Large-Scale Diversity Estimation Through Surname Origin Inference

Bulletin de Méthodologie Sociologique
2018, Vol. 139 59–73

DOI: 10.1177/0759106318778828
journals.sagepub.com/home/bms
SAGE

Antoine Mazières
Centre Marc Bloch, Germany and UMR-LISIS, INRA, France

Camille Roth
Sciences Po, France and Centre Marc Bloch, Germany

Abstract
The study of surnames as both linguistic and geographical markers of the past has proven valuable in several research fields spanning from biology and genetics to demography and social mobility. This article builds on the existing literature to conceive and develop a surname origin classifier based on a data-driven typology. This enables us to explore a methodology to describe large-scale estimates of the relative diversity of social groups, especially when such data is scarcely available. We subsequently analyze the representativeness of surname origins for 15 socio-professional groups in France.

Résumé
L'étude des noms de famille comme marqueurs linguistiques et géographiques du passé s'est avérée pertinente dans des contextes variés allant de la biologie et la génétique, à la démographie et la mobilité sociale. Cet article construit, en s'inspirant d'éléments de la littérature existante, un classifieur des origines des noms de famille basé sur une typologie empirique. Cela nous permet d'explorer ainsi une méthodologie destinée à estimer, à grande échelle, la diversité relative des groupes sociaux, en particulier lorsque les données nécessaires sont difficilement accessibles. L'article enfin propose une première analyse de la représentativité des origines de noms de famille de 15 groupes socio-professionnels en France.

Corresponding Author:
Antoine Mazières, Centre Marc Bloch Berlin Friedrichstr. 191, Berlin 10117, Germany
Email: antoine.mazieres@gmail.com

Keywords
Onomastics, machine learning, diversity, representativeness, geographical origins

Mots clés
Onomastique, apprentissage machine, diversité, représentativité, origines géographiques

Introduction

Surnames have the objective property of designating a path in the ancestry tree, up to a point in time and space where the name was first coined and made hereditary. While they are usually distant markers of a historical and geographical context, surnames still exhibit connections with present features and have thus been considered as a valuable proxy in population studies. For one, surnames correlate with genetic proximity within populations (Jobling, 2001; King et al., 2006; Lasker, 1985) and have been diversely used to analyze human population biology (Lasker, 1980), identify cohorts of ethnic minority patients in bio-medical studies (Choi et al., 1993; Polednak, 1993; Shah et al., 2010), improve research in genealogy (King and Jobling, 2009) or describe the migration rates of human populations (Piazza et al., 1987). Social sciences more recently made use of surnames to statistically and indirectly appraise the composition of populations in various situations (Mateos, 2007, 2014), including the demography of online (Chang et al., 2010; Mislove et al., 2011) and research (Wu et al., 2014) communities, or the history of social mobility (Clark, 2014; Güell et al., 2015). The purpose of the present article is twofold. First, it aims at assessing the possibility of building a general-purpose, worldwide surname origin classifier. Our approach combines elements which are already available in the literature, and endeavors at both enhancing the learning data quality and broadening the geographical breadth and universality of surname origin typology. Second, we use this classifier to show that, despite its limitations at the individual level, it nonetheless enables simple and pertinent applications to the estimation of representation biases in origins in populations where no such data is explicitly available. We further illustrate its potential relevance for discrimination studies by comparing surname origin distributions for various sets of occupational groups and exam candidates in France.

Statistically Inferring a Surname Origin

Surname Origin vs. Ethnicity

Our approach relies essentially on the notion of surname *origin* rather than ethnicity. Indeed, ethnicity is often defined (Barth, 1998; Tonkin et al., 2016; Weber, 1978) as a *subjective* feeling of membership to one or several groups or self-defined identities, composed of linguistic, national, regional and religious criteria. A quick glance at the present article's bibliography reveals how much the academic literature aimed at inferring information from surnames relies on ethnicity to put names and individuals into groups, and derive subsequent analyses.

By contrast, a surname objectively corresponds to a genealogical and traditionally patrilineal path whose origin coincides with the first appearance of this socially hereditary property in the family tree. These moments vary much from one region to another, spanning from about 5,000 years ago in China to less than a century ago in Turkey.

Over 20 generations, the unique path of a name is one among more than a million (for about double the ancestors). Thus, in a randomly mating population, i.e. without any kind of endogamy, this marker would assuredly carry extremely little information: given these figures, someone bearing a surname of a specific origin would not be more likely to exhibit characteristics found in other bearers of a surname of the same origin. However, the existence of a strong endogamy among humans – albeit probably decreasing (Rosenfeld, 2008) – entails a correlation between surnames and the preferences that characterize this endogamy: geographical proximity, social and economic status, languages, political, genetic, regional and religious criteria. Put simply, as a result of, say, geographical endogamy, the correlation between the geographical origins of the father and the mother of a person induces a correlation between the geographical origins of their surnames, whereby the father name partly informs on the geographical origin of the mother. This phenomenon is likely the common cause behind the significance of the results found in the above-cited studies.

With this in mind, ethnicity appears as a potentially uncertain detour through a context-dependent and highly subjective matter, while the reference to an origin offers a more objective description of the variations in features extracted from surnames. To speak of origins nonetheless demands that we make a decision on how we partition the world into distinct regions. At the very low level, to make matters simple and comparable, we first decided to use the present-day list of countries, acknowledging that no spatial or temporal partition of the world would be likely to take into account the wide diversity of overlaps between territories and populations at various points in time.

Crafting the Learning Data

How could we, humans, be able to form an intuition on the origin of some surnames? If one has never encountered the name ‘Toriyama’, one might still correctly make a guess on its Japanese origin, for instance because of the way it sounds when being pronounced, or the pattern of letter ordering. This admittedly hints at the existence of a second, closely-related proxy: surnames were originally coined (and have also been modified) by speakers belonging to a given linguistic space. Some structural and recurrent linguistic properties are more likely to be found in surnames of the same origin.

Thus, we aim at creating a classifier able to infer sufficiently well the probable origin of a surname from its spelling. To take a simple example, the distribution of letters in a text usually yields a good prediction of its language, assuming sufficiently many words and prior knowledge of empirical distributions for a set of languages. While it would be ambitious to expect a decent precision from surname single letter distributions, the use of *subsets* of letters, including morphemes, appears much more promising. To define *learning features*, we thus decompose all surnames into various subsets of letters of size n, or “n-grams”. This eventually constitutes the feature set for the whole dataset. We then describe a given surname by its distribution on these features.

Building a statistical model able to reproduce the above intuition at large scale for all origins means that we must first fit the model by using a large and diversified number of surnames labeled with their origins, or training dataset. To gather such learning examples, previous works relied on a variety of explicitly labeled sources including census data (Mislove et al., 2011), Olympic Games participant records (Lee et al., 2017), phone books (Mateos, 2014) or even Wikipedia data (Amberkar, 2009).

Another study used the PubMed search engine to extract scientific bibliographical records (Torvil and Agarwal, 2016). We follow a similar approach since this open data source[1] enables easy reproducibility of our research and provides an extensive volume of references with more than 25 million publications. For each record, we extracted author surnames and their affiliations when they were related to one of the 176 countries of the Natural Earth dataset.[2]

We assume that surnames whose affiliation distribution is heavily peaked for a given country are more likely to originate from that country. However, using PubMed data suffers from several biases, among which:

- The increased nomadism of the scientific population, lowering the quality of the affiliation as a reliable origin.
- The heterogeneous academic activity of countries, over-sampling the most productive ones at the expense of others.
- The potential bias of medical publication databases in favor of Anglo-Saxon publication venues (Nieminen and Isohanni, 1999), under-sampling the rest of the world.

A first obvious step for counterbalancing these biases consists in considering surname frequencies, i.e. normalizing surname occurrences in a given country by the total number of occurrences for that country. Then, in an effort to restrain our training dataset to true positives, we use a measure of statistical dispersion, the Herfindahl–Hirschman Index (HHI) (Herfindahl, 1950; Hirschman, 1945), to identify names whose presence is highly concentrated in one country only. We request a HHI of at least 0.8 as well as a maximal frequency over all countries of at least 0.0001%. Even though this method eliminates some of the most common names, for they are susceptible to have spread all over the world, it narrows our focus to a set of about 650k surnames which we call "core names" and which we assign to the country where frequency is maximal.

A Data-Driven Typology of Surname Origins

Nonetheless, the number of these core names remains unevenly distributed across countries, partly as a result of the above-mentioned under-sampling. It goes from 163 names for Montenegro to 41k names for Spain, with an overall average of 5,145. Before training our model, we thus need to introduce coarser categories to achieve a minimal significance for each geographic area.

Keeping in mind the eventual goal of appraising over- and under-representation of origins in socio-professional groups, we conservatively decide to categorize countries into a relatively small number of world regions. To do so, we first cluster countries

according to the training features. More precisely, we created a large "country/n-gram" matrix whose rows are countries and columns are n-grams of core names: a cell indicates the frequency of a given n-gram among the core names of a given country. We then performed hierarchical clustering on this matrix using Ward linkage (Ward, 1963). This yields the dendrogram shown in Figure 1 from which we may extract 7 rough categories of surname origins. We concretely aggregate countries by following the dendrogram in a monotonous manner from the bottom to the top while avoiding to merge categories belonging to strongly unrelated geographical areas. This process creates what appears to be an interpretable regionalization of the world at the cost of a very limited number of inconsistencies.

We re-label the original 'surname-country' associations according to these clusters. We eventually train a classifier on this new 'surname-world region' dataset, using the same learning features. Broadly, a classifier is a model (and, in practice, a function) which takes as inputs the learning features for a given observation (in our case, a surname and its letter subsets) and outputs a guessed label (in our case, an origin under the form of a world region).

The state of the art features a variety of methods such as Hidden Markov Models (HMM) and decision trees (Ambekar et al., 2009), recurrent neural networks (Lee et al., 2017) or logistic regressions (Torvik and Agarwal, 2016). We focus on one of the most classical classifiers, called Naive Bayes, which in our case yielded the best overall results among a variety of other traditional approaches available in the Scikit-learn (Pedregosa et al., 2011), the python classification algorithm library we used.[3] Naive Bayes is a simple classifying technique consisting in estimating the probability that an object belongs to a certain class given a set of observed features. It applies the Bayes theorem on the probabilities that surnames exhibit certain features knowing that they belong to some origin. It additionally relies on the assumption that these features are statistically independent, i.e. the contributions of each of these features to the target probability are independent from one another, hence the 'naive' qualification. In practice, we train the model on about 85% of the core name dataset while keeping aside about 15% of the core name dataset to evaluate model performance.

Classification performance is shown in Table 1 and is expressed in terms of *precision* and *recall*, along with the corresponding set sizes. For instance, the model achieves a precision of 61% for Asian and a recall of 77%, meaning that 61% of names guessed as 'Asian' belong to the Asian cluster, while 77% of names belonging to the Asian cluster are correctly guessed (recalled by the model) as 'Asian'. Success differs significantly from one class to another, with very satisfying results for the Central/South European and Slavic clusters and quite moderate performance for the African cluster. How much of this error is due to the lack of academic data in certain areas or the difficulty to identify pattern in surnames of a specific area is yet to be determined.

Notwithstanding, since we are interested in comparing the over- and under-representation of surname origins *between* two socio-professional populations of a given country, we contend that this type of error does not significantly jeopardize our aim. We first postulate that classification errors for a given surname origin remain homogeneous from one dataset to the other, i.e. that the names of a given origin are globally going to be classified (and misclassified) with the same success in both datasets. In other words,

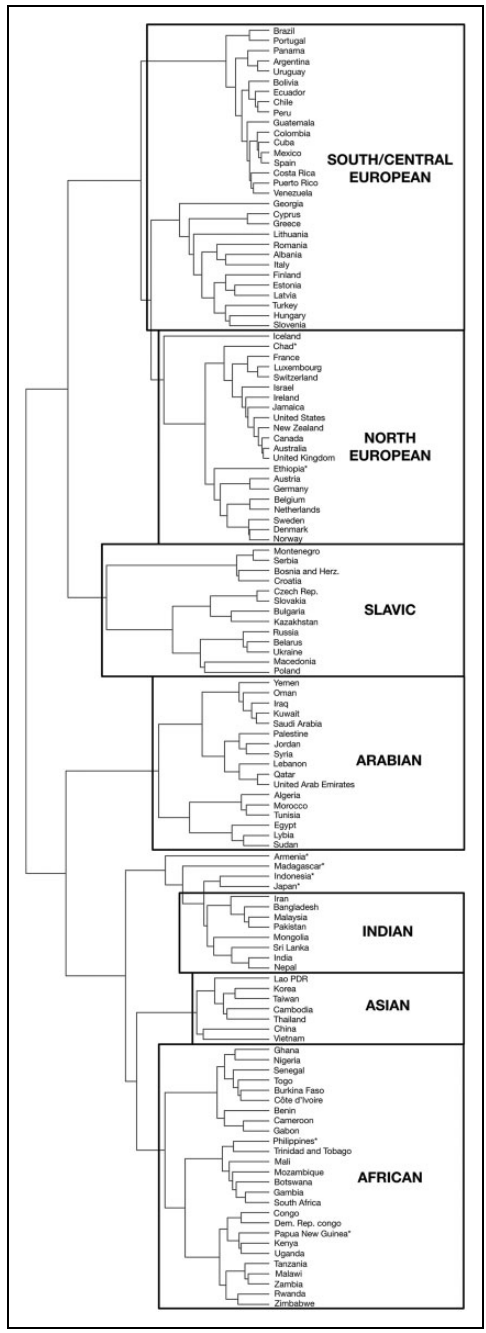

Figure 1. Cluster of surname origins. Countries marked by a star (*) are interpreted as misclassified and reassigned in the following manner: Philippines, Japan and Indonesia are reassigned to the Asian cluster; Ethiopia to Africa; Papua New Guinea , Madagascar, Jamaica, Chad and Armenia are deleted from the dataset as they represented a very low number of initial observations.

Table 1. Number of core names (totals, while around 15% are used for the evaluation) and classifier performances for each cluster in terms of precision and recall

	Core names		Class. Perf.	
Cluster	*Total*	*Evaluation*	*Precision*	*Recall*
African	30 748	4 529	0.43	0.61
Arabian	31 272	4 596	0.52	0.72
Asian	44 658	6 754	0.61	0.77
CS-European	189 624	28 668	0.81	0.71
Indian	68 145	10 067	0.63	0.72
N-European	216 465	32 469	0.78	0.62
Slavic	65 259	9 843	0.64	0.84
Total	*646 171*	*96 926*		

irrespective of their proportion within a given dataset, we assume that all surnames of, say, Indian origin, will be as often correctly recalled by our algorithm as Indian in all datasets, i.e. 71.8% of the time (and errors will be distributed across other origins in similar proportions for all datasets). Put differently, we suppose that names which pose inference problems w.r.t. our model are roughly distributed homogeneously and are not biased across datasets (for example, if 'Toriyama' is misclassified, we assume that it is no more or less present among Asian names in one dataset than in another one).

We nonetheless have to consider that classification errors vary across origins. This is shown by the confusion matrix on Table 2. Here, names of Arabian origin are guessed as Asian 2.46% of the time, while it is about 7.04% for names of African origin. Even if the above assumption enables us to use the same confusion matrix for all datasets, we still have to adjust guesses knowing that the algorithm exhibits some propensity to over-/under-estimate depending on the origin. In other words, knowing that a proportion of names which *actually* belong to a given origin are *guessed* as belonging to another origin, we correct guesses to infer back the probability of actual origin for a given guess $P(\text{actual} = j \,|\text{guessed} = i)$. In practice, we multiply guessed numbers of surname origins by this probability which we extract from C by Bayesian inference.[4]

Estimating Origin-Based Discrimination in France

Datasets and Estimation Methodology

We now illustrate the method on 15 datasets representing various areas of French society (see Table 3). Three datasets are linked to political functions (Mayors, Parliament Members and Senators), five of them represent various types of occupations (Pharmacists, Lawyers, Accountants, Veterinarians, Researchers), and six are made of lists of candidates to various state exams (*Brevet*, *Baccalauréat*, BEP, CAP, BTS, Professional Baccalauréat). The *Ecole Polytechnique* dataset lists students at one of the most highly-ranked engineering schools in France.

From the list of surnames of each dataset, we apply the classifier to obtain vectors of values representing the guessed distributions of surname origins according to our

Table 2. Confusion matrix C. This matrix shows the number of names from the evaluation sets (see Table 1) of an actual origin (in Columns) which are guessed as belonging to a given origin (in rows). The first sub row indicates total numbers, the second sub row refers to proportions within an actual origin

Guessed origin	Actual origin						
	African	Arabian	Asian	CSE	Indian	NE	Slavic
African	**2763**	165	381	1081	460	1441	157
%	***61.0***	*3.59*	*5.64*	*3.77*	*4.60*	*4.44*	*1.60*
Arabian	159	**3292**	84	577	598	1549	77
%	*3.51*	***71.6***	*1.24*	*2.01*	*5.94*	*4.77*	*0.78*
Asian	319	113	**5200**	831	716	1147	174
%	*7.04*	*2.46*	***77.0***	*2.90*	*7.11*	*3.53*	*1.77*
CS-European	258	128	274	**20364**	299	3535	324
%	*5.70*	*2.79*	*4.06*	***71.0***	*2.97*	*10.9*	*3.29*
Indian	273	487	420	991	**7226**	1862	191
%	*6.03*	*10.6*	*6.22*	*3.46*	***71.8***	*5.73*	*1.94*
N-European	643	351	315	3254	609	**20183**	670
%	*14.2*	*7.64*	*4.66*	*11.4*	*6.05*	***62.2***	*6.81*
Slavic	114	60	80	1570	159	2752	**8250**
%	*2.52*	*1.31*	*1.18*	*5.48*	*1.58*	*8.48*	***83.8***
	4529	4596	6754	28668	10067	32469	9843
	100	*100*	*100*	*100*	*100*	*100*	*100*

typology. Note that this approach works by construction at the level of groups and may not be used at the level of individuals: to take an example from a distinct context, if we know that the given name 'Camille' is about 80% of the time a female name, we are not able to draw a precise conclusion on the gender of a given Camille, while we can say that a group of 100 Camille is likely to be around 80% female.

In order to show how the diversity in terms of surname origins of certain subgroups of the population departs from that of a common reference point, we rather focus on dataset-to-dataset comparisons rather than raw distributions. In other words, comparing surname origin distribution across datasets enables us to assess the extent and magnitude of the divergence in the representativeness of groups of people with a given surname origin and, more broadly, the fact that some datasets and some origins exhibit the same pattern of divergence, likely indicative of similar underlying processes. There is no public and unbiased source of data which covers surnames of the French population in order to perform such comparisons. Therefore, we chose the *Brevet* dataset as a point of comparison since it represents the most widely passed exam in France and therefore, a wide sample of people who lived in France and were generally aged 14–15 as of 2008, hence 23–24 as of 2017. As such, it is also likely to exhibit a bias towards younger people.

Table 3. List of datasets along with the corresponding numbers of observations

Name	List of Surnames of all . . .	Nb. Obs.
Brevet	Candidates to *Diplôme national du brevet* in 2008[5]	562,952
Baccalauréat	Candidates to the nationwide *Baccalauréat* (*général* and *technologique*) in 2008	435,645
BEP	Candidates to *Brevet d'Etudes Professionnelles* in 2008	116,814
CAP	Candidates to *Certificat d'aptitude professionnelle* in 2008	98,364
BTS	Candidates to *Brevet de technicien supérieur* in 2008	87,917
Professional *Baccalauréat*	Candidates to *Baccalauréat* professionnel in 2008	80,672
Pharmacists	Pharmacists registered in their *Ordre professionnel* in 2017[6]	73,422
Mayors	Mayors of French cities (*communes*) in 2014[7]	36,628
Parisian lawyers	Layers registered in the Parisian bar association in 2017[8]	32,021
Ecole Polytechnique	Students at *Ecole polytechnique* (1958-2016)[9]	23,058
Accountants	Accountants registered in their *ordre professionnel* in 2017[10]	20,946
Veterinarians	Veterinary physicians registered in their *ordre professionnel* in 2017[11]	15,710
Researchers	Researchers at the *Centre national de la recherche scientifique* in 2017[12]	12,657
Parliament Members	Parliament members of *Assemblée nationale* (1958-2016)[13]	8,326
Senators	Senators of the French Fifth Republic (1958-2017)[14]	1,710

Simply calculating the ratio between each target dataset and *Brevet* yields the results shown in Figure 2, which enables the observation of several profiles of representativeness among the datasets described in Table 3. As such, values higher (resp. lower) than 1 correspond to surname origins which are over-represented (resp. under-represented) compared with their presence in *Brevet* (logically, *Brevet* exhibits a flat profile where all origins have a ratio of 1). Of course, these ratios do not render the fact that some categories are significantly more populated than others: this is typically the case for 'North European', which is the most common surname origin found in these French datasets. As a result, large under- or over-representation of less populated categories may have a relatively marginal effect on the over- or under-representation of the most populated category. The graphs of Figure 2 should be read with this provision in mind: because of their sheer presence in all datasets, North European surnames ratio are often grouped around 1, while other categories may vary significantly below or above 1. In other words, these ratios tend to emphasize the over- or under-representation of minority categories, rather than the strong presence of the majority category – this can prove useful in the context of discrimination studies.

Besides, datasets and origins can be grouped according to the similarity of their divergence profiles, using for instance a simple hierarchical clustering based on the Canberra distance. Graphs in Figure 2 have been organized according to this proximity in order to display and emphasize datasets or origins behaving in a comparable manner.

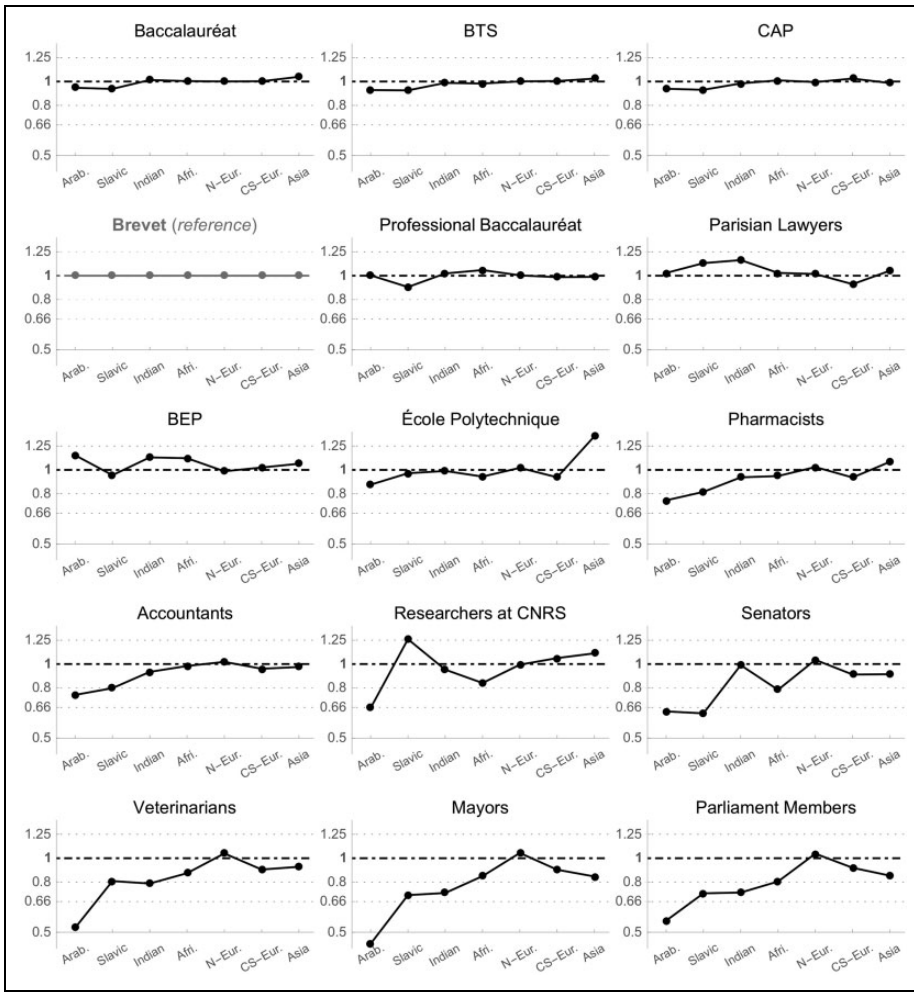

Figure 2. Over-/under-representation of surname origins among all datasets. Each graph shows the ratios between the target dataset and the reference dataset (*Brevet*) for each origin category. A logarithmic scale is used to depict equivalent over- or under-representation ratios at equal distance from the y=1 reference line.

Preliminary Results: Observations

While we do not aim at discussing in detail the implications of such and such bias in some dataset, we may emphasize a few trends to illustrate the interpretation of the results.

All elective political functions (Mayors, Parliament Members and Senators) together with Veterinarians, exhibit a marked over-representation of Northern European surnames. On the other hand under-representation, when it appears in these four datasets, is much more pregnant than in other datasets. It is actually spread among the

remaining origins by following a comparable pattern across all four datasets, with Arabian names being the most significantly affected, closely followed by Slavic, Indian and African surnames.

State exams show an overall smoother profile, with *Baccalauréat*, BTS and CAP having the almost exact same distributions while professional *Baccalauréat* and BEP display slightly different configurations. Interestingly, some datasets exhibit specific over-representation peaks for a single surname origin, such as Asian for *Ecole Polytechnique* and Slavic for Researchers.

Some additional patterns emerge by examining these results along origins rather than datasets. For one, the under-representation of Arabian names is constant across all datasets, to the exception of BEP. Surnames of Asian origin are generally under-represented in elective functions, while their strongest over-representation occurs for two groups related to higher education, *Ecole Polytechnique* and Researchers. As said above, North European surnames represent, in absolute numbers, the bulk of inferred origins. Their representativeness is generally close to 1, indicating no remarkable *relative* variation across datasets, from *Brevet* to Parliament Members, even though the ratio rises slightly above 1 for the last four datasets, possibly as a result of the strong under-representation of other origins.

Contribution Scope and Future Work

There is an undefined leap between the statistical observation of representativeness and fairness, between under-representation and discrimination and between over-representation and privilege. For instance, while we can say that a discrimination often implies an under-representation, the inverse is not necessarily true (Jobard and Nevanen, 2007) and discrimination is usually evaluated on multiple complementary dimensions (Delattre et al., 2013), both qualitative and quantitative. Moreover, our method does not yet take into account other socio-demographic variables to control for the existence of common causes to the results. This would make it possible to say that, for instance, there is more of such surname origin in a given dataset because such origin is over-represented in such socio-demographic segment, which is itself over-represented in the population of the said dataset. Taking for example the most under-represented case in our results – mayors with Arabian surnames – one may conclude that it illustrates a well-known discrimination in France (Cediey and Foroni, 2007; Foroni et al., 2016) towards people of Arabian origins in elective functions. However, it is unclear how much of this ratio may be explained by discrimination or, for instance, by an uneven geographical presence of a given group of immigrants, or descendants thereof (Brutel, 2016).

The results presented here should be further examined and perhaps challenged both for their statistical significance and historical relevance. In this respect, our article simply acknowledges that the study of representativeness is part of discrimination studies, whereby methods for large-scale estimation of the former contribute to the latter. Our results could be nuanced with expertise of the demography of each socio-professional group considered here together with in-depth knowledge of the history of colonization and immigration in France (Beauchemin et al., 2016).

Finally, while we applied our methodology to datasets representative of certain groups in French society, comparisons with other contexts (countries, world regions, transnational entities) with the help of relevant surname datasets could yield fruitful insights on both our method and on possible interpretations.

Concluding Remarks

The aim of this article lies in demonstrating the feasibility of a technique of estimation of representativeness based on a combination of open data sources, in contexts where data explicitly documenting individual origins may be difficult to process. We endeavored at showing that these methods can work in the absence of public data and/or data specifying distribution priors (Chang et al., 2010) or *a priori* ethnic taxonomies (Ambekar et al., 2009).

By making the model available to anyone and relying on data open sources, we hope to encourage further exploration and improvements of such techniques, especially in the context of discrimination studies and the discussion of the specific biases corresponding to present and future datasets.

Acknowledgements

The authors would like to thank Telmo Menezes, Mikaela Keller, Elian Carsenat, Jean-Philippe Cointet, Élise Marsicano, Fabien Jobard, Jérémy Levy and Mélanie Bourgeois for their help with this research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article has been partially supported by the 'Algodiv' grant (ANR-15-CE38-0001) funded by the ANR (French National Agency of Research).

Notes

1. Using the query 1800:2020[dp] on https://www.ncbi.nlm.nih.gov/pubmed/
2. Natural Earth Data, 1:110m Cultural Vectors, http://www.naturalearthdata.com/downloads/110m-cultural-vectors/
3. We concretely apply a multinomial naive Bayes model with an additive (Laplace/Lidstone) smoothing parameter of 0.1. A programming notebook is available to observe and reproduce all steps described here: https://namograph.antonomase.fr/
4. More precisely, we compute P (actual = i|guessed = j) as $C_{ij}/\sum_{j'} C_{ij'}$, Moreover, since the confusion matrix is computed using prior proportions of surname origins extracted from Pubmed, it is likely to be based on priors which very significantly diverge from the average proportions of surname origins in the "general" French population. To accommodate for the Pubmed bias as much as possible, we adjust the priors of the confusion matrix so that they match a distribution guessed initially by the uncorrected classifier on the *Brevet* dataset. This uncorrected distribution yields respectively 4.8, 8.3, 3.1, 20.7, 3.4, 57.1 and 2.6 % for each of

the origins: African, Arabian, Asian, Central SE, Indian, NE, and Slavic. We thus correct the original confusion matrix of Table 2 by making column sample sizes proportional to these figures. In other words, the confusion matrix that we eventually use exhibits a structure more similar to that of the initially guessed *Brevet* proportions than the Pubmed ones.

5. Source for all 2008 exams: http://www.bankexam.fr/resultat/2008
6. Source: Online directory of the Ordre National des Pharmaciens: http://www.ordre.pharmacien.fr/annuaire/pharmacien
7. Source: French government open data repository: https://www.data.gouv.fr/fr/datasets/liste-des-maires-au-17-juin-2014/
8. Source: Online directory of the Parisian Bar Association: http://www.avocatparis.org/annuaire
9. Source: Alumni online directory of *Ecole Polytechnique*, https://www.polytechnique.org/search
10. Only independent, salaried and honorary accountants. Source: Online directory of the *Ordre National des Experts-Comptables*: http://www.experts-comptables.fr/annuaire
11. Source: Online directory of the *Ordre National des Vétérinaires*: https://www.veterinaire.fr/outils-et-services/trouver-un-veterinaire.html
12. Only tenured researchers. Source: CNRS Online directory: https://annuaire.cnrs.fr/l3c/owa/annuaire.recherche/index.html
13. Source: French National Assembly online database: http://www.assemblee-nationale.fr/sycomore/liste_legislature.asp?legislature=48
14. Source: Bibliographical notices of French Senators, https://www.senat.fr/elus.html

References

Ambekar A, Ward C, Mohammed J, Male S and Skiena S (2009) Name-Ethnicity Classification from Open Sources. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 49-58.

Barth F (1998) *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. Illinois: Waveland Press.

Beauchemin C, Hamel C and Simon P (2016) *Trajectoires et origines: enquête sur la diversité des populations en France*. Paris: Ined Editions.

Brutel C (2016) La localisation géographique des immigrés: une forte concentration dans l'aire urbaine de Paris. *INSEE Première*. Numéro 1591.

Cediey E and Foroni F (2007) Les discriminations à raison de 'l'origine' dans les embauches en France. *Bureau International du Travail*.

Chang J, Rosenn I, Backstrom L and Marlow C (2010) Epluribus: Ethnicity on Social Networks. *ICWSM* 10: 18-25.

Choi BC, Hanley J, Holowaty EJ and Dale D (1993) Use of Surnames to Identify Individuals of Chinese Ancestry. *American Journal of Epidemiology* 138(9): 723-734.

Clark G (2014) *The Son Also Rises: Surnames and the History of Social Mobility*. Princeton: Princeton University Press.

Delattre É, Leandri N, Meurs D and Rathelot R (2013) Trois approches de la discrimination: évaluations indirectes, expérimentation, discriminations ressenties. *Économie et statistique* 464(1): 7-13.

Foroni F, Ruault M and Valat E (2016) Discrimination à l'embauche selon 'l'origine': que nous apprend le testing auprès de grandes entreprises? *Ministère du Travail*.

Güell M, Mora JVR and Telmer CI (2015) Intergenerational Mobility and the Informational Content of Surnames. *Review of Economic Studies* 82: 693-735.

Herfindahl OC (1950) *Concentration in the Steel Industry*. PhD Thesis, New York, Columbia University.

Hirschman AO (1945) *National Power and the Structure of Foreign Trade*. Volume 105. California: University of California Press.

Jobard F and Nevanen S (2007) The Color of Judgment – Discrimination in Rulings on Cases Involving Offenses Against French Police Officers, 1965–2005. *Revue Francaise de Sociologie* 48(2): 243-272+439–441+445.

Jobling MA (2001) In the Name of the Father: Surnames and Genetics. *TRENDS in Genetics* 17(6): 353-357.

King TE, Ballereau SJ, Schürer KE and Jobling MA (2006) Genetic Signatures of Coancestry Within Surnames. *Current Biology* 16(4): 384-388.

King TE and Jobling MA (2009) What's in a Name? Y Chromosomes, Surnames and the Genetic Genealogy Revolution. *Trends in Genetics* 25(8): 351-360.

Lasker GW (1980) Surnames in the Study of Human Biology. *American Anthropologist* 82(3): 525-538.

Lasker GW (1985) *Surnames and Genetic Structure*. Volume 1. Cambridge: Cambridge University Press.

Lee J, Kim H, Ko M, Choi D, Choi J and Kang J (2017) Name Nationality Classification with Recurrent Neural Networks. In: *Proceedings of the 26th IJCAI International Joint Conference on Artificial Intelligence*.

Mateos P (2007) A Review of Name-Based Ethnicity Classification Methods and their Potential in Population Studies. *Population, Space and Place* 13(4): 243-263.

Mateos P (2014) *Names, Ethnicity and Populations – Tracing Identity in Space. Advances in Spatial Science*. New York: Springer.

Mislove A, Lehmann S, Ahn YY, Onnela JP and Rosenquist JN (2011) Understanding the Demographics of Twitter Users. *ICWSM* 11: 5th.

Nieminen P and Isohanni M (1999) Bias Against European Journals in Medical Publication Databases. *The Lancet* 353(9164): 1592.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(Oct): 2825-2830.

Piazza A, Rendine S, Zei G, Moroni A and Cavalli-Sforza LL (1987) Migration Rates of Human Populations from Surname Distributions. *Nature* 329(6141): 714-716.

Polednak AP (1993) Estimating Cervical Cancer Incidence in the Hispanic Population of Connecticut by Use of Surnames. *Cancer* 71(11): 3560-3564.

Rosenfeld MJ (2008) Racial, Educational and Religious Endogamy in the United States: A Comparative Historical Perspective. *Social Forces* 87(1): 1-31.

Shah BR, Chiu M, Amin S, Ramani M, Sadry S and Tu JV (2010) Surname Lists to Identify South Asian and Chinese Ethnicity from Secondary Data in Ontario, Canada: A Validation Study. *BMC Medical Research Methodology* 10(1): 42.

Tonkin E, McDonald M and Chapman MK (2016) *History and Ethnicity*. Volume 27. Oxford: Routledge.

Torvik VI and Agarwal S (2016) Ethnea: An Instance-Based Ethnicity Classifier Based on Geocoded Author Names in a Large-Scale Bibliographic Database. In: *International Symposium on Science of Science*.

Ward JH Jr (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58(301): 236-244.

Weber M (1978) *Economy and Society: An Outline of Interpretive Sociology*. Volume 1. California: University of California Press.

Wu Z, Yuan D, Treeratpituk P and Giles CL (2014) Science and Ethnicity: How Ethnicities Shape the Evolution of Computer Science Research Community. *arXiv preprint arXiv:1411.1129*