Predicting plant endemicity based on herbarium data : application to French data

Jessica Tressou, Liliane Bel, Thomas Haevermans

Predicting plant endemicity based on herbarium data: application to French data

Jessica Tressou [1], Liliane Bel[1] & Thomas Haevermans [2]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*

[2] *UMR 7205 CNRS - MNHN - École Pratique des Hautes Études – Université Pierre et Marie Curie, Sorbonne Universités, CP39, 75231 Paris Cedex 05, France*

Résumé. L'évaluation formelle de l'état de conservation global des espèces végétales et animales est un travail indispensable mais extrêmement long qui prendra au rythme actuel de nombreuses années. Nous proposons une approche permettant de prédire rapidement et de manière reproductible le niveau de menace d'une espèce pour les 360,000+ espèces végétales recensées. La probabilité qu'une espèce soit menacée est estimée pour chaque plante par l'analyse des données émanant de la digitalisation complète de l'herbier du Muséum National d'Histoire Naturelle, plus grande collection au monde, à partir d'algorithmes d'apprentissage. Dans un premier temps, les noms scientifiques des plantes tels que recensés dans l'herbier de Paris, sont reliés aux noms scientifiques tels que décrits dans la liste internationale "Plant List" grâce à un moteur de recherche open source appelé Terrier. Cette première étape permet d'estimer que les 6+ millions d'enregistrements de l'herbier représentent 167 355 noms acceptés de plantes, soit près de 47% des 356 106 plantes vasculaires connues. Une série de statistiques liées au nom accepté de la plante ont ensuite été calculées et intégrées comme prédicteurs dans un modèle de classification binaire : les deux modalités sont "Préoccupation Mineure" (LC pour 'Least Concern') et "Potentiellement menacée" ('not LC'). Le jeu de données d'entrainement comprend les 15 824 espèces déjà évaluées et référencées dans la liste rouge mondiale des espèces menacées de l'International Union for Conservation of Nature (IUCN). L'algorithme des forêts aléatoires uniformes a été retenu car il permet de traiter des données comprenant un grand nombre de valeurs manquantes , il est peu sensible au choix de paramètres de réglage et fournit l'estimation de l'erreur de prédiction. Pour chaque nom accepté de plante, la probabilité de "menace" est estimée avec un intervalle de confiance, avec un taux global de mauvaise classification proche de 20%. Les résultats sont présentés sur une carte du monde selon certaines caractéristiques des plantes.

Mots-clés. Apprentissage et classification ; Environnement, climat ; Grande dimension, données massives.

Abstract. Evaluating formal threat criteria for every organism on earth is a tremendously resource-consuming task which will need many more years to accomplish at the

actual rate. We propose here a method allowing for a faster and reproducible threat prediction for the 360,000+ known species of plants. Threat probabilities are estimated for each known plant species through the analysis of the data from the complete digitization of the largest herbarium in the world using machine learning algorithms, allowing for a major breakthrough in biodiversity conservation assessments worldwide. First, the full scientific names from Paris herbarium database were matched against all the names from the international plant list using a text mining open source search engine called Terrier. The 6+ millions of records represent 167,355 species level accepted names, i.e. 47% of a total of 356,106 known vascular plants. A series of statistics related to the accepted names of each plant were computed and served as predictors in a statistical learning model with a binary output: 'Least Concern' (LC) versus 'not Least Concern'. The training data contained the 15,824 usable entries from the International Union for Conservation of Nature global Redlisting plants assessments. Random uniform forests were selected for their ability to deal with numerous missing values, the included estimation of the generalization error and tuning parameters default settings robustness. For each accepted name, the probability to be LC was estimated with a confidence interval and a global misclassification rate of 20%. Results are presented on the world map and according to different plant traits.

Keywords. Statistical learning; Classification; Environment; Big data.

1 Matériels et méthodes

1.1 Description des données

Les données utilisées proviennent de 3 sources distinctes : les données collectées dans l'Herbier du Muséum National d'Histoire Naturelle à Paris (description détaillée dans Lebras et al, 2017) ; les données publiques internationales issues de The Plant List (TPL, `http://www.theplantlist.org/`) et les données de l'IUCN (`http://www.iucnredlist.org`, IUCN, 2016a et 20016b) qui constitue la liste rouge des espèces menacées.

Les données de l'Herbier L'extraction initiale de l'Herbier comprend 6,104,130 enregistrements, associés à 5,318,001 pages d'herbier physiques ou "parts" distinctes. Chacune de ces pages est identifiée par un code barre qui sert d'identifiant à l'image numérisée de la page d'herbier. Les informations minimales associées à un code barre sont au moins un nom scientifique (et jusqu'à 8, par le jeu des synonymes), la classification et un secteur géographique large propre à l'Herbier (ASI, AME, EUR, etc.). Certaines pages d'herbier comportent de plus des informations liées à la collecte de la plante comme le nom du récolteur, l'année, le code ISO du pays où elle a été collectée. Au total 613,313 récoltes sont décrites, couvrant 1,463,754 des enregistrements (24.0%). Pour cette analyse, la base a été enrichie de la surface en km2 correspondant au code ISO quand celui ci est disponible.

Les données internationales L'extrait de la base TPL comprend 1,298,042 enregistrements: chaque enregistrement a un identifiant, un nom scientifique (famille, genre, espèce, auteurs), le nom accepté associé (ANID dans la suite), et l'année de publication. 393,585 noms sont reconnus comme des noms acceptés. Cette base a été complétée par des informations géographiques, sur le climat et les formes de vie des plantes de la Royal Botanical Gardens Kew, World Checklist of Selected Plant Families (`http://apps.kew.org/wcsp/`) (un total de 684,477 enregistrements). L'information géographique (9 codes continent, 53 codes région, 388 codes de sous région ou "area" appelé code TDWG et utilisé dans la Figure 1) du lieu de collecte de la plante est disponible pour 168,725 plantes (dont 130,726 noms acceptés).

Les données d'entrainement : la liste rouge de l'IUCN La liste rouge de l'IUCN est constituée de 19,200 évaluations que l'on peut extraire par pays (IUCN, 2016a) ou par taxon (IUCN, 2016b), classant les plantes comme LC pour *Least Concern* (28.3%), NT pour *Near Threatened* (9.1%), VU pour *Vulnerable* (27.0%), EN pour *Endangered* (16.4%), CR pour *Critically Endangered* (10.8%), EW pour *Extinct in the Wild* (0.2%), EX pour *Extinct* (0.5%), ou DD pour *Data Deficient* (7.7%). Ces 19,200 classements correspondent à 18,826 noms acceptés: quand plusieurs évaluations se rapportent à la même plante au sens du nom accepté, le classement le plus "élevé" a été retenu, en considérant l'ordre suivant LC>NT>VU>EN>CR>EW>EX>DD.

1.2 Analyse de texte / text mining

Appariemment des bases de données L'appariemment des 3 bases de données principales décrites dans la section précédente a été faite manuellement pour ce qui concerne IUCN et TPL et à partir d'un moteur de recherche open source appelé Terrier (Macdonald et al, 2012) pour l'Herbier et TPL. Pour chaque ligne des bases de données, un "document" est créé en concaténant le texte (en minuscule) des champs famille, genre, espèce, et des différents champs auteurs. Ensuite, un score de similarité est calculé entre chaque "document" Herbier et la liste des "documents" TPL qui constituent notre référence, ce qui permet d'identifier pour chaque enregistrement de l'Herbier l'enregistrement le plus proche dans TPL. Pour assurer une certaine efficacité de la procédure, les enregistrements de l'herbier comportant de trop nombreuses valeurs manquantes ou indéterminées ont été supprimés, laissant tout de même 5,589,233 enregistrements à apparier à la référence TPL. La qualité de ce matching a été évaluée en calculant le taux de concordance de différents champs.

Traitement des synonymes Lorsque plusieurs noms (synomymes) coexistent pour une même plante, on retient l'un de ces synonymes comme le nom accepté de la plante qui servira d'identifiant (ANID pour "accepted name identifier"). Chaque ligne de TPL correspond soit à un nom accepté, soit à un synonyme pointant vers un nom accepté, soit à un "irrésolu" (ie le nom n'est pas un nom accepté et aucune indication de nom accepté associé n'est fournie). Les 1,298,052 lignes correspondent à 393,585 ANID distincts,

(356,106 ANID en excluant les plantes non vasculaires). Dans l'Herbier, les 5,589,233 lignes correspondent finalement à 167,891 ANID, (167,355 en excluant les plantes non vasculaires). Pour la base IUCN, les 19,200 lignes correspondent à 17,098 ANID distincts (15,824 en excluant ceux classés DD). Les plantes non vasculaires sont exclues de l'analyse car absentes de notre base d'apprentissage.

1.3 Modèle d'apprentissage

Dans un premier temps nous avons considéré le problème binaire consistant à prédire si une espèce fait l'objet d'une préoccupation mineure (LC) ou non. Une extension naturelle est de prédire chacun des 9 statuts ou au moins de travailler avec 3 groupes en isolant les 3 catégories d'espèces menacées d'extinction que sont CR, EN et VU.

Choix de l'algorithme Plusieurs approches ont été testées. L'approche la plus classique est la régression logistique (cas binaire) ou multinomiale (pour classer en 3 catégories ou plus). Elle est bien connue et très populaire chez les botanistes mais ne permet pas d'obtenir une prédiction dès lors qu'une partie des données est manquante ce qui est un gros inconvénient pour des données très incomplètes. Notre choix s'est alors orienté vers une méthode basée sur les arbres de régressions (Breiman et al, 1984) de type CART qui gère la présence de données manquantes et permet l' interprétation des règles de décision sous forme graphique. Toutefois, les approches les plus simples de cette famille sont en général trop proches des données d'entrainement et présentent un fort risque de sur-apprentissage. Des méthodes où les individus et/ou variables sont rééchantillonnées aléatoirement sont plus robustes, d'où le recours aux boosting (Friedman et al, 2000) ou aux forêts aléatoires (Breiman, 2001). Nous avons retenu les forêts aléatoires uniformes (Ciss, 2015) du fait de leur faible sensibilité aux paramètres de réglage, et de la possibilité d'inclure différents traitements des valeurs manquantes (FastImpute, AccurateImpute). De plus le package R associé fournit le calcul de l'erreur de généralisation (prédiction OOB pour "out of bag"), et le graphique montrant l'influence des différents prédicteurs. On s'y réfère dans la suite comme l'algorithme RUF. Le principe de cet algorithme est de combiner les réponses de plusieurs arbres de régression, très peu corrélés entre eux car obtenus en choisissant aléatoirement les variables à inclure dans chaque arbre et en choisissant selon la loi uniforme la valeur qui déterminera les branches de l'arbre.

Construction des variables prédictives Nous avons construit les variables prédictives en résumant l'information disponible au niveau du nom accepté. Par exemple, une variable compte le nombre d'enregistrements de l'herbier en lien avec l'ANID (N_LINE), le nombre de codes barre lié à un ANID (N_CB), le nombre de synonymes liés à un ANID (NB_SYN_SONNERAT), le nombre de secteurs géographiques (N_SECTOR), le nombre de codes ISO distincts (N_ISO), etc. Dans TPL, en plus de l'année de publication de l'ANID (YEAR_PUBLICATION), on conserve l'année de publication minimale et maximale, ainsi que l'écart entre les deux (DELTA_YEAR_TPL) associées à l'ANID via les dates de publication des synonymes. Au total, nous obtenons 38 variables quantita-

tives et 31 variables qualitatives (SUPER_ORDER, 5 sur le climat, et 25 les formes de vie). D'autres variables n'ont pas été incluses dans le modèle mais construites pour la présentation des résultats comme le nombre d'ANID associés à un code TDWG ou un code ISO, ou la forme de vie et le climat les plus fréquemment associés à un ANID.

2 Principaux résultats

Le travail d'analyse de texte sur les noms des plantes a permis d'identifier que l'Herbier de Paris couvre 42.7% des espèces de plantes en termes de noms acceptés, et même 47% si on exclut les plantes non vasculaires.

Nous avons choisi de garder le modèle le plus simple avec les paramètres par défaut de l'algorithme RUF et l'ensemble des variables (69 au total). La figure 1 liste les variables les plus influentes pour la prédiction. Nous obtenons une erreur de prédiction OOB de

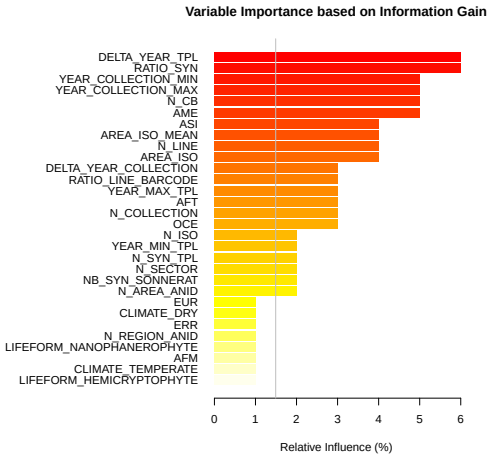

Figure 1: Importance des variables

19.8% sur le jeu de données d'entrainement de taille 15,824. Pour chacune des 356,106 plantes vasculaires de TPL, nous pouvons à partir de ce modèle prédire sa classe en tant que LC ou non LC ainsi que la probabilité et l'intervalle de confiance associé, basés sur la distribution des votes des différents arbres de la forêt aléatoire. Nous avons également estimé le modèle avec 3 réponses (LC-NT, CR-EN-VU, EX-EW).

En agrégeant les résultats au niveau des codes TDWG, on obtient en figure 2 une carte d'endémicité (carte principale), les zones plus rouges comportant le plus d'espèces menacées, à mettre en lien avec la carte du nombre d'espèces connues par polygone (carte de l'encadré).

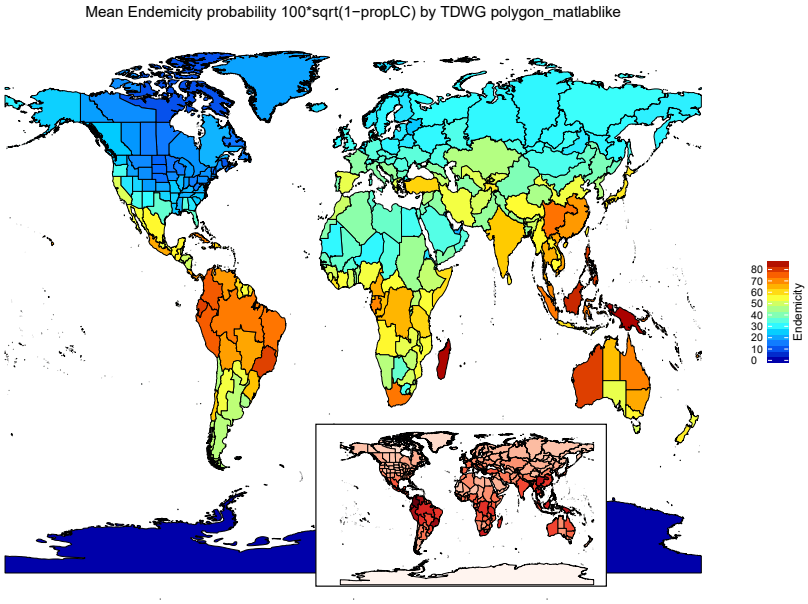

Figure 2: Cartes de probabilité de danger

Bibliographie

[1] Le Bras G., Pignal M., Jeanson M.L., Muller S., Aupic C., Carré B., Flament G., Gaudeul M., Gonçalves C., Invernón V.R., Jabbour F., Lerat E., Lowry P.P., Offroy B., Pimparé E.P., Poncy O., Rouhan G., Haevermans T. (2017) The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. Scientific Data, 4, 170016.

[2] International Union for Conservation of Nature. (2016a) Table 6b: Red List Category summary country totals (Plants).

[3] International Union for Conservation of Nature. (2016b) Table 4b: Red List Category summary for all plant classes and families.

[4] Macdonald, C., McCreadie, R., Santos, R. L. T. & Ounis, I. (2012). From puppy to maturity: Experiences in developing Terrier. in Proceedings of the SIGIR 2012 Workshop in Open Source Information Retrieval, 60–63.

[5] Friedman, J.H., Hastie, T., Tibshirani, R. (2000). Additive Logistic Regression: a Statistical View of Boosting, Annals of Statistics, 28(2):337-374.

[6] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). Classification and regression trees. Chapman and Hall.

[7] Breiman, L. (2001) Random Forests. Machine Learning, 5–32.

[8] Ciss, S. (2015) RandomUniformForest: Random Uniform Forests for classification, regression and unsupervised learning. R package version 1.1.5.