Estimation de diversité par l'inférence de l'origine des noms de famille

Antoine Mazieres, Camille Roth

Estimation de diversité par l'inférence de l'origine des noms de famille

Antoine Mazières[1]
Centre Marc Bloch e.V. (UMIFRE CNRS 14),
Computational Social Science team, Berlin, Allemagne
UMR-LISIS, INRA, Marne-la-Vallée, France

Camille Roth[2]
Centre Marc Bloch e.V. (UMIFRE CNRS 14),
Computational Social Science team, Berlin, Allemagne
Sciences Po Paris, France

https://namograph.antonomase.fr/

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. En nous appuyant en partie sur des éléments de la littérature existante, nous avons construit un classifieur des origines des noms de famille. Pour ce faire, nous avons extrait de l'ensemble des articles référencés sur PubMed environ 25 millions d'affiliations liant des auteurs à des pays. Nous nous sommes ensuite intéressés aux noms dont la concentration était particulièrement forte dans un pays donné afin de définir un ensemble de noms de références pour ce pays.

Chaque nom est découpé en *n-grammes,* c'est-à-dire en sous-ensembles de taille variable de lettres successives. Afin de définir un nombre de catégories d'origines plus restreint que les 176 pays considérés, nous avons opéré un regroupement hiérarchique sur l'ensemble de ces *n-grammes* rassemblés par pays. Le résultat (cf. Figure 1) permet, en suivant simplement la structure de l'arbre, de reconstruire un découpage intelligible des régions du monde au prix d'un très petit nombre de classifications qualitativement surprenantes, qui sont corrigées individuellement. Cette typologie issue des données nous permet de nous écarter du concept d'*ethnicité* (subjectif) généralement à l'honneur dans la littérature, et de faire appel au fondement plus objectif de l'*origine géographique* des noms de famille. L'endogamie permet alors d'expliquer en quoi les noms sont encore aujourd'hui des variables intermédiaires pertinentes dans de nombreux domaines de recherche.

Nous obtenons ainsi une base de données de noms d'auteurs associés à une région du monde, qui nous permet de construire un modèle *via* une procédure d'apprentissage supervisé simple et, ainsi, de pouvoir inférer l'origine géographique de noms qui n'étaient pas présents dans les données initiales. Nous améliorons les performances de ce classifieur en prenant en compte ses taux d'erreur dans l'évaluation de la distribution des origines des noms d'une population donnée.

Ce modèle nous permet d'explorer une méthodologie pour estimer à grande échelle la diversité relative des groupes sociaux. Plus précisément, en comparant la distribution des origines d'une base de données de noms de référence pour une population et en la comparant à une base de données cible, nous pouvons estimer les sur-/sous-représentations de chaque origine dans cette dernière. Cette méthode peut se montrer utile notamment lorsque les données sur les origines ne sont pas couramment prise en compte ou peu disponibles, comme c'est le cas en France.

Enfin, nous appliquons cette méthode pour mesurer la représentativité des origines de noms de famille parmi 15 groupes socio-professionnels en France (cf. Tableau 1). Les résultats (cf. Figure 2) montrent des similarités fortes entre certains types de groupes. Par exemple, les fonctions électives montrent un profil comparable de diversité avec une sur-représentation de l'origine nord-européenne et une sous-représentation plus prononcée que dans les autres bases de données pour les autres origines.

En conclusion, nous discutons plusieurs biais possible dans l'observation des sous-représentations permise par cette méthode, et des éléments nécessaires à ce qu'elle puisse contribuer à l'étude des discriminations liées à l'origine.

[1] antoine.mazieres@gmail.com
[2] camille.roth@sciencespo.fr

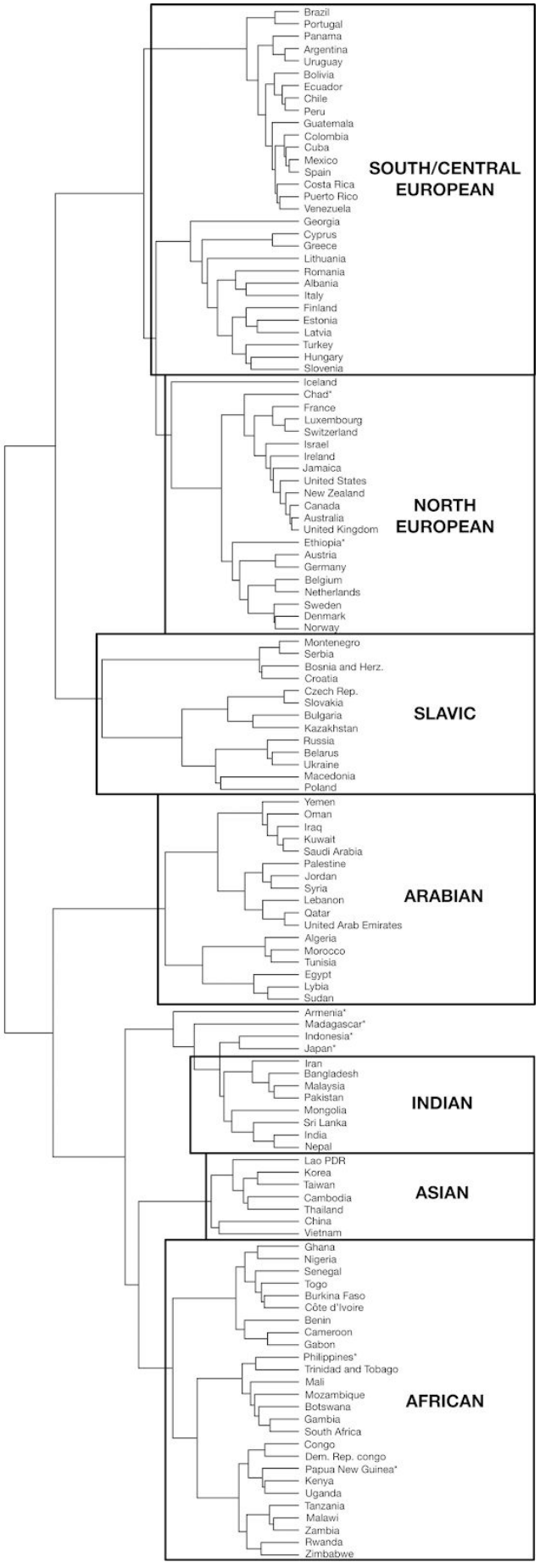

Figure 1: Clusters d'origines des noms de famille.

Regroupement hiérarchique utilisant la méthode de Ward, appliqué à la matrice "pays / n-grammes" où les lignes représentent les pays et les colonnes les n-grammes de tous les noms qui y sont associés.

Les pays marqués d'une astérisque (*) sont interprétés comme mal classifiés et sont réassignés à la main de la manière suivante:

- Philippines, Japon et Indonésie sont assignés au cluster *Asian*.
- Ethiopie au cluster *African.*

En outre, Papouasie Nouvelle Guinée, Madagascar, Jamaique, Tchad et Arménie sont supprimés de la base de données étant donné qu'ils représentaient un très faible nombre d'observations initiales.

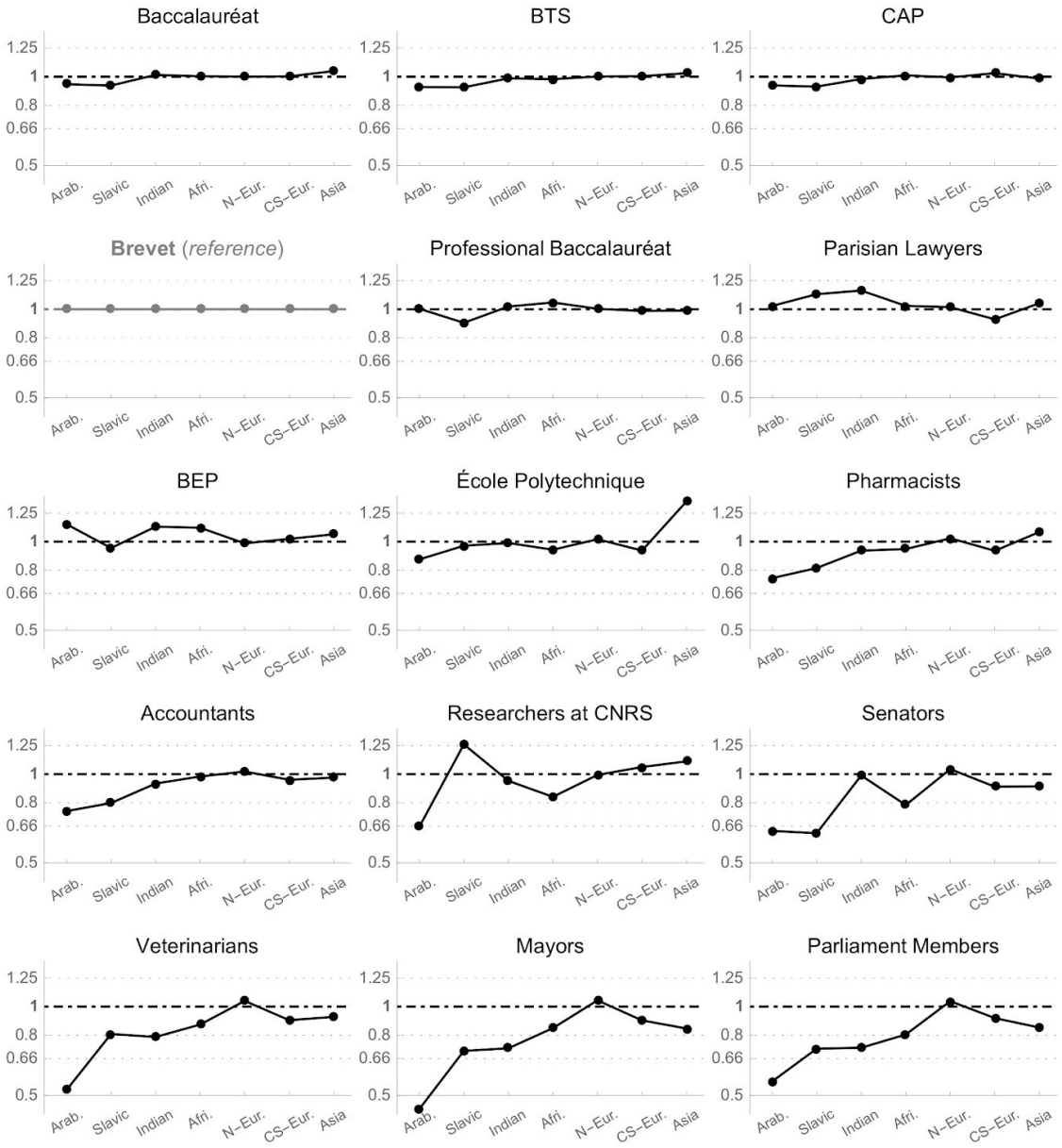

Figure 2 : Sur-/Sous-représentation des origines des noms de famille pour toutes les bases de données (cf. Annexe 1).

Chaque graphique montre le ratio entre la base de donnée cible et celle de référence (Brevet), pour chaque catégorie d'origine. Une échelle logarithmique est utilisée pour illustrer les sur-/sous-représentations par rapport à la ligne de référence, y =1.

En l'absence de liste complète ou non-biaisée des noms de famille de la population Française, nous avons décidé d'utilisé comme distribution de référence les noms des candidats au Brevet des Collèges de 2008. Le Brevet étant le diplome le plus passé en France, cette liste représente probablement un échantillon représentatif des personnes vivant en France agées aujourd'hui de 23-24, mais implique donc un biais d'âge par rapport à la population globale.

Nom	**Liste des noms de famille des...**	**Nb. Observations**
Brevet	Candidats au Diplôme National du Brevet de 2008[3]	562 952
Baccalauréat	Candidats au Baccalauréat National (Général et Technologique) de 2008	435 645
BEP	Candidats au Brevets d'Études Professionnelles de 2008	116 814
CAP	Candidats aux Certificats d'Aptitude Professionnelle de 2008	98,364
BTS	Candidats aux Brevets de Technicien Supérieur de 2008	87 917
Bac Pro	Candidats aux Baccalauréats Professionnels de 2008	80 672
Pharmaciens	Pharmaciens inscrits à leur ordre professionnel en 2017[4]	73 422
Maires	Maires des communes Françaises en 2014[5]	36,628
Avocats parisiens	Avocats inscrits au Barreau de Paris en 2017[6]	32 021
Polytechniciens	Étudiants à l'École Polytechnique (1958-2016)[7]	23 058
Comptables	Experts-Comptables inscrits à leur ordre professionnel en 2017[8]	20 946
Vétérinaires	Vétérinaires inscrits à leur ordre professionnel en 2017[9]	15 710
Chercheurs CNRS	Chercheurs au Centre National de la Recherche Scientifique en 2017[10]	12 657
Députés	Députés à l'Assemblée Nationale (1958-2016)[11]	8,326

Tableau 1 : Liste de toutes les bases de données avec les nombres d'observations correspondants.

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

Financement
Ce travail a été partiellement soutenu par le projet "Algodiv" (ANR-15-CE38-0001) financé par l'Agence Nationale de la Recherche.

[3] Source pour tous les examens de 2008: http://www.bankexam.fr/resultat/2008
[4] Source: http://www.ordre.pharmacien.fr/annuaire/pharmacien
[5] Source: https://www.data.gouv.fr/fr/datasets/liste-des-maires-au-17-juin-2014/
[6] Source: http://www.avocatparis.org/annuaire
[7] Source: https://www.polytechnique.org/search
[8] Source: http://www.experts-comptables.fr/annuaire
[9] Source: https://www.veterinaire.fr/outils-et-services/trouver-un-veterinaire.html
[10] Source: https://annuaire.cnrs.fr/l3c/owa/annuaire.recherche/index.html
[11] Source: http://www.assemblee-nationale.fr/sycomore/liste_legislature.asp?legislature=48