I- Train occurrences

I.A- Specific treatments per data source

I.A-(i) Pl@ntNet queries

Pl@ntNet is a smartphone app using machine learning to identify plant species from pictures submitted by a broad public of users. For each submission, also called a **query**, the Pl@ntNet algorithm answers a distribution of probability values across the targeted taxonomic referential. If the users allows it, the query's geolocation is also stored.

For this dataset, we used all the geolocated queries in Western Europe from the beginning of 2017 to november 2018. Occurrences with more than 30 meters geolocation uncertainty were removed. The field "**scName**" present in the final Pl@ntNet datasets provides the original data source taxon name of the occurrence.

I.A-(ii) GeoLifeClef 2018 (GBIF)

Train and test occurrences datasets from the previous year edition (https://www.imageclef.org/node/229) were merged to feed the current challenge. Those plants occurrences were extracted from the Global Biodiversity Information Facility (GBIF) The protocol note of last edition (downloadable on the same page) explains in details their extraction and post-treatments. Note that the original fields from the GBIF were kept, so that participants can use them, even though they will not be given for prediction in the test set. The field "**scName**" present in the final dataset provides the original GBIF taxon name ("**scientificname**") of the occurrence.

I.A-(iii) Non-plants (GBIF)

This data source is made of species that are not plants, but interact somehow with plants, and are likely to carry interesting correlations with plant species presences. Remember that **those species will not be in the list of species to predict in the test set**. Those occurrences have also been extracted from the GBIF (https://www.gbif.org/occurrence/search). We extracted occurrences from 7 non-plant taxonomic groups:

Chordata/ Aves (8,000,000)
Chordata/ Mammalia (1,300,000)
Chordata/ Amphibia (300,000)
Chordata/ Reptilia (200,000)
Arthropoda/ Insecta (3,250,000)
Arthropoda/ Arachnida (70,000)
Fungi/ Basidiomycota (50,000)

3 others filters were applied :

- Basis of record: Human

- Location : include coordinates
- Country or area : France

The 7 extracted files were merged into 1 file and the field "**scName**" present in the final dataset provides the original GBIF taxon name ("**scientificname**") of the occurrence.

I.B- Taxonomic and geographic filters applied to all datasets

Because scientist doesn't name species the same way in all regions of the world, many official lists of names that may include a single species, and sometimes with spelling variations. The distinct data sources don't use the same referentials (list of admissible names). Also, distinct names might be considered as redundant (synonyms) in another referential. GBIF uses its own referential made of several taxonomic referentials, and GBIF occurrences may not be species, but sub-species, or genus, etc. Pl@ntNet uses several plants taxonomic referentials (like bdtfx, GRIN, etc.), so that each occurrence taxon name correspond to one referential.

Thus, for attributing species identifiers in GeoLifeCLEF, it was important to first match all occurrences names to a single taxonomic referential adapted to the French Flora. We chose to use **Taxref v12 referential**. We only kept names matching Taxref v12 according to an exact matching algorithm (see the R script below). Some true species might have been lost due to distinct spelling between the GBIF taxonomy and Taxref.

We only kept points falling inside the French territory (Polygon from https://gadm.org/) or inside a 30 meters buffer zone, to account for geolocation uncertainty.

Finally, occurrences were randomly shuffled to avoid any bias introduced by their order.

A commented R script is provided for this procedure :
https://github.com/maximiliense/GLC19/blob/master/GITHUB_taxonomic_and_spatial_filtering.R

I.C- Common structure of final datasets

All final train occurrence datasets are provided as ";" separated CSV files with a header. Each CSV has at least 4 fields :

- **glc19SpId**: The GLC19 reference identifier for the species name. The correspondence between **glc19SpId** and the textual scientific name in Taxref is given by table **taxaName_glc19SpId.csv.**
- **Longitude**: decimal longitude in the WGS84 coordinate system.
- **Latitude:** decimal latitude in the WGS84 coordinate system.
- **scName:** Original taxon name in the data source.

I.D- Listing of provided datasets

I.D-(i) Pl@ntNet complete dataset

Applying filters B to Pl@ntNet queries described in A-(i), we end up with the Pl@ntNet complete dataset. This file "**PL_complete.csv**" contains 2,377,610 occurrences covering 3,906 species. In this CSV, the field **FirstResPLv2Score** gives the confidence score of the automatically identified species. This dataset is very heterogene in species identification quality. The field **accuracy** gives the coordinate uncertainty in meters mostly computed by smartphone devices.

I.D-(ii) Pl@ntNet trusted dataset

An identification confidence filter was applied to "**Pl@ntNet complete**" dataset. We only kept the occurrences for which the first species probability value was above 0.98. This score has been determined by expert to give a reasonable degree of identification confidence. It removed around 90% of the occurrences. This set of 237,087 occurrences covering 1,364 species with accurate geolocation and identification has never been used in biogeography before. It is provided in the file "**PL_trusted.csv**".

I.D-(iii) GLC 2018 dataset

Applying filters B to GeoLifeCLEF 2018 occurrences described in A-(ii), we end up with the GLC_2018 dataset (file "**GLC_2018.csv**"). It contains 281,952 occurrences covering 3,231 species. With this dataset, occurrences are often aggregated on a same geographic point, which denotes uncertain or degraded geolocation. The field **coordinateuncertaintyinmeters** also informs on the location uncertainty when filled up.

I.D-(iv) Non-Plants dataset

Applying filters B to Non-Plant occurrences described in A-(iii), we end up with the Non-Plants dataset (file "**noPlant.csv**"). It contains 5,771,510 occurrences covering 23,893 taxons. None of these species will appear in the test set, but this complementary dataset could be used for improving models predictive power by using strong correlations between plants species and other taxas.

II- Test Occurrences

We have chosen an independent source dataset of occurrences for the test set. It is extracted from the SILENE database maintained by the Conservatoire Botanique Mediterranéen. Those observations come from various providers including mainly the conservatory himself, but also national parcs, botanical associations or impact study consultants. We removed species not present globally in the train set, fragile species according to the SINP referential "espèces sensibles" and species that are at least vulnerable according to the IUCN red list. This dataset have a high identification certainty, the geolocation certainty is under 50 meters, and it should contain species that are harder to detect or identify.
We used random weighted selection scheme to draw 25,000 test occurrences among the 700,000 initially contained in SILENE occurrences set noted **S**. We compute for each

occurrence **s_i** in **S** a weight **w_i**. We define a spatial scale d=2 kilometers, the scale at which we want the concentration of occurrences to be constant and . We first compute **r_i**, the number of species among occurrences of **S** in the radius of **d** around **s_i.** Secondly, we compute **n_i,** the number of occurrences in the same radius. Then, we define **w_i = 1/(n_i*r_i).** With these weights and the following drawing algorithm, we guaranty in expectation that **(i)** test occurrences are uniformly distributed in the geographic space at scale 2d **(ii)** there is as many occurrences of each present species on neighborhoods of radius 2d.
We use the following algorithm for drawing without replacement occurrences from S:

(0) we initialize the bag of occurrences **S':=S** and the set **T** of retained occurrences for the test set, initially empty.

(i) we randomly draw an occurrence in **S',** say the occurrence **i**

(ii) we uniformly draw a scalar **z~U(0, max(w_1,...,w_|S|))**

(iii) if z<w_i, we remove **s** from **S'** and add it to **T,** otherwise leave it in **S'**

(iv) We stop if |T|=25 000, otherwise we go back to step **(i)**

Note that the test occurrences only cover the French mediterranean region. Thus, their spatio-environmental distribution is biased compared to the train set. The map of test occurrences distribution over the mediterranean region is shown on **Figure 1**.

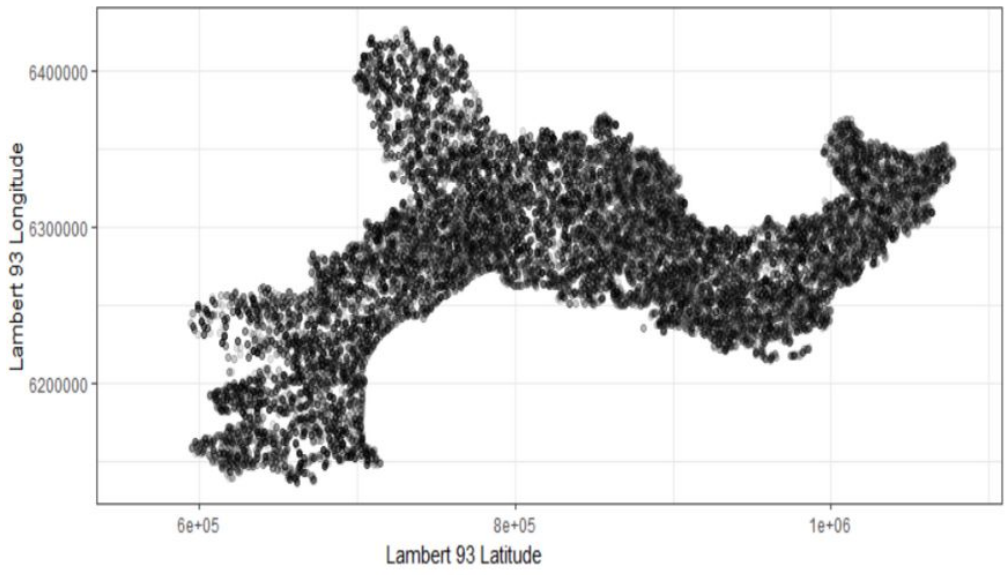

Figure 1. Spatial distribution of the test set GeoLifeCLEF 2019. Those 25 000 species occurrences were drawn with a random weighted selection scheme from the SILENE database to correct their heterogeneous spatial distribution and species representation.

III - Environmental data

Specimens of a plant species tend to aggregate spatially due to dispersal phenomenons. However, a very important filter for their survival is their adequacy to the abiotic environment.

Thus, in addition to know geolocations of species observations, information about the nature of the environment is very useful for modeling species presence or abundance. Those approaches, establishing statistical links between environmental descriptors and species presence, are widely used and usually called Species Distribution Models (SDM). In this task, we supply the participants with a set of 33 Environmental Variables (EV) in the form spatial rasters (TIF files) whose values cover at least the French metropolitan territory. General informations concerning the nature and range of provided EVs are described in **Table 1**. This describe the source data of the provided environmental rasters and the post-processing steps. For a guidance on easy extraction of tensor data from the rasters, using the supplied Python code, please refer the the Github page:
https://github.com/maximiliense/GLC19/blob/master/README.md

III-A Sources and production method of the original environmental data

-- Chelsea Climate data 1.1: those are raster data with worldwide coverage and 1km resolution. A mechanistic climatic model is used to make spatial predictions of monthly mean-max-min temperatures, mean precipitations and 19 bioclimatic variables, which are downscaled with statistical models integrating historical measures of meteorologic stations from 1979 to today. The exact method is explained in the reference paper Karger et al. [1]. The data is under Creative Commons Attribution 4.0 International License and downloadable at (http://chelsa-climate.org/downloads/).

-- The ESDB v2 - 1kmx1km Raster Library (Panagos [2],Van Liedekerke et al. [3]): The library contains multiple soil pedology descriptor raster layers covering Eurasia at a resolution of 1km. We selected 11 descriptors from the library. More precisely, those variables have ordinal format, representing physico-chemical properties of the soil, and come from the PTRDB. The PTRDB variables have been directly derived from the initial Soil Geographical Data Base of Europe (SGDBE) using expert rules. SGDBE was a spatial semantic data base relating spatial units to a diverse pedological attributes of categorical nature, which is not useful for our purpose. For more details, see Panagos et al. [4]. The data is maintained and distributed freely for scientific use by the European Soil Data Centre (ESDAC) at http://eusoils.jrc.ec.europa.eu/content/european-soil-database-v2-raster....

-- Corine Land Cover 2012, version 18.5.1, 12/2016: It is a raster layer describing soil occupation with 48 categories across Europe (25 countries) at a resolution of 100 meters. This classification is the result of an automated interpretation process applied to earth surface high resolution satellite images. This data base of the European Union is freely accessible online for all use at http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012.

Name	Description	Nature	Values
CHBIO_1	Annual Mean Temp.	quanti.	[-10.7,18.4]
CHBIO_2	Mean of monthly max(temp)-min(temp)	quanti.	[7.8,21.0]
CHBIO_3	Isothermality (100*2/7)	quanti.	[41.1,60.0]
CHBIO_4	Temp. Seasonality (std.dev.*100)	quanti.	[302.7,777.8]
CHBIO_5	Max Temp. of Warmest Month	quanti.	[6.1,36.6]
CHBIO_6	Min Temp. of Coldest Month	quanti.	[-28.3,5.4]
CHBIO_7	Temp. Annual Range (5- 6)	quanti.	[16.7,42.0]
CHBIO_8	Mean Temp. of Wettest Quarter	quanti.	[-14.2,23.0]
CHBIO_9	Mean Temp. of Driest Quarter	quanti.	[-17.7,26.5]
CHBIO_10	Mean Temp. of Warmest Quarter	quanti.	[-2.8,26.5]
CHBIO_11	Mean Temp. of Coldest Quarter	quanti.	[-17.7,11.8]
CHBIO_12	Annual Precip.	quanti.	[318.3,2543.3]
CHBIO_13	Precip. of Wettest Month	quanti.	[43.0,285.5]
CHBIO_14	Precip. of Driest Month	quanti.	[3.0,135.6]
CHBIO_15	Precip. Seasonality (Coef. of Var.)	quanti.	[8.2,26.5]
CHBIO_16	Precip. of Wettest Quarter	quanti.	[121.6,855.6]
CHBIO_17	Precip. of Driest Quarter	quanti.	[19.8,421.3]
CHBIO_18	Precip. of Warmest Quarter	quanti.	[19.8,851.7]
CHBIO_19	Precip. of Coldest Quarter	quanti.	[60.5,520.4]
etp	Potential Evapo Transpiration	quanti.	[133,1176]
alti	Elevation	quanti.	[-188,4672]
awc_top	Topsoil available water capacity	ordinal	$\{0, 120, 165, 210\}$
bs_top	Base saturation of the topsoil	ordinal	$\{35, 62, 85\}$
cec_top	Topsoil cation exchange capacity	ordinal	$\{7, 22, 50\}$
crusting	Soil crusting class	ordinal	$[\|0, 5\|]$
dgh	Depth to a gleyed horizon	ordinal	$\{20, 60, 140\}$
dimp	Depth to an impermeable layer	ordinal	$\{60, 100\}$
erodi	Soil erodibility class	ordinal	$[\|0, 5\|]$
oc_top	Topsoil organic carbon content	ordinal	$\{1, 2, 4, 8\}$
pd_top	Topsoil packing density	ordinal	$\{1, 2\}$
text	Dominant surface textural class	ordinal	[|0,5|]
proxi_eau_fast	<50 meters to fresh water	bool.	$\{0, 1\}$
clc	ground occupation	categ.	[|1,48|]

Table 1: The environmental variables supplied for the task.

III.A- Details on the source and production method of the originaldata

-- CGIAR-CSI ETP data: The CGIAR-CSI distributes this worldwide monthly potential-evapotranspiration raster data. It is pulled from a model developed by Antonio

Trabucco (see Zomer et al. [5,6]). Those are estimated by the Hargreaves formula, using mean monthly surface temperatures and standard deviation from WorldClim 1:4 (http://www.worldclim.org/version1), and radiation on top of atmosphere. The raster is at a 1km resolution, and is freely downloadable for a nonprofit use at http://www.cgiar-csi.org/data/global-aridity-and-pet-database#description.

-- USGS Digital Elevation data: The Shuttle Radar Topography Mission achieved in 2010 by Endeavour shuttle managed to measure digital elevation at 3 arc second resolution over most of the earth surface. Raw measures have been post-processed by NASA and NGA in order to correct detection anomalies. The data is available from the U.S. Geological Survey, and downloadable on the Earthexplorer (https://earthexplorer.usgs.gov/). See https://lta.cr.usgs.gov/SRTMVF for more informations.

-- BD Carthage v3: BD Carthage is a spatial semantic database holding many informations on the structure and nature of the french metropolitan hydrological network. For the purpose of plants ecological niche, we focus on the geometric segments representing watercourses, polygons representing hydrographic fresh surfaces and seas. The data has been produced by the Institut National de l'information Géographique et forestière (IGN) from an interpretation of the BD Ortho IGN. It is maintained by the SANDRE under free license for non-profit use and downloadable at http://services.sandre.eaufrance.fr/telechargement/geo/ETH/BDCarthage/FX....

III.B- Building environmental rasters from source datasets

For reproducibility, we explain the treatments applied to the original data. As a first treatment, we crop the source layers as short as possible, to make calculations faster, still including the extent of the metropolitan French territory and Corsica. Then, as the original coordinate system of the layer vary among sources, we change it to WGS84 using the R package **rgdal**, which is the occurrences coordinate system of the GBIF occurrences. Additional processing were necessary to get **proxi_eau**: Its raster have been made from a vector shapefile according to the following procedure. We use qgis to rasterize to a 12.5 meters resolution, with a buffer of 50 meters, the shapefile **COURS_D_EAU.shp** on one hand, and the polygons of **SURFACES_HYDROGRAPHIQUES.shp** with attribute NATURE="Eau douce permanente" on the other hand. We then create the maximum raster of the previous ones (So the value of 1 correspond to an approximate distance of less than 50 meters to a watercourse or hydrographic surface of fresh water). The database is in the form of a folder of 33 subfolders, one per environmental raster, named with the variable name as given in **Table 1**. As the raster themselves are standard .tif, they lack geographic informations. So we introduced in the folder of every environmental raster a file **GeoMetaData.csv** which informs raster's geographic extent, spatial resolution as shown in **Table 2**.

proj4	xmin	ymin	nrows	ncols	resolutionX	resolutionY

+proj=longlat +datum=WGS84 +ellps=WGS84 +towgs84=0,0,0	-7.018	40.042	1313	1357	0.0129	0.00895

Table 2. Table of the file **GeoMetaData.csv** for the **bs_top** variable.

For the particular case of ordinal and categorical environmental rasters, including the 10 pedologic variables coming from the ESDB v2, CORINE land cover (**clc**), and proximity to fresh water (**proxi_eau_fast**) we added an attribute table called **attrib_variablename.csv.** This table gives a correspondence between the .tif value (column **storage_8bit**), the descriptor or code from the original database whose description may be found on the website given above (column **variablename**) and a quantitative value interpretation (column **QUANTI**) for pedologic variables.

III.C- Extraction of an environmental patch from a WGS84 location

Python code is provided for an easy extraction of the environmental data. It enables to automatically extract a tensor of equal size images, called tensor, for a user defined set of environmental variables directly from the provided environmental rasters database, with the option to store it on the disk for a later access by another programmation language than Python. All this procedure is explained in the **GeoLifeCLEF 2019 Github repository**, please look up at the readme on : https://github.com/maximiliense/GLC19.
Figure 1 gives a representation of all the channels of a complete environmental patch.

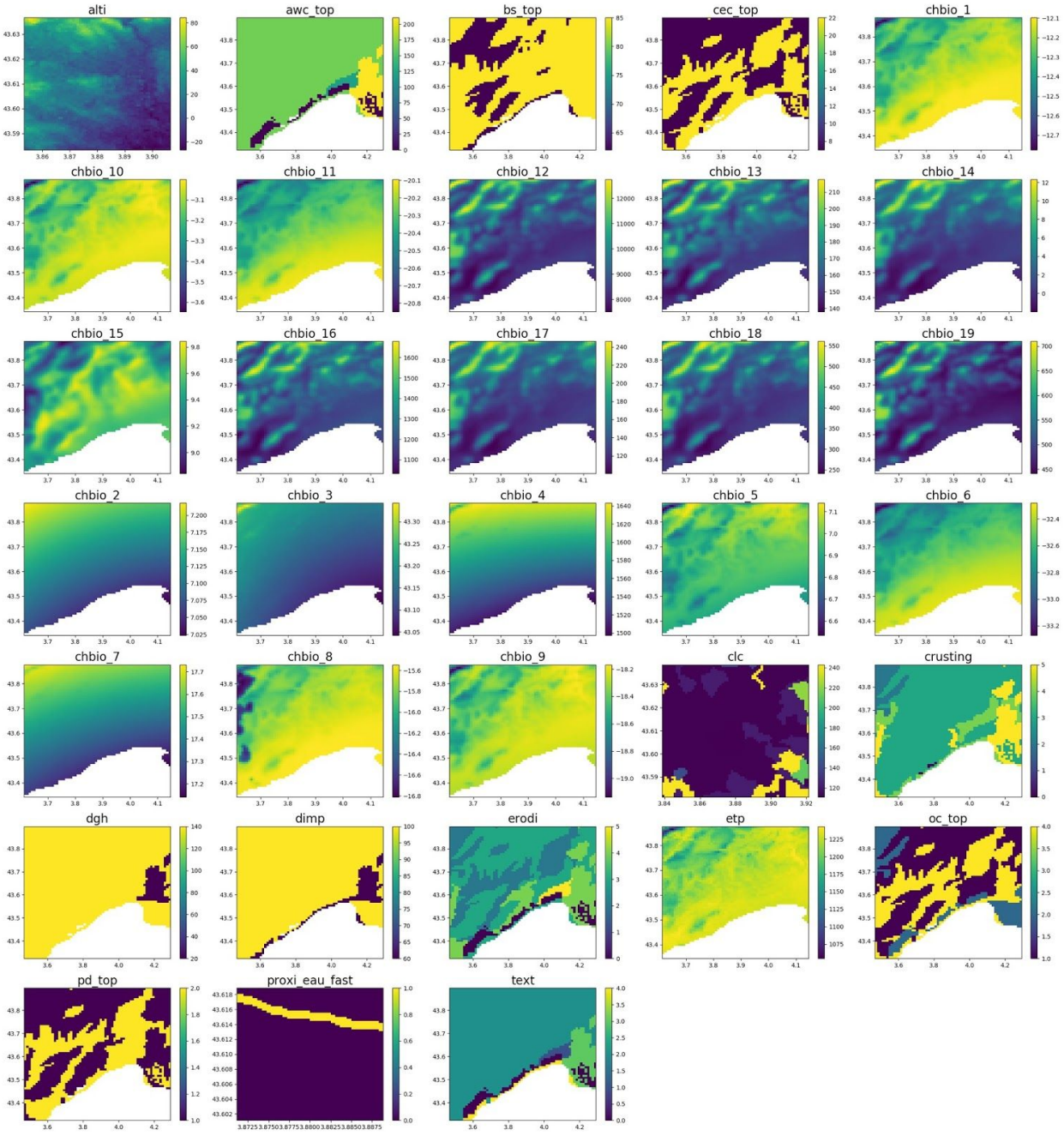

Figure 2. Representation of the 33 channels of a complete environmental patch at position (longitude:3.88,latitude:43.61), i.e. in the surroundings of Montpellier. This tensor is extracted and plotted with the Python code provided in the **GLC19** Github repository.