HAL open science

Overview of GeoLifeCLEF 2021: Predicting species distribution from 2 million remote sensing images

Titouan Lorieul, Elijah Cole, Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Alexis Joly

To cite this version:

Titouan Lorieul, Elijah Cole, Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, et al.. Overview of GeoLifeCLEF 2021: Predicting species distribution from 2 million remote sensing images. CLEF 2021 - Conference and Labs of the Evaluation Forum, Guglielmo Faggioli; Nicola Ferro; Alexis Joly; Maria Maistro; Florina Piroi, Sep 2021, Bucarest, Romania. pp.1451-1462. hal-03353487

HAL Id: hal-03353487

https://hal.inrae.fr/hal-03353487v1

Submitted on 24 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Overview of GeoLifeCLEF 2021: Predicting species distribution from 2 million remote sensing images

Titouan Lorieul[1], Elijah Cole[2], Benjamin Deneu[1], Maximilien Servajean[3], Pierre Bonnet[4] and Alexis Joly[1]

[1]*Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France*
[2]*Department of Computing and Mathematical Sciences, Caltech, USA*
[3]*LIRMM, AMI, Univ Paul Valéry Montpellier, Univ Montpellier, CNRS, Montpellier, France*
[4]*CIRAD, UMR AMAP, Montpellier, France*

Abstract

Understanding the geographic distribution of species is a key concern in conservation. By pairing species occurrences with environmental features, researchers can model the relationship between an environment and the species which may be found there. To advance research in this area, a large-scale machine learning competition called *GeoLifeCLEF 2021* was organized. It relied on a dataset of 1.9 million observations from 31K species mainly of animals and plants. These observations were paired with high-resolution remote sensing imagery, land cover data, and altitude, in addition to traditional low-resolution climate and soil variables. The main goal of the challenge was to better understand how to leverage remote sensing data to predict the presence of species at a given location. This paper presents an overview of the competition, synthesizes the approaches used by the participating groups, and analyzes the main results. In particular, we highlight the ability of remote sensing imagery and convolutional neural networks to improve predictive performance, complementary to traditional approaches.

Keywords

LifeCLEF, evaluation, benchmark, biodiversity, presence-only data, environmental data, remote sensing imagery, species distribution, species distribution models

1. Introduction

In order to make informed conservation decisions, it is essential to understand where different species live. Citizen science projects now generate millions of geo-located species observations every year, covering tens of thousands of species. But how can these point observations be used to predict what species might be found at a new location?

A common approach is to build a *species distribution model* (SDM) [1], which uses a location's *environmental covariates* (e.g. temperature, elevation, land cover) to predict whether a species is likely to be found there. Once trained, the model can be used to make predictions for any location where those covariates are available.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
titouan.lorieul@inria.fr (T. Lorieul); ecole@caltech.edu (E. Cole); benjamin.deneu@inria.fr (B. Deneu); servajean@lirmm.fr (M. Servajean); pierre.bonnet@cirad.fr (P. Bonnet); alexis.joly@inria.fr (A. Joly)
0000-0001-5228-9238 (T. Lorieul); 0000-0001-6623-0966 (E. Cole); 0000-0003-0640-5706 (B. Deneu); 0000-0002-9426-2583 (M. Servajean); 0000-0002-2828-4389 (P. Bonnet); 0000-0002-2161-9940 (A. Joly)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-2936/paper-124.pdf

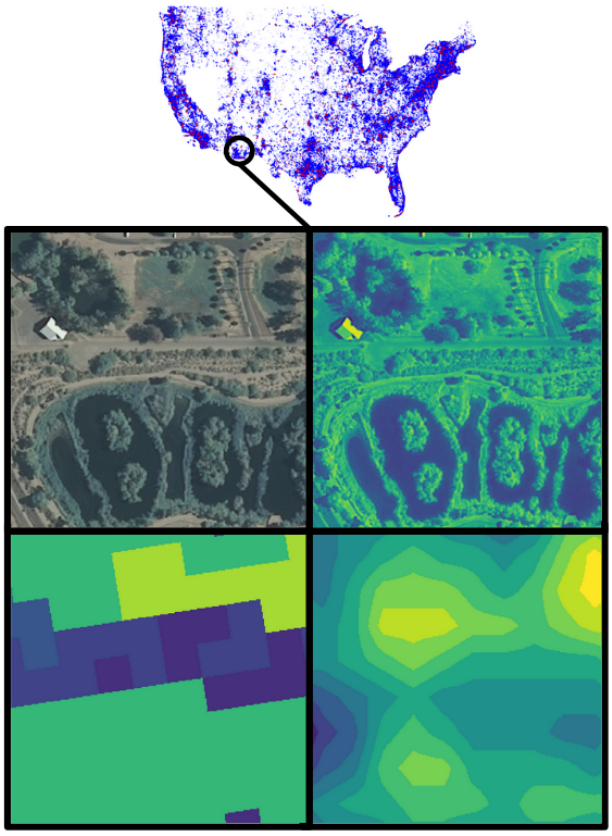


Figure 1: Each species observation is paired with high-resolution covariates (clockwise from top left: RGB imagery, IR imagery, altitude, land cover).

Developing an SDM requires a dataset where each species observation is paired with a collection of environmental covariates. However, many existing SDM datasets are both highly specialized and not readily accessible, having been assembled by scientists studying particular species or regions. In addition, the provided environmental covariates are typically coarse, with resolutions ranging from hundreds of meters to kilometers per pixel.

In this work, we present the results of the GeoLifeCLEF 2021 competition which is part of the LifeCLEF evaluation campaign [2] and co-hosted in Eighth Workshop on Fine-Grained Visual Categorization (FGVC8)[1] at CVPR 2021. This competition is the fourth GeoLifeCLEF challenge. In the first two editions, GeoLifeCLEF 2018 [3] and GeoLifeCLEF 2019 [4], each observation was associated only with environmental features given as vectors or patches extracted around the observation. Like last year's competition, GeoLifeCLEF 2020 [5], GeoLifeCLEF 2021 is aimed at bridging the previously mentioned gaps by (i) sharing a large-scale dataset of observations paired with high-resolution covariates and (ii) defining a common evaluation methodology to measure the predictive performance of models trained on this dataset. The dataset is based on over 1.9 million observations of plant and animal species. Each observation is paired with

[1]https://sites.google.com/view/fgvc8

high-resolution remote sensing imagery – see Figure 1 – as well as traditional environmental covariates (e.g. climate, altitude and soil variables). To the best of our knowledge, this is the first publicly available dataset to pair remote sensing imagery with species observations. Our hope is that this analysis-ready dataset and associated evaluation methodology will (i) make SDM and related problems more accessible to machine learning researchers and (ii) facilitate novel research in large-scale, high-resolution, and remote-sensing-based species distribution modeling.

2. Dataset and evaluation protocol presentation

Data collection. The data for this year's challenge is the same as last year reorganized in a more easy-to-use and compact format. A detailed description of the GeoLifeCLEF 2020 dataset is provided in [6]. In a nutshell, it consists of 1,921,123 observations covering $31,435$ species (mainly plants and animals) distributed across US (1,097,640) and France (823,483), as shown in Figure 2. Each species observation is paired with high-resolution covariates (RGB-IR imagery, land cover and altitude) as illustrated in Figure 1. These high-resolution covariates are resampled to a spatial resolution of 1 meter per pixel and provided as 256×256 images covering a 256m $\times$ 256m square centered on each observation. RGB-IR imagery come from the 2009-2011 cycle of the National Agriculture Imagery Program (NAIP) for the US[2], and from the BD-ORTHO® 2.0 and ORTHO-HR® 1.0 databases from the IGN for France[3]. Land cover data originates from the National Land Cover Database (NLCD) [7] for the US and from CESBIO[4] for France. All elevation data comes from the NASA Shuttle Radar Topography Mission (SRTM)[5]. In addition, the dataset also includes traditional coarser resolution covariates: 19 bio-climatic rasters (30arcsec^2/pixel, i.e., 1km^2/pixel, from WorldClim [8]) and 8 pedologic rasters (250m^2/pixel, from SoilGrids [9]). The details of these rasters are given in Table 1.

Train-test split. The full set of occurrences was split in a training and testing set using a spatial block holdout procedure as illustrated in Figure 2. This limits the effect of *spatial auto-correlation* in the data [10]. Using this splitting procedure, a model cannot perform well by simply interpolating between training samples. The split was based on a global grid of 5km $\times$ 5km quadrats. 2.5% of these quadrats were randomly sampled and the observations falling in those formed the test set. 10% of those observations were used for the public leaderboard on Kaggle while the remaining 90% allowed to compute the private leaderboard providing the final results of the challenge. Similarly, another 2.5% of the quadrats were randomly sampled to provide an official validation set. The remaining quadrats and their associated observations were assigned to the training set.

Evaluation metric. For each occurrence in the test set, the goal of the task was to return a candidate set of species likely to be present at that location. Due to the *presence-only* [11] nature of the observation data used during the evaluation of the methods, for each location in the test set, we only have the knowledge of the presence of one species – the one observed – among the

[2]https://www.fsa.usda.gov
[3]https://geoservices.ign.fr
[4]http://osr-cesbio.ups-tlse.fr/~oso/posts/2017-03-30-carte-s2-2016/
[5]https://lpdaac.usgs.gov/products/srtmgl1v003/

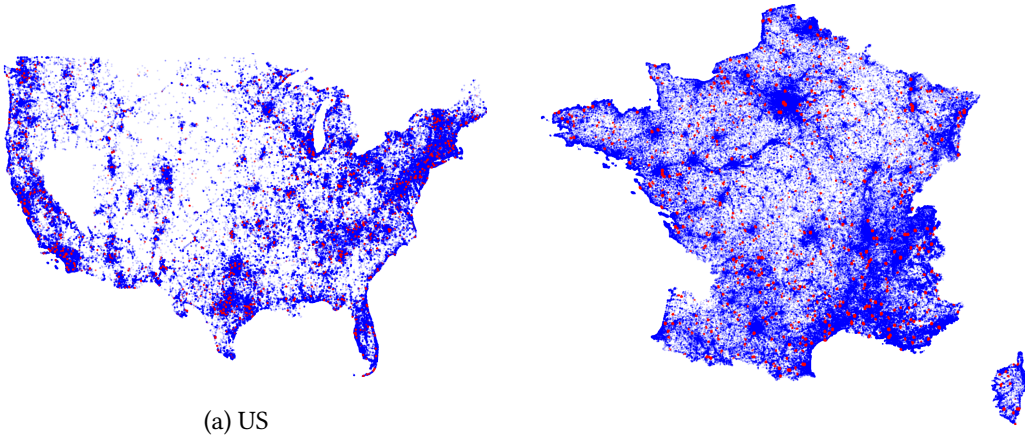
Table 1
Summary of the low-resolution environmental variable rasters provided. The first 19 rows correspond to the bio-climatic variables from WorldClim [8]. The last 8 rows correspond to the pedologic variables from SoilGrid [9].

Name	Description	Resolution
bio_1	Annual Mean Temperature	30 arcsec
bio_2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	30 arcsec
bio_3	Isothermality (bio_2/bio_7) (* 100)	30 arcsec
bio_4	Temperature Seasonality (standard deviation *100)	30 arcsec
bio_5	Max Temperature of Warmest Month	30 arcsec
bio_6	Min Temperature of Coldest Month	30 arcsec
bio_7	Temperature Annual Range (bio_5-bio_6)	30 arcsec
bio_8	Mean Temperature of Wettest Quarter	30 arcsec
bio_9	Mean Temperature of Driest Quarter	30 arcsec
bio_10	Mean Temperature of Warmest Quarter	30 arcsec
bio_11	Mean Temperature of Coldest Quarter	30 arcsec
bio_12	Annual Precipitation	30 arcsec
bio_13	Precipitation of Wettest Month	30 arcsec
bio_14	Precipitation of Driest Month	30 arcsec
bio_15	Precipitation Seasonality (Coefficient of Variation)	30 arcsec
bio_16	Precipitation of Wettest Quarter	30 arcsec
bio_17	Precipitation of Driest Quarter	30 arcsec
bio_18	Precipitation of Warmest Quarter	30 arcsec
bio_19	Precipitation of Coldest Quarter	30 arcsec
orcdrc	Soil organic carbon content (g/kg at 15cm depth)	250 m
phihox	Ph x 10 in H20 (at 15cm depth)	250 m
cecsol	cation exchange capacity of soil in cmolc/kg 15cm depth	250 m
bdticm	Absolute depth to bedrock in cm	250 m
clyppt	Clay (0-2 micro meter) mass fraction at 15cm depth	250 m
sltppt	Silt mass fraction at 15cm depth	250 m
sndppt	Sand mass fraction at 15cm depth	250 m
bldfie	Bulk density in kg/m3 at 15cm depth	250 m

different ones which can actually be found all together at that point. To measure the precision of the predicted sets while accommodating with this limited knowledge, a simple *set-valued classification* [12] metric was chosen as the main evaluation criterion: top-30 error rate. Each observation i is associated with a single ground-truth label y_i corresponding to the observed species. For each observation, the submissions provided 30 candidate labels $\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,30}$. The top-30 error rate is then computed using

$$\text{Top-30 error rate} = \frac{1}{N}\sum_{i=1}^{N} e_i \quad \text{where} \quad e_i = \begin{cases} 1 & \text{if } \forall k \in \{1, \ldots, 30\},\ \hat{y}_{i,k} \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

Note that this evaluation metric does not try to correct the sampling bias [13] inherent to present-only observation data (linked to the density of population, etc.). The absolute value of the resulting figures should thus be taken with care. Nevertheless, this metric does allow to compare the different approaches and to determine which type of input data and of models are



(a) US

(b) France

Figure 2: Observations distribution over the US and France. Training observation data points are shown in blue while test data points are shown in red.

useful for the species presence detection task.

Course of the challenge. The training and test data were publicly shared in early March 2021 through the Kaggle platform[6]. Any research team wishing to participate in the evaluation could register on the platform and download the data. Each team could submit up to 3 submissions per day to compete on the public leaderboard. A submission (also called a *run* in the next sections) takes the form of a CSV file containing the top-30 predictions of the method being evaluated for all observations in the test set. For each submission, the top-30 error rate was first computed only on a subset of the test set to produce the public leaderboard which was visible to all the participants while the competition was still running. Once the submission phase was closed (mid-May), only 5 submissions per team were retained to compute the private leaderboard using the rest of the test set. These submissions were either hand-picked by the team or automatically chosen as the 5 best performing submissions on the public leaderboard. The participants could then see the final scores of all the other participants on the private leaderboard as well as their final ranking. Each participant was asked to provide a *working note*, i.e. a detailed report containing all technical information required to reproduce the results of the submissions. All LifeCLEF *working notes* were reviewed by at least two members of the LifeCLEF organizing committee to ensure a sufficient level of quality and reproducibility.

3. Baseline methods

Based on last year's competition, three baselines were provided by the organizers of the challenge to serve as comparison references for the participants while developing their own methods. They consisted in:

- **Top-30 most present species:** a constant predictor returning always the same list of

[6]https://www.kaggle.com/c/geolifeclef-2021/

the most present species, i.e. the ones having the most occurrences in the training set.

- **RF on environmental variables:** a Random Forest model trained on environmental feature vectors only, i.e. on the 27 climatic and soil variables extracted at the position of the observation.
- **CNN on 6-channels patches:** this method [14] is the one that obtained the best result during GeoLifeCLEF 2020 competition [5]. It is based on a Convolution Neural Network trained on all the high-resolution image covariates, i.e. on 6-channel tensors composed of RGB-IR images, land cover image, and altitude image.

4. Participants and methods

Seven teams participated to the GeoLifeCLEF 2021 challenge and submitted a total of 26 submissions: *University of Melbourne*, *DUTH* (Democritus University of Thrace), *CONABIO* (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad), *UTFPR* (Federal University of Technology – Paraná) as well as three participants for which we could not identify the affiliation and which we denote here, respectively, as *Team Alpha*, *Team Beta* and *Team Gamma*. The results are shown in Figure 3. Only the winning team (*University of Melbourne*) submitted a working notes paper with a detailed description of the used methodology and models trained [15]. We summarize hereafter the two main models they employed:

- **CNN on RGB patches** (*University of Melbourne - Run 1*): this method is based on a Convolution Neural Network (ResNet50) trained on the RGB remote sensing images only. The model was first trained in an unsupervised way using MOCO on the dataset, a contrastive representation learning framework (for 20 epochs). The resulting model was then supervisely fine-tuned entirely (i.e., end-to-end) using 7 epochs of stochastic gradient descent (SGD) on the whole training set (validation set included). Three types of data augmentation were used to reduce overfitting: (i) random horizontal flip, (ii) random vertical flip, and (iii) RandAugment [16], an automated data augmentation optimizer.
- **Bi-modal CNN on RGB patches & altitude** (*University of Melbourne - Run 2*): The CNN on RGB patches was combined with another CNN on the altitude patches (using the same ResNet50 architecture). This other CNN was also pre-trained in an unsupervised way similarly using MOCO as for the RGB-based model. The two models are then combined by concatenating the final bottleneck layer of the ResNet50 (i.e. the new layer contains 4096 neurons instead of 2048 in the mono-modal CNN on RGB patches) and the global model was fine-tuned as before on fewer epochs. Most data augmentation was removed during training.

5. Global competition results analysis

The global results of the GeoLifeCLEF 2021 are shown in Figure 3. Generally speaking, CNN-based models trained on high-resolution patches used in runs by *University of Melbourne* and *Team Alpha* as well as in the baseline *CNN (high resolution patches)* are very competitive and efficient compared to the traditional model (*RF (environmental vectors)*). This observation tends

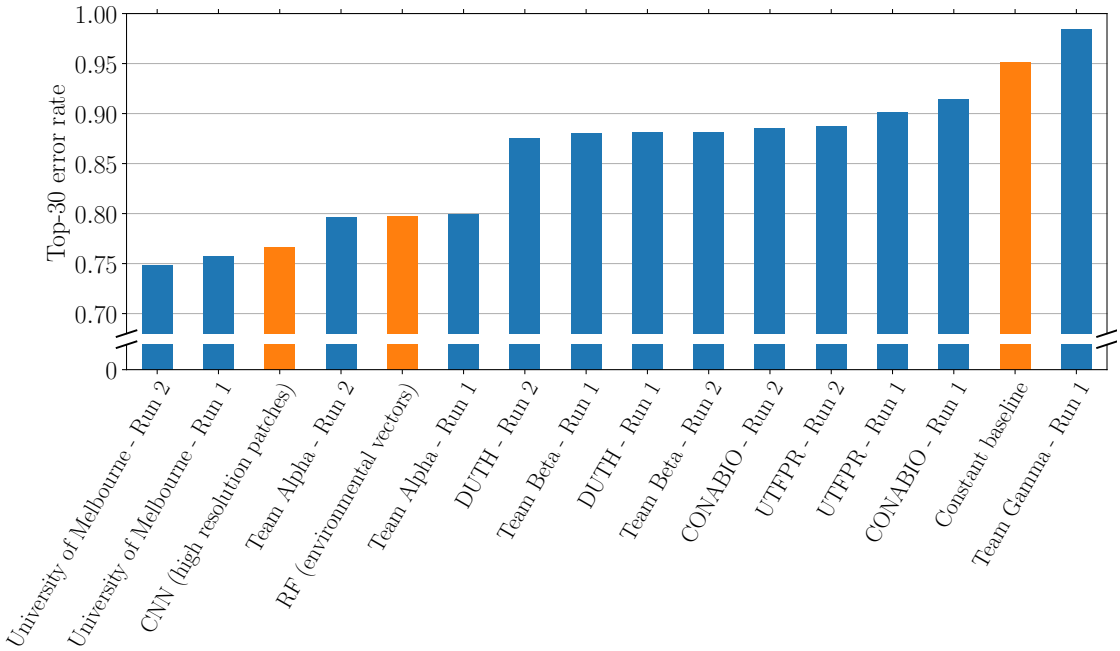


Figure 3: Results of the GeoLifeCLEF 2021 task. The top-30 error rates of the submissions of each participant are shown in blue. The provided baselines are shown in orange.

to show that (i) important information explaining the species composition is contained in the high-resolution patches, and, (ii) convolutional neural networks are able to capture and exploit this information.

One question raised by the challenge is how to properly aggregate the different variables provided as input. Adding altitude data to the model (*University of Melbourne - Run 2*) provides an improvement in prediction accuracy backing the intuition that this variable is informative of the species distribution. However, aggregating all the variables does not mechanically lead to higher performance: *CNN (high resolution patches)* makes use of the additional land cover data but its performance is not as good as the two runs from *University of Melbourne*. It seems that it is important not to aggregate the features representation of those variables too early in the architectures of the networks: concatenation of higher-level features (*University of Melbourne - Run 2*) is more efficient than early aggregation (*CNN (high resolution patches)*).

6. Complementary analysis

In this section, we provide complementary analyses of the submitted results. These analyses are based on the complete predicted scores (the logits for every class of every test occurrence) gently provided by the *University of Melbourne* team [15] after the competition for both of their models. As the dataset used this year is the same as for GeoLifeCLEF 2020, some of the conclusions of last year's overview paper [5] still hold and we refer the reader to it. Here, we focus on other aspects of the dataset not considered previously.

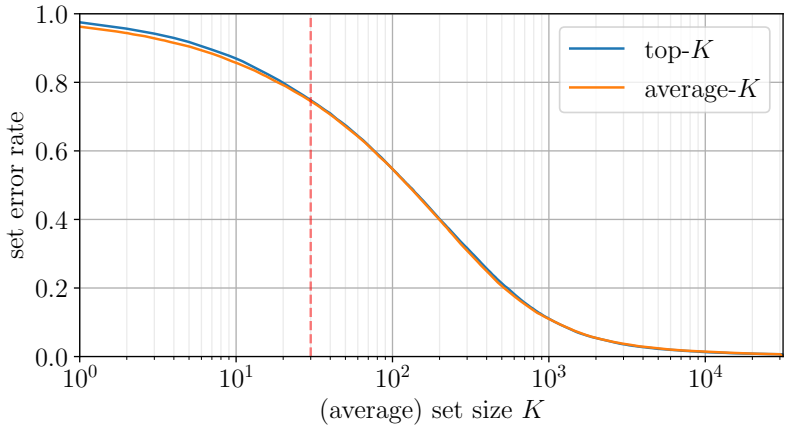
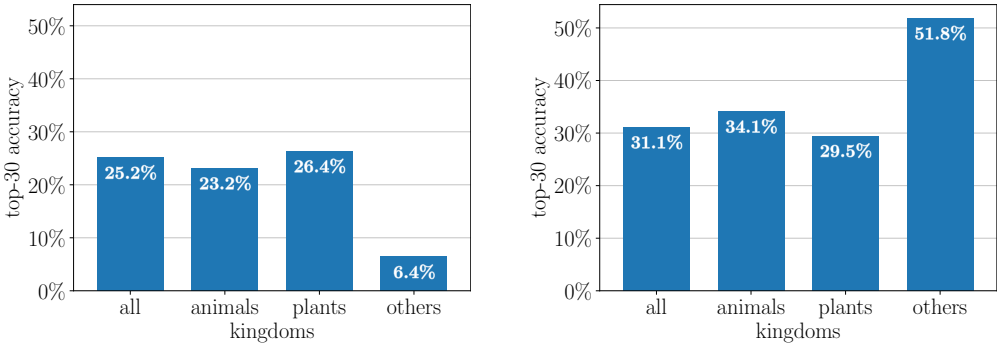


Figure 4: Comparison of the top-K error rate (blue) and average-K error rate (orange) metrics on the predictions of the bi-modal CNN of [15]. In dashed red, the vertical line $K = 30$ corresponding to the value of K used in this year's challenge is displayed. Both metrics are nearly indistinguishable with the provided predicted scores.

Comparison of set-valued classification metrics. As stated previously, the task considered here fits in the framework of set-valued classification [12]. A set of species is present at each location, the model tries to predict this set, however, we are only given a single of those present species to evaluate the performance of the model. This year, the top-30 error rate was chosen as the main evaluation metric. This metric however considers that, at each location, the same number of species will be present, implicitly assuming that the *species richness* is uniform over the geographical extent considered. To surpass this limitation, we can relax this constraint of predicting always sets of size K by allowing to predict adaptive sets which size in average is equal to K. This type of set-valued classification is known as *average-K* classification [17]. In Figure 4, we compare top-K error rate to average-K error rate for all values of K. Average-K error rate is computed by finding the appropriate threshold on the scores which will result in sets of average size K on the test set. We refer the reader to [17, 12] for more details. On the comparison plot, average-K error rate can only improve upon top-K error rate. However, in our case, both curves are indistinguishable. Meaning that either, average-K does not capture the proper heterogeneity of the samples between each location [17] or that the predicted scores are not good enough to estimate it properly. We tried additional probability calibration procedures such as temperature scaling [18] to improve the average-K predictions but it did not provide a significant edge. Further investigation is required to understand why average-K set classifiers do not perform better on this dataset.

Kingdoms analysis. Most species are animals (16,328) and plants (13,035). Due to an omission during the construction of the dataset, other kingdoms are also represented in the US, accounting for 2,072 species (mainly Fungi) but representing only 1.5% of the observations. They are thus negligible. In Figure 5, we detail the performance of the bi-modal CNN of [15] for these different kingdoms. The main evaluation metric consists of all the kingdoms mixed together. For the separate kingdoms metrics, only the test points which observed species belonging to this specific kingdom are considered. If we directly compute the metrics on the top-30 sets



(a) Single top-30 sets used for all kingdoms.

(b) Separate top-30 sets for each kingdom.

Figure 5: Comparison of the top-30 accuracy on the different kingdoms of species present in the dataset based on the predictions of the bi-modal CNN of [15]. In Figure 5a, the top-30 sets are predicted for each test observation and the error rate is then computed separately for all the observations belonging to the different kingdoms. In Figure 5b, for each test observation, only the species belonging to the kingdom of the observed species are used to build the top-30 sets resulting in different top-30 sets for each kingdoms.

predicted as in Figure 5a, it would seem that the task is harder for animals than for plants. However, this gives a biased image of the difficulty of the task. A better approach is to compute separate top-30 sets for each kingdom by retaining only the species belonging to it. In this case, the predicted top-30 sets at the same location are thus different for each kingdom. The resulting error rates are shown in Figure 5b. In that case, animals, although containing more species and fewer observations, seem easier to predict than plants. Surprisingly, although the dataset contains few (28,573) observations from other kingdoms, they seem to be easier to predict. This is actually likely to be simply an artefact of the metric: as they are relatively few species compared to the other kingdoms, taking the top-30 most probable species is probably a too large and not very informative set. This analysis suggests that predicting separate top-K sets for each kingdom and use these sets to derive a new evaluation metric is worth considering. Note that, this might not be as direct as it seems. Indeed, it could be important to use different values of K for the different kingdoms depending of the number of species they contain and the likelihood to find several of them present at the same location. This would thus require some tweaking.

Late-fusion of models. We test the complementarity of the predictions made by the different methods by aggregating their output (the logits) in a late-fusion manner: their outputs are averaged before computing the top-30 most confident species. Here, we focus on the runs submitted by team *University of Melbourne* as well as the baseline *CNN (high resolution patches)*. The results are shown in Figure 6. As explained previously, one important difference between the baseline and the runs submitted is the way the variables are aggregated. Surprisingly, fusing the predictions of the baseline method with either the uni-modal or bi-modal model from *University of Melbourne* actually results in top-30 sets of significantly similar accuracy. This seems to suggest that, beyond the aggregation method, the unsupervised pre-training used by *University of Melbourne* might be particularly helpful for RGB data but less for altitude data

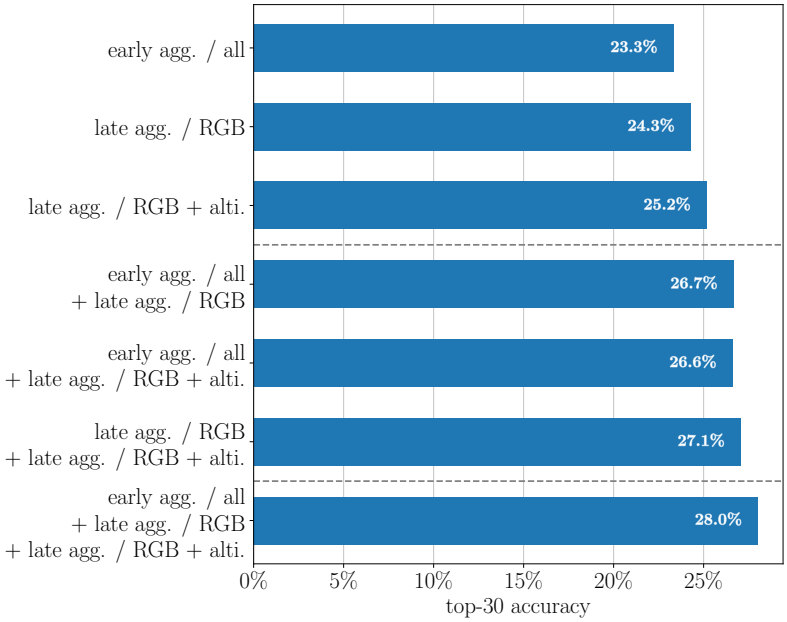


Figure 6: Comparison of top-30 accuracy of the late-fusion of the CNN-based models using different covariates and different aggregation methods.

which is already fairly well exploited by the baseline. Another interesting observation is that combining the uni-modal and bi-modal models from [15] provides a little additional gain of about 2% This might come up as a surprise: the bi-modal model consists essentially of the uni-modal model with additional altitude data. This additional gain might come from three factors:

- the slight differences in the training of uni-modal and bi-modal models, i.e. data augmentation and training time;
- the late aggregation of the altitude data with RGB images might not be perfectly suited and adding a second layer of aggregation via late-fusion might allow to recover it additional complementary information;
- the variance reduction due to ensembling of similar models with high bias.

Finally, late-fusing all three models provides yet another non-negligible gain, highlighting the complementary of those different models. Further investigation could understand better where this complementary originates and help build better models.

7. Conclusion

The main challenges raised by this year's challenge are two folds:

- How to properly aggregate the different covariates provided?
- How much complementary or redundant is the information contained in the high-resolution patches to the one captured from the bioclimatic and soil variables?

Moreover, there remains considerable room for improvement on this challenge as the winning solution does not make use of all the different patches provided and its top-30 error rate is still high, near 75% error rate.

Acknowledgement

This project has received funding from the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 and from the European Union's Horizon 2020 research and innovation program under grant agreement No 863463 (Cos4Cloud project).

References

[1] J. Elith, J. R. Leathwick, Species Distribution Models: Ecological Explanation and Prediction Across Space and Time, Annual Review of Ecology, Evolution, and Systematics (2009).

[2] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, I. Bolon, H. Glotin, R. Planqué, R. Ruiz De Castañeda, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Overview of lifeclef 2021: an evaluation of machine-learning based species identification and species distribution prediction, in: Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021), 2021.

[3] C. Botella, P. Bonnet, F. Munoz, P. Monestiez, A. Joly, Overview of GeoLifeCLEF 2018: location-based species recommendation, CLEF: Conference and Labs of the Evaluation Forum (2018).

[4] C. Botella, M. Servajean, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences, CLEF: Conference and Labs of the Evaluation Forum (2019).

[5] B. Deneu, T. Lorieul, E. Cole, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, A. Joly, Overview of LifeCLEF location-based species prediction task 2020 (GeoLifeCLEF), in: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece., 2020.

[6] E. Cole, B. Deneu, T. Lorieul, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, A. Joly, The GeoLifeCLEF 2020 dataset, arXiv preprint arXiv:2004.04192 (2020).

[7] C. Homer, J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. Herold, J. Wickham, K. Megown, Completion of the 2011 national land cover database for the conterminous united states – representing a decade of land cover change information, Photogrammetric Engineering & Remote Sensing 81 (2015) 345–354.

[8] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas, International Journal of Climatology: A Journal of the Royal Meteorological Society 25 (2005) 1965–1978.

[9] T. Hengl, J. M. de Jesus, G. B. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, et al., Soilgrids250m: Global gridded soil information based on machine learning, PLoS one 12 (2017).

[10] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J.

Lahoz-Monfort, B. Schröder, W. Thuiller, et al., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (2017) 913–929.
[11] J. L. Pearce, M. S. Boyce, Modelling distribution and abundance with presence-only data, Journal of applied ecology 43 (2006) 405–412.
[12] E. Chzhen, C. Denis, M. Hebiri, T. Lorieul, Set-valued classification–overview via a unified framework, arXiv preprint arXiv:2102.12318 (2021).
[13] S. J. Phillips, M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, S. Ferrier, Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data, Ecological applications 19 (2009) 181–197.
[14] B. Deneu, M. Servajean, A. Joly, Participation of LIRMM / Inria to the GeoLifeCLEF 2020 challenge, in: CLEF working notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece., 2020.
[15] S. Seneviratne, Contrastive representation learning for natural world imagery: Habitat prediction for 30,000 species, in: CLEF working notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania., 2021.
[16] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 702–703.
[17] T. Lorieul, Uncertainty in predictions of deep learning models for fine-grained classification, Ph.D. thesis, Université de Montpellier (UM), FRA., 2020.
[18] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.