movement (Fryer & Levitt, 2004) has also presumably increased the average uniqueness of US first names.

• **Cultural openness to first name neologisms**: Countries containing cultures in which it is acceptable to invent first names would have a greater number of unique first names. This seems to apply to the USA (see below).
• **Cultural openness to name changes**: Name changes in one country may serve to make its names unique. Immigrants to the USA sometimes Americanised their last names, for example, and in the UK people with Germanic names may have changed them around the time of the last two world wars for patriotism or self-preservation.

• **Mass media cultural reach**: Awareness of first names may spread through various media, leading to the adoption of pop star names internationally (e.g., Kylie started to increase in popularity as an academic first name in the UK from about 2010, which is possibly connected to singer/actor Kylie Minogue, whose first number 1 UK single was in 1987).

• **National laws**: Iceland has laws that restrict the names that can be given to children to those approved for the Personal Names Register (https://island.is/mannanofn), increasing homogeneity.

### 3.4 RQ4: Do first and last names that occur disproportionately often in a country tend to be associated with that country, even for researchers in other countries?

To test the assumption that names occurring more frequently in a country were more likely to be from people with a national origin in that country, additional datasets were constructed as above, but specifying that the country name also had to occur in the title, abstract or keywords. For example, the Nigeria sets summarised the national frequency of first (or last) names of authors of articles mentioning Nigeria. An author was recorded when any of their articles mentioned the country name. Assuming that (a) a name that occurs in a higher proportion in one country is more likely to be from someone with a national origin in that country and (b) people with a national origin in a country are more likely to research that country, then the relationship between the proportion of authors with a name from a country and the proportion of authors with that name researching the country should be linear with a positive slope, even considering only researchers not based in the country.

Consistent with the above assumptions, and therefore partially validating them, there were positive Pearson correlations between national percentage and percentage of non-national authors mentioning the country for the 50 countries with the most authors. The average Pearson correlation was 0.50 for first names (min: 0.14, max: 0.86) and 0.52 for last names (min: 0.09, max: 0.89). Broadly linear shapes can also be seen in the graphs (e.g., Figure 5, others available in the online supplement: https://doi.org/10.6084/m9.figshare.21956795). For example, 0% of the world’s first authors called Rajesh are in Greece and authors outside Greece with a first name similarly rare in Greece have a 0.1% chance of mentioning Greece in any of their
articles (abstract, keywords or titles). In contrast, 82% of the world’s authors called Vassiliki have a Greek affiliation but the Vassilikis outside Greece (and the 5 other first names that are 81% in Greece) have an 8.4% chance of mentioning Greece in any of their Scopus articles 2001-2021 (abstract, keywords or titles). This is consistent with the Vassilikis outside Greece tending to have a far above average interest in Greece, such as through a family connection or move. The figure of 8.4% is not low given that the vast majority of research does not mention a country because it is general, theoretical, or mentions a narrower location.

Figure 5. The proportion of articles mentioning Greece and written by authors with an affiliation outside Greece against Greece’s share of the world’s authors with the author’s first name. Qualification: at least ten authors with the same first name. Percentage points without data have no names with at least 100 researchers affiliated outside Greece. The y co-ordinates are smoothed by a 3-point moving average (e.g., the 4% y coordinate is the average of the data from the 3%, 4%, and 5% points). Points are annotated with an illustrative first name in the correct percentage point.

3.5 RQ5: From which countries/regions is it easiest to track diaspora researchers through first or last names?

The countries with first or last names that tend to be a high percentage national (Table 1) are the main candidates for supporting an analysis of the international
spread of diaspora researchers (mobile or embedded), although special care should be taken with larger countries (mainly China) because names that coincidentally appear in small and large countries at similar rates would appear to be national to the larger country only. More generally, for all pairs of countries, a popular name might coincidentally appear in a few other countries (e.g., Lee in South Korea is also an Anglophone last name), creating a spurious impression of national variation. An attempt to track international spread of ethnicities through first or last names would also be insensitive to historical factors. For example, many US names originated from the UK centuries ago, and it has little meaning to think of US culture as having “spread from” the UK (and elsewhere). The same is true for long-established Indian and Chinese communities in other countries, such as Singapore, Malaysia, and Indonesia.

This section focuses on the distribution of national name percentages for a country as a key factor suggesting whether a sample of predominantly national names can be extracted from it.

3.5.1 Countries with national unimodal name distributions

Countries with few international names and a unimodal distribution of name frequency against national percentage are straightforward to use for name-based identification of diaspora researchers, at least from a surface examination. In such cases, a percentage point can be selected from the graph, above which names are likely to be associated with that country. For example, in the case of Turkey, 60% seems to be a reasonable percentage, at the start of the large peak for both first and last authors, but above the more international (but still majority Turkish) Mustafa, which would otherwise contribute a substantial number of false matches (Figure 6). A higher percentage point could be selected to increase the purity of the results if an examination of names associated with, say 60-70%, found them to be often not Turkish but from a culture within Turkey that overlapped with other countries (e.g., Kurdish names, international Islamic names). Other countries with similarly national unimodal distributions include Iran, Russia (last names only), Thailand (first names only), India, Japan, and China.

An understanding of historical context and cultural overlaps can influence the choice of percentage cut-off. For the unimodal distribution of China (Figure 7), a cut-off of 85% might be chosen rather than about 60% because names in the small peak in the range of 60% to 85% might represent a partial overlap with Taiwan or Hong Kong, China, with their overlapping names (e.g., Cantonese language/culture in Hong Kong, China also appears in Macao, China, South China and diaspora communities). Manual checking of a sample of names would be needed to verify this.
Can first or last name uniqueness help to identify diaspora researchers from any country?

Figure 6. The number of authors with an affiliation in Turkey against the national percentage for the name. Qualification: at least ten authors with the same name. Percentages without data have no names with at least 10 researchers. Points are annotated with an illustrative name at the correct percentage, but most columns represent multiple names.

Figure 7. As Figure 6 for China.

3.5.2 Countries with bimodal name distributions

Countries with bimodal first or last name distributions – many international names
and many national names, appearing as separate peaks on a name frequency graph – have the potential for name-based identification of diaspora researchers. The name distribution of each such country needs to be examined separately to detect the appropriate range of percentages to classify as national names. For example, in the case of Greece, it seems that the Greek first and last names tend to occur above about 45% (Figure 8), reflecting wide international mobility of Greek researchers. Thus, a subset of apparently Greek first names could be generated from this set, perhaps with manual checks of a random sample to test this assumption, or with comprehensive manual checks of these names to remove obviously non-Greek occurrences.

Other countries with bimodal distributions include Finland, Brazil (although fairly flat), Italy (different peaks for first and last names), France (but fairly flat), Germany (weak second peak), Hong Kong, China (but the “national” peak is low and weak, presumably due to overlaps with China), Ukraine (first names only), Saudi Arabia (but very weak national peak), Romania, Indonesia, Israel, Malaysia, Nigeria (with stronger last names), Denmark (but weak flat national peak), Hungary, and the Czech Republic. In some cases, the bimodal distribution may hide other countries within the national peak.

Figure 8. As figure 6 for Greece. There are many more first names than last names because fewer last names met the minimum criterion, partly due to the use of initials instead of full names.

3.5.3 Countries with unimodal name distributions of international names

Some countries have unimodal distributions strongly dominated by international names, and these hold little prospect for name-based diaspora researcher identification. If the few mainly national names of countries like the UK (Figure 9)
with mainly international names are used to find UK diaspora researchers then the results would be extremely partial due to the small numbers of results as well as overrepresentation of distinctive ethnic groups (e.g., the Welsh for the UK, Māori for New Zealand). Other countries with unimodal distributions and few national names include Argentina, Australia, Austria, Canada, Chile, Columbia, Ireland, New Zealand, Portugal, and South Africa. In the case of South Africa, there are also a few African national names, such as Mthiyane and Tshepo.

Figure 9. As Figure 6 for the United Kingdom.

3.5.4 Countries with relatively uniform name distributions

The USA has a slightly bimodal distribution but it is relatively flat, with no obvious cut-off point for US names (Figure 10). Names with high percentages seem to be partly last names that have been anglicized from a European name or that were Central European Jewish names before the Holocaust, perhaps from extended families migrating to the USA where those left behind didn’t survive. There are also first names that are relatively unique to the USA, such as Elwood and Seth. Choosing names above a high percentage (e.g., 80%) might give the potential to identify a small minority of US diaspora researchers internationally, although this would need to be validated by checking a sample of non-US researchers with a high percentage of US names (e.g., Seth, Vogelstein).

Pakistan’s similarly flat name distribution (Figure 11) also makes identifying national names difficult. This is particularly the case because there are few names that are more than 80% from Pakistan. The flat distribution may be due to the large cultural overlap with India, some Islamic names, and large diaspora populations in the UK and elsewhere. Spain also has a relatively flat distribution.
4 Discussion

The results are limited by the restriction to a single database (Scopus) and a range of years (2001-2021). They are also limited by the information processing assumptions made. For example, the Spanish results would be different if double last names had been allowed. The definition of diaspora researcher used here is vague (all humans are ancestrally African) and in practice relates to the eras in which a
country’s names were standardized. The biggest limitation is a lack of validation of the tentative conclusions to the last research question based on frequencies alone. In particular, information about local cultures and histories is needed to evaluate whether a name is strongly associated with a country, irrespective of its relative frequency in that country. Another major limitation is the bypassing of a discussion of the meaningfulness of countries as a unit of analysis, given that many contain partially distinct cultures (e.g., Scottish, Welsh, Irish, English in the UK) and many cultures substantially overlap national boundaries (e.g., Hausa in Nigeria and Niger). The name extraction heuristic will also have misidentified some names with multiple parts that are rarely used on an international level, such as “van ‘t” (as in “van ‘t Schip”) or “op den” (as in “op den Winkel”) in The Netherlands and Belgium, which were only caught by examining the top results. Finally, alternative name-based heuristics could be developed for other countries exploring their unique features, such as the double names of Spain or the distinctive name endings of Russia (Karaulova et al., 2019).

The answer to RQ1, about which countries have the most distinctive researcher first and last names, was a list of countries with this information. No previous academic article seems to have reported national name uniqueness information for the world, whether for researchers or all citizens, although there are many online services that associate first and last names with countries (and genders) based on social media profiles (e.g., https://gender-api.com/en/api-docs/v2/country-of-origin). The supplementary files associated with the current article (https://doi.org/10.6084/m9.figshare.21954467) seem to be the first relatively comprehensive set, albeit using a fixed time and the country names of Scopus. The results confirm that Greece is not an exception or outlier for the potential to track Greek researchers internationally by name (Ioannidis et al., 2021).

For RQ2, it was shown that last names tend to be more unique than first names for countries, although there are some countries for which the reverse is true. Again, this issue does not seem to have been the subject of previous academic publications, although this fact is probably implicit in the online name databases mentioned above.

For RQ3, several factors affecting the uniqueness of researcher first and last names in a country were listed, building on previous systematic (Mateos, 2007) and ad-hoc descriptions. No list can claim to be comprehensive and there are likely to be many rarer exceptions not listed.

In answer to RQ4, there is an almost universal tendency for names that occur disproportionately often in a country to tend to be associated with that country, even for researchers affiliated with other countries. This has not been hypothesized explicitly before on an international scale, although it is implicit in prior research for
individually countries (e.g., Ioannidis et al., 2021) and all prior name-based ethnicity studies (Mateos, 2007).

Finally, for RQ5, an argument was presented that names could, in principle, be used to help identify a subset of a country’s diaspora researchers when certain name distributions occur. In particular, bimodal distributions with international and national peaks, and unimodal distributions with national peaks both lend themselves to identifying a subset of names likely to associate with a country’s researchers. The distributions are available online for checking (https://doi.org/10.6084/m9. figshare.21956795). This extends the argument previously for Greece (Mateos, 2007) and extends and modifies most previous analysis of names to associate with ethnicities (e.g., Ambekar et al., 2009; Lakha et al., 2011; Mateos, et al., 2011) by being international and fully country-based. Combining first and last names into a heuristic is likely to give the most accurate national predictions (e.g., Ye et al., 2017) for countries with a degree of distinctiveness for both. Some of the results will be embedded rather than mobile researchers and so can be added to a set of mobile diaspora researchers derived from bibliometric affiliation information or used to help cross-check it.

5 Conclusion

There are substantial international differences in the extent to which it is possible to use first or last names to find diaspora researchers: people with ancestral origins in one country that are working as researchers in other countries, whether due to researcher mobility or as part of an embedded diaspora. This seems easiest for Thailand and Japan, which have relatively distinctive names, and impossible for some Anglophone countries due to largely shared names.

For most of the 200 largest countries or regions listed in Scopus it is theoretically possible to identify a subset of first and last names from a country that occurs disproportionately often in that country and lists of these names are provided online (https://doi.org/10.6084/m9.figshare.21954467) together with graph shapes for the first 50 (https://doi.org/10.6084/m9.figshare.21956795). Individual names could be manually checked for false matches (e.g., substantial overlaps with another country) and the remainder used to track the international spread of a country’s researchers from this subset. For the most accurate results, the lists might be combined with heuristics that exploit both first and last name information for individual researchers (e.g., Ye et al., 2017). For more comprehensive lists of diaspora researchers, bibliometric databases could be used to identify internationally mobile researchers through their author IDs, and the results of both approaches could then be combined. Conversely, to specifically identify embedded diaspora researchers, author ID
information could be used to remove the mobile diaspora researchers from the initial diaspora researcher set.

The results of any name-based approach to identify the national origin of researchers are necessarily incomplete since, for example, name changes due to marriage affect the results and minority cultures within a country may be less represented. This article has only partially validated (RQ4) the name-heritage association, although a previous article has demonstrated that first and last information alone can give high precision and reasonable recall for researchers of Greek heritage outside Greece (Ioannidis et al., 2021).

References


Can first or last name uniqueness help to identify diaspora researchers from any country?


Copyright: © 2023 Mike Thelwall. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

**Appendix**

**Supplement to: Can national researcher mobility be tracked by first or last name uniqueness?**

*RQ101: Are there international differences in the distribution of relative uniqueness of first and last names*

There are huge differences between countries in the distribution of national first name uniqueness, even for the countries with the most researchers (Figure 4). For example, whilst only 20% of UK authors have first names that are at least 20% British, over 80% of Japanese authors have first names that are at least 90% Japanese. The cumulative distribution shapes also differ, as evidenced most sharply by lines crossing. Countries with fewer authors tend not to have many authors with unique names. For example, only 20% of Puerto Rican authors have first names that are more than 1% Puerto Rican (Figure 5).
Figure S1. The proportion of author **first** names in a country against the percentage of each first name in that country (ten countries with the most authors with first names).

Figure S2. The proportion of author first names in a country against the percentage of each first name in that country (ten countries with moderate numbers of authors with first names).
Can first or last name uniqueness help to identify diaspora researchers from any country?

The situation for last author distributions (Figure 6, 7) is similar to that for first author distributions (Figure 3, 4) but the shapes are not the same.

Figure S3. The proportion of author last names in a country against the percentage of each last name in that country (ten countries with the most authors with last names).

Figure S4. The proportion of author last names in a country against the percentage of each last name in that country (ten countries with moderate numbers of authors with last names).