Detailed data sources and treatments

1) Species occurrences data

Occurrences data were extracted from the Global Biodiversity Information Facility platform (GBIF: https://www.gbif.org/). To achieve precise species prediction from a geolocation, the geolocations in question must be as precise as possible. However, a high number of occurrences from the GBIF have a spatially degraded geolocation for conservation reasons. Thus, we have chosen source datasets with undegraded geolocations in France, which are :

- **Carnet en ligne** (https://www.gbif.org/dataset/baa86fb2-7346-4507-a34f-44e4c1bd0d57) from **Tela Botanica.**
- **Cartographie des Leguminosae (Fabaceae) en France** (https://www.gbif.org/dataset/cbd241aa-a115-4856-af66-fac5cb90f2cc) from **Tela Botanica**
- **Naturgucker** dataset (https://www.gbif.org/dataset/6ac3f774-d9fb-4796-b3e9-92bf6c81c084)
- **iNaturalist Research-grade Observations** (https://www.gbif.org/dataset/50c9509d-22c7-4a22-a47d-8c48425ef4a7).

In order to maximize the size of data while ensuring homogeneous and rich environmental data (see next part), we restricted the domain of study to **the French metropolitan territory**, as can be seen of **Figure 1**. Thus, observations falling outside of it were removed. As some occurrences submitted to the GBIF have been generated from species checklists, several observations of a same species can be at the exact same location. To avoid model bias due to the heterogeneity of the datasets on this matter, **we removed the redundancies of a single species at a single location**. We chose to work **at the species level**, while GBIF occurrences can be at any taxonomic level, even inferior to the species (sub-species, varieties), with the accepted name of the **bdtfx** taxonomic referential for western Europe plant species. We eliminated observations for which the taxonomic matching with the GBIF backbone taxonomy was fuzzy or with higher rank than the species (genus, family etc.). For the remaining observations, we added a field **espece_retenue_bdtfx** which is the name of the species in the referential bdtfx 4.01(from 15/03/2017). The matching was done from the field **scientificname** with an internal exact matching algorithm. If the rank was inferior to the species, we took the corresponding species accepted name. If there was no exact match, the observation was excluded, but this case was rare. The field **species_glc_id** attributes an id for each possible value of **espece_retenue_bdtfx** and will be **our class identifier for the task**.

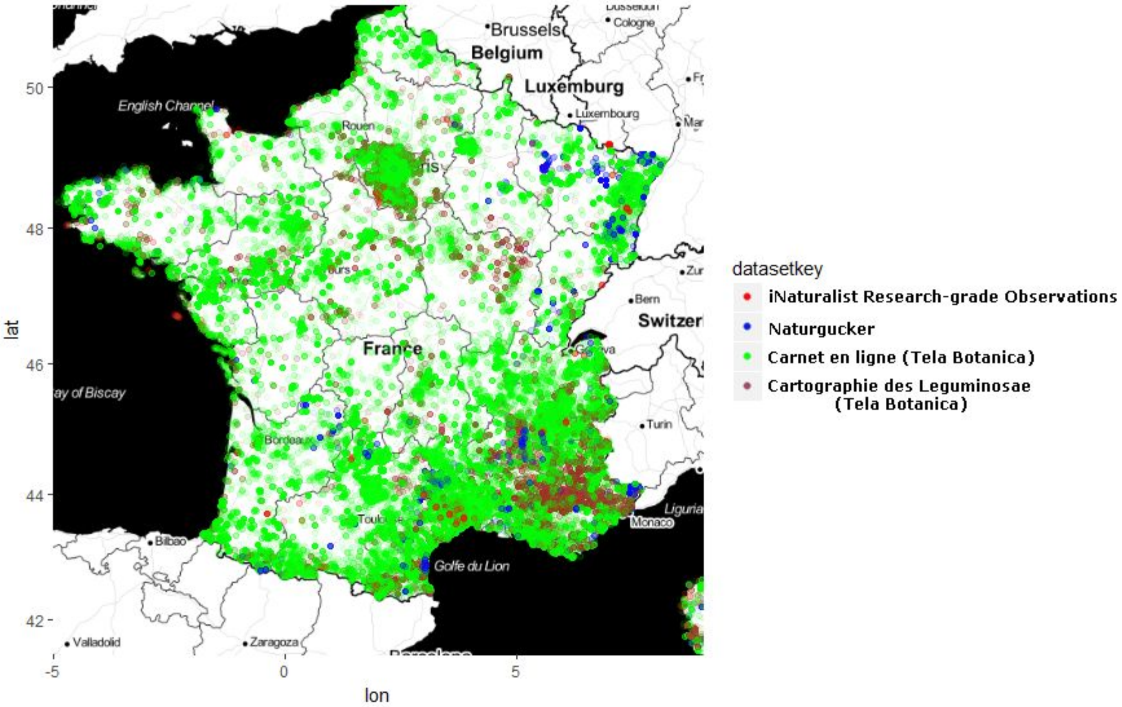


Figure 1: Spatial distribution and source dataset of the occurrences

2) 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 Environmental Variables (EV) along with every observation. General informations concerning the nature and range of provided EVs are described in **Table 1**. Each observation **o** is associated a **TIFF** file containing 33 spatial rasters (centered on the location of **o** and of identical dimensions). Each raster is associated with an EV **j**, and must be understood **as a spatial map of the values of j** around the location of **o.** More precisely, it covers the extent:

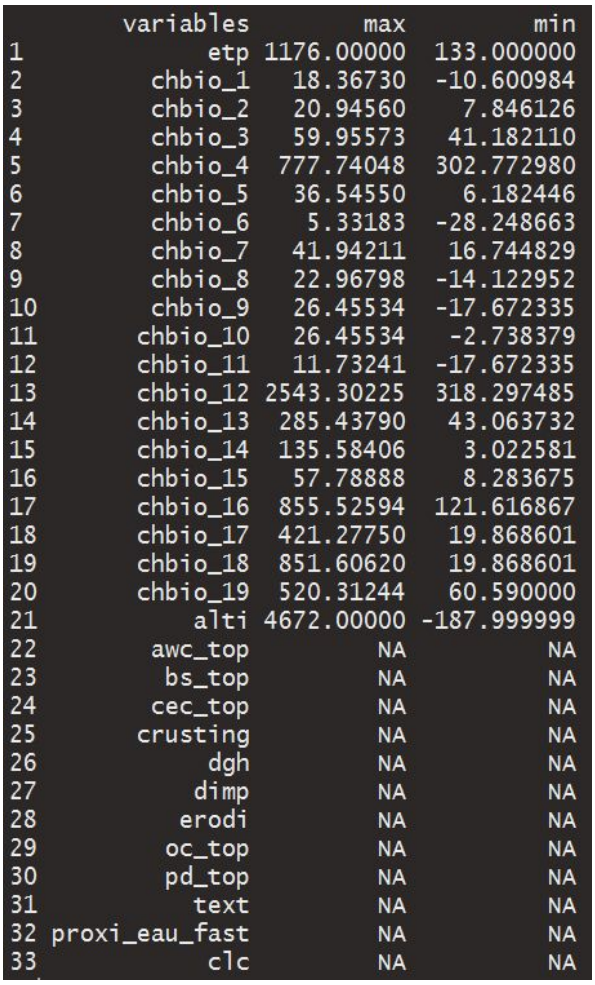
(x_o-res_x_j*32 , x_o+res_x_j*32 , y_o-res_y_j*32 , y_o+res_y_j*32)

Where **x_o** is the longitude of **o, y_o** is its latitude, **res_x_j** is the initial resolution in longitude of EV **j,** and **res_y_j,** its resolution in latitude. The TIFF file of **o** is located in the directory **“patchTrain” if o is a train observation,** or in **“patchTest”** otherwise. Inside each of those directories, there are many subdirectories with names “256”, “512” etc. The field **patch_dirname** gives the name of the subdirectory where is the file of **o**. Then, the file is identified with the field **patch_id**. Values stored in each pixel of a TIFF file are between 0 and 255, thus for **"awc_top", "bs_top", "cec_top", "crusting", "dgh", "dimp", "erodi", "oc_top", "pd_top", "text", "clc",** and **"proxi_eau"** there is no transformation, so **the pixel value corresponds exactly to the original value of the EV.** However, for

"chbio_xx","alti" and **"etp"** one will recover the original value by applying the following operation :

$$\text{original\_value} = \min + (\max-\min) \left[(\text{pixel\_value}/255) -0.1 \right] /0.8$$

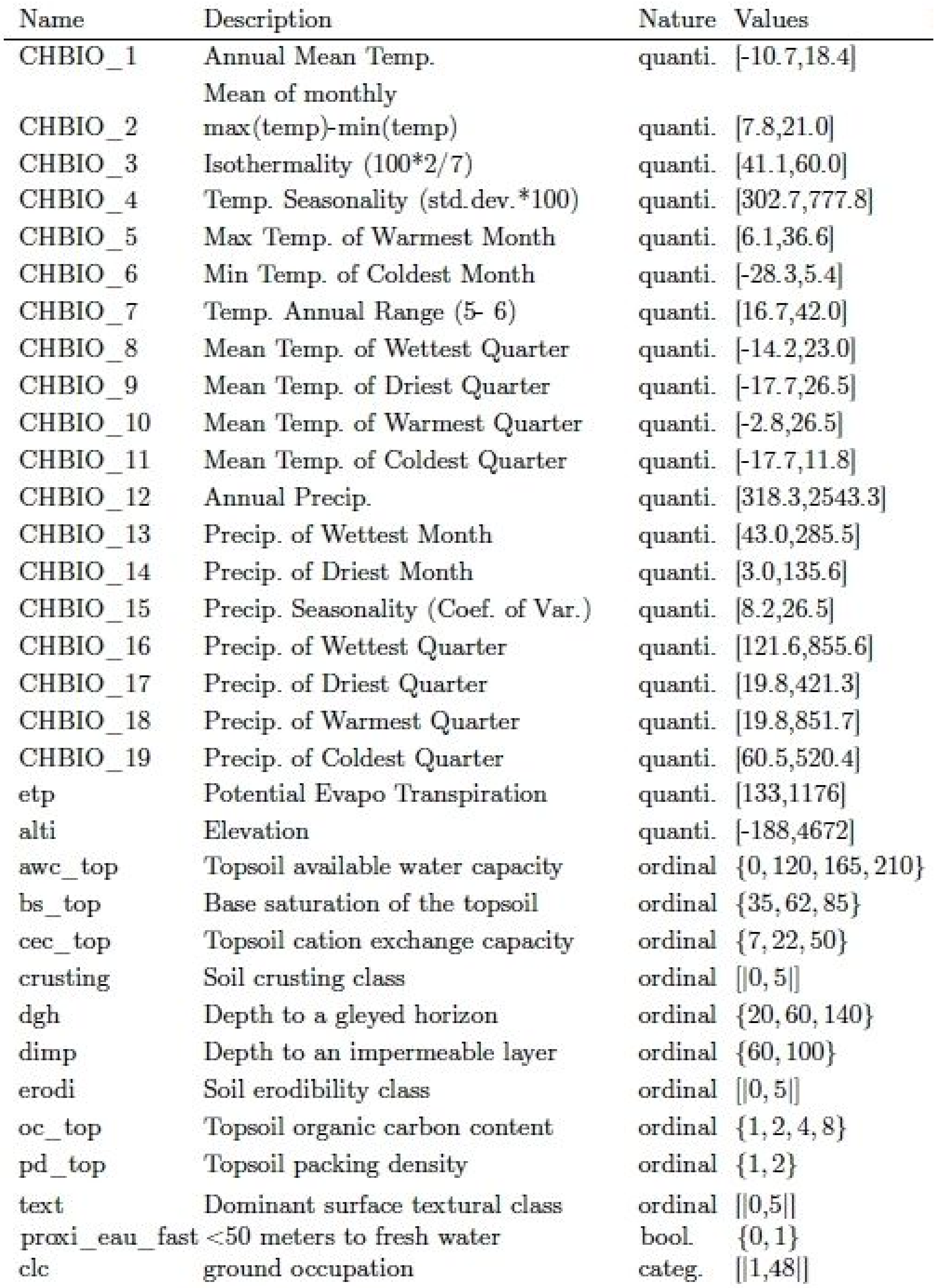
Here are the max/min values of the EVs and their order in the TIFF:



```
        variables        max        min
1             etp 1176.00000 133.000000
2         chbio_1   18.36730 -10.600984
3         chbio_2   20.94560   7.846126
4         chbio_3   59.95573  41.182110
5         chbio_4  777.74048 302.772980
6         chbio_5   36.54550   6.182446
7         chbio_6    5.33183 -28.248663
8         chbio_7   41.94211  16.744829
9         chbio_8   22.96798 -14.122952
10        chbio_9   26.45534 -17.672335
11       chbio_10   26.45534  -2.738379
12       chbio_11   11.73241 -17.672335
13       chbio_12 2543.30225 318.297485
14       chbio_13  285.43790  43.063732
15       chbio_14  135.58406   3.022581
16       chbio_15   57.78888   8.283675
17       chbio_16  855.52594 121.616867
18       chbio_17  421.27750  19.868601
19       chbio_18  851.60620  19.868601
20       chbio_19  520.31244  60.590000
21           alti 4672.00000 -187.999999
22        awc_top         NA         NA
23         bs_top         NA         NA
24        cec_top         NA         NA
25       crusting         NA         NA
26            dgh         NA         NA
27           dimp         NA         NA
28          erodi         NA         NA
29         oc_top         NA         NA
30         pd_top         NA         NA
31           text         NA         NA
32 proxi_eau_fast         NA         NA
33            clc         NA         NA
```

Note: Some participants could be only interested in the punctual value of environmental variables at the occurrence geolocation, as it is often done in ecological modeling. To facilitate the use of the data, we stored those values as columns in the occurrences CSV files. All these values are on the variable original scale.

Note: patch_dirname = -1 * euclidian_division_quotient(-1*patch_id / 256) * 256.



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.

Note: To open the TIFF files correctly and access to all the 33 channels, we advise participants to use a TIFF dedicated library like **tifffile** for **python** or **TIFF** for **R**.

3) Details on the source and production method of the original 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.

-- 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....

Note about variables building: 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, making sure the extent contains all the metropolitan French territory. Then, as the original coordinate system of the layer vary among sources, we change it to WGS84, 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).

4) Train & test sets

The total of 291,392 occurrences were randomly splitted in a train (218,543) and test (72,849) set with the constraints that :

- For each species in the test set, there is at least one observation of it in the train set.
- An observation of a species in the test set is distant of more than 100 meters from all observations of this species in the train set to avoid major reporting dependances.

Train observations are listed in the file **occurrences_train.csv**, with every field from the GBIF, plus the ones described previously. Test observations are in the file **occurrences_test.csv** whose taxonomic informations were removed.

5) Task description

Data for the task can be downloaded at http://otmedia.lirmm.fr/LifeCLEF/GeoLifeCLEF2018/.
Given the test set of plants observations, the goal of the task is to return for each observation a ranked list of plant species sorted according to the likelihood that they might have been observed at that location. The test set as well as the list of the 3,203 candidate species will be provided within the registration/submission system (see main page of LifeCLEF 2018 for registration).
Each participating group is allowed to submit several run files built from different methods. Each run file has to be named as "teamname_runX.run" where X is the identifier of the run (i.e. 1, 2, 3 or 4). The run file has to contain as much lines as the total number of species recommendations, with at least one recommendation per occurrence of the test set and a maximum of XX recommendations per occurrence (3,203 being the total number of species).
Each recommendation (i.e. each line of the run file) has to respect the following format:
< glc_id;species_glc_id;probability;rank>
where **glc_id** is the identifier of an occurrence in the test set, **species_glc_id** is the identifier of one of the 3,203 possible species, probability is a real value in [0;1] decreasing with the confidence in that recommendation, rank is the rank of that recommendation among all recommended species for the test occurrence **glc_id**.
Here is a short fake run example respecting this format on only 3 test occurrences:

```
2777702;3152583;0.614310483913869;2
```

```
2777702;2888740;0.955044784350321;1
2777702;2808330;0.0201473159249872;4
2777702;2926110;0.345950950868428;3
2775920;3152583;0.255595623515546;2
2775920;2888740;0.109065826749429;3
2775920;2808330;0.99270333093591;1
2775920;5415039;0.772525980370119;1
2775920;8324121;0.462680039694533;2
```

For each submitted run, please give within the submission system a short description of the run in the dedicated text area (in addition to the working note to be written later on).

References

[1] Karger, Dirk Nikolaus, Conrad, Olaf, Böhner, Jürgen, Kawohl, Tobias, Kreft, Holger, Soria-Auza, Rodrigo Wilber, Zimmermann, Niklaus, Linder, H Peter, & Kessler, Michael. 2016. Climatologies at high resolution for the earth’s land surface areas. arXiv preprint arXiv :1607.00217.

[2] Panagos, Panos. 2006. The European soil database. GEO : connexion, 5(7), 32–33.

[3] Panagos, Panos, Van Liedekerke, Marc, Jones, Arwyn, & Montanarella, Luca. 2012. European Soil Data Centre : Response to European policy support and public data requirements. Land Use Policy, 29(2), 329–338.

[4] Van Liedekerke, M, Jones, A, & Panagos, P. 2006. ESDBv2 Raster Library-a set of rasters derived from the European Soil Database distribution v2. 0. European Commission and the European Soil Bureau Network, CDROM, EUR, 19945.

[5] Zomer, Robert J, Bossio, Deborah A, Trabucco, Antonio, Yuanjie, Li, Gupta, Diwan C, & Singh, Virendra P. 2007. Trees and water : smallholder agroforestry on irrigated lands in Northern India. Vol. 122. IWMI.

[6] Zomer, Robert J, Trabucco, Antonio, Bossio, Deborah A, & Verchot, Louis V. 2008. Climate change mitigation : A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agriculture, ecosystems & environment, 126(1), 67–80.