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

Monestiez Pascal\textsuperscript{1}, Christophe Botella\textsuperscript{2}

\textsuperscript{1} INRA, BioSP, Avignon, France
\textsuperscript{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 de-
rive, 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 imple-
mented 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 back-
ground points ([Phillips et al., 2009] (called TGB), a way of correcting for sam-
pling 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 con-
stant 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 cer-
tainty 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\textsubscript{1}, chbio\textsubscript{5}), annual precipitations (chbio\textsubscript{12}), potential evapo-transpiration (etp), elevation (alti), slope (slope), available water capacity of the soil (awc\textsubscript{top}), a soil pH proxy (bs\textsubscript{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\textsubscript{1}, chbio\textsubscript{5}, chbio\textsubscript{12}-etp, etp, alti and slope. We combined annual precipitations chbio\textsubscript{12} and potential evapotranspiration etp into chbio\textsubscript{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.

$$
\sim \text{etp} + \text{I}(\text{etp}^2) + \text{I(chbio\textsubscript{12}-etp}) + \text{I((chbio\textsubscript{12}-etp)^2)} \\
+ chbio\textsubscript{1} + \text{I(chbio\textsubscript{1}^2)} + chbio\textsubscript{5} + \text{I(chbio\textsubscript{5}^2)} + alti + \text{I(alti}^2) \\
+ slope + \text{I(slope}^2) + awc\textsubscript{top} + \text{I(awc\textsubscript{top}^2)} + bs\textsubscript{top} + \text{I(bs\textsubscript{top}^2)} \\
+ spht + slope:\text{I(chbio\textsubscript{12}-etp)}
$$

(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/](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 iden-
tification $\geq 0.98$) 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 in-
tercept $\alpha_i$ 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 $Z_{\text{test}}$ 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 el-
ementary 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 en-
vironmental 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
$z \to \exp(\sum_{j=1}^{N_i} \gamma_j 1_{z \in c_j}) \exp(\beta_i^T x(z))$. 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 struc-
ture 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 includ-
ing 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/ltHjVfFr5fSGL1S4hXqqTzly/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](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


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

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