Predicting Plant Threat Based on Herbarium Data: Application to French Data

Jessica Tressou, Thomas Haevermans, Liliane Bel

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

Jessica Tressou*, Thomas Haevermans and Liliane Bel

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. A series of statistics related to the accepted names of each plant were computed and served as predictors in a statistical learning algorithm with binary output. The training data was build based on the International Union for Conservation of Nature (IUCN) global Redlisting plants assessments. For each accepted name, the probability to be of least concern (LC, not threatened) 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.

J. Tressou *(corresponding author)*
UMR MIA-Paris, AgroParisTech - INRA - Université Paris-Saclay, 16 rue Claude Bernard, 75231, Paris Cedex 05, France, e-mail: jessica.tressou@inra.fr

T. Haevermans
UMR 7205 CNRS - MNHN - École Pratique des Hautes Études Université Pierre et Marie Curie, Sorbonne Universités, CP39, 75231 Paris Cedex 05, France, e-mail: thomas.haevermans@mnhn.fr

L. Bel
UMR MIA-Paris, AgroParisTech - INRA - Université Paris-Saclay, 16 rue Claude Bernard, 75005, Paris, France, e-mail: liliane.bel@agroparistech.fr

Introduction

In the last years, the French National Museum of Natural History (MNHN) started a huge work of digitalization of its herbarium, one of the largest collections in the world. Based on this and the increasing development of machine learning in all fields, botanists from the MNHN met the statisticians from INRA to develop an algorithm allowing to predict plant endemicity and constitute a list of threatened species similarly to what is currently performed in the IUCN Red List. These assessments require a lot of time and resources and many plants still have not been assessed (20,000 out of nearly 400,000 have been assessed).

We propose in this work an original approach combining the French Paris (P) herbarium data to international public data to classify plants according to their threat level, first in a binary model (threatened vs not threatened) and then in a multinomial model (merging some of the IUCN threat categories). A huge part of the work concerned the matching of the different databases and the construction of predictors based on the available data and the knowledge of what is actually used for performing red list assessments for the IUCN. Then based on the 20, 000 plants already classified by IUCN, a uniform random forest algorithm is trained to be able to predict the threat category of all 400,000 plants. The end goal of the present work is to provide a tool that can rapidly and at a less cost predict roughly the threat level for a large amount of plants so that it may help botanists prioritize which plant should be assessed in detail next. A side result is also the analysis of the features determining whether a plant endangered or not.

The paper is organized as follows. First, we give a brief description of the available databases that were combined in the analysis. Then we describe the proposed methodology with a focus first on the matching between the different databases and then on the modeling approach based on random uniform forests. In the last section, we expose some of the results and discuss the perspectives of this work.

1 Data description

The data used come from 3 main different sources: the data collected by the Herbarium of the National Museum of Natural History in Paris (MNHN, available at `https://science.mnhn.fr/institution/mnhn/collection/p/item/search/form`) ; international public data from The Plant List (TPL, `http://www.theplantlist.org/`) and previously assessed plants from the IUCN redlisting data (`http://www.iucnredlist.org`).

1.1 Herbarium

Specimen stored at Paris MNHN Herbarium were fully digitized constituting one of the largest collection in the world, see [5] for the construction of this huge database. For this study, we extracted the following information: taxonomic information (family, genus, species, names of authors), geographic sector information, and collection information when available (ISO code of the country where the plant was collected, year when collected). External data was used to add the area of each country with respect to its ISO code. The raw data is constituted of 6,104,130 records, associated with 5,318,001 physical distinct herbarium sheets. Each of these sheets is identified by a barcode. Each barcode is associated with at least one plant name (and up to 8 due to synonymy), a geographic sector (ASI, AME, EUR, FRA, etc.). A total of 613,313 collections were described covering 1,463,754 of the records (24.0%).

1.2 TPL and international databases

The Plant List is a working list of all known plant species. It aims to be comprehensive for species of Vascular plant (flowering plants, conifers, ferns, and their allies) and of Bryophytes (mosses and liverworts). It was created jointly by the Royal Botanic Gardens, Kew and Missouri Botanical Garden. It provides the Accepted Latin name for most species, with links to all synonyms by which that species has been known. Around 20% of names are unresolved indicating that the data sources included provided no evidence or view as to whether the name should be treated as accepted or not, or there were conflicting opinions that could not be readily resolved. See `http://www.theplantlist.org/` for summary statistics by family of plants. Our extract from the TPL database contained 1,298,042 records: each record has an identifier, a scientific name (family, genus, species, authors), the associated accepted name (ANID in the following), and the year of publication. 393,585 names are recognized as accepted names, 356,106 if we narrow the database to vascular plants only. This database has been supplemented with geographical, climate and plant life information from the Royal Botanical Gardens Kew, World Checklist of Selected Plant Families `http://apps.kew.org/wcsp/`, 684,477 records). Geographical information (9 continent codes, 53 region codes, 388 sub-region codes or "area" called TDWG code and used in Figure 1) of the collection site of the plant is available for 168,725 distinct plants (among which 130,726 have accepted names). Lifeform data was available for 126,730 plants (among which 113,264 have accepted names) and climate information for 136,783 plants (among which 122,346 have accepted names). The 258 described lifeforms were summarized into 25 lifeform binary criteria (such as phanerophyte, epiphyte, annual, climbing, hydrophyte, ...). Similarly, the 23 described climates were summarized into 5 climate binary criteria (tropical, aquatic, temperate, dry, altitude). To group plants at a more aggregated level than the family level (about 500 distinct categories), the order (around 70 categories) was considered as well as the "super order" (8 distinct categories:

Gymnosperms, Magnoliids, Monocots, other Angiosperms, other Eudicots, Pteridophyta, Superasterids, Superrosids).

1.3 IUCN redlist

The IUCN Red List consisted of 19,200 assessments that can be extracted by country or by taxon ranking plants as LC for *Least Concern* (28.3%).), NT for *Near Threatened* (9.1 %), VU for *Vulnerable* (27.0 %), EN for *Endangered* (16.4 %), CR for *Critically Endangered* (10.8 %), EW for *Extinct in the Wild* (0.2 %), EX for *Extinct* (0.5 %), or DD for *Data Deficient* (7.7 %). These 19,200 rankings correspond to 18,826 accepted names (all vascular plants): when several evaluations relate to the same plant in the sense of the accepted name, the "highest" ranking was retained, considering the following order LC $>$ NT $>$ VU $>$ EN $>$ CR $>$ EW $>$ EX $>$ DD. In the following, the plants classified as DD are excluded from the training data, yielding a training sample of size 15,824.

2 Methods

The main idea of the proposed methodology was to predict the red list status of each plant based on training data (IUCN data) and the available information from the French herbarium and general information (TPL mainly). A first step to this approach is to match the different databases which all have taxonomic information but no common identifier. Then the information available in the Herbarium and TPL had to be summarized at the accepted name level, that is each and every synonym of a plant will have the same red list status prediction. We first considered the binary problem of predicting whether a species is of least concern (LC) or not. A natural extension is to predict each of the 9 statuses or at least to work with some groupings of these, isolating the 3 categories of endangered species that are CR, EN, and VU in a group.

2.1 Text mining

The matching of the 3 main databases described in the previous section was done manually concerning IUCN and TPL and from an open source search engine called Terrier ([6]) for the Herbarium and TPL. For each row of the databases, a "document" is created by concatenating the text (in lower case) of the family, genus, species, and different fields of authors. Then, a similarity score is calculated between each "Herbarium" document and the list of "TPL" documents that constitute our reference/dictionary. This allows us to identify for each record of the Herbarium the

closest record in TPL. To ensure a certain efficiency of the procedure, the records of the herbarium with too many missing or indeterminate values were deleted, leaving 5,589,233 records to be matched to the TPL reference. The quality of this matching was evaluated by calculating the concordance rate of different fields. Some random checks were also performed by the botanists to ensure the number of errors resulting from this approach remains low.

2.2 Dealing with synonyms

When several names (synonymy) coexist for the same plant, one of these synonyms is retained as the accepted name of the plant that will serve as an identifier (ANID for "accepted name identifier"). Each TPL line is either an accepted name or a synonym pointing to an accepted name or an "unresolved" (ie the name has not been critically evaluated yet and is thus neither an accepted name nor a synonym). The 1,298,052 lines correspond to 393,585 separate ANIDs, (356,106 ANID excluding non-vascular plants). In the Herbarium, after matching the names to the TPL reference, the 5,589,233 records finally correspond to 167,891 ANID, (167,355 excluding non-vascular plants). For IUCN, the 19,200 lines correspond to 17,098 distinct ANIDs (15,824 excluding those classified as DD). Non-vascular plants are excluded from the analysis because they are absent from our learning base (IUCN).

2.3 Construction of the predictors

Predictors were constructed by summarizing the available information at the accepted name id level. For example, for each ANID, the variable `N_LINE` counts the number of herbarium records related to the ANID, `N_CB` counts the number of barcodes linked to the ANID, `NB_SYN_SONNERAT` counts the number of synonyms linked to the ANID, `NB_SECTOR` counts the number of distinct geographical sectors, the number of occurrences in each sector being stored in the variables `ASI` fro Asia, `AME` for America, `EUR` for Europe, `AFT` for Tropical and South Africa, `AFM` for Africa and Madagascar, `OCE` for Oceania, etc... `N_ISO` counts the number of distinct ISO codes. In TPL, in addition to the year of publication of ANID (`YEAR_TPL`), the minimum and maximum year of publication associated with the ANID via the dates of publication of the synonyms were calculated, as well as the difference between the two (`DELTA_YEAR_TPL`). The number of synonyms in TPL was also calculated for each ANID (`NB_SYN_TPL`) and used to compute the ratio of the number of synonyms in the Herbarium to the number of synonyms in TPL (`RATIO_SYN`). The number of distinct continents, regions and areas (`N_CONTINENT`,`N_REGION`, `N_AREA`) from the checklist data were computed for each ANID including the synonyms or not (suffix `_ANID` added when synonyms are not included).

A total of 38 quantitative variables and 31 qualitative variables (SUPER_ORDER, 5 on climate, and 25 on lifeforms) were constructed following this principle. Other variables were not included in the model but constructed for the presentation of results such as the number of ANIDs associated with a TDWG code or ISO code, or the lifeform and climate most frequently associated with a given code. Due to the predictor construction process, a large number of data is missing, some are missing from the original database and some are inherently missing due to the fact for example that some ANID do not appear in the French herbarium.

2.4 Random uniform forests

Several approaches have been tested. The most classic approach is logistic regression (binary case) or multinomial regression (to classify into 3 or more categories). They are well known and very popular methods among botanists but they lack of robustness when dealing with a high number of covariates or/and factors with many levels. Our choice then turned to a method based on regression trees ([2]) of the CART type that allows a non-parametric modeling of the link between predictors and response and the interpretation of the decision rules in a graphical form. However, the simplest approaches in this family are generally too close to the training data and present a high risk of overlearning. Methods where individuals and/or variables are randomly resampled are more robust, hence the use of boosting ([4]) or random forests ([1]). Missing values can be dealt with using imputation. We have retained uniform random forests because of their low sensitivity to tuning parameters, the possibility of including/comparing different methods for the imputation of missing values (FastImpute, AccurateImpute), and its native handling of categorical variables using a randomization mechanism at the node level (see [3] for details). Furthermore, the associated R package includes the calculation of the generalization error (OOB prediction for "out of bag"), and the graph showing the influence of the different predictors. It is referred to hereinafter as the RUF algorithm. The principle of this algorithm is to combine the responses of several regression trees, presenting very low correlation because obtained by randomly choosing the variables to be included in each tree and by choosing from the uniform distribution the cut-points which determine the branches of the tree. Each tree is grown on a random subsample of the training observations, the rest of them is used to evaluate the generalization error (OOB) similarly to what cross-validation allows to do. The missing value imputation can either be performed within the R package by FastImpute (missing values are replaced with the median value of the observed) or AccurateImpute (after initialization with FastImpute, a RUF learning algorithm is run on the observed values of each variable using the remaining ones as predictors).

3 Main results

Text mining

The text analysis of plant names allowed to determine that the Herbarium of Paris covers about 42.7% of the plant species in terms of accepted names, and even 47% if we exclude non-vascular plants.

Generalization error

We run the RUF algorithm using the default parameters with the 69 predictors (31 categorical variables). We obtained an OOB prediction error of 19.8% on the training dataset of size 15,824. This OOB prediction error was compared to the misclassification error obtained by cross-validation: from the 15,824 training observations, we built 40 test sets (sampled with replacement) of size 1,000 (or 5,000) and used the remaining observations to train the model. For each test set, the misclassification error is calculated: it varies from 17.9% to 22.3% for the 1,000 size, and from 19.1% to 21.3% for the 5,000 size. The OOB prediction error is therefore a good proxy of the generalization error.

Tuning the RUF algorithm

Missing imputation method	`ntree`	`mtry`	OOB error (%)	Time
Fast	100	69	19.8	2.5 mins
Accurate	100	69	5.9	6.6 mins
Fast	200	69	19.7	5.1 mins
Fast	500	69	19.6	12.8 mins
Fast	1000	69	19.7	37.6 mins
Fast	50	69	20.5	1.3 mins
Fast	100	50	20.3	1.9 mins
Fast	100	100	19.9	3.2 mins

Table 1 OOB prediction errors for different tuning parameters of the RUF model (`ntree` is the number of trees to grow, `mtry` is the number of variables randomly sampled with replacement as candidates at each split).

Table 1 illustrates how the OOB prediction error varies when modifying the tuning parameters that are the missing value imputation method, `ntree`, the number of trees to grow, `mtry`, the number of variables randomly sampled with replacement as candidates at each split, and the nested missing values treatment. Modifying `ntree` and `mtry` does not reduce the OOB error but can substantially increase the running time. Using accurateImpute rather than fastImpute reduces the OOB error (from 20% to 6%). However, further tests should be performed as the proportion of

missing values is high and the risks of overlearning by accurately imputing missing values here are also high as a consequence.

Important variables

We chose to keep the simplest model with the default parameters of the RUF algorithm (Fast, `ntree`= 100, `mtry`= 69) and the full set of variables (69 in total). The figure 1 lists the most influential variables for the prediction.

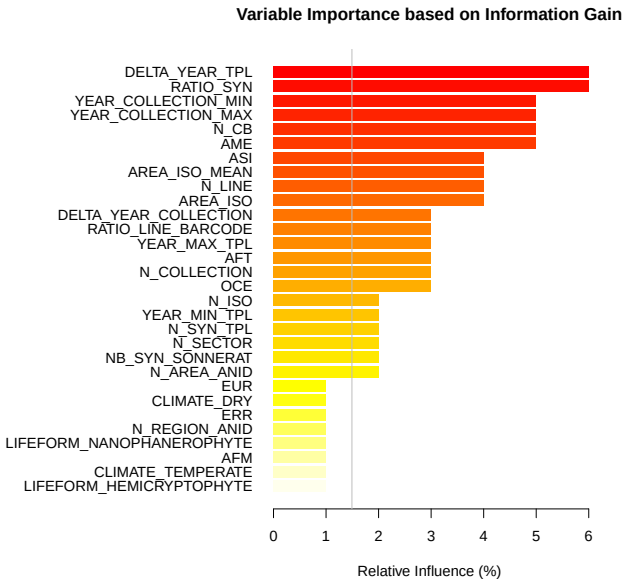

Fig. 1 Variable importance in predicting the binary response LC vs not LC. (see the construction of the predictors section for the meaning of variables names).

We observe that the variables that are the most influential are `DELTA_YEAR_TPL` and `RATIO_SYN` as well as the collection year (min or max) or the number of herbarium sheets stored at the herbarium (`N_CB`). Graphics showing the links between these most influential variables and the response (probability to be LC were drawn (not shown here). For example, we observed that the larger `DELTA_YEAR_TPL`, the larger the probability to be LC, meaning that plants with synonyms having very different dates of publication in TPL tend to not be threatened. The important conclusion is that, as expected, the more specimens of a plant were collected, the greater the probability to be of least concern, but this relationship is highly nonlinear as some point rare plants tend to be specifically searched for while more frequent

plants may be ignored. These results will be further detailed and commented in a publication aiming the botanists' audience.

Vizualization of the results

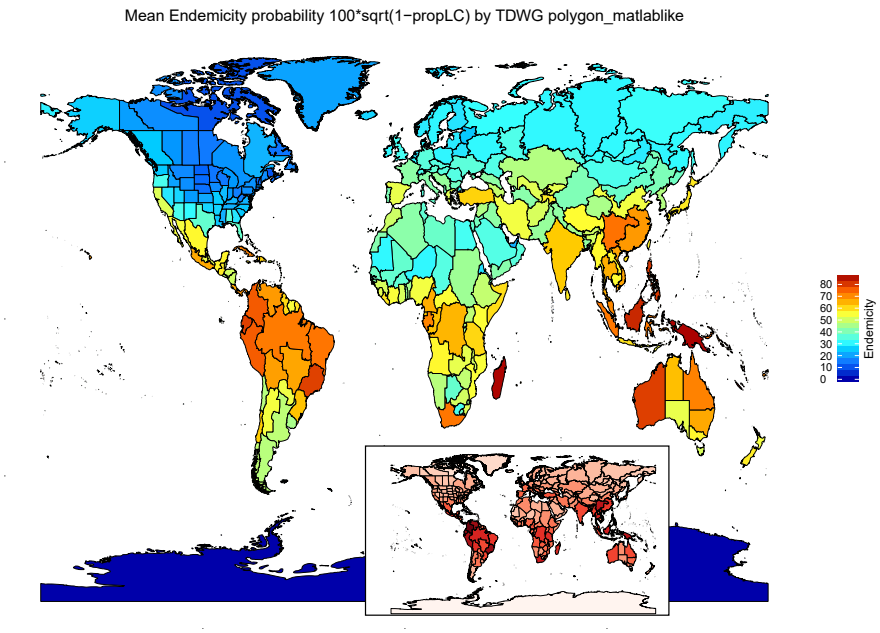

Fig. 2 Threat map (main map) with enclosed richness map. The endemicity/threat is calculated as the square root of the mean probability to be "not LC" among plants within each polygon; the richness is the number of plants per polygon in our database.

For each of the 356,106 vascular plants of TPL, we can use this model to predict whether the plant is LC or non-LC as well as the probability and associated confidence interval, based on the distribution of the votes of the different trees of the forest. We also trained the model with a 3 class response (LC-NT, CR-EN-VU, EX-EW), yielding an OOB prediction error of 26.5% with the default tuning parameters (and a running time of 2.9 minutes).

Globally, in the binary model, we find a mean probability to be of least concern (LC) of 29.1%, ranging from 19.4% for Magnoliids to 39.8% for Monocots. In the 3 class model, we find a mean probability to be LC/NT of 38.2%, ranging from 28.6% for Magnoliids to 50.1% for Gymnosperms. More detailed results will be published soon at the family level or at the ANID level.

By aggregating the results at the level of the TDWG codes (based only on the 130,408 ANID for which the information was available), we obtain in figure 2 a

threat map (main map), the red zones containing the most endangered species, to be linked to the map of the number of species per polygon (small map called richness map).

4 Perspectives

The statistical learning approach presented in this paper is quite innovative in the field of plant threat assessment. It gives interesting results which could help botanists choose what plant they should assess in details next. It is nevertheless only a first attempt at tackling this difficult question and several research directions merit further study. The initial matching step needs further validation and due to the way the predictors are built, we should assess further the role of missing value imputation. Although descriptive statistics were compared to rule out the representativeness bias that could exist between the training data set and the full data set, this could definitely be studied further. Overall more than the representativeness itself it has to be checked that the relationship between the outcome and the covariates is still well estimated even if the training set is not totally representative of the whole set. In addition, other machine learning methods (e.g. deep learning) should be tested to confirm the obtained results. An alternative approach would be a direct modeling of the phenomenon as a spatiotemporal process, allowing to capture quantities such as the area covered by the convex hull of the locations of the specimens of a plant or the evolution of the density of points along time, which are some of the main determinants of the IUCN classification. This type of approach would eliminate the aggregating step in the data preparation.

Acknowledgements The authors would like to thank Dr. Alan Paton, head of Science (Collections) at Royal Botanic Gardens, Kew for providing the TPL data and for his insights about the use of the data in this work.

References

1. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
2. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression trees (chapman y hall, eds.). Monterey, CA, EE. UU.: Wadsworth International Group (1984)
3. Ciss, S.: randomuniformforest-package: Random uniform forests for classification, regression and unsupervised learning (2015)
4. Friedman, J., Hastie, T., Tibshirani, R., et al.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics **28**(2), 337–407 (2000)
5. Le Bras, G., Pignal, M., Jeanson, M.L., Muller, S., Aupic, C., Carré, B., Flament, G., Gaudeul, M., Gonçalves, C., Invernón, V.R., et al.: The french muséum national dhistoire naturelle vascular plant herbarium collection dataset. Scientific data **4**, 170,016 (2017)
6. Macdonald, C., McCreadie, R., Santos, R., Ounis, I.: From puppy to maturity: Experiences in developing terrier. Open Source Information Retrieval **60** (2012)