

Differentiating Emigration from Return Migration of Scholars Using Name-Based Nationality Detection Models

Faeze Ghorbanpour^{1,3,4*}, Thiago Zordan Malaguth², Aliakbar Akbaritabar²

¹School of Computation, Information and Technology, TU Munich

²Max Planck Institute for Demographic Research, Rostock, Germany

³Center for Information and Language Processing, LMU Munich

⁴Munich Center for Machine Learning (MCML)

faeze.ghorbanpour@tum.de, zordanmalaguth@demogr.mpg.de, akbaritabar@demogr.mpg.de

Abstract

Most web and digital trace data do not include information about an individual’s nationality due to privacy concerns. The lack of data on nationality can create challenges for migration research. It can lead to a left-censoring issue since we are uncertain about the migrant’s country of origin. Once we observe an emigration event, if we know the nationality, we can differentiate it from return migration. We propose methods to detect the nationality with the least available data, i.e., full names. We use the detected nationality in comparison with the country of academic origin, which is a common approach in studying the migration of researchers. We gathered 2.6 million unique name-nationality pairs from Wikipedia and categorized them into families of nationalities with three granularity levels to use as our training data. Using a character-based machine learning model, we achieved a weighted F1 score of 84% for the broadest- and 67%, for the most granular, country-level categorization. In our empirical study, we used the trained and tested model to assign nationality to 8+ million scholars’ full names in Scopus data. Our results show that using the country of first publication as a proxy for nationality underestimates the size of return flows, especially for countries with a more diverse academic workforce, such as the USA, Australia, and Canada. We found that around 48% of emigration from the USA was return migration once we used the country of name origin in contrast to 33% based on academic origin. In the most recent period, 79% of scholars whose affiliation has consistently changed from the USA to China, and are considered emigrants, have Chinese names in contrast to 41% with a Chinese academic origin. Our proposed methods in addressing left-censoring issues are beneficial for other research that uses digital trace data to study migration.

Introduction

Sociological and demographic research has increasingly used web and digital trace data (Kashyap et al. 2023; Albrez-Gutierrez et al. 2019). The nationality of an individual is not included in most digital trace data to preserve privacy. If we only use observational data for migration research, we cannot be certain about the country of origin of a

migrant. Hence, the migration trajectories based on the digital traces could be prone to a left-censoring issue because we do not know if the first emigration event observed in digital traces is from the actual country of origin. This emigration event could have been recorded in one of the next locations along the migration trajectory. To address this, we need to identify an individual’s nationality. Here, we proposed methods for this identification that use the least available information, i.e., full names. If we accurately identify nationalities using names, we can complement observational data by adding a baseline for comparison to resolve uncertainties in migration trajectories.

Bibliometric data, which is information extracted from scientific publications, is increasingly used for demographic research (Kashyap et al. 2023). The literature highlights its value as a resource for analyzing the migration patterns of actively publishing researchers by examining consistent changes in the countries of affiliation (Zhao et al. 2023; Akbaritabar, Theile, and Zagheni 2024; Sanliturk et al. 2023). Bibliometric databases do not provide information about a scholar’s country of origin or nationality. As a result, researchers use various methods and data to determine the country of origin for these individuals, such as census data (Lewison et al. 2016; Grilli and Allesina 2017). Using the country of affiliation where a scientist published their first paper as their country of “academic birth” is a common practice (Kashyap et al. 2023; Zhao et al. 2023; Thelwall 2023). Here, we aimed to provide an alternative method for identifying scholars’ nationalities through bibliometric data by using their names. By recognizing the nationalities of scholars, we can re-identify the composition of the population of scholars per country. We can recalculate migration rates with this alternative measurement of the country of origin. Furthermore, the detected nationality helps differentiate emigration from *return migration* of the graduate students to their home countries who were identified as emigrants based on their academic origin (Sanliturk et al. 2023).

Names are often culturally specific and passed down through generations. They may reflect the nationality, race, or ethnicity of one’s parents. Consequently, full names can sometimes serve as indicators to infer these attributes (Newman et al. 2018; Schaefer and Grusky 2013). However, we acknowledge that identifying someone’s nationality, race, or

*This work was conducted while the author was visiting the Max Planck Institute for Demographic Research. Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ethnicity, based solely on their name is not always reliable, particularly among second-generation migrants, children of mixed-ethnic families, and those who experience changes in names. Additionally, the nationality identified using names could not be considered an equivalent for citizenship (Amit and Dolberg 2023; Vathi and Vathi 2015). Nevertheless, in the context of migration research, when we have already observed a consistent movement event, if we identify the nationality using names, we can differentiate whether this is more likely to be an emigration or a return migration. Language models have shown remarkable proficiency in text-based tasks, including identifying patterns in names to predict nationality. ByT5 models are considered one of the most advanced language models for character-based text processing (Xue et al. 2022).

This paper aims to differentiate between the emigration and return migration of actively publishing researchers using Scopus bibliometric data, comprising 28+ million publications from 8.2 million researchers between 1996 and 2020. The dataset includes the names and affiliation addresses of scholars but lacks information on the countries of birth or nationalities. We use character-based language models to predict the likely origin of scholars' names. We train these models using 2.6 million unique name-nationality pairs from individuals with Wikipedia pages, which is independent of Scopus data. Our training data is more comprehensive than prior studies that relied on publicly available data. Following the methodology of Ye et al. (2017) and Ye and Skiena (2019), we categorized nationalities into three hierarchical levels and customized them based on our training data, ranging from broad nationality or ethnic groups (12 classes) to more specific countries (175 classes). After analyzing the composition of selected countries based on the country of name origin, we will explore the impact of using the origin of scholars' names as a proxy for their nationality. Specifically, we aim to assess whether this approach would alter the size and direction of return migration flows based on academic origin. To support replication and further research, we publicly released the pre-trained models, country categorizations, and training data developed in this study. The next sections are structured as follows: after reviewing the related works, we describe nationality detection, which covers data processing, categorization, model training, and evaluation. Following this, there is a section on our empirical study of Scopus bibliometric data that provides insights into the emigration and return migration of researchers.

Related Work

Bibliometric data is increasingly used to analyze scholars' migration. These studies consider consistent changes in affiliation addresses spanning more than one year as an indicator of a migration event. They used the country where a scientist published their first paper as the country of academic birth (Kashyap et al. 2023; Zhao et al. 2023; Sanliturk et al. 2023; Akbaritabar, Theile, and Zagheni 2024). Thelwall (2023) used academic origins and studied the uniqueness and distribution of first and last names. Sanliturk et al. (2023), and Akbaritabar, Theile, and Zagheni (2024) used academic origin to study the relationship between schol-

ars' migration and economic development, and to investigate the global bilateral flows and rates of international migration, respectively. It is shown that a significant proportion of researchers move from developing to developed countries, a phenomenon often described as "brain drain" (Dodani and LaPorte 2005). However, evidence of return migration suggests a "brain circulation," where scholars contribute to their home countries after gaining international experience (Crescenzi, Holman, and Orru' 2017). The literature also highlights the role of collaborative networks and institutional policies in facilitating or hindering scholars' mobility (Moed, Aisati, and Plume 2013; Walk and Akbaritabar 2024).

Recent studies have explored the return migration of scholars and the factors influencing it. Zhao et al. (2022) used Scopus data from 1996 to 2020 to analyze the departure and return of German-affiliated researchers—based on country of academic origin—and found that gender, academic cohort, and discipline significantly impact return migration. Male researchers were more likely to return to Germany compared to their female counterparts. Similar patterns were observed in a study on the return migration of academics by Gaulé (2014), which identified career opportunities, personal reasons, and institutional support as key drivers for scientists deciding when and if to return to their home country.

The literature also highlights the complexity of migration decisions beyond simple return scenarios. For example, scholars may engage in circular or onward migration (similar to migration pathways proposed by King and Skeldon (2010)), where they continue to move between different countries rather than permanently returning to their home country, which could be driven by the global dynamics of academic opportunities and knowledge exchange (Constant 2021, 2020). Hence, it is important to consider individual and contextual factors in the study of global academic mobility (Cattaneo and Robinson 2020). Furthermore, Zhao et al. (2023) have reported gender disparities in scholarly migration and showed that female researchers are less likely to migrate internationally, which could often be due to institutional, social, or familial constraints. It is important to consider demographic factors' effect on academic mobility.

Researchers studying the distribution of scholars' nationalities and migration often use census data to determine the nationality. Census data provides the most common or popular names for each country. For example, this approach is used by Lewison et al. (2016) to determine the nationality of lung cancer researchers based on the countries where their last names are most prevalent. A similar approach is taken by Grilli and Allesina (2017) for analyzing mobility, gender imbalance, and nepotism within academic systems. Ioannidis et al. (2021) examine scientists with common Greek surnames to identify Greek scientists in Greece and in the diaspora.

In recent years, several methods have been developed to analyze the text of an author's name and estimate their nationality. For instance, Le et al. (2021) conducted an analysis of disparities in scientific society honors using machine learning to determine the origins of authors' names, and

Jain, Enamorado, and Rudin (2022) used machine learning to predict ethnicity to analyze political donations from various ethnic groups. However, they applied a machine learning method to a smaller amount of training data.

Machine learning (ML) models have demonstrated exceptional performance in text-based tasks, such as identifying patterns in names to predict nationality, ethnicity, or race. Numerous studies have used ML for name-based classification. For instance, Treeratpituk and Giles (2012) developed a name-ethnicity classifier using multinomial logistic regression, while Lee et al. (2017), Le et al. (2021), and Jain, Enamorado, and Rudin (2022) used recurrent neural networks, and Kang (2020); Chaturvedi and Chaturvedi (2024) used convolutional neural networks for name-nationality classification. These approaches, however, are often constrained by limited training datasets, leading to potential biases in the results. To mitigate such biases, it is crucial to train models on large datasets and evaluate them on distinct data from the training set. One example of a system addressing this is NamePrism (Ye et al. 2017; Ye and Skiena 2019), which uses Naive Bayes to create name embeddings for classifying nationalities and ethnicities, revealing relationships between names across different groups. Despite its strength in leveraging extensive data, the proprietary nature of its dataset and lack of publicly accessible code restricts further exploration of its method by other researchers.

Following the presented literature, we aimed to provide an alternative method for identifying scholars' country of origin through bibliometric data by assigning a nationality using ML models. Our analysis uses authors' first and last names and a substantially larger amount of training data than previous research. We ask whether part of the high emigration rates reported in the literature (Akbaritabar, Theile, and Zagheni 2024; Sanliturk et al. 2023; Zhao et al. 2023), from countries such as the USA to destinations such as China, could be the return migration of graduate students and scholars to their country of origin.

Data and Methods

To develop an ML model to assign a country of origin to names, we need a training dataset containing pairs of names and nationalities. We used this to train our classification models. The full names served as our textual data, while the nationalities served as the classes. Obtaining such data proved challenging due to the private nature of personally identifiable information.

Public Wikipedia introduction pages for individuals are a valuable resource as these pages often include a biography outlining the nationality, birthplace, or citizenship details. We gathered publicly available datasets based on Wikipedia, typically created for other purposes like text generation for ML. We combined seven different datasets, resulting in 2.6 million unique name and nationality pairs for training our classification models (see Table 1). Additionally, we used DBpedia's Application Programming Interface (API)¹ to fill in missing information from Wikipedia. We focused on

¹https://dbpedia.org/page/Web_API

pages that provided at least one of the following information: nationality, birthplace, or citizenship.

We evaluated our model's performance on two test datasets from different sources than the Wikipedia dataset used for training (see the bottom rows of Table 1, also see more information about data collection and pre-processing in Appendix).

We developed a hierarchical nationality categorization framework with three levels of granularity. Each level builds on the lower level, allowing for a clearer understanding of name similarities among individuals. The first level has 12 classes, the second has 30, and the third, country level, has 175. To do this, we used the frameworks developed by Ye et al. (2017) and Le et al. (2021) and adjusted them to fit our dataset. For instance, we added countries from Oceania and the Caribbean. This ensures a more comprehensive representation of naming patterns across diverse regions. We eliminated countries where we had fewer than 100 instances, as they were considered small classes. We also excluded countries with multicultural populations, such as the US, Canada, Australia, New Zealand, and South Africa, from the training and testing. However, consistent with prior frameworks, we included them in the empirical analysis, i.e., to assign nationality to names from Scopus data. In addition, we present comparative results on the composition of scholars in these countries of immigration to show their diverse scientific workforce. The UK, while a multicultural country, was chosen to represent English-origin names because it strikes a balance between diversity and classification performance, as reflected in its relatively lower entropy and higher F1 score compared to other predominantly English-speaking nations. For the rationale regarding the inclusion and exclusion of countries, please refer to Appendix.

Figure 1 presents our classification as a treemap, where the first column shows broader first-level groupings (e.g., "German" and "Dutch" as "German"), the middle column groups the countries by second-level classes (e.g., Germany, Austria, and Switzerland as "German"), and the third column corresponds to the main classification at the country-, i.e., most granular, level. The classes are ordered by size within each category. This figure also shows the distribution of classes in the training dataset at each of the three levels. At the country level, Germany is the largest class, followed by the UK, France, and Italy. At the second level, the German class is predominant, followed by English, Spanish, and French. At the broadest level, Germanic leads, followed by Hispanic, Romance, English, and Slavic. Although class sizes are influenced by the number of Wikipedia pages per country, we mitigated this by collecting more data and adjusting our methods to balance small and large classes.

We also used two separate test sets to test the robustness of our trained models (see the bottom rows of Table 1). It is important to note that these test sets have different class distributions; for example, in the IUSSP data, the major classes are India, Brazil, France, and Nigeria, and in the athletes' data, the largest classes are Indonesia, South Korea, and China (See figures with the composition of classes in the Appendix).

Letter patterns are important in full-name processing. We

	Name	Description	Volume
Training data	Wikipedia notable people’s mobility (Lucchini 2019)	This dataset contains mobility data of notable people whose biography is registered in Wikipedia.	2,190,457
	Nationality Estimate (Le et al. 2021)	This dataset is constructed with name-nationality pairs by parsing the English-language Wikipedia.	696,136
	WikiBio (Wikipedia Biography Dataset) (Lebret, Grangier, and Auli 2016)	This dataset gathers biographies from English Wikipedia.	418,978
	Wikipedia Person Dataset (Wang et al. 2018)	This dataset gathers the infobox with the description of the person’s page.	355,468
	Politicians on Wikipedia (Wagner 2017)	This dataset contains information about politicians from Wikipedia.	54,877
	Pantheon 1.0 dataset (Yu et al. 2016)	This dataset includes the biographies present in more than 25 languages in Wikipedia.	16,071
	Wikidata	Processed by our colleague.	131,731
		SUM	3,874,776
		SUM after removing duplicates	2,634,445
		SUM after removing multi-cultural countries and small classes	2,245, 768
Test data	Athletes (Kang 2020)	Data collected from the sports sites.	42,413
	International Union for the Scientific Study of Population (IUSSP) public members’ directory ²	Information on IUSSP member’s self-reported nationality, field of specialty, spoken languages, affiliation, etc.	1,547

Table 1: Datasets used for training (top) and testing (bottom) the ML models.

used character-based ML methods to detect these patterns. We used three neural network approaches. First, we used FastText, a skip-gram model that represents each word as a collection of character n-grams (Bojanowski et al. 2017). It is well-suited for name processing due to its ability to capture subword information, making it robust to rare or unseen names. Second, we used a character-based convolutional neural network (char-CNN), which includes an embedding layer, k CNN layers (with k as a hyperparameter), and layer normalization (Zhang, Zhao, and LeCun 2015; Wu 2017). CNNs are effective for processing names because they excel at capturing local patterns within character sequences, which is essential for distinguishing variations in names. Similarly, a character-based long short-term memory (char-LSTM), consisting of an embedding layer, k BiLSTM layers, and layer normalization, is particularly suited for handling sequential data like names (Verwimp et al. 2017; Graves and Graves 2012). It captures long-range dependencies between characters, enabling better handling of name structure and variations. The best practices that are publicly available, and used in previous studies, include char-LSTM by Lee et al. (2017), Le et al. (2021), and Jain, Enamorado, and Rudin (2022), and char-CNN by Kang (2020) and Chaturvedi and Chaturvedi (2024).

Third, we used a byte-level transformer model based on the T5 architecture (ByT5) in three versions: small, base, and large (Xue et al. 2022). ByT5 processes text at the byte level, making it inherently character-based, which is crucial for handling names with diverse structures and spellings. ByT5’s character-level processing enables it to learn subtle orthographic and phonetic patterns that are often specific to different nationalities. This capability is particularly advantageous in name classification tasks (Lee et al. 2017).

Furthermore, ByT5’s token-free design enables it to handle noisy inputs and to generalize to rare or previously unseen names, which are common challenges in name-based classification tasks. Huang et al. (2023) used ByT5 to study character representations within subword-based language models, and Stankevičius et al. (2022) have used ByT5 for tasks such as diacritic restoration and typo correction. We trained all three versions from scratch on our Wikipedia dataset without altering the model architecture.

To test the performance of our models, we divided our dataset into training, validation, and test subsets, comprising 65%, 15%, and 20% of the data, respectively (see Table 1). Models were trained on the training set, fine-tuned using the validation set with Optuna (Akiba et al. 2019), and then re-trained on the combined training and validation sets before being evaluated on the test set using 5 random seeds. Due to computational constraints, this process was limited for larger models like ByT5. We also assessed performance on two additional test datasets (athletes and IUSSP data). Given the data imbalance, we used weighted F1 score, macro F1 score, and accuracy as evaluation metrics. The weighted and macro F1 score balances precision and recall, emphasizing minority class performance while checking accuracy to provide a fair measure of overall performance. Please refer to Appendix for more information about the hyperparameters and training details.

Table 2 provides a comparison of the performance of the models on the test set derived from the Wikipedia dataset as our first model assessment. The results demonstrate that ByT5 models, on average, slightly outperform other character-based methods in terms of accuracy and F1 scores across all three levels of classification. Based on the evaluations presented in this table, we used the ByT5 model,

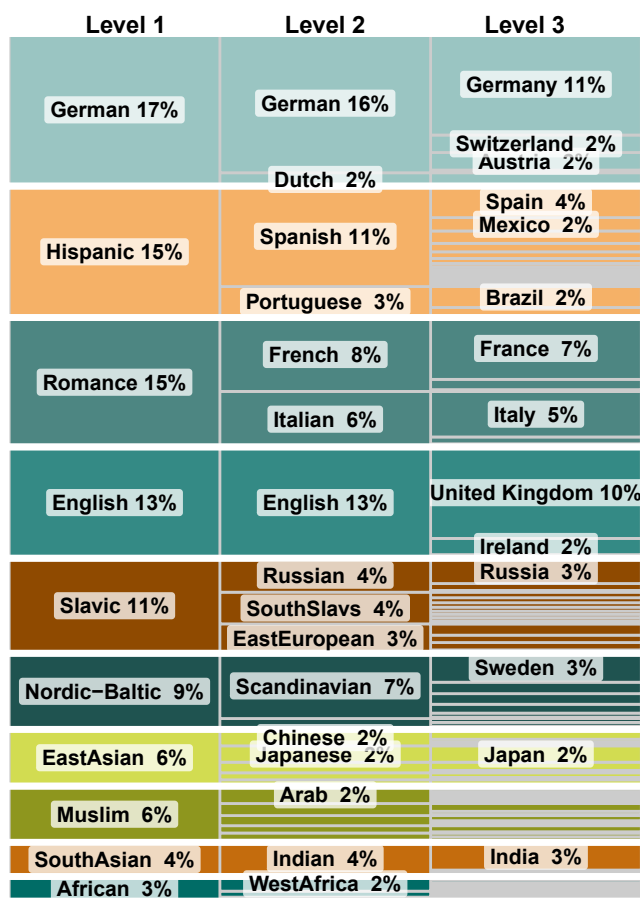


Figure 1: Distribution of nationality classes in the training dataset. Each column sums up to 100% and represents one of the three classification levels, from the left (broader) to the right (more detailed). All classes are shown, but only those bigger than 1.5% of the total are labeled.

which performed better. At the first level of classification, the model achieves an accuracy of approximately 84%. At the second level, the accuracy remains high at around 81%. However, as the number of classes increases, the model’s performance decreases to about 68% at the third level. Nevertheless, considering the complexity arising from the 175 classes, this performance can still be considered acceptable, although it could be improved in future research. Accuracy at the third level varies considerably between classes. Despite an average F1-weighted score of about 68%, some countries, such as Japan, have F1 scores above 90%, while others have F1 scores below 10. The four countries for which the model is most accurate belong to the “East Asian” at the first level, while 6 of the 10 countries with the lowest F1 score belong to the Hispanic class (see Figure 8 in Appendix).

These findings highlight the effectiveness of the ByT5 models in handling the classification task in comparison with other character-based neural networks. FastText outperforms both char-CNN and char-LSTM models. Addition-

ally, since FastText’s core code is written in C++ rather than Python, it is the fastest model among those we used (Refer to Appendix for more information).

As a second model assessment, we used two out-of-domain and separate test sets from different sources. While our model has not seen these datasets, and these data have different origins and distributions (see Figure 5 in Appendix), the metrics are almost the same as the performance on the Wikipedia test set (see Table 3), demonstrating the robustness of the performance of the trained ByT5 models. The reasons for this are twofold: First, we used a sufficiently large training dataset to ensure that we have a reasonable number of samples, even for minor classes. Second, we used a character-based large language model (ByT5) that includes several layers of transformer and attention modules, which allows it to extract subtle and hidden patterns in names.

Results

We aimed to compare the emigration flows of researchers based on the country of academic origin—country of affiliation in the first publication—with nationality predicted using machine learning. We used a dataset from Scopus containing 8.2 million full author names and countries of affiliation throughout an author’s career.

First, we investigated the composition of a country’s population of scholars based on ML-assigned nationalities (level 2 regions). The top row in Figure 2 shows the three countries that had the largest populations in Scopus among those countries included in the training dataset. It shows that 97% of scholars affiliated with institutions in China have Chinese names. 73% of scholars in Germany have German names, and 90% of scholars in India have Indian names. These percentages indicate that China has a smaller share of scholars with non-Chinese names, and Germany, in comparison to China and India, has a higher share of scholars with non-German names. Despite the difference between these three countries, they have a more homogeneous population compared to the countries excluded from the training, which are shown in the bottom row in Figure 2. These countries were excluded from training data due to their multicultural population and the figure shows our model’s capability to identify the diversity of origin of names in these countries. Most authors in Australia, Canada, and the USA have English names with about 52%, 33%, and 38%, respectively. In all these countries, the second most frequent origin of names is Chinese, with about 13%, 14%, and 14%, respectively. The third most common origin differs between these countries, with Indian names being the third most common in Australia (6%), and French (12%) in Canada, probably due to the predominance of French-speaking individuals in Canada. In the USA, German names are the third most common origin. So far, our results do not differentiate between second-generation migrant scholars in a country who might have names that are different from the popular names in the host country, which will be discussed next.

In the context of migration research, and once we have already observed a movement based on a consistent change in the affiliation addresses of scholars, we can use the ML nationality to further probe into aggregated emigration flows

	Level 1			Level 2			Level 3		
	F1-macro	F1-weighted	Accuracy	F1-macro	F1-weighted	Accuracy	F1-macro	F1-weighted	Accuracy
Char-LSTM	24.191	37.892	27.712	05.493	15.103	30.124	03.118	05.980	16.536
Char-CNN	36.432	53.929	53.929	53.929	53.929	53.929	18.906	06.785	18.906
Fasttext	80.887	83.598	83.619	69.132	78.487	78.956	33.205	63.445	65.139
ByT5 small	74.566	83.854	83.759	72.737	80.971	81.121	41.666	66.789	68.356
ByT5 base	74.247	83.557	83.453	72.491	80.673	80.828	42.044	66.453	67.672
ByT5 large	79.765	82.987	82.944	72.051	80.209	80.349	41.723	65.799	66.822

Table 2: Evaluation results based on the Wikipedia test set.

	Level 1			Level 2			Level 3		
	F1-macro	F1-weighted	Accuracy	F1-macro	F1-weighted	Accuracy	F1-macro	F1-weighted	Accuracy
Athletes	79.276	83.423	83.222	75.739	79.889	79.889	40.916	66.433	66.003
IUSSP	71.360	84.669	84.223	74.597	82.243	82.119	41.729	66.523	66.262

Table 3: Evaluation results based on the two out-of-domain test datasets using the best-performing ByT5 models.

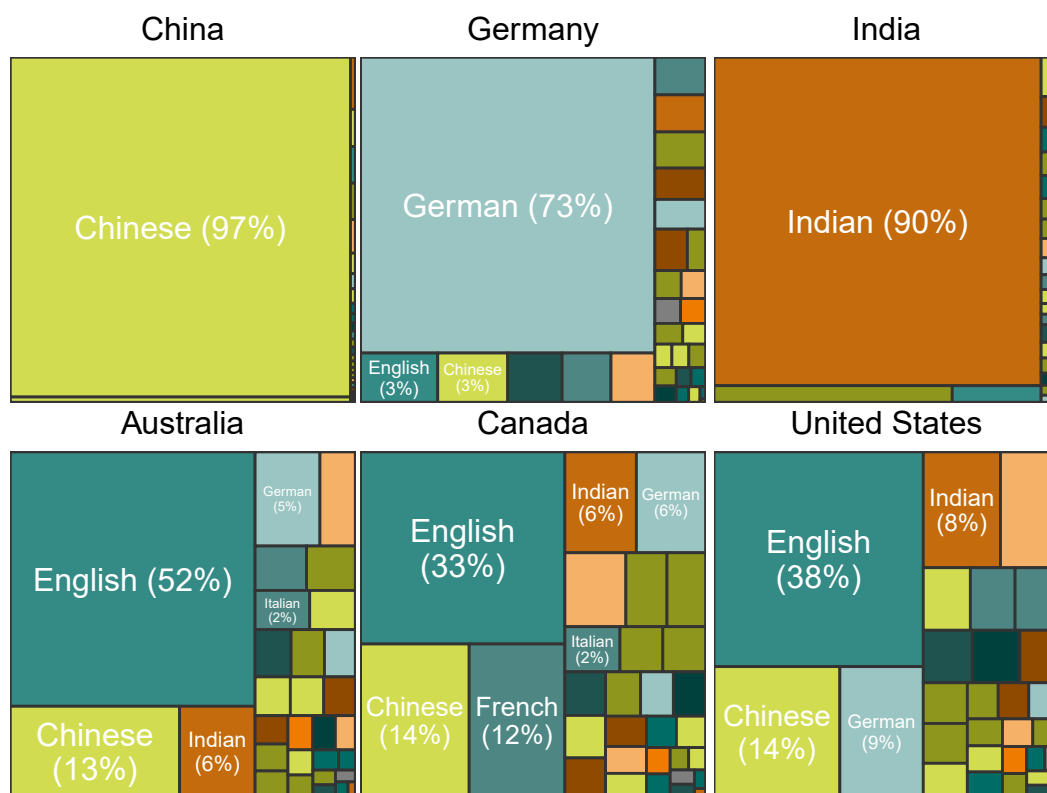


Figure 2: Composition of scholars affiliated with countries (names printed above panels) included in (top row) and excluded from the training data (bottom row) based on the country assigned using ML model (Level 2 regions, printed inside tree map).

at the country level and ask if a subset of these emigrations could be considered returns to the country of origin. Here we wanted to identify the top destinations for scholars who emigrate versus those who return, based on their country of 1) academic origin and 2) ML-assigned name origin. We chose the USA for this analysis because it has the largest population of researchers in the Scopus database and is one of the

countries with the most diverse population of researchers in terms of nationality (see Figure 2). From all scholars affiliated with the USA, Figure 3 represents the proportion of movements to the five most common destinations for emigration (left) and the five most common destinations for return movements (right) classified based on academic origin (top) and ML name origin (bottom). Based on academic

origin, the most common emigration destinations from the USA were the UK, Canada, Germany, China, Japan, and South Korea. The most common destinations for return migration were similar but with differences in the order and size of the flows.

ML name origin shows a different pattern and enables us to differentiate emigration from return migration. Using the ML name origin, Japan and South Korea are no longer among the top five most common destinations for emigration, and China appears only in the last period as the fifth most common destination; while it was the first destination based on the academic origin. This shift in the list of most common destinations for emigration and return migration suggests that using the academic origin might underestimate the size of return flows. When considering academic origin, the top 5 destinations concentrated less than 50% of emigrations. However, when considering the ML-name origin, this tendency was more than 50% of return movements for all considered time periods, with a higher share for East Asian and South Asian countries. These findings align with previous research indicating that Chinese and West Asian scholars have a higher tendency to return to their home countries after an international migration experience (Marini and Yang 2021; Hu 2021).

Figure 4 shows the proportion of scholars who return to their home countries after migrating, compared to the total number of scholars who migrate across six continents and genders. To identify the gender of scholars, we used the method introduced by Zhao et al. (2023). The mean proportion of return migration based on ML name origin was higher than that found based on academic origin (see the shift in points toward the right in the bottom panel compared to the top panel.) While similar patterns were observed across most continents, return migration proportions were much higher based on name origin for Oceania, North America, and Europe. Nearly half of the scholars leaving the USA return to their country of ML name origin. This indicates that these countries have a much higher share of foreign graduate students and scholars who start their publishing careers in these countries and later return to their countries of origin.

Based on the averages for men and women (vertical dashed and solid bars), male scholars tend to return to their home countries more often than female scholars, except in Africa, where the trend is reversed when considering ML name origin. This tendency is particularly strong for male scholars who were located in North America compared to scholars from other continents. This could signal differences in the rate of success in being hired permanently in the host country among genders and nationalities that require further research.

Discussion and Conclusions

Scholars contribute to knowledge production and innovation. Research has shown that migrant scholars are among those with the highest productivity and impact (Sugimoto et al. 2017). Studying the migration of scholars, which can be traced using a consistent change in affiliation addresses in publication records (Kashyap et al. 2023), enables identifying the countries that send scholars from those that receive

them. However, without information about scholars' nationality or country of origin, it is difficult to interpret the migration patterns (Sanliturk et al. 2023) due to left-censoring issues in digital trace data (Akbaritabar, Theile, and Zagheni 2024). Particularly when assessing whether scholars are emigrating or returning to their country of origin.

This left-censoring problem could affect all migration research that relies solely on observational digital trace data. To address this problem, we used the most recent Machine Learning (ML) methods and language models to detect the nationality of scholars based on their full names. We used substantially larger data than previous studies to train models. We tested these models with various datasets and selected the best-performing ones. In our empirical study, and after testing our model's performance using different sets of test data, we used the model to assign names to 8.2 million scholars from Scopus bibliometric data. This enabled us to compare migration trends based on the country of academic origin and the country assigned using the ML model. It further allowed us to differentiate between emigration and return migration in the case of graduate students or more junior scholars who start their publication career in the host country and later return to their country of origin.

We observed substantial changes in emigration and return migration flows. For instance, the literature reports high rates of emigration from the USA to China (using the country of academic origin) (Zhao et al. 2023; Sanliturk et al. 2023; Akbaritabar, Theile, and Zagheni 2024). Here, we asked whether part of this emigration could be the return migration of Chinese scholars who graduated in the USA and returned to their country of origin. It is evident that some of the scholars with Chinese names in the USA could be second generation migrants. Nevertheless, in the context of migration research, and when we have already observed a migration event, we can investigate with higher certainty if the movement is a return migration rather than emigration.

As we observed in the most recent period, 79% of those emigrating from the USA to China have Chinese names and are categorized as returnees using the nationality assigned by machine learning in contrast to 41% based on academic origin. Our observation aligns with China's government-initiated "Thousand Talents Plan"³, launched in 2008. This program aimed to attract top international talent to work in mainland China (Jia 2018) and is shown to have succeeded in motivating the return of Chinese scholars with international migration experience (Shi, Liu, and Wang 2023). This could also be associated with some recent policies, such as the US-China Initiative. Launched in 2018 to counter economic espionage, this initiative has been criticized for disproportionately targeting Chinese-origin academics, creating a chilling effect on collaboration and increasing return migration to China (Jia et al. 2024). Studies report a significant rise in Chinese-born scientists leaving the USA since the initiative's inception, with many returning to China (Xie et al. 2023). While the temporal scope of our data from 1998 to 2017 (we exclude two years at the start and end of the data from 1996 to 2020 to minimize left- and right censoring and

³https://en.wikipedia.org/wiki/Thousand_Talents_Plan

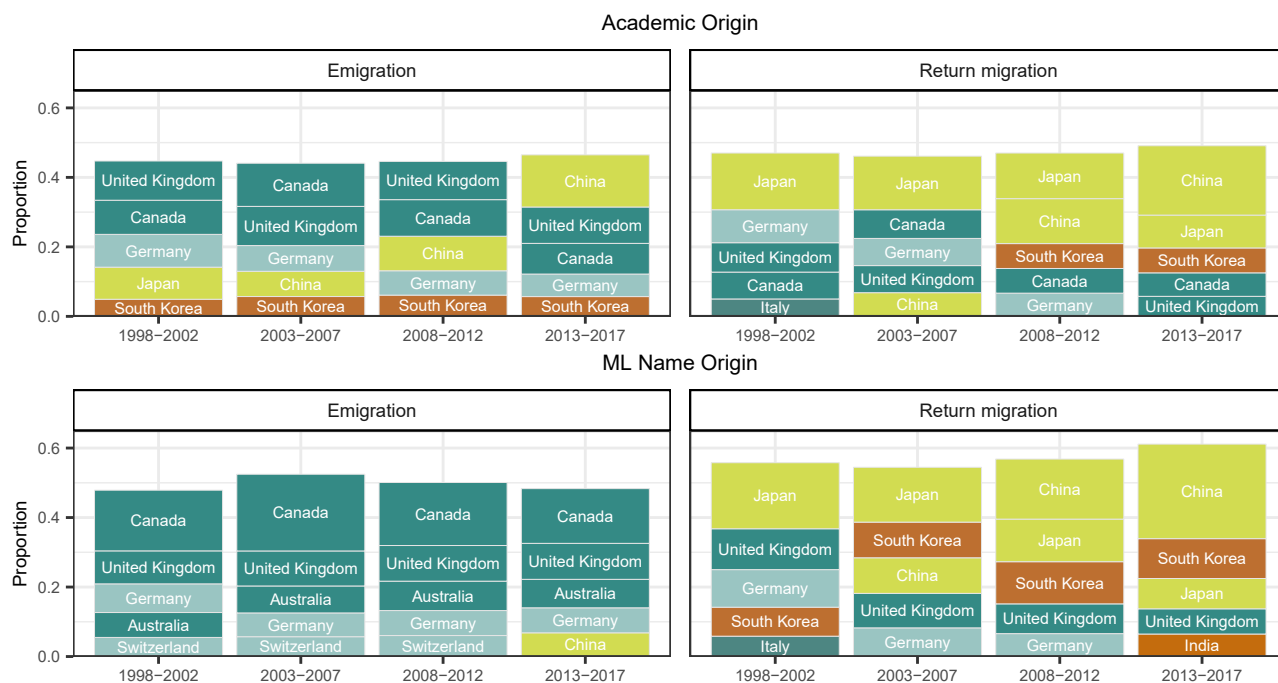


Figure 3: Top 5 countries with the highest proportion of scholars among those affiliated to institutions in the United States divided into emigration (left bars) and return migration (right bars) based on academic origin (top bars) versus the country of ML-name origin (bottom bars).

truncation issues) is not exactly overlapping with this recent policy change, however, the literature emphasizes that scholars are among the subsets of the population that have credentials and networks to react faster to upcoming changes.

We publicly share our model’s code and training data to enable future research. Follow-up studies using digital trace data could use our model to extend the application of language models to analyze the migration patterns beyond the case of actively publishing researchers, for instance, to other data from social media, LinkedIn, etc. Furthermore, surveys of scholars’ actual behavior could enable further validation of the model’s ability to reliably predict scholars’ migration trajectories, including their return or continued residence in host countries. Additionally, our models could be further refined to account for potential cultural, regional, gender, and disciplinary differences in migration behavior. Future research could investigate the possibility of long-term stays and integration in host countries that might be different between subsets of migrants depending on their nationalities and countries of origin.

To clarify the ethical terms and best practices in machine learning research, we emphasize that our models were trained using publicly available datasets that had been collected from Wikipedia. Our use of these datasets was solely for research purposes and completely complies with the licensing and use terms of this platform. Since the datasets only include names and nationalities of individuals from public Wikipedia pages, they do not contain private infor-

mation. We gain access to Scopus data through a license⁴ to do our empirical study on migration trends. This data is accumulated using publicly accessible information from scientific publications and does not constitute private information.

Our research has certain limitations that we would like to acknowledge. In preprocessing the data, we removed non-ASCII and redundant characters. However, this process could also eliminate information, such as language-specific umlauts and characters, that might be useful for classification. While we evaluated the effect of this preprocessing step, future work could conduct more experiments beyond the ones presented here to assess the impact of the preprocessing. Additionally, ML and language models are valuable tools for detecting nationality, but they have uncertainty. While our model has a relatively high accuracy rate, this performance primarily applies to nationality groups with the highest representation in the training data. Predictions for nationality groups with a small number of observations have lower F1 scores (please see confusion matrices in the Appendix). While we use weighted measures to circumvent such problems, and these issues are less prevalent in migration corridors with the highest count of scholars such as those presented here, some errors can still occur and be explained by factors such as second-generation migrants, name or citizenship changes, or the diverse ways people choose names (please refer to the in the Appendix). To address this, we collected a large dataset and used one

⁴Please see the acknowledgment section for details.

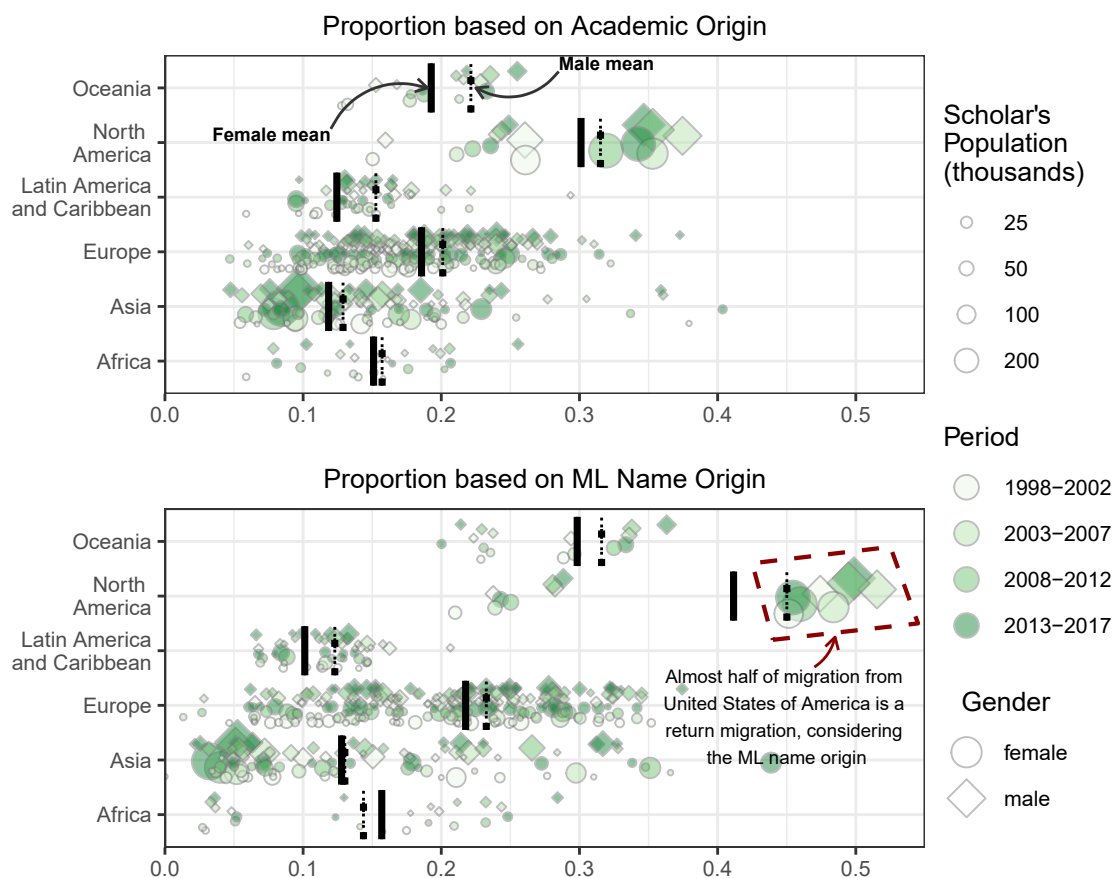


Figure 4: The proportion of return migration based on academic origin (top panel) versus the country of ML-name origin (bottom panel) divided by female scholars (diamonds) and male scholars (circles), relative to the size of the population of scholars (point size).

of the most advanced language models available. However, there are still errors, particularly for nationalities and groups with lower representation in training data, that remain unexplained. These inaccuracies can impact analyses of emigration and return migration among scholars and could be improved in future research.

Despite these limitations, our methods and publicly shared model could help reduce left-censoring issues in migration research using digital trace data. Our results highlight that a substantial part of the emigration, as reported in the literature using observational data, could be considered return migration, which has implications for migration research that uses digital trace data.

Availability of Model Code and Training Data

The training dataset used in this study is publicly available on Zenodo (Ghorbanpour, Malaguth, and Akbaritabar 2025). The scripts to replicate our analysis and figures are accessible via the following GitHub repository: <https://github.com/FaezeGhorbanpour/NameBasedNationalityDetection>.

Acknowledgements

We thank Tom Theile for lending us one of the datasets presented in the text based on Wikidata. This study has received access to the bibliometric data through the project “Kompetenznetzwerk Bibliometrie” (Schmidt et al. 2024), and we acknowledge their funder Bundesministerium für Bildung und Forschung (grant number 16WIK2101A).

References

Akbaritabar, A.; Theile, T.; and Zagheni, E. 2024. Bilateral Flows and Rates of International Migration of Scholars for 210 Countries for the Period 1998-2020. *Scientific Data*, 11(1): 816.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

Alburez-Gutierrez, D.; Zagheni, E.; Aref, S.; Gil-Clavel, S.; Grow, A.; and Negraia, D. V. 2019. Demography in the digital era: new data sources for population research. *SocArXiv*.

- Amit, K.; and Dolberg, P. 2023. Who do you think I am? Immigrant's first name and their perceived identity. *Comparative Migration Studies*, 11(1): 6.
- Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.
- Cattaneo, A.; and Robinson, S. 2020. Multiple moves and return migration within developing countries: A comparative analysis. *Population, Space and Place*, 26(7): e2335.
- Chaturvedi, R.; and Chaturvedi, S. 2024. It's all in the name: A character-based approach to infer religion. *Political Analysis*, 32(1): 34–49.
- Constant, A. F. 2020. Time-space dynamics of return and circular migration. *Handbook of labor, human resources and population economics*, 1–40.
- Constant, A. F. 2021. *Return, circular, and onward migration decisions in a knowledge society*. Springer.
- Crescenzi, R.; Holman, N.; and Orru', E. 2017. Why do they return? Beyond the economic drivers of graduate return migration. *The Annals of Regional Science*, 59: 603–627.
- Dodani, S.; and LaPorte, R. E. 2005. Brain drain from developing countries: how can brain drain be converted into wisdom gain? *Journal of the Royal society of Medicine*, 98(11): 487–491.
- Gaulé, P. 2014. Who comes back and when? Return migration decisions of academic scientists. *Economics Letters*, 124(3): 461–464.
- Ghorbanpour, F.; Malaguth, T.; and Akbaritabar, A. 2025. Data for the training section of "Differentiating Emigration from Return Migration of Scholars Using Name-Based Nationality Detection Models" paper. Data set, DOI: 10.5281/zenodo.15103505.
- Graves, A.; and Graves, A. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37–45.
- Grilli, J.; and Allesina, S. 2017. Last name analysis of mobility, gender imbalance, and nepotism across academic systems. *Proceedings of the National Academy of Sciences*, 114(29): 7600–7605.
- Hu, D. 2021. Understanding transnational academic migration: US doctorate recipients' choices of returning to Chinese academe. *International Journal for Educational and Vocational Guidance*, 1–17.
- Huang, J.; Wu, Z.; Mahowald, K.; and Potts, C. 2023. Inducing Character-level Structure in Subword-based Language Models with Type-level Interchange Intervention Training. In *Findings of the Association for Computational Linguistics: ACL 2023*, 12163–12180. Toronto, Canada: Association for Computational Linguistics.
- Ioannidis, J. P.; Koutsioumpa, C.; Vakka, A.; Agoranos, G.; Mantsiou, C.; Drekolia, M. K.; Avramidis, N.; Contopoulos-Ioannidis, D. G.; Drosatos, K.; and Baas, J. 2021. Comprehensive mapping of local and diaspora scientists: A database and analysis of 63,951 Greek scientists. *Quantitative Science Studies*, 2(2): 733–752.
- Jain, V.; Enamorado, T.; and Rudin, C. 2022. The Importance of Being Ernest, Ekundayo, or Eswari: An Interpretable Machine Learning Approach to Name-Based Ethnicity Classification. *Harvard Data Science Review*, 4(3). <https://hdsr.mitpress.mit.edu/pub/wgss79vu>.
- Jia, H. 2018. China's plan to recruit talented researchers. *Nature*, 553(7688): S8–S8.
- Jia, R.; Roberts, M. E.; Wang, Y.; and Yang, E. 2024. The impact of US–China tensions on US science: Evidence from the NIH investigations. *Proceedings of the National Academy of Sciences*, 121(19): e2301436121.
- Kang, Y. 2020. Name-nationality classification technology under Keras deep learning. In *Proceedings of the 2020 2nd international conference on big data engineering*, 70–74.
- Kashyap, R.; Rinderknecht, R. G.; Akbaritabar, A.; Alburez-Gutierrez, D.; Gil-Clavel, S.; Grow, A.; Kim, J.; Leasure, D. R.; Lohmann, S.; Negraia, D. V.; Perrotta, D.; Rampazzo, F.; Tsai, C.-J.; Verhagen, M. D.; Zagheni, E.; and Zhao, X. 2023. Digital and Computational Demography. In *Research Handbook on Digital Sociology*, chapter Research Handbook on Digital Sociology, 48–86. Edward Elgar Publishing. ISBN 978-1-78990-676-9.
- King, R.; and Skeldon, R. 2010. 'Mind the Gap!' Integrating Approaches to Internal and International Migration. *Journal of Ethnic and Migration Studies*, 36(10): 1619–1646.
- Le, T. T.; Himmelstein, D. S.; Hippen, A. A.; Gazzara, M. R.; and Greene, C. S. 2021. Analysis of scientific society honors reveals disparities. *Cell Systems*, 12(9): 900–906.
- Lebret, R.; Grangier, D.; and Auli, M. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Lee, J.; Kim, H.; Ko, M.; Choi, D.; Choi, J.; and Kang, J. 2017. Name Nationality Classification with Recurrent Neural Networks. In *IJCAI*, volume 17, 2081–2087.
- Lewison, G.; Roe, P.; Webber, R.; and Sullivan, R. 2016. Lung cancer researchers, 2008–2013: their sex and ethnicity. *Scientometrics*, 106: 105–117.
- Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Lucchini, L. 2019. Wikipedia notable people's mobility.
- Marini, G.; and Yang, L. 2021. Globally Bred Chinese Talents Returning Home: An Analysis of a Reverse Brain-Drain Flagship Policy. *Science and Public Policy*, 48(4): 541–552.
- Moed, H. F.; Aisati, M.; and Plume, A. 2013. Studying scientific migration in Scopus. *Scientometrics*, 94: 929–942.
- Newman, L. S.; Tan, M.; Caldwell, T. L.; Duff, K. J.; and Winer, E. S. 2018. Name norms: A guide to casting your next experiment. *Personality and Social Psychology Bulletin*, 44(10): 1435–1448.
- Sanliturk, E.; Zagheni, E.; Daňko, M. J.; Theile, T.; and Akbaritabar, A. 2023. Global patterns of migration of scholars with economic development. *Proceedings of the National Academy of Sciences*, 120(4): e2217937120.

- Schaefer, H. G.; and Grusky, D. B. 2013. The relationship between names and ethnicity in the United States. *Social Forces*, 91(3): 919–944.
- Schmidt, M.; Rimmert, C.; Stephen, D.; Lenke, C.; Donner, P.; Gärtner, S.; Taubert, N.; Bausenwein, T.; and Stahlschmidt, S. 2024. The Data Infrastructure of the German Kompetenznetzwerk Bibliometrie: An Enabling Intermediary between Raw Data and Analysis.
- Shi, D.; Liu, W.; and Wang, Y. 2023. Has China’s Young Thousand Talents program been successful in recruiting and nurturing top-caliber scientists? *Science*, 379(6627): 62–65.
- Stankevičius, L.; Lukoševičius, M.; Kapočiūtė-Dzikiene, J.; Briedienė, M.; and Krilavičius, T. 2022. Correcting diacritics and typos with a ByT5 transformer model. *Applied Sciences*, 12(5): 2636.
- Sugimoto, C. R.; Robinson-Garcia, N.; Murray, D. S.; Yegros-Yegros, A.; Costas, R.; and Larivière, V. 2017. Scientists have most impact when they’re free to move. *Nature*, 550(7674): 29–31.
- Thelwall, M. 2023. Can first or last name uniqueness help to identify diaspora researchers from any country? *Journal of Data and Information Science*, 8(2): 1–25.
- Treeratpituk, P.; and Giles, C. L. 2012. Name-ethnicity classification and ethnicity-sensitive name matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, 1141–1147.
- Vathi, Z.; and Vathi, Z. 2015. Identities of the first and second generation: the role of ethnicity. *Migrating and Settling in a Mobile World: Albanian Migrants and Their Children in Europe*, 37–72.
- Verwimp, L.; Pelemans, J.; Wambacq, P.; et al. 2017. Character-word LSTM language models. *arXiv preprint arXiv:1704.02813*.
- Wagner, C. 2017. Politicians on Wikipedia and DBpedia. GESIS - Leibniz Institut für Sozialwissenschaften. Datenfile Version 1.0.0, <https://doi.org/10.7802/1515>.
- Walk, E.; and Akbaritabar, A. 2024. Resolving the Paradox between Migration and Collaboration of Scholars Worldwide.
- Wang, Q.; Pan, X.; Huang, L.; Zhang, B.; Jiang, Z.; Ji, H.; and Knight, K. 2018. Describing a Knowledge Base. In *Proceedings of the 11th International Conference on Natural Language Generation*, 10–21. Tilburg University, The Netherlands: Association for Computational Linguistics.
- Wu, J. 2017. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23): 495.
- Xie, Y.; Lin, X.; Li, J.; He, Q.; and Huang, J. 2023. Caught in the Crossfire: Fears of Chinese–American Scientists. *Proceedings of the National Academy of Sciences*, 120(27): e2216248120.
- Xue, L.; Barua, A.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; and Raffel, C. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10: 291–306.
- Ye, J.; Han, S.; Hu, Y.; Coskun, B.; Liu, M.; Qin, H.; and Skiena, S. 2017. Nationality classification using name embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1897–1906.
- Ye, J.; and Skiena, S. 2019. The secret lives of names? name embeddings from social media. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3000–3008.
- Yu, A. Z.; Ronen, S.; Hu, K.; Lu, T.; and Hidalgo, C. 2016. Pantheon 1.0, A Manually Verified Dataset of Globally Famous Biographies.
- Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Zhao, X.; Akbaritabar, A.; Kashyap, R.; and Zagheni, E. 2023. A Gender Perspective on the Global Migration of Scholars. *Proceedings of the National Academy of Sciences*, 120(10): e2214664120.
- Zhao, X.; Aref, S.; Zagheni, E.; and Stecklov, G. 2022. Return migration of German-affiliated researchers: analyzing departure and return by gender, cohort, and discipline using Scopus bibliometric data 1996–2020. *Scientometrics*, 127(12): 7707–7729.

Ethics Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, see the Discussion.**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes.**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, see the empirical study of migration patterns in Scopus Data.**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, see Appendix**
 - (e) Did you describe the limitations of your work? **Yes, see the Discussion section.**
 - (f) Did you discuss any potential negative societal impacts of your work? **We believe this work does not have negative societal impacts.**
 - (g) Did you discuss any potential misuse of your work? **We believe this work can not be misused.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **We will release the model documentation, data, and code for nationality detection and will explain how to reproduce the findings.**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes.**

2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? NA
 - (b) Have you provided justifications for all theoretical results? NA
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
 - (e) Did you address potential biases or limitations in your theoretical framework? NA
 - (f) Have you related your theoretical results to the existing literature in social science? [Yes](#).
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? [Yes](#).
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? NA
 - (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes](#)
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes, see Appendix](#).
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No, while we did run some experiments with different random seeds, however, for larger models due to computational resource limitation, we weren't able to do so.](#)
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes, see Appendix](#)
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? [Yes](#).
 - (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? [Yes, see Appendix](#).
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [Yes, we have cited the models, datasets, and libraries. See also the Appendix](#)
 - (b) Did you mention the license of the assets? [Yes, the models, datasets, and libraries are all licensed for research and non-profit use and are publicly available.](#)
 - (c) Did you include any new assets in the supplemental material or as a URL? [Yes](#).
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA, we use publicly available data.
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes, see the Discussion](#).
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? [Yes, see the Discussion](#).
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? [We created the Hugging Face Hub ⁵ for models and datasets, making them publicly available.](#)
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots? NA
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
 - (d) Did you discuss how data is stored, shared, and de-identified? NA

Appendix

Data Preprocessing

The following preprocessing steps were applied to standardize and clean names:

- **Removal of Extraneous Information:** Nicknames, pronunciations, and metadata (e.g., content in parentheses, quotation marks, or following delimiters like commas or slashes) in Wikipedia introduction pages were removed. These elements are alternative spellings in other languages or pronunciation phonetics, and they introduce noise and are not part of the core name.
- **Normalization and Standardization:** Punctuation (e.g., periods, commas, quotes) was replaced with spaces, and common prefixes (e.g., “Dr”, “Sir”) and suffixes (e.g., “Jr”, “III”) were removed. Additionally, mojibake and Unicode problems were fixed by using the `ftfy`⁶ library, and Unicode normalization was performed using `unicodedata`⁷ to standardize encoding and ensure compatibility.
- **Validation and Cleaning:** Names containing invalid characters (e.g., symbols, digits) were removed, and whitespace was standardized.

During the preprocessing steps, only 0.15% of the names were identified as invalid and removed, while all remaining non-English characters were converted to their corresponding English equivalents or removed. Table 4 illustrates some examples of the changes made during the process.

⁵<https://huggingface.co/docs/hub/en/index>

⁶<https://ftfy.readthedocs.io/en/latest/>

⁷<https://docs.python.org/3/library/unicodedata.html>

Before Preprocessing	After Preprocessing
Jürgen Håholdt	jurgen huholdt
Román Montañez	roman montanez
Søren Hess-Olesen	sren hess olesen
Mónika Remsei	monika remsei
Rupert König	rupert konig
Miladin Ševarić	miladin sevaric

Table 4: Examples of names that were normalized and edited during preprocessing.

For country names, we standardized all country names based on the ISO 3166 country codes⁸. These steps ensure the dataset contains clean and consistent names.

Rationale for Country Inclusion and Exclusion

To justify the inclusion and exclusion of multicultural countries, we used two metrics to guide our decisions: (1) Entropy, which measures the uniformity of character n-grams in a country’s names, where higher values indicate greater naming diversity and overlap with other countries, and (2) F1 score, which reflects the model’s classification performance, with lower scores indicating poorer performance.

Table 5 shows that excluded countries such as the USA, Canada, Australia, New Zealand, and South Africa have higher entropy values and lower F1 scores, which negatively impact the classification performance of other countries, particularly the UK. In contrast, included countries like the UK, Germany, and France exhibit lower entropy and higher F1 scores, justifying their retention. Based on these findings, we excluded South Africa and New Zealand due to their high entropy and low F1 scores, while retaining the UK to balance diversity with classification performance. Additionally, we needed one country to represent the English category in our classification and selected the UK for this task. (Normalized entropy is the scaled entropy value, adjusted to a range between 0 and 1.)

Training Details

We implemented our work in Python, using various libraries for different models (which are listed in a “requirements.txt”). For FastText, we used its official Python library⁹ and trained it from scratch. For the neural networks (char-LSTM and char-CNN), we used PyTorch and TorchText. ByT5 was implemented using the Transformers and Datasets libraries. We used Optuna (Akiba et al. 2019) for hyperparameter tuning for all models except FastText, which has its own parameter-tuning function.

The best hyperparameters for char-CNN and char-LSTM include a learning rate of 0.005, weight decay of 1e-4, dropout of 0.2, embedding size of 100, and the AdamW optimizer (Loshchilov 2017) with a linear learning rate scheduler. For ByT5, the optimal learning rate is 3e-5, weight decay is 1e-7, and the AdamW optimizer is used without a learning rate scheduler. For FastText, we set the learning rate

⁸<https://www.iso.org/iso-3166-country-codes.html>

⁹<https://fasttext.cc/>

to 0.1, the embedding size to 100, and used an n-gram size of 5. FastText and ByT5 have their own tokenizer and character encoder, while for char-LSTM and char-CNN, we generated character embeddings using TorchText with a maximum length of 64.

Computational Resources

We used “NVIDIA GeForce GTX 1080 Ti” GPU servers for our model training experiments. To accommodate the large models, we used “gradient accumulation” and “mixed precision” techniques. Gradient accumulation helps manage memory usage by accumulating gradients over several mini-batches before performing an update, while mixed precision training optimizes performance and reduces memory consumption by using both 16-bit floating-point numbers.

Test Datasets

Figure 5 shows the distribution of classes at each of the three levels for the Athletes (top) and IUSSP (bottom) datasets. At the first level, East Asian and Hispanic are the largest classes in the Athletes and IUSSP datasets, respectively. The class distribution in all three levels differs between the two test sets and is distinct from the Wikipedia data used to train and test the model. Despite the differences, the evaluation metrics are similar in all three datasets (presented in the main text), indicating the robustness of our model.

Error Analysis

Here we clarify the errors and inaccuracies in our ML classification. We know that the ML models used have errors in classification and could still be improved, especially in the third level. However, we faced some misclassifications that could be due to other reasons. Below, we clarify examples of the observed errors. We explain potential reasons for these errors and also present our reasoning as to why the errors do not significantly affect our paper’s conclusions:

- **Misclassification of names from countries with cultural/historical similarity:** In some cases, the classification models confuse names from culturally or linguistically similar countries. For example, the names listed in Table 6 illustrate that Flemish individuals who, according to Wikipedia, have Belgian nationality might have names that are similar to those in the Netherlands. This could lead to the misclassification of their nationality to the Netherlands rather than Belgian, which is the true label based on Wikipedia data. Countries with historical or cultural relationships, such as Spain and Mexico, often have overlapping name structures, causing misclassifications. Another example is individuals of different nationalities who share the same name i.e., homonyms, such as “Alexander Stein¹⁰.” There are four Wikipedia pages with this name for distinct individuals with various nationalities: German, American, Latvian, and Russian.
- **Second generation migrants with names from the country of origin or changed nationality:** The second generation migrants sometimes have names that are

¹⁰[https://de.wikipedia.org/wiki/Alexander_Stein](https://de.wikipedia.org/wiki/Alexander__Stein)

	United States	Canada	Australia	New Zealand	South Africa	United Kingdom	Germany	France	average
Normalized Entropy	0.96	0.96	0.94	0.93	1.00	0.90	0.86	0.78	0.72
F1 score (Before excluding)	0.65	0.34	0.22	0.23	0.45	0.54	0.75	0.73	0.60
F1 score (After excluding)	-	-	-	-	-	0.77	0.78	0.76	0.63

Table 5: Normalized entropy values and F1 scores (before and after exclusion) for a few countries.

based on their parents’ country of origin. They might obtain the citizenship of the host country, which is reflected in the Wikipedia introduction page, while they maintain the same names. This could lead to misclassifications in the nationality detection models, as shown in the examples in Table 7.

- **Error in datasets based on Wikipedia:** We combined multiple data sources collected from Wikipedia, which included the names and nationalities of individuals. Upon reviewing the data, we can assert that most of the name-nationality pairs were correct; however, minor errors did exist. On the one hand, the Nationality Estimate dataset (Le et al. 2021) extracts information from Wikipedia descriptions to assign a country of origin to individuals whose pages lack explicit nationality details. However, this method can lead to errors when descriptions mention the country where the individual studied or lived instead of their actual country of origin. Table 8 illustrates examples of the inaccuracies in which the nationality was incorrectly stated in the data. On the other hand, about 50 percent of the nationality information in the Politicians dataset (Wagner 2017) was missing. To address this, we used the DBpedia API to fill in the gaps. We checked if the pages contained nationality or country information. If it was not available, we used the birthplace of the politicians as a substitute for their nationality. While this approach helped solve problems, it also introduced some errors, as shown by the examples in Table 9.
- **Imbalances and classes with a small number of members:** Due to the imbalanced nature of our training dataset (please see Figure 1 in the main text with the composition of classes in three levels), the model often predicts major classes when presented with data from classes with a small number of members, as it has not had sufficient exposure to learn to recognize them.

Name	Nationality in Wikipedia	Predicted nationality
Bob van Laerhoven	Belgium	Netherlands
Marc van den Abeelen	Belgium	Netherlands
Jan van Hembyze	Belgium	Netherlands

Table 6: Examples of misclassification of Flemish people with Belgian nationality to the Netherlands.

Regarding the impact of the errors on our findings, we used a large corpus of 2.6 million name-nationality pairs, which ensured that the model was trained on diverse data, which is substantially larger than previous studies. Despite

Name	Nationality in Wikipedia	Predicted nationality
Bhavna Limbachia	United Kingdom	India
Mahmoud Dahoud	Germany	Egypt
Xu Pei	Germany	China

Table 7: Misclassification of second-generation migrants due to names similar to country of origin or changes in nationality in Wikipedia.

Name	Nationality	Nationality in our dataset
Arthur Glasser	USA	China
Daniel Bayerstorfer	Germany	China
Christopher Ludwig Eisgruber	USA	India

Table 8: Errors in the training dataset because of incorrect nationality extraction from Wikipedia page descriptions.

Name	Nationality	Birthplace
Georg Heinrich Schnell	Germany	China
Mathangi Arulpragasam	Sri Lanka	UK

Table 9: Errors in the training dataset because of missing nationality information and using Birthplace instead.

errors in certain country pairs, the model’s performance remained consistent, as demonstrated by the stable results on both the training and out-of-domain test sets.

Our main finding in the Scopus analysis predominantly highlights specific origin-destination country pairs (or migration corridors) with the highest number of scholars moving between them, such as the USA-China corridor. For these highly travelled corridors, our model has a very reliable performance. Additionally, our model demonstrated strong performance at the first and second levels. However, future studies aiming to differentiate names from culturally and linguistically similar countries, such as Spanish-speaking countries or Flemish regions, should incorporate additional information, as name-based nationality detection alone may not be sufficient.

Assessment Per Class

We examined the model’s performance in classifying at levels 1 (see Figure 6) and 2 (see Figure 7). The assessment criteria used are precision, recall, and F1 score. It is evident that the model demonstrates higher precision compared to recall in both levels. This indicates that the model accurately predicts each class. Moreover, for classes with substantial

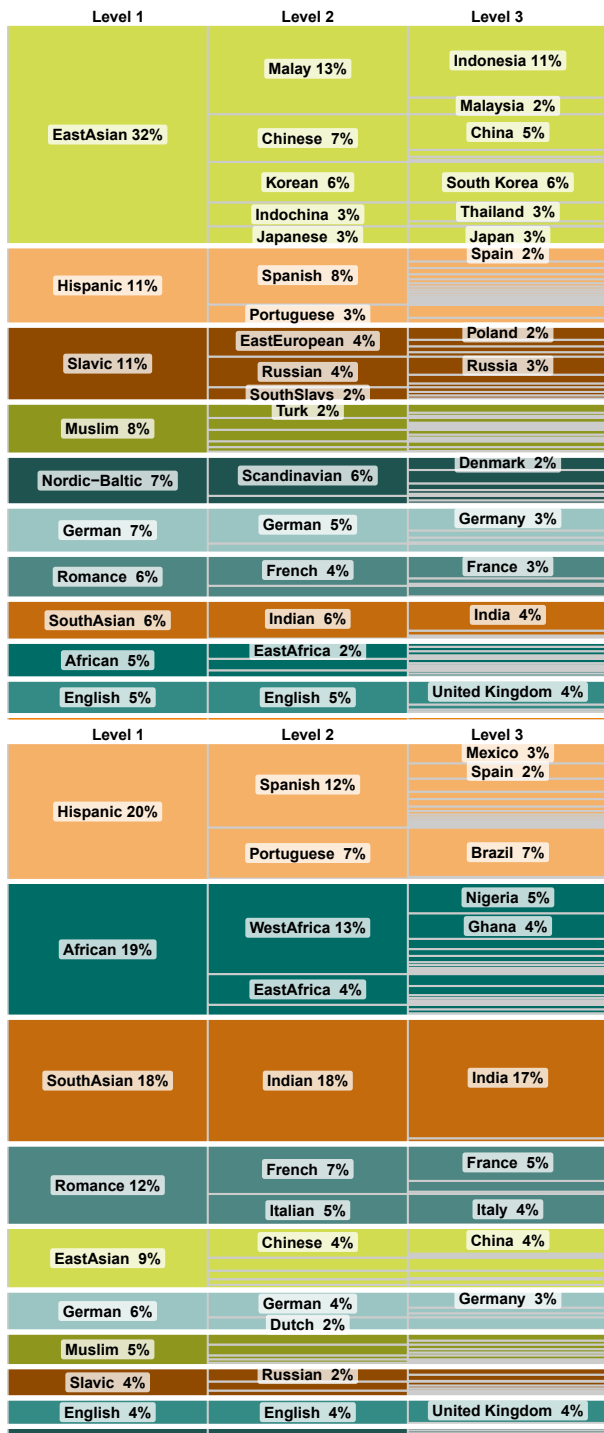


Figure 5: Distribution of nationality classes in the Athletes (top) and IUSSP (bottom) out-of-domain test datasets. Each column sums up to 100% and represents one of the three classification levels, from the left (broader) to the right (more detailed). All classes are shown, but only those bigger than 2% of the total are labeled.

data, such as European countries and regions, the average F1 score reached approximately 80%. Conversely, classes with limited data have an F1 score of about 60%. Although Korea, Indochina, and Greece have less data, they achieved a precision rate of around 80%, demonstrating the model’s ability to recognize names from these countries, perhaps due to distinct patterns in names from these countries. This assessment shows for which regions and groups of countries our model is the most reliable.

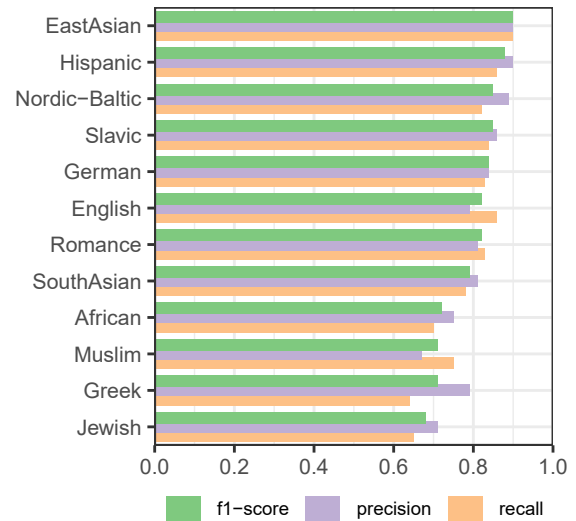


Figure 6: Evaluation metrics for the first level classification

We also examined the countries for which our model is most and least reliable. Considering only countries with at least 200 names in the Wikipedia test subset, the F1 score ranges from 05% for Montenegro to 94% for Japan. This high variability across countries suggests that the model is appropriate for analyzing some migration corridors but may not be reliable for some origin and destination pairs. Figure 8 shows the 10 countries with the highest and lowest F1 scores. The variability in the performance of the model at the country level exists even between countries within the same class at the first and second levels. While some first-level classes are only present in the top 10, such as East Asian with 4 out of 10 countries, or in the bottom 10, such as Hispanic with 6 out of 10, other countries with Slavic names appear in the top and bottom 10.

Figure 9 shows the confusion matrices for the first (top left), second (top right), and third levels (bottom). At the first level, about 87% of the names for which the true value is Hispanic were predicted as Hispanic, showing that the model can identify names of Hispanic origin. At the second level for Hispanic names, the model correctly predicted more than 80% of Spanish and Portuguese names, but the predicted values for countries with Hispanic Southern African names, i.e., Angola and Mozambique, are distributed across Spanish, Portuguese, and Hispanic Southern African classes.

Our model can correctly identify a name with a Hispanic pattern and even distinguish between Spanish and Por-

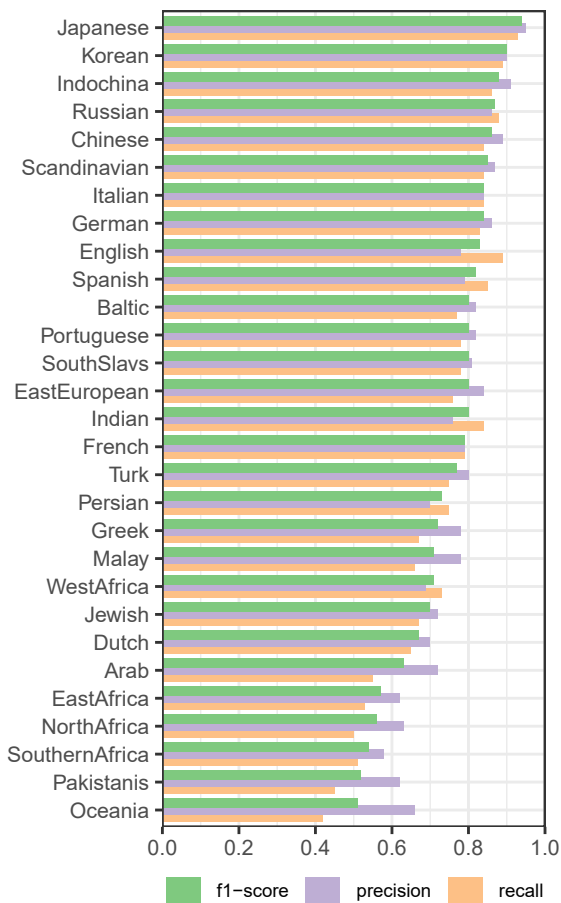


Figure 7: Evaluation metrics for the second level classification

tuguese names, but the model cannot correctly identify the country within the second-level class. The same applies to English names; the model can determine the English origin but not the specific country. In both examples, the model cannot correctly identify the countries because they have similar names, and there are imbalances and classes with a small number of cases at the country level. On the one hand, the misclassification of names from countries with cultural/historical similarity leads to a spread of predicted values across multiple countries; this pattern is clear for most of the Spanish countries in the country-level confusion matrix (see the first rows of the bottom matrix in Figure 9). On the other hand, the imbalance and minor classes lead the model to predict the same country for different true values, and other countries are not predicted even when the name is from that country. (See the first and last two columns of the Hispanic class in the bottom matrix in Figure 9).

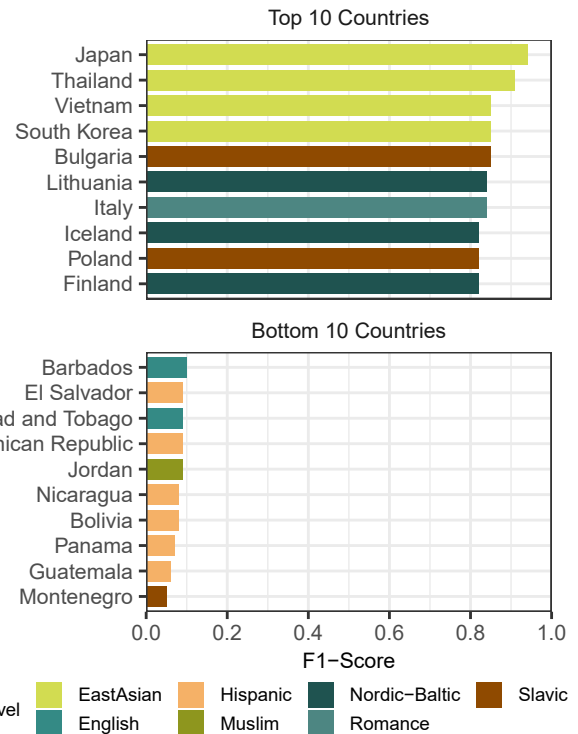


Figure 8: F1 score for the 10 countries with the highest scores (top) and 10 countries with the lowest scores (bottom). Only countries with more than 200 names in the Wikipedia test subset were considered.

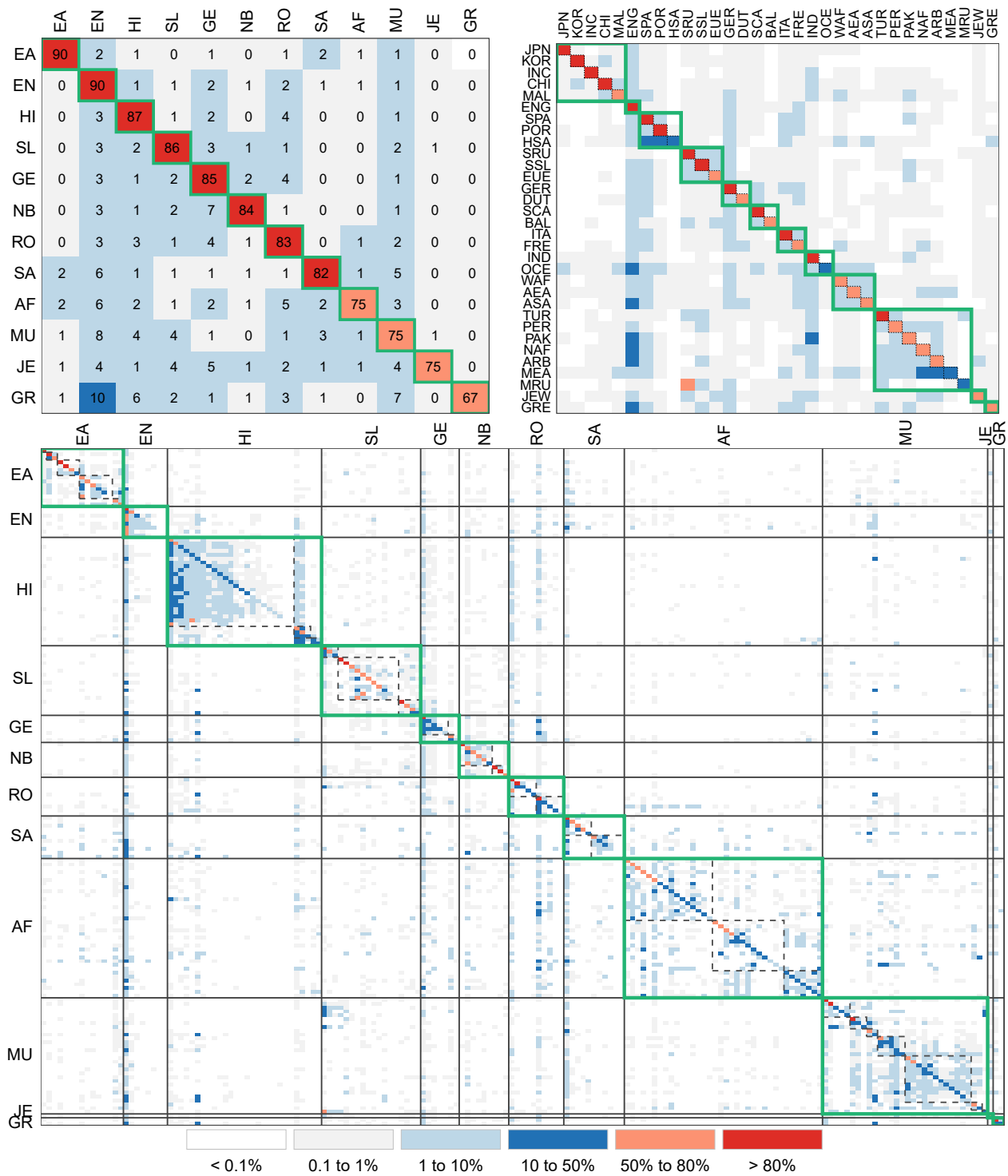


Figure 9: Confusion matrices for the first- (top left), second- (top right), and third-level classes (bottom). The rows represent the true and the columns show the predicted values. Colors represent the percentage of observations in each cell, and each row sums up to 100%. In all matrices, the green rectangle highlights the cells for which the true value is equal to the predicted value at the first level, and the dashed rectangle highlights the classes at the second level. Labels used for the first level: African (AF), East Asian (EA), English (EN), German (GE), Greek (GR), Hispanic (HI), Jewish (JE), Muslim (MU), Nordic-Baltic (NB), Romance (RO), Slavic (SL), and South Asian (SA). Second level labels: Arab (ARB), Baltic (BAL), Chinese (CHI), Dutch (DUT), East Africa (EAF), East European (EUE), English (ENG), French (FRE), German (GER), Greek (GRE), Indian (IND), Indochina (INC), Japanese (JPN), Jewish (JEW), Korean (KOR), Malay (MAL), North Africa (NAF), Oceania (OCE), Pakistanis (PAK), Persian (PER), Portuguese (POR), Russian (RUS), Scandinavian (SCA), Southern Africa (SAF), South Slavs (SSL), Spanish (SPA), Turk (TUR), West Africa (WAF)