Species recommendation using intensity models and sampling bias correction (GeoLifeCLEF2019: Lof of Lof team)

Pascal Monestiez, Christophe Botella

Species recommendation using intensity models and sampling bias correction (GeoLifeCLEF 2019: Lof_of_Lof team)

Monestiez Pascal[1], Christophe Botella[2]

[1] INRA, BioSP, Avignon, France
[2] INRA, UMR AMAP, Montpellier, France

Abstract. This paper presents three algorithms for species spatial recommendation in the context of the GeoLifeCLEF 2019 challenge. We submitted three runs to this task, all based on the estimation of species environmental intensities through Poisson processes models: The first is directly derived from MAXENT method used for species distribution models. The second method is a modification that uses sites were species observed as background points in MAXENT to correct for spatial sampling bias due to heterogeneous sampling in the training occurrences. The last method jointly estimates species and sampling intensities to correct for sampling bias. The best run was the MAXENT method which was ranked 14 over 44 runs with a top30 accuracy of 0.111 on the test set while the worst performing method was LOF with an accuracy of 0.086 (ranked 19).

1 Introduction

Predicting the species most likely to be observed from a location participate to build better biodiversity identification systems by reducing the list of candidate species that are observable at a given location. This may unlock the participation of citizen masses in biodiversity monitoring. It also helps experts with the burden of data quality check. Last but not least, it might serve educational purposes thanks to biodiversity discovery applications providing innovative features such as contextualized educational pathways. For this purpose, the GeoLifeCLEF 2019 ([Botella et al., 2019b]) task aims at evaluating species spatial recommendation algorithms to stimulate their improvement. It is one of the three tasks of the LifeCLEF 2019 evaluation campaign ([Alexis Joly, 2019]).
This challenge is highly related to the problem known as Species Distribution Modeling (SDM) in ecology [Elith and Leathwick, 2009]. SDM have become increasingly important in the last few decades for the study of biodiversity, macro

ecology, community ecology and the ecology of conservation. Concretely, the goal of SDM is to infer the spatial distribution of a given species, and they are often based on a set of geo-localized occurrences of that species (collected by naturalists, field ecologists, nature observers, citizen sciences project, etc.). No standard methods for species distribution method have been implemented in last year edition of GeoLifeCLEF. We implemented three methods based on Poisson point processes models ([Diggle, 2003], [Renner et al., 2015]) that estimate the relative environmental intensity of species from geolocated occurrences. In all our approaches, we estimate an absolute species intensity over space, and derive, for any location, the relative probability of any species from normalization of its intensity by the sum of intensities over all species.
The first run (MAXENT) submitted is a simple loglinear Poisson process implemented with `maxnet` R package ([Phillips et al., 2017]) for 141 species of the test set with a selection of environmental variables globally suited for modeling plant species distribution. We then fit a modified version with Target-Group background points [Phillips et al., 2009] (called TGB), a way of correcting for sampling selection bias. Finally, we used another sampling bias correction method (LOF). We implemented a joint estimation of all species intensities along with the spatial sampling effort over a regular mesh of squares, that is then set constant for prediction.

2 Data pre-processing

The dataset description may be found on the challenge overview
[Botella et al., 2019b].

Species selection - We first filtered species occurrences whose identification certainty score (field `FirstResPLv2Score`) was above 0.85 in the `PL_complete` dataset. Then, we kept only the **300** species with highest number of occurrences to prevent over-fitting (list of species L_0). We then made the intersection of those species with species of GeoLifeCLEF test set, which gave the list of 141 species (L_1) included in our predictions, so our predictions lack 703 test set species. L_1 is shown in the table `SpeciesTable.csv` of the Github repository [3]. MAXENT and TGB runs used only occurrences with identification score superior to 0.98, while LOF run used occurrences with identification score superior to 0.85, but they were fitted on the same list of species L_1. In all cases, we finally kept only the occurrences which had valid values for the selected environmental variables (described below).

Environmental features selection We selected a set of 9 environmental variables to model the environmental intensity of species included in the model. Following the recommendations of [Mod et al., 2016] on environmental variables for

[3] `https://github.com/ChrisBotella/GLC19runs`

modeling macro ecological species niches, we included mean and annual variation of temperature (chbio_1, chbio_5), annual precipitations (chbio_12), potential evapo-transpiration (etp), elevation (alti), slope (slope), available water capacity of the soil (awc_top), a soil pH proxy (bs_top) and a simplified plant habitat type descriptor (based on clc). Even though, the land cover category and elevation are not directly linked to species eco-physiological requirements, they have strong empirical links with species distributions as described by [Mod et al., 2016] and have a much sharper spatial grain, with a resolution around 100 meters. Then, we defined features derived from those environmental variables that would constitute the linear predictor of the species intensity. For continuous environmental variables we chose to model the intensity response with a Gaussian density function, which means that we kept the original variable and added a quadratic transformation of it to the linear predictor. This concerns variables: chbio_1, chbio_5, chbio_12-etp, etp, alti and slope. We combined annual precipitations chbio_12 and potential evapotranspiration etp into chbio_12-etp, called the water balance, which is commonly used in plants SDM [Mod et al., 2016]. We included categorical pedologic variables representing physico-chemical properties categories. However, for clc variable, we aggregated the 48 initial land cover categories into 5 to avoid inflating the number of parameters for the land cover effect. Indeed, we defined a Simplified Habitat Typology (spht) with types: cultivated, forest, grasslands, urban and other. Each type includes several CORINE Land Cover 2012 categories as shown in Table 1. We also included an interaction effect between water balance and slope in the model. As a summary, equation 1 shows the R formula of the linear predictor of any species intensity. It yields 18 features parameters.

$$\begin{aligned}\sim\ & \texttt{etp} + \texttt{I(etp}^2\texttt{)} + \texttt{I(chbio\_12-etp)} + \texttt{I((chbio\_12-etp)}^2\texttt{)} \\ & + \texttt{chbio\_1} + \texttt{I(chbio\_1}^2\texttt{)} + \texttt{chbio\_5} + \texttt{I(chbio\_5}^2\texttt{)} + \texttt{alti} + \texttt{I(alti}^2\texttt{)} \\ & + \texttt{slope} + \texttt{I(slope}^2\texttt{)} + \texttt{awc\_top} + \texttt{I(awc\_top}^2\texttt{)} + \texttt{bs\_top} + \texttt{I(bs\_top}^2\texttt{)} \\ & + \texttt{spht} + \texttt{slope:I(chbio\_12-etp)}\end{aligned} \tag{1}$$

3 Methodology

The R code for fitting models and producing the runs can be found on the dedicated github repository https://github.com/ChrisBotella/GLC19runs/.

run 27124: MAXENT - We used the R package maxnet ([Phillips et al., 2017]) to fit independently each species intensity from its occurrences. This package implements the method MAXENT. We constrained the features to be those of the previous paragraph **Environmental features selection**. This method requires to provide quadrature points, which are meant to represent the distribution of environmental variables in the spatial domain where the species is observed. We drew a set Z_0 of background points uniformly over the French territory until there was

at least 3 points per 4x4km square cells of a regular grid. With this setting, the method approximately fits a L1 penalized Poisson Process for each species with the given loglinear intensity model over environmental features. For each species of L_1, we ran the `maxnet` function with the occurrences of the species (score identification ≥ 0.98) and background points Z_0. For more implementation details on this part, one must refer to the file `make_maxent_and_tgb_models.R` of the github repository. Once the `maxnet` model was fitted for each species $i \in [1, 141]$, its features parameters $\beta_i \in \mathbb{R}^p$ were stored. Then, we estimated a posteriori the intercept α_i of the linear predictor because `maxnet` package doesn't provides it. We can compute it with the following formula $\alpha_i = \log(n_i / \sum_{z \in Z_{\text{test}}} \exp(\beta_i^T x(z)))$, where n_i is the number of training occurrences for i and Z_{test} is the set of test occurrences locations. Then we compute the probability of species i at location z as $\exp(\alpha_i + \beta_i^T x(z)) / \sum_{j=1}^{141} \exp(\alpha_j + \beta_j^T x(z))$. This second step is implemented in the script file `make_maxent_and_tgb_runs.R` of the github repository. We note that this two steps procedure is equivalent to a one step estimation of features and intercept parameters.

run 27123: TGB - This run implements the Target-Group Background method introduced in [Phillips et al., 2009]. It is the same as the MAXENT run (27124) except that the background points are selected differently. We first defined elementary sites where it is assume that the sampling effort is constant across sampled sites. We used the cells of the raster grid of the `alti` variable as sites because it is the highest resolution variable of our model, so the environment is roughly constant in those sites. We took a background point for each site where lies at least an occurrence of L_0. We removed points were non-valid environmental features were found. Once the model were fitted, we used the same procedure as for MAXENT to derive run predictions. The model fitting and run building are implemented in the script files `make_maxent_and_tgb_models.R` and `make_maxent_and_tgb_runs.R` of the github repository.

run 27063: LOF - We fitted a marked point process where the marks are the species identifiers. For any species i, its occurrences intensity is decomposed as $z \to \exp(\sum_{j=1}^{Q} \gamma_j 1_{z \in c_j}) \exp(\beta_i^T x(z))$. The first factor models the sampling effort which is shared between all species, while the second is the species environmental intensity, i.e. its abundance. Each species intensity has exactly the same structure has the ones estimated in previous runs. The sampling effort is constructed as a step function constant over units of a spatial mesh. The mesh is a regular spatial grid of squares of 4km side over the French metropolitan territory including Corsica, and restricted it to squares whose center was closer than 4km from the border or coast. We kept only the squares that had more than 5 occurrences inside and excluded the occurrences from the other squares. Thus, we ended up for this model with a set of 475,138 occurrences from L_0, distributed over 15,556 spatial squares covering around 40% of the French territory. This model and fitting procedure is entirely described in details in the submitted article [Botella et al., 2019a] which preprint is already downloadable [4] (the submitted

[4] `https://filedn.com/ltHjVlFrSfSJL1S4hXqqTzy/botella_MEE_2019_Main.pdf`

run are directly extracted from the real data illustration) and the R implementation of this method can be found in the script file `plantnet_effort.R` at `https://github.com/ChrisBotella/SamplingEffort`. Once we fitted this model, we extracted for each species of L_1 its environmental intensity component, and then built the run predictions as for other runs. This last step is implemented in the script file `make_lof_run.R` of the repository.

4 Results and discussion

The top30 results of our three runs on the test set are represented in the graph of Figure, which was pulled from the GeoLifeCLEF 2019 overview working note [Botella et al., 2019b], with dark yellow bars.

The best run was the MAXENT method which was ranked 14 over 44 runs with a top30 accuracy of 0.111 on the test set while the worst performing method was LOF with an accuracy of 0.086 (ranked 19). TGB made a score of 0.098. Thus, both bias correction methods failed to improve the model performance compared to the standard MAXENT approach. Theoretically, it was expectable that bias correction would not change the performance. Indeed, global sampling bias in the training data, which similarly multiplies the intensity of all species, doesn't affect the relative probabilities across species at a given place. However, the loss of performance is surprising. In the case of LOF, this performance gap might be due to the training data that include more occurrences of each species with less identification reliability, or to a model variance problem due to the high number of sampling effort parameters (around 15,000). The performance gap between MAXENT (27124) and TGB (27123) shows that the background points selection scheme have an impact on model predictive power.

Despite the small number of features in the MAXENT model and the fact that we only included 141 of the 844 test species in our models prediction, MAXENT method had better performances than purely spatial machine learning algorithms (runs 26988, 27102), artificial neural networks (runs 26875) and some CNN methods (runs 27004, 27005) learnt on all species. The few environmental features selected for the species model based on expert knowledge might have enabled to capture an important part of species abundance variance while avoiding to fall in the trap of model over-fitting.

5 Perspectives

Further work include investigating why bias correction methods failed compared to MAXENT, and study the generalisation power, at long spatial range, of MAXENT predictions compared to other more complex models like ANN and CNN, more likely to overfit.

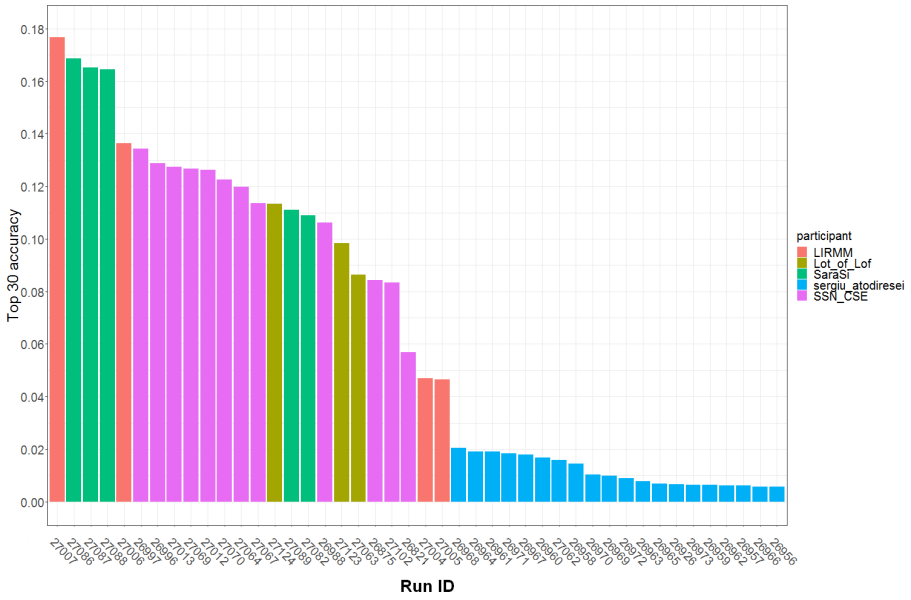

Fig. 1. Top30 accuracy results per run and participant for the GeoLifeCLEF 2019 challenge

References

Alexis Joly, 2019. Alexis Joly, Hervé Goëau, C. B. S. K. M. S. H. G. P. B. W.-P. V. R. P. F.-R. S. H. M. (2019). Overview of lifeclef 2019: a new snapshot of the performance of species identification and prediction algorithms. In *Proceedings of CLEF 2019*.

Botella et al., 2019a. Botella, C., Joly, A., Bonnet, P., Munoz, F., and Monestiez, P. (2019a). [under review] jointly estimating ecological niches and spatial sampling effort from multiple species occurrences. In *Methods in Ecology and Evolution*.

Botella et al., 2019b. Botella, C., Servajean, M., Bonnet, P., and Joly, A. (2019b). Overview of geolifeclef 2019: plant species prediction using environment and animal occurrences. In *CLEF working notes 2019*.

Diggle, 2003. Diggle, P. (2003). Statistical analysis of spatial point patterns: Oxford university press. *New York*.

Elith and Leathwick, 2009. Elith, J. and Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697.

Mod et al., 2016. Mod, H. K., Scherrer, D., Luoto, M., and Guisan, A. (2016). What we use is not what we know: environmental predictors in plant distribution models. *Journal of Vegetation Science*, 27(6):1308–1322.

Phillips et al., 2017. Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., and Blair, M. E. (2017). Opening the black box: an open-source release of maxent. *Ecography*.

Phillips et al., 2009. Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only

distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197.

Renner et al., 2015. Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

CLC category description	spht category name	Raster code
Non-irrigated arable land	cultivated	12
Permanently irrigated land	cultivated	13
Vineyards	cultivated	15
Fruit trees and berry plantations	cultivated	16
Complex cultivation patterns	cultivated	20
agriculture, with areas of natural vegetation	cultivated	21
Agro-forestry areas	cultivated	22
Pastures	grasslands	18
Natural grasslands	grasslands	26
Moors and heathland	grasslands	27
Sclerophyllous vegetation	grasslands	28
Broad-leaved forest	forest	23
Coniferous forest	forest	24
Mixed forest	forest	25
Transitional woodland-shrub	forest	29
Continuous urban fabric	urban	1
Discontinuous urban fabric	urban	2
Industrial or commercial units	urban	3
Road and rail networks and associated land	urban	4
Airports	urban	6
Green urban areas	urban	10
Sport and leisure facilities	urban	11
Port areas	other	5
Mineral extraction sites	other	7
Dump sites	other	8
Construction sites	other	9
Rice fields	other	14
Olive groves	other	17
Annual crops associated with permanent crops	other	19
Beaches, dunes, sands	other	30
Bare rocks	other	31
Sparsely vegetated areas	other	32
Burnt areas	other	33
Glaciers and perpetual snow	other	34
Inland marshes	other	35
Peat bogs	other	36
Salt marshes	other	37
Salines	other	38
Intertidal flats	other	39
Water courses	other	40
Water bodies	other	41
Coastal lagoons	other	42
Estuaries	other	43
Sea and ocean	other	44
No data	other	48
Unclassified land surface	other	49
Unclassified water bodies	other	50

Table 1. spht (Aggregated land cover) categories correspondance with Corine Land Cover 2012.