Spatiotemporal Ensemble ML in R: with examples / R tutorial

Sept 6, 2021: 13:30 - 15:00

Tom Hengl

tom.hengl@opengeohub.org

https://opengeohub.org

Carmelo Bonannella

carmelo.bonannella@opengeohub.org

https://opengeohub.org

Co-financed by the Connecting Europe Facility of the European Union

Outline

- Ensemble ML: rationale
- Ensemble ML in R: SuperLearner and mlr packages
- Examples with real datasets:
  - Daily temperatures (meteo);
  - Cookfarm dataset 3D+T (landmap);
  - Fagus Sylvatica (eumap);

Why Ensemble ML?

- We can **potentially increase accuracy** (usually not a lot but sometimes even few percent counts);
- EML prediction system is more robust: there is at the order of magnitude more training involved and this helps with the extrapolation problems;
- We can derive model-free prediction errors;
- By doing Ensemble ML, we can combine fine-tuning, model selection, feature selection etc all at once;

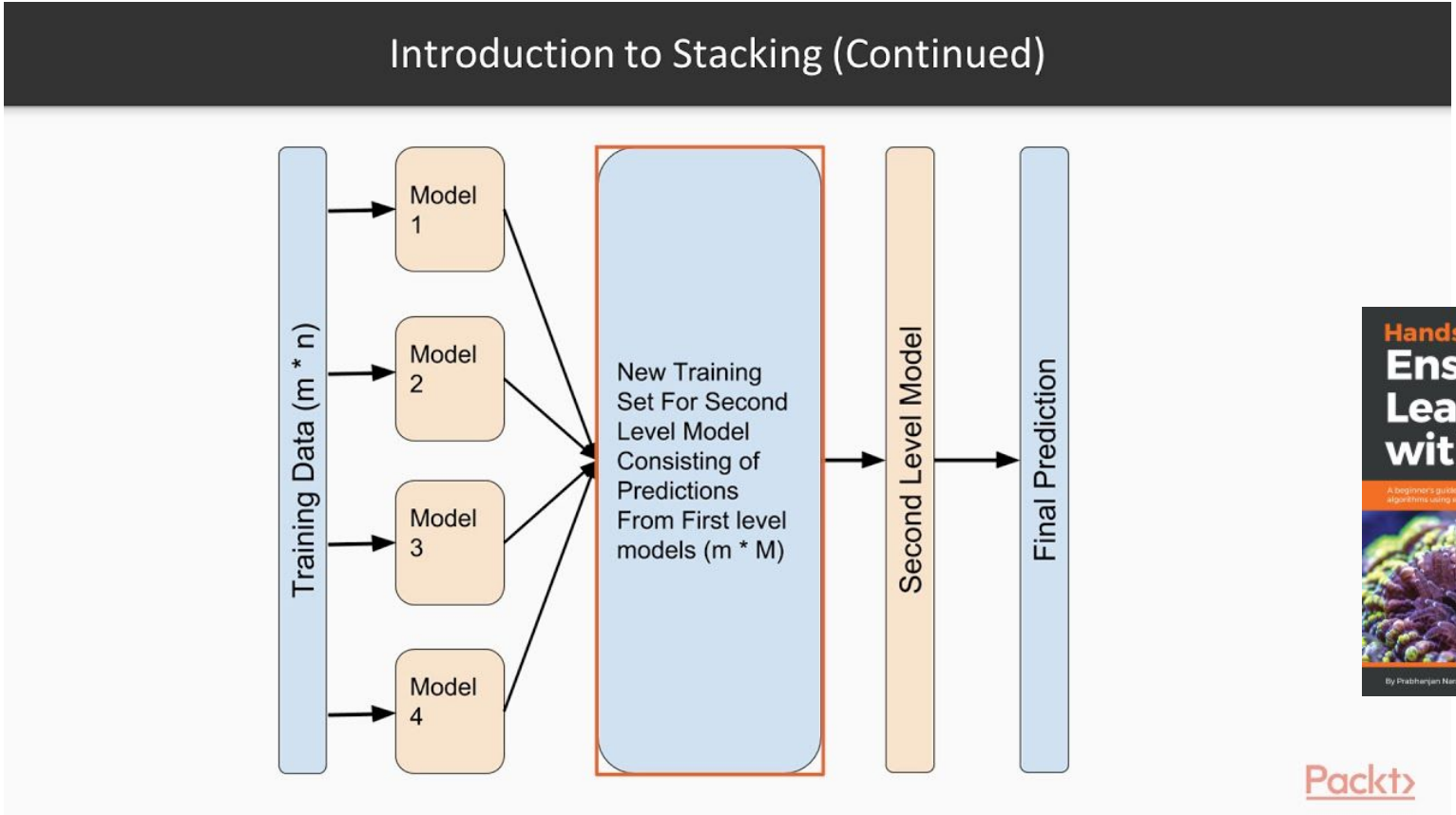
The stacking approach to Ensemble ML

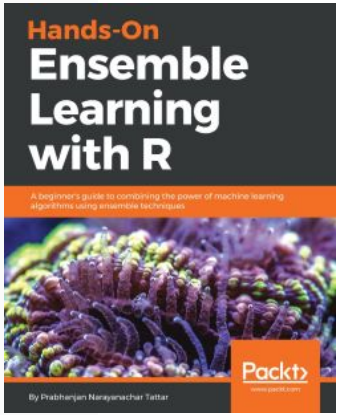
Chapter 6 Stacking

https://koalaverse.github.io/machine-learning-in-R/stacking.html

OpenDataScience

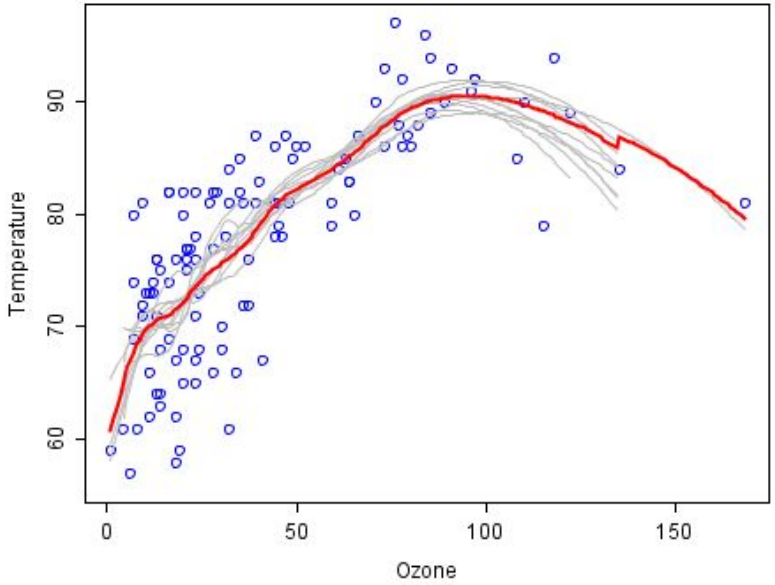
The stacking approach to Ensemble ML





Packt
OpenDataScience

SuperLearner



University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 266

Super Learner In Prediction

Eric C. Polley* Mark J. van der Laan†

*Division of Biostatistics, University of California, Berkeley, eric.polley@nih.gov
†University of California - Berkeley, laan@berkeley.edu

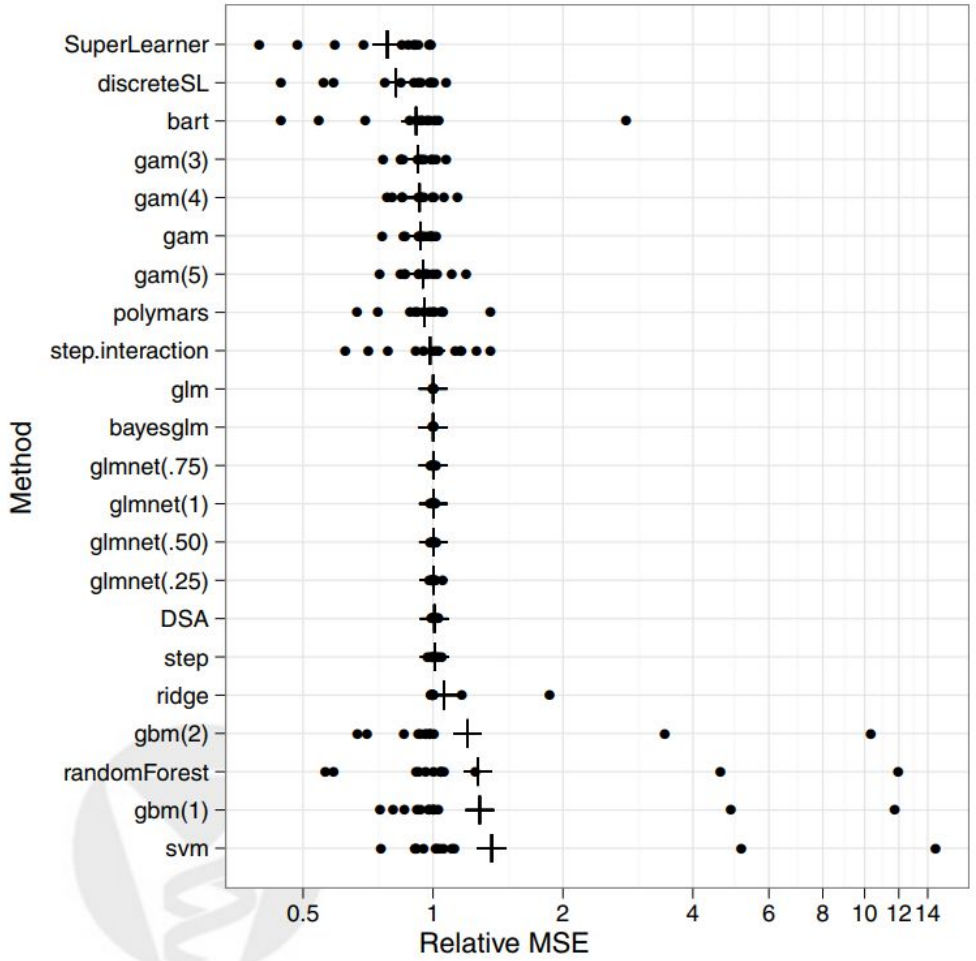
This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper266

Copyright ©2010 by the authors.

OpenDataScience

Figure 3: 10-fold cross-validated relative mean squared error compared to glm across 13 real datasets. Sorted by the geometric mean, denoted with the plus (+) sign.



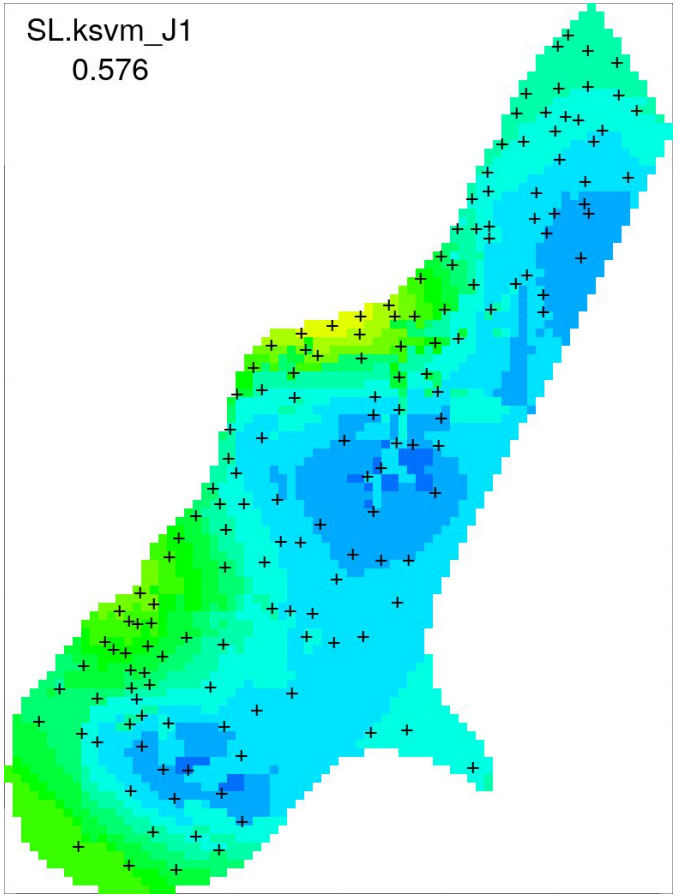
"Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance** (bagging), **bias** (boosting), or **improve predictions** (stacking)."
https://blog.statsbot.co/ensemble-learning-d1dcd548e936

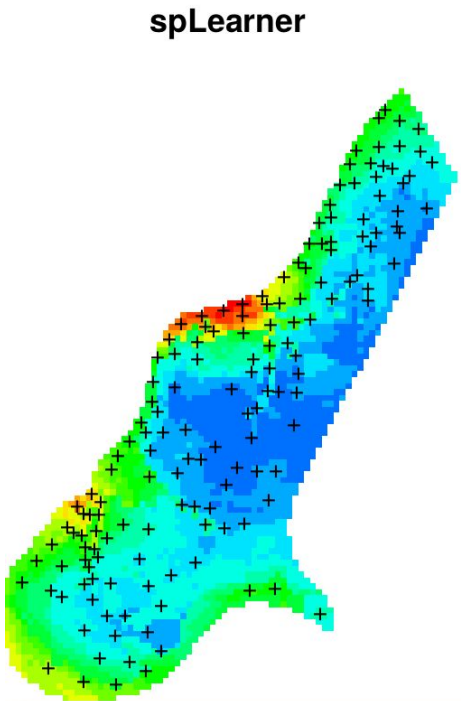
This however comes at costs:

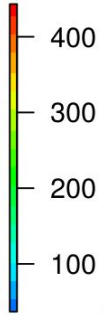
- higher computational load,
- higher RAM requirements,

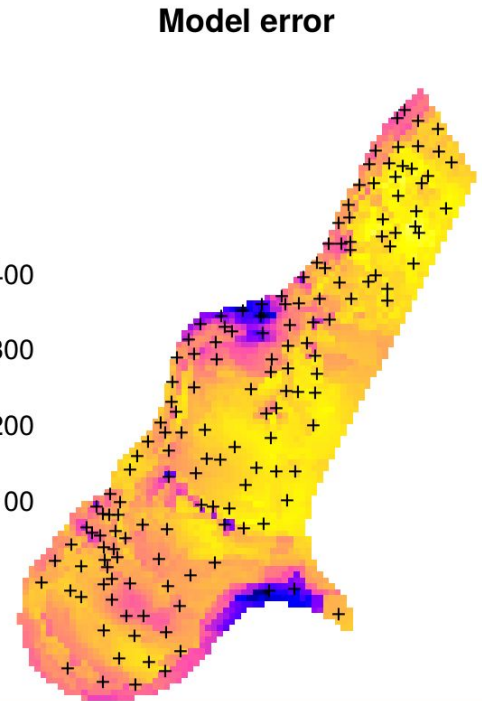
OpenDataScience

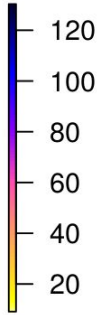
Model uncertainty











OpenDataScience

Machine Learning in R

build passing CRAN 2.14.0 - a month ago CRAN WARN downloads 15K/month stackoverflow mlr lifecycle stable dependencies 9/38

- CRAN release site
- Online tutorial
- Cheatsheet
- Changelog

We are actively working on mlr3 as a successor of *mlr*. This implies that we have less time to reply to *mlr* issues.

- Stackoverflow: mlr
- Slack
- Blog.

Installation

Release

```
install.packages("mlr")
```

Development

```
remotes::install_github("mlr-org/mlr")
```

Links

Download from CRAN at https://cloud.r-project.org/package=mlr

Browse source code at https://github.com/mlr-org/mlr

Report a bug at https://github.com/mlr-org/mlr/issues

Cheatsheet at https://github.com/mlr-org/mlr/blob/master/addon/cheatsheet/MlrCheatsheet.pdf

License

BSD_2_clause + file LICENSE

Citation

Citing mlr

Developers

Bernd Bischl
Author

Michel Lang
Author

Regression (59)

Additional learner properties:

- *se*: Standard errors can be predicted.

Class / Short Name / Name	Packages	Num.	Fac.	Ord.	NAs	Weights	Props	Note
regr.bartMachine *bartmachine* Bayesian Additive Regression Trees	bartMachine	X	X		X			`use_missing_data` has been set to `TRUE` by default to allow missing data support.
regr.bcart *bcart* Bayesian CART	tgp	X	X				se	
regr.bgp *bgp* Bayesian Gaussian Process	tgp	X					se	
regr.bgpllm *bgpllm* Bayesian Gaussian Process with jumps to the Limiting Linear Model	tgp	X					se	
regr.blm *blm* Bayesian Linear Model	tgp	X					se	

Contents

Classification (82)
Regression (59)
Survival analysis (12)
Cluster analysis (10)
Cost-sensitive classification
Multilabel classification (3)

makeStackedLearner

From mlr v2.13
by Bernd Bischl

99.99th Percentile

Create A Stacked Learner Object.

A stacked learner uses predictions of several base learners and fits a super learner using these predictions as features in order to predict the outcome. The following stacking methods are available:

`average` Averaging of base learner predictions without weights.

`stack.nocv` Fits the super learner, where in-sample predictions of the base learners are used.

`stack.cv` Fits the super learner, where the base learner predictions are computed by crossvalidated predictions (the resampling strategy can be set via the `resampling` argument).

`hill.climb` Select a subset of base learner predictions by hill climbing algorithm.

`compress` Train a neural network to compress the model from a collection of base learners.

Usage

```
makeStackedLearner(base.learners, super.learner = NULL,
  predict.type = NULL, method = "stack.nocv", use.feat = FALSE,
  resampling = NULL, parset = list())
```

Arguments

Learn R at work
Try it free

Guide to SuperLearner

Chris Kennedy, University of California, Berkeley

March 16, 2017

- 1 Background
- 2 Software requirements and installation
- 3 Setup dataset
- 4 Review available models
- 5 Fit individual models
- 6 Fit multiple models
- 7 Predict on new data
- 8 Fit ensemble with external cross-validation
- 9 Customize a model hyperparameter
- 10 Test algorithm with multiple hyperparameter settings
- 11 Multicore parallelization

README.md

landmap package for R

Package provides methodology for automated mapping i.e. spatial interpolation and/or prediction using Ensemble Machine Learning (extends functionality of the subsemble and the SuperLearner packages). Key functionality includes:

- `train.spLearner` --- train a spatial prediction and/or interpolation model using Ensemble Machine Learning (works with numeric, binomial and factor-type variables),
- `buffer.dist` --- derive buffer (geographical) distances that can be used as covariates in spLearner,
- `spc` --- derive Principal Components using stack of spatial layers,
- `tile` --- tile spatial layers so they can be used to run processing in parallel,
- `download.landgis` --- access and download LandGIS layers from www.openlandmap.org,

Warning: most of functions are optimized to run in parallel by default. This might result in high RAM and CPU usage.

Spatial prediction using Ensemble Machine Learning with geographical distances is explained in detail in:

- Hengl, T., MacMillan, R.A., (2019). Predictive Soil Mapping with R. OpenGeoHub foundation, Wageningen, the Netherlands, 370 pages, www.soilmapper.org, ISBN: 978-0-359-30635-0.
- Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B. (2018). Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518.

Installing

Install development versions from github:

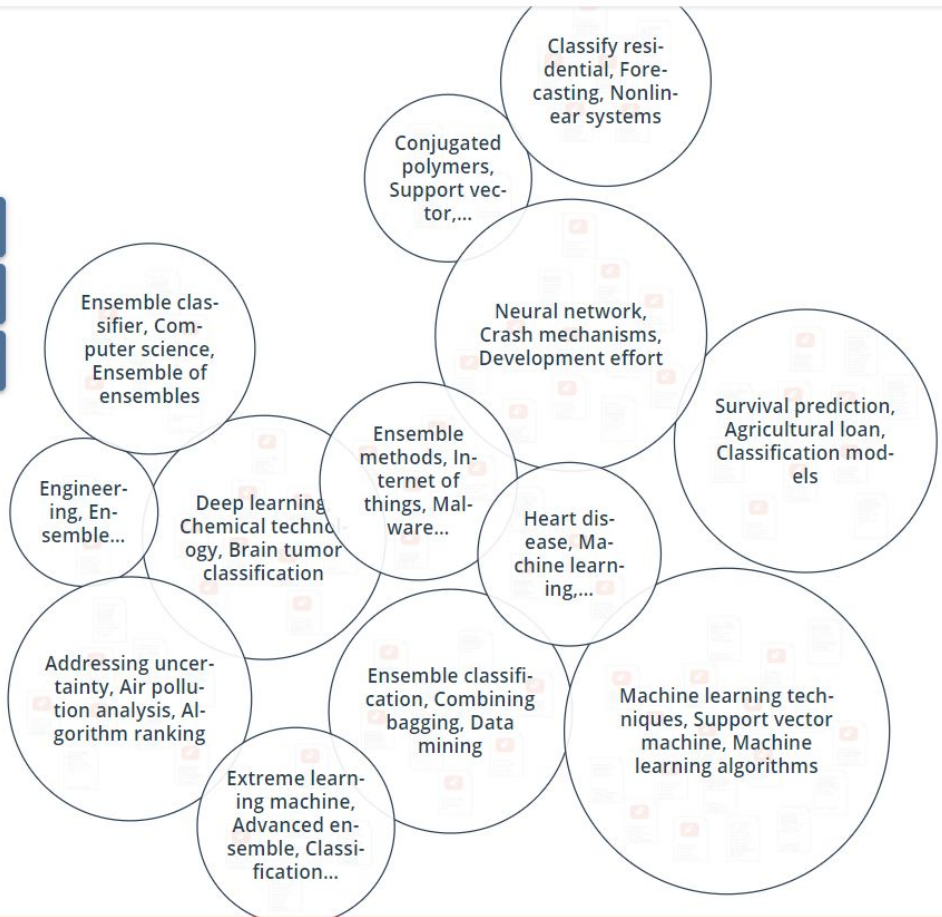
```
library(devtools)
install_github("envirometrix/landmap")
```

Under construction. Use for testing purposes only.

Functionality

Why Ensemble ML?

- We can potentially increase accuracy (usually not a lot but sometimes even few percent counts);
- Prediction system is more robust: there is at the order of magnitude more training involved;
- We can derive model-free prediction errors;
- By doing Ensemble ML, we can combine fine-tuning, model selection, feature selection etc all at once;



A Novel Ensemble Machine Learning for Robust Microarray Data Classification.

LINK

Yonghong Peng **(2006)**
[link]: http://hdl.handle.net/10454/3688

No ; Microarray data analysis and classification has demonstrated convincingly that it provides an effective methodology for the effective diagnosis of diseases and cancers. Although much research has been performed on applying machine learning techn...

Area: Ensemble classification, Combining bagging, Data mining

open access

MRI-Based Brain Tumor Classification Using Ensemble of Deep Features and Machine Learning Classifiers

PDF

Jaeyong Kang, Zahid Ullah, Jeonghwan Gwak in *Sensors, Vol 21, Iss 2222, p 2222 (2021)* **(2021-03-01T00:00:00Z)**
[doi]: https://doi.org/10.3390/s21062222

Brain tumor classification plays an important role in clinical diagnosis and effective treatment. In this work, we propose a method for brain tumor classification using an ensemble of deep features and machine learning classifiers. In our proposed fr...

Area: Deep learning, Chemical technology, Brain tumor classificat...

open access

Dropout prediction in e-learning courses through the combination of machine learning techniques

PDF

Spatiotemporal interpolation of daily temperatures using EML

Case study #1

Theor Appl Climatol (2012) 107:265–277
DOI 10.1007/s00704-011-0464-2

ORIGINAL PAPER

Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images

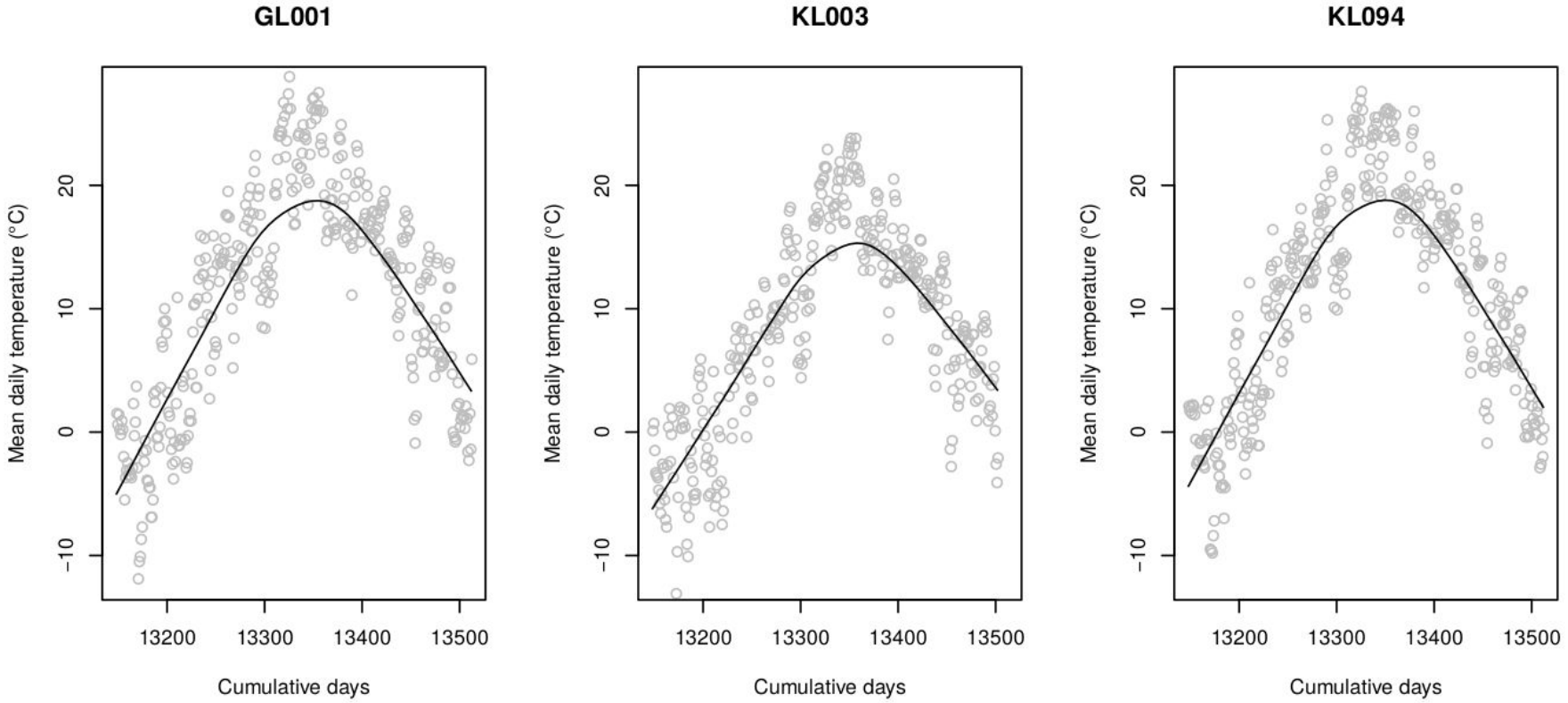
Tomislav Hengl · Gerard B. M. Heuvelink · Melita Perčec Tadić · Edzer J. Pebesma

Received: 28 October 2010 / Accepted: 25 May 2011 / Published online: 8 July 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract A computational framework to generate daily temperature maps using time-series of publicly available MODIS MOD11A2 product Land Surface Temperature (LST) images (1 km resolution; 8-day model can explain a significant part of the variation in station-data (84%). MODIS LST 8-day (cloud-free) images are unbiased estimator of the daily temperature, but with relatively low precision (±4.1°C); however

OpenDataScience

Meteo station data



OpenDataScience

Spatiotemporal prediction of soil moisture (VW) in 3D+T

Cookfarm dataset

Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The Cook Agronomy Farm data set

Caley K. Gasch [a,*], Tomislav Hengl [b], Benedikt Gräler [c], Hanna Meyer [d], Troy S. Magney [e], David J. Brown [a]

[a] Department of Crop and Soil Sciences, Washington State University, USA
[b] ISRIC – World Soil Information/Wageningen University and Research, The Netherlands
[c] Institute of Geoinformatics, University of Münster, Germany
[d] Department of Geography/Environmental Informatics, Philipps-Universität Marburg, Germany
[e] College of Natural Resources, University of Idaho, USA

ARTICLE INFO

Article history:
Received 1 November 2014
Accepted 1 April 2015
Available online xxxx

Keywords:
Digital soil mapping
Random forests algorithm
Regression-kriging
Soil sensor network

ABSTRACT

The paper describes a framework for modeling dynamic soil properties in 3-dimensions and time (3D + T) using soil data collected with automated sensor networks as a case study. Two approaches to geostatistical modeling and spatio-temporal predictions are described: (1) 3D + T predictive modeling using random forests algorithms, and (2) 3D + T kriging model after detrending the observations for depth-dependent seasonal effects. All the analyses used data from the Cook Agronomy Farm (37 ha), which includes hourly measurements of soil volumetric water content, temperature, and bulk electrical conductivity at 42 stations and five depths (0.3, 0.6, 0.9, 1.2, and 1.5 m), collected over five years. This data set also includes 2- and 3-dimensional, temporal, and spatio-temporal covariates covering the same area. The results of (strict) leave-one-station-out cross-validation indicate that both models accurately predicted soil temperature, while predictive power was lower for

OpenDataScience

Case study: Spatiotemporal distribution of Fagus sylvatica

Species Distribution Modeling (SDM) - basics

Correlative SDMs

1. Survey location of species occurrence
2. Associate values of predictor variables to locations
3. Fit model to estimate similarity between sites
4. Predict in the Region of Interest (ROI) in spacetime

Species Distribution Modeling (SDM) - basics

$$Probability\ of\ occurrence\ =\ f(Climate,\ Terrain,\ Reflectance,\ Competition)$$

Climate = Temperature, precipitation, snow cover, cloud fraction, water vapor, wind speed ecc.

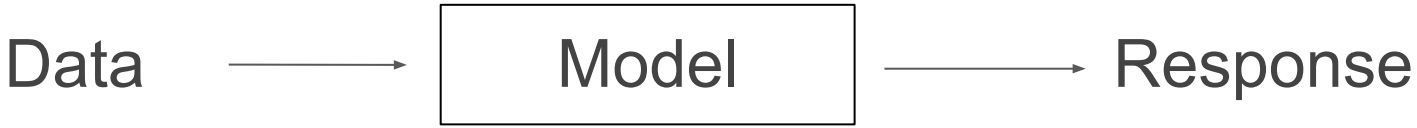
Terrain = DTM, CHM, lithology, slope, aspect, and other DTM-derivatives

Spectral reflectance = Landsat data, FAPAR

Competition = European Atlas of Forest Tree Species

Species Distribution Modeling (SDM) - basics

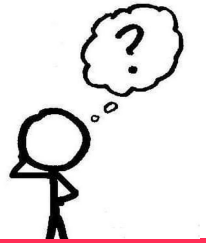
Modeling is easy!



Species Distribution Modeling (SDM) - basics

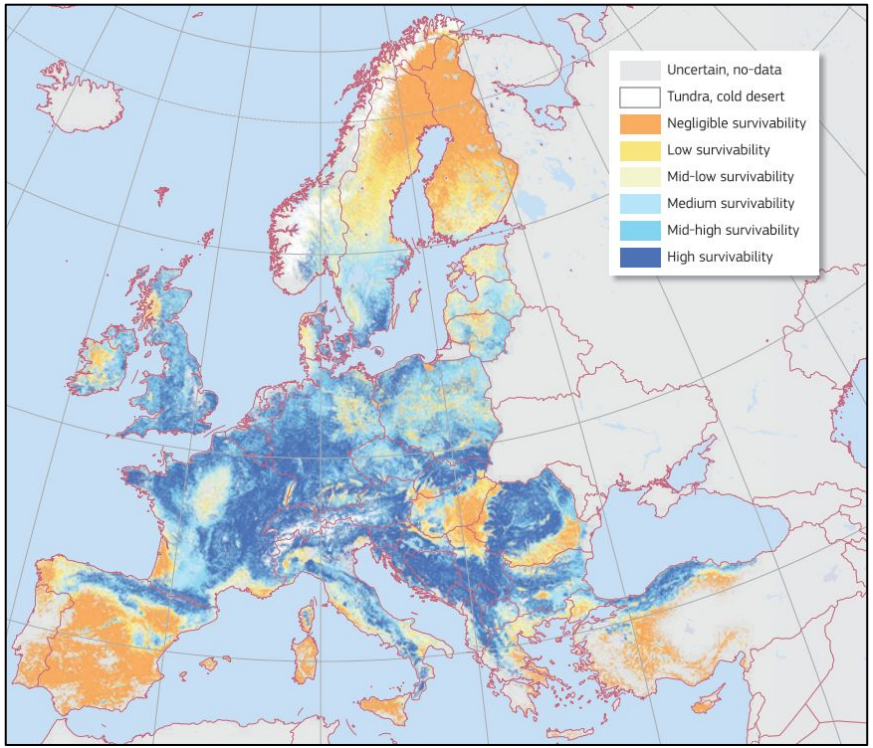
...or is it?

- Understand the data
  (presence only vs presence/absence vs presence/pseudo-absence)
- Clean the data
  (missing values, rescaling, skewed distribution etc.)
- Understand the problem
  (suitability, potential niche, realized niche etc.)
- Choose the proper tools
  (which algorithms, which features, which parameters)



Species Distribution Modeling (SDM) - basics

Maximum Habitat Suitability
(*Fagus sylvatica*)



Chorological map
(*Fagus sylvatica*)

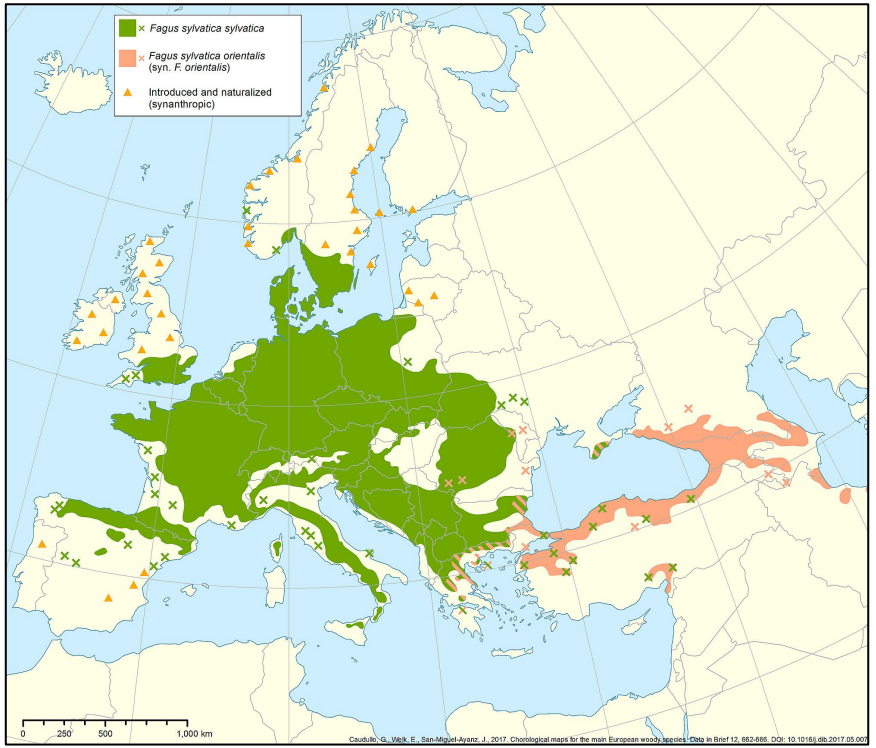


Image credits: [San-Miguel-Ayanz, J. et al. (2016) - European Atlas of Forest Tree Species]

OpenDataScience

Previous EU distribution maps

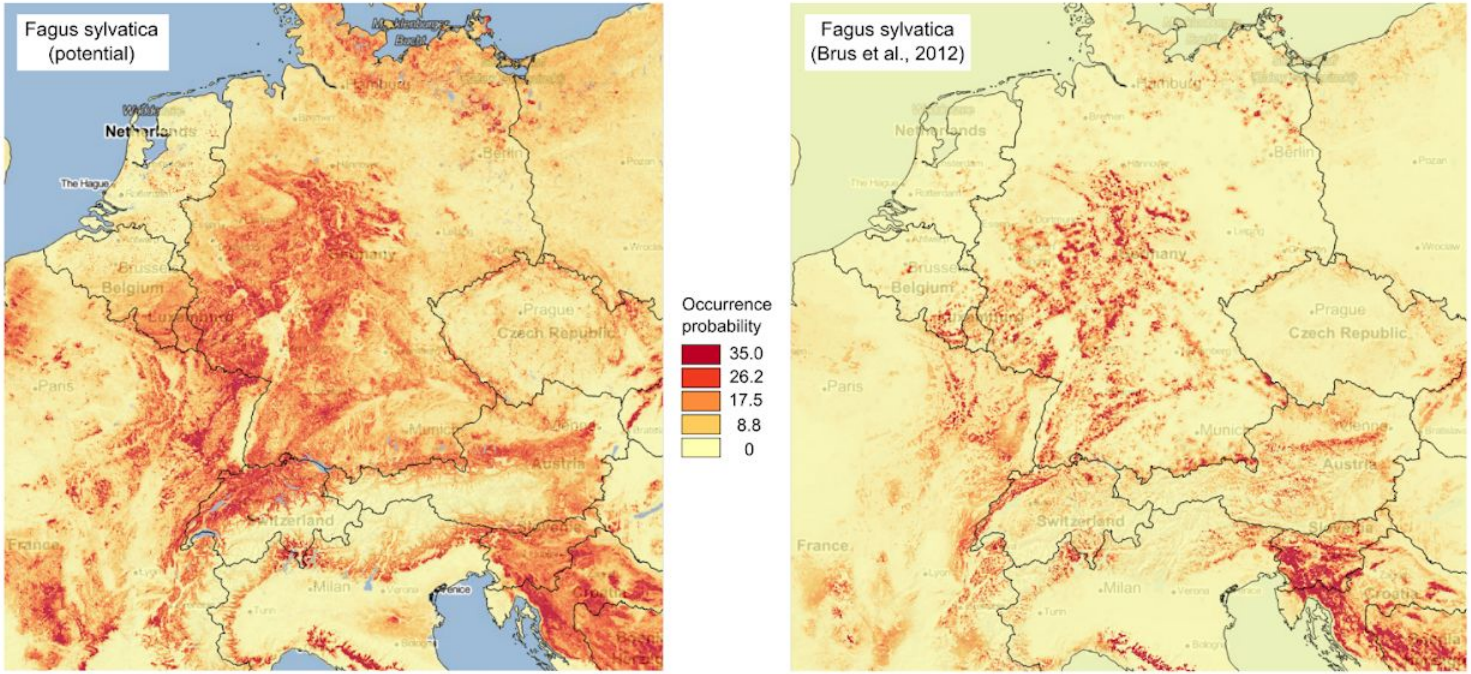


Figure 10. Comparison between predicted PNV distribution for *Fagus sylvatica* based on our results, and based on the maps generated by Brus et al. (2012) i.e. showing assumed actual distribution of the tree species.

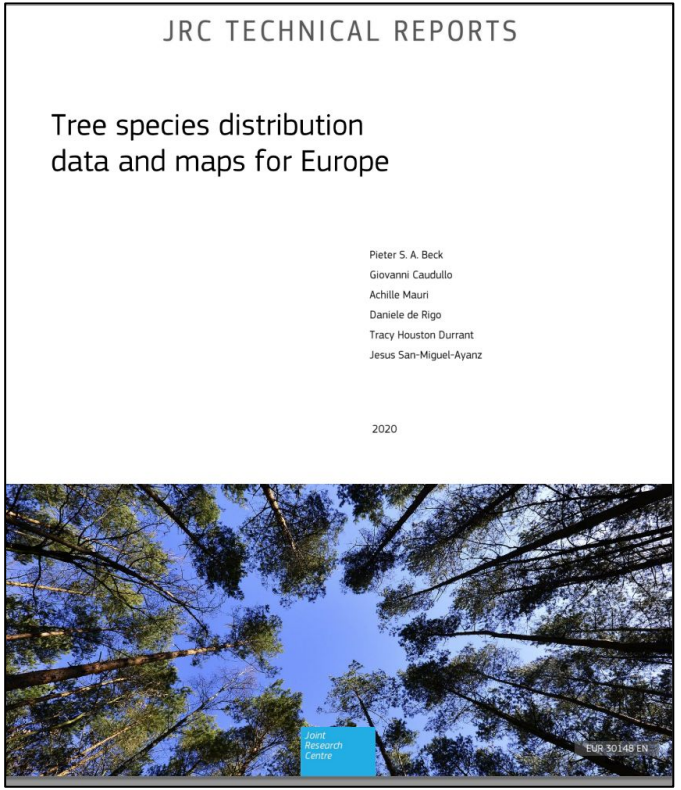
Image credits: Hengl et al., (2018)

OpenDataScience

Scientific datasets - state of the art

Datasets mentioned:

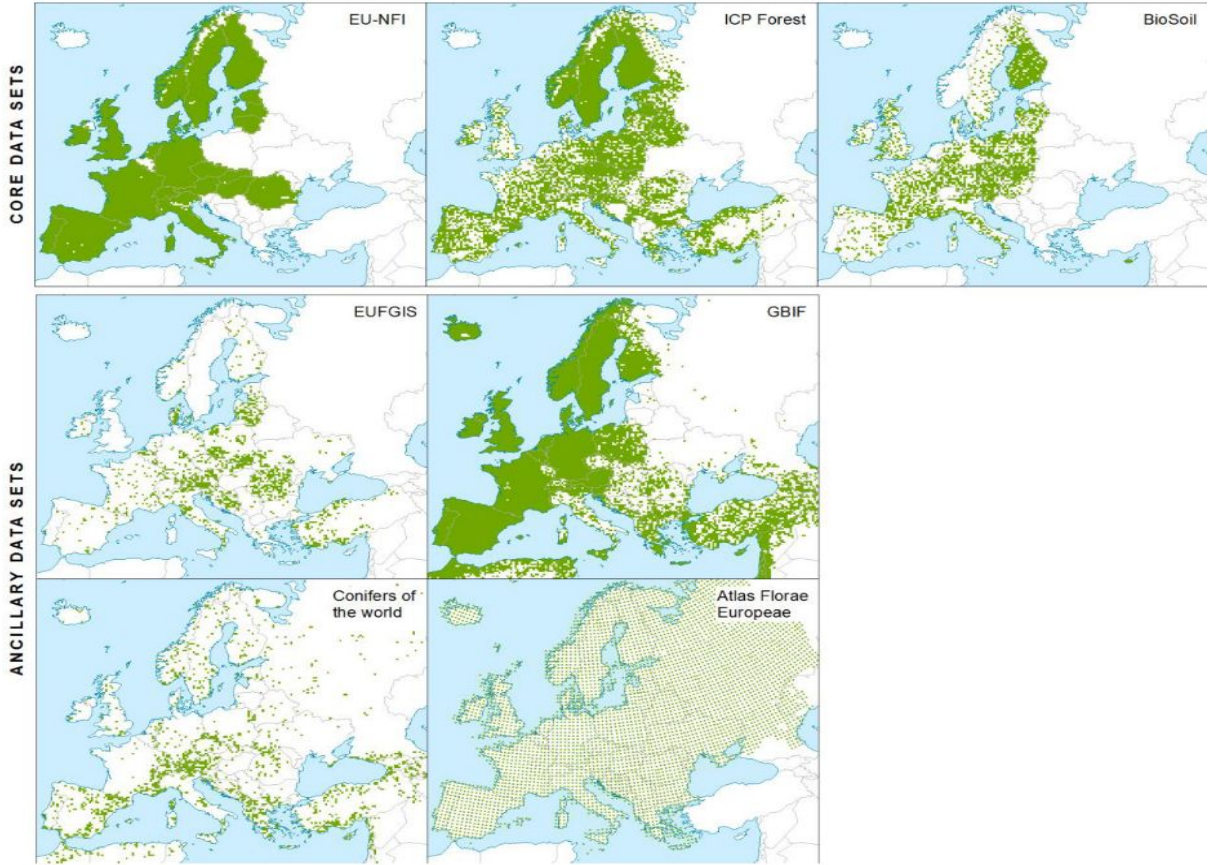
- Core datasets
  - EU-NFI data
  - ICP-Forest database
  - BioSoil
- Ancillary datasets
  - EUFGIS
  - GBIF
  - Conifers of the World
  - Atlas Florae Europaeae



Beck et al., (2020)

OpenDataScience

Scientific datasets - state of the art



- Freely available
- High density of points
- Constantly updated
- Easy to infer absence/pseudo-absence
- Wide range of species

OpenDataScience

Harmonized tree species occurrences for EU

OpenLandMap › EU forest tree point data

EU forest tree point data
Project ID: 20617763

☆ Star 0

20 Commits | 1 Branch | 0 Tags | 5.1 MB Files | 5.1 MB Storage

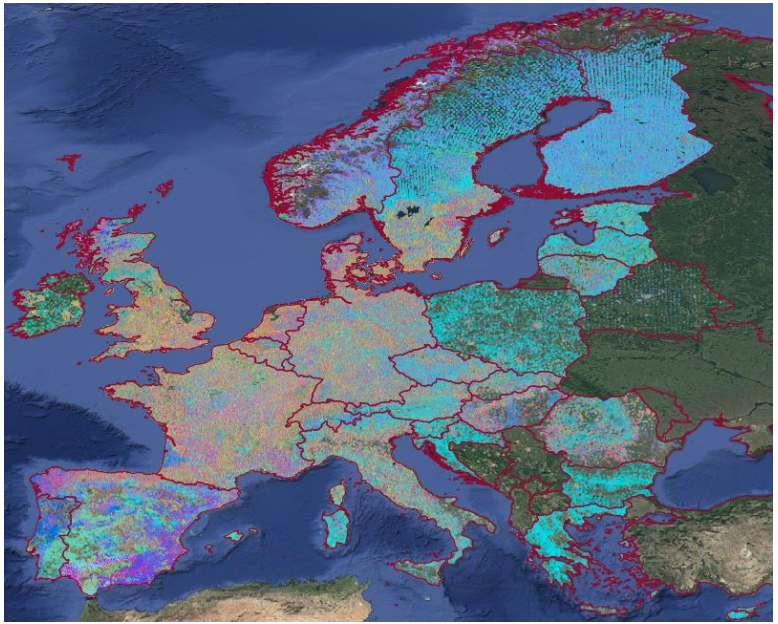
A compilation of analysis-ready point data for the purpose of vegetation and Potential Natural Vegetation mapping (EU coverage only)

master | eu-forest-tree-point-data | History | Find file | Clone

Update README.md
Johannes Heisig authored 4 months ago
ef31c893

README | No license. All rights reserved

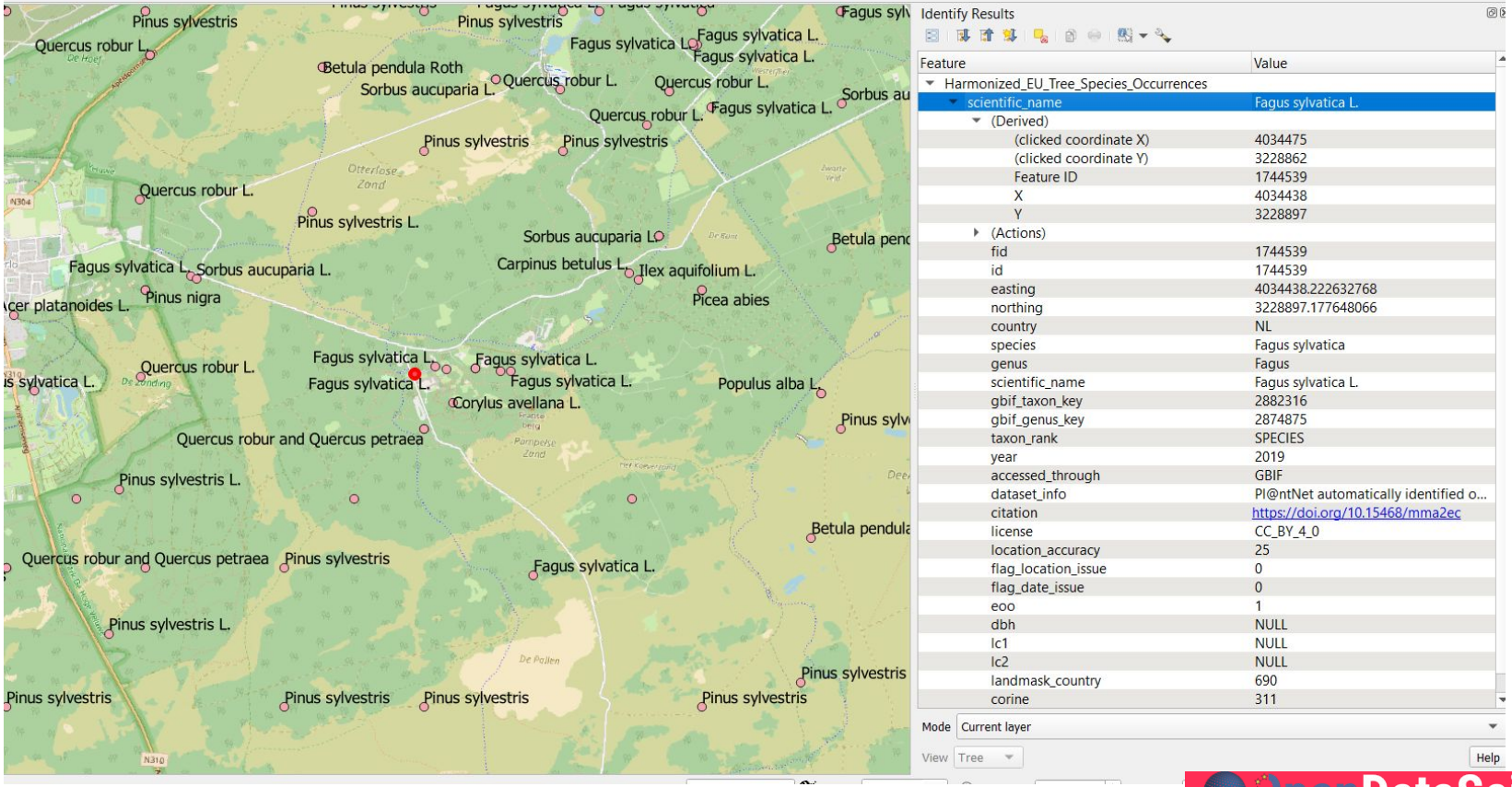
Name	Last commit	Last update
001_preview_treespeciespoints....	Upload New File	4 months ago
01_download.Rmd	upload for presentation	5 months ago
01_download.html	upload for presentation	5 months ago
02_assemble.Rmd	Replace 02_assemble.Rmd	4 months ago
02_assemble.html	Replace 02_assemble.html	4 months ago
03_explore.Rmd	Upload New File	4 months ago
03_explore.html	Upload New File	4 months ago
README.md	Update README.md	4 months ago



Datasets available on Zenodo: Heisig & Hengl (2020)
Code available on Gitlab: EU forest tree point data

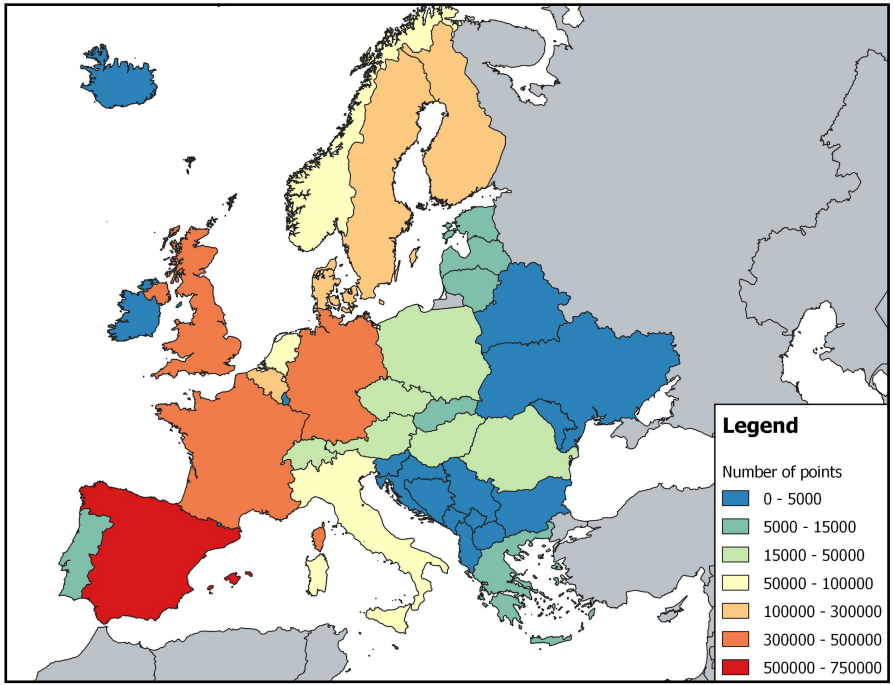
OpenDataScience

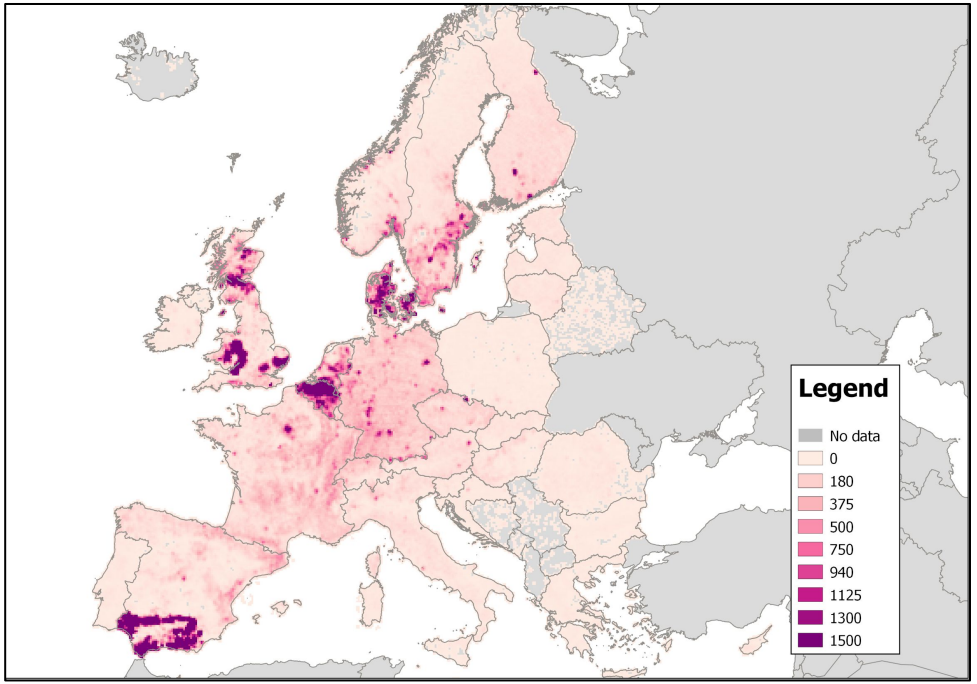
Harmonized tree species occurrences for EU



OpenDataScience

Dataset structure



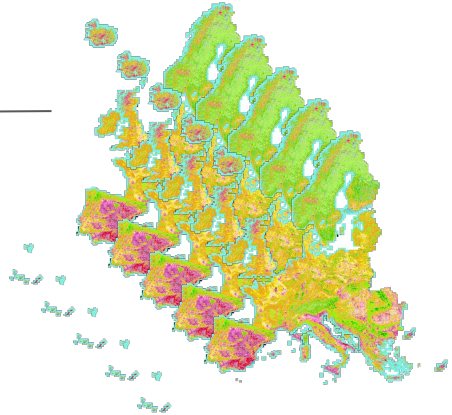


GBIF JRC EUROPEAN COMMISSION Land Monitoring Service

OpenDataScience

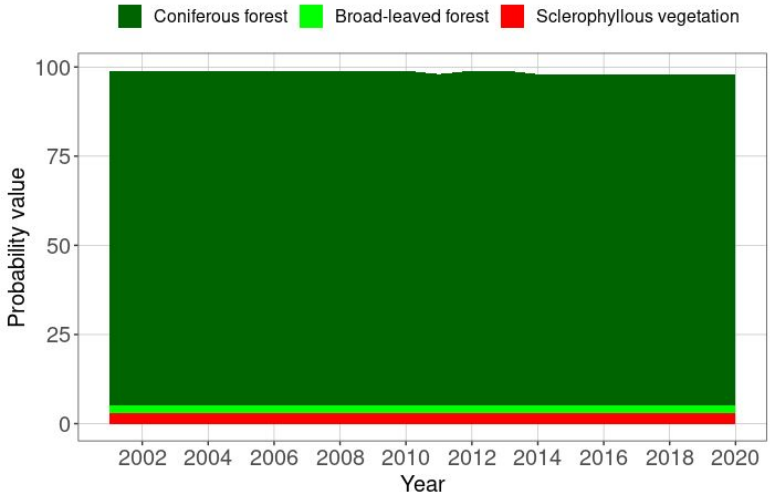
Data cleaning

Tree species occurrence points



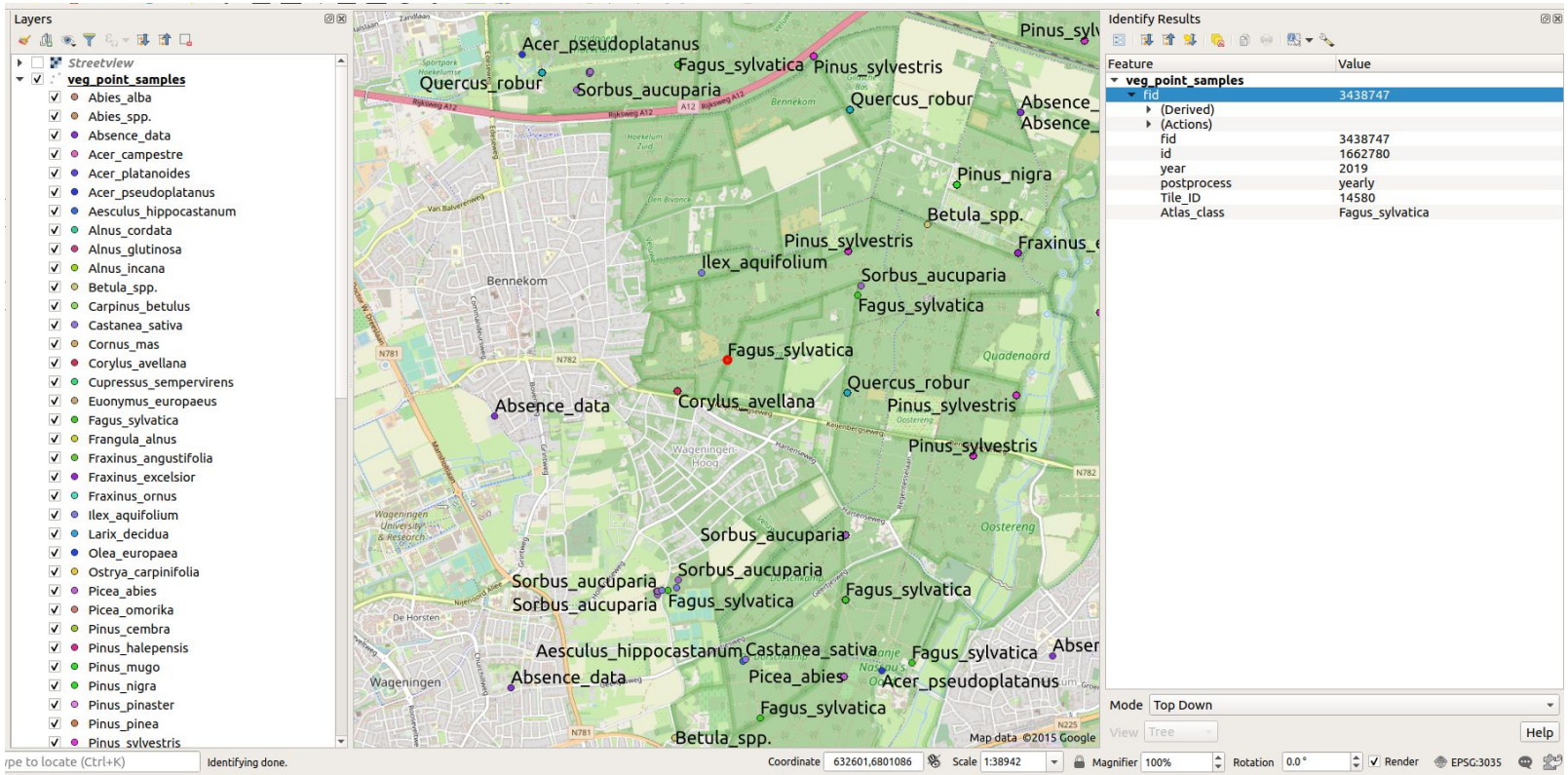
Spatio-temporal overlay

Land cover classification (2000 - 2019)
(probabilities)



OpenDataScience

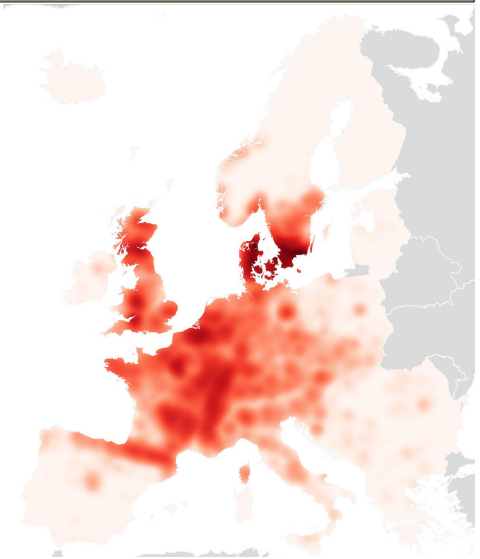
Dataset structure



OpenDataScience

Covariates in use

662 Covariates

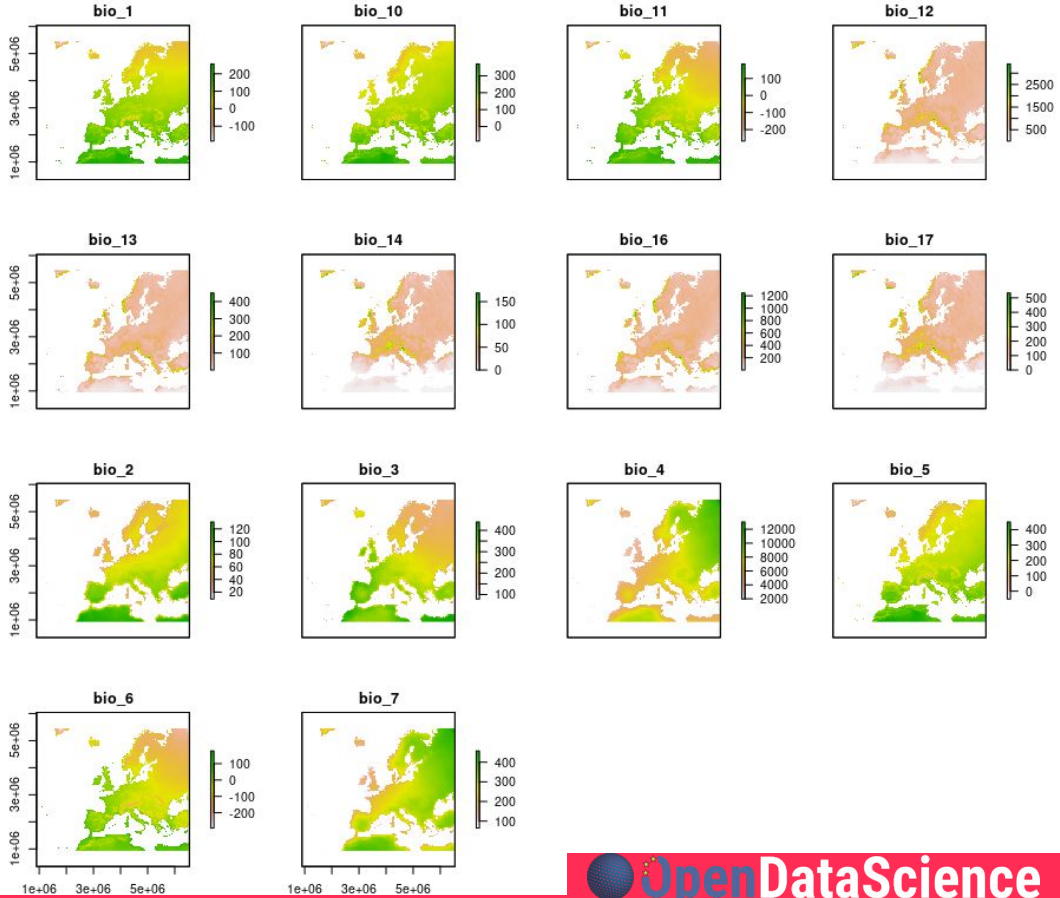
293 Climate	190 Terrain	140 Reflectance	49 Other tree species occurs.
Monthly precipitation (sum, std, avg 5 yrs)	**Slope and elevation**	**Seasonal Landsat bands** (Blue, Green, Red, etc.)	
Monthly temperature (mean, std, avg 5 yrs)	**Lithology and landform**	**Spectral indices** (NDVI, NDWI, NBR, etc.)	
Water vapor (long-term monthly)	**Roughness**	**FAPAR from PROBA-V**	
Wind speed (long-term monthly)	**Flood map**		
Solar irradiance (2016)	**Water occurs.** (long-term)		
Others...	Others...		

OpenDataScience

Covariates in use - Bioclim dataset

Bioclim dataset from Worldclim (Hijmans et al., 2005)

- Bioclimatic variables derived from the monthly temperature and rainfall value
- Used as long-term trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors
- Computed at 1 km resolution

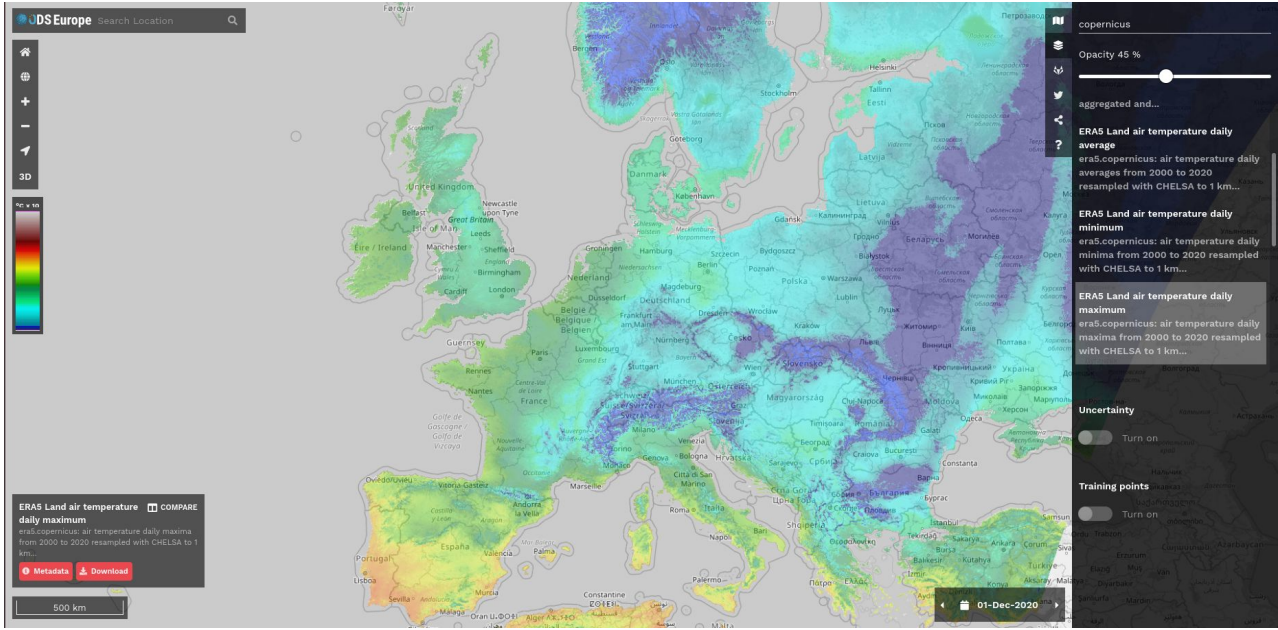


OpenDataScience

Covariates in use - Copernicus ERA5

ERA5 is the latest climate reanalysis produced by ECMWF, providing hourly data on many atmospheric, land-surface and sea-state parameters together with estimates of uncertainty.

- 3 Variables in use:
  - Air temperature (2m above the surface)
  - Precipitation
  - Surface temperature
- Daily values aggregated by month
- Monthly average computed on previous 5 years
- 1 km resolution

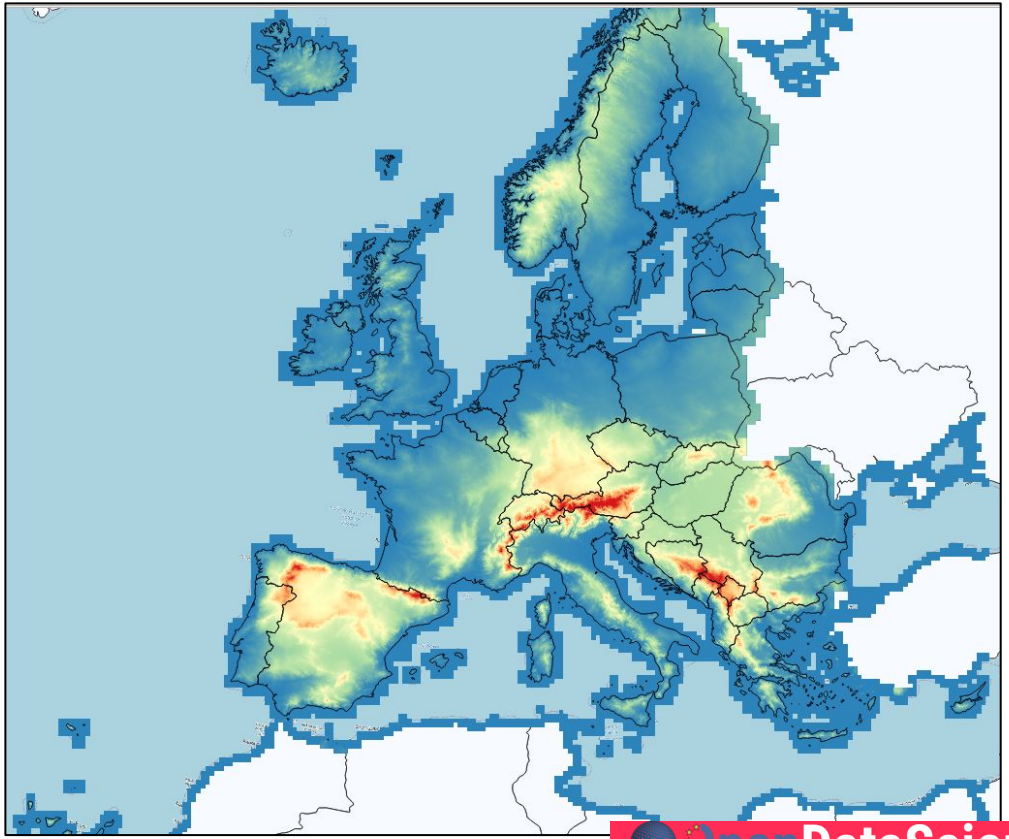


OpenDataScience

Covariates in use

Slope-based cost distance from coastline

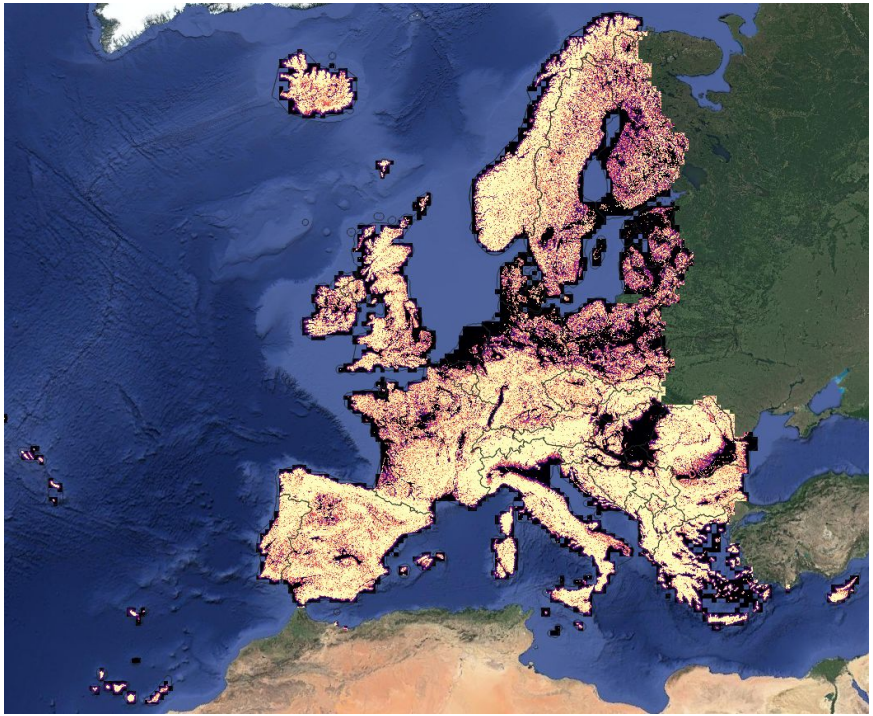
- Used as proxy for continentality
- Computed through GRASS at 30m resolution
- Cost function takes minimum cumulative slope from the coast as input



OpenDataScience

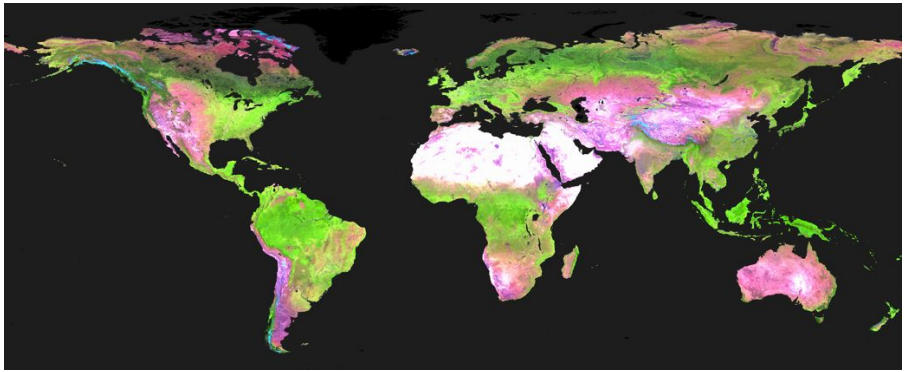
Covariates in use - DTM and DTM derivatives

- Geomorphometry derivatives based on the DTM for Continental Europe
- Available at 30 m, 100 m and 250 m resolution
- Variables available:
  - devmean = deviation from mean value
  - downlocal / down = downslope local and general curvature
  - hillshade = hillshading
  - mnr = Module Melton Ruggedness Number
  - northerness/easterness
  - openp / openn = openness positive negative,
  - slope = slope in percent
  - topidx = a topographic index (wetness index)
  - tpi = Topographic Wetness Index,
  - vbf = Multiresolution Index of Valley Bottom Flatness



Covariates in use - GLAD Landsat ARD

- Globally consistent analysis ready data (ARD) for multi-decadal LCLU monitoring
- 16-day time-series composites from Landsat 5, 7 and 8 (TM, ETM+ and OLI)
- Per-pixel observation quality flag
- MODIS (MOD44C) surface reflectance calibrated
- Product organized by 1 × 1 degree tiles
- Automatically download through HTTP API
- Product under Creative Commons Attribution License



UNIVERSITY OF MARYLAND

OpenDataScience

Covariates in use - GLAD Landsat ARD

Aggregate 16-days composites in 4 seasons:

1. Winter
2. Spring
3. Summer
4. Fall

Interval ID	Start	End	Composite
1	1-Jan	16-Jan	
2	17-Jan	1-Feb	
3	2-Feb	17-Feb	1
4	18-Feb	4-Mar	
5	5-Mar	20-Mar	
6	21-Mar	5-Apr	
7	6-Apr	21-Apr	
8	22-Apr	7-May	2
9	8-May	23-May	
10	24-May	8-Jun	
11	9-Jun	24-Jun	
12	25-Jun	10-Jul	
13	11-Jul	26-Jul	
14	27-Jul	11-Aug	
15	12-Aug	27-Aug	3
16	28-Aug	12-Sep	
17	13-Sep	28-Sep	
18	29-Sep	14-Oct	
19	15-Oct	30-Oct	
20	31-Oct	15-Nov	
21	16-Nov	1-Dec	4
22	2-Dec	17-Dec	
23	18-Dec	31-Dec	

→

Remove cloud and cloud shadow pixels
(According to quality assessment band)

Perc. 25 | **Median** | **InterQuantile Range**

Mosaicking, Reproject and Rescale
(For the whole EU)

→ Gap filling
(temporal moving window median)

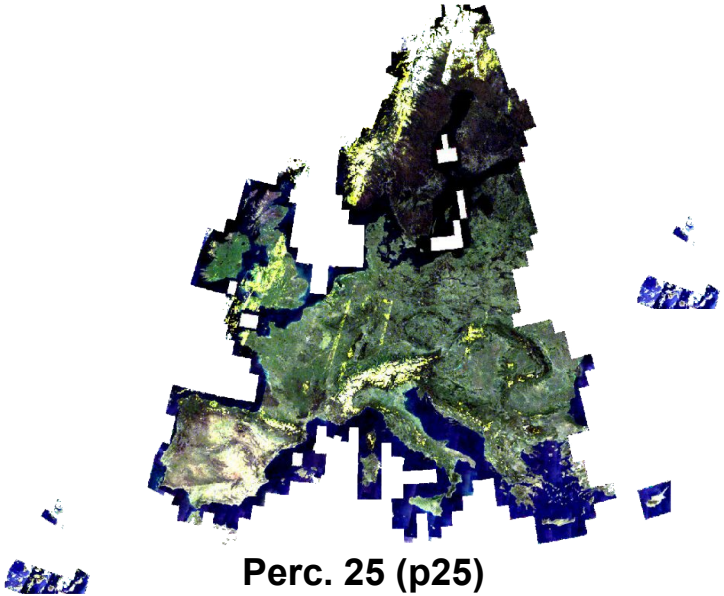
OpenDataScience

Covariates in use - GLAD Landsat ARD

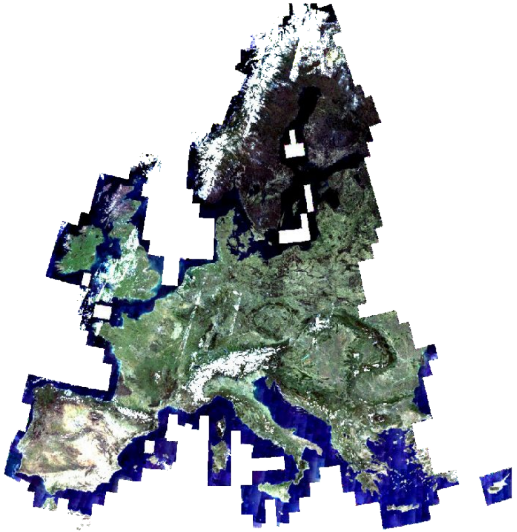
Red, Green, Blue

2018-03-21 to 2018-06-24 (spring)

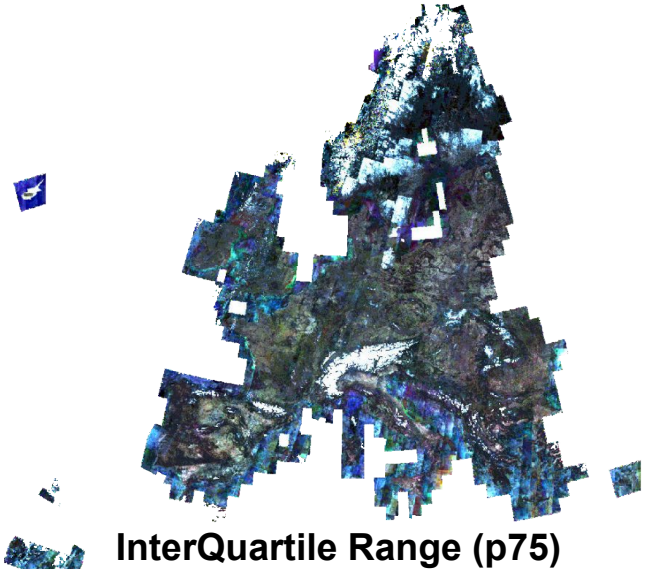
188,000 x 151,000 pixels



Perc. 25 (p25)



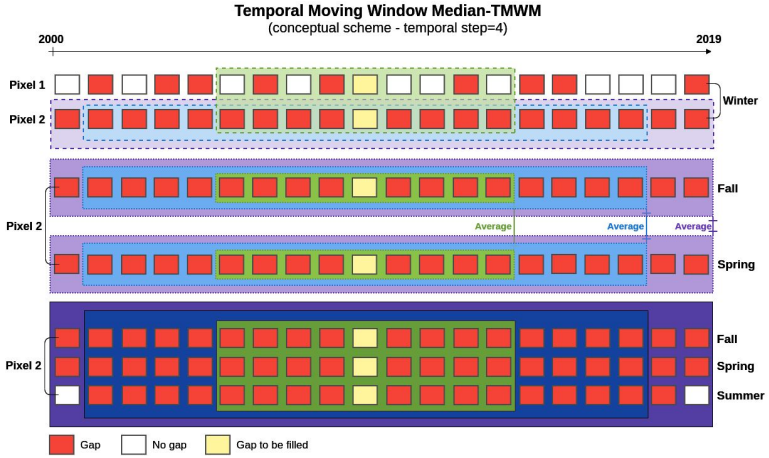
Median (p50)



InterQuartile Range (p75)

OpenDataScience

Covariates in use - GLAD Landsat ARD

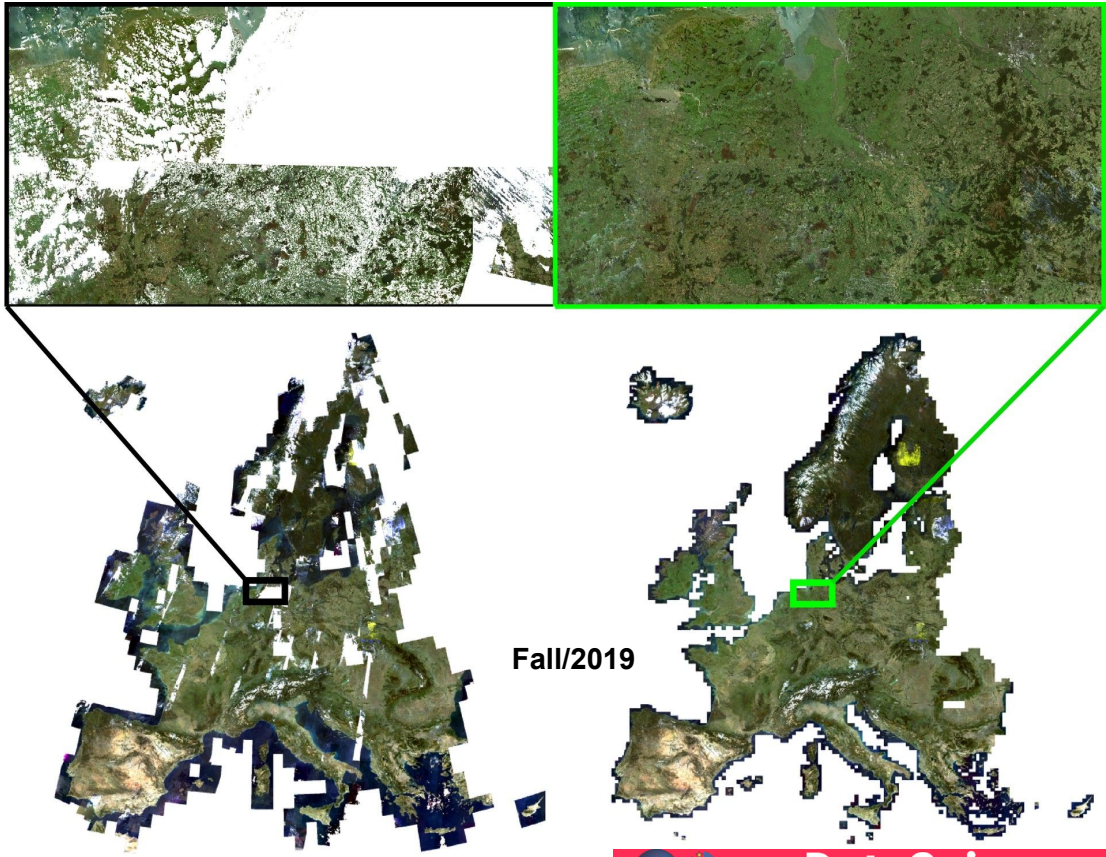


Total amount of yearly covariates: **91**

(7 bands X 3 quantiles X 4 seasons) +

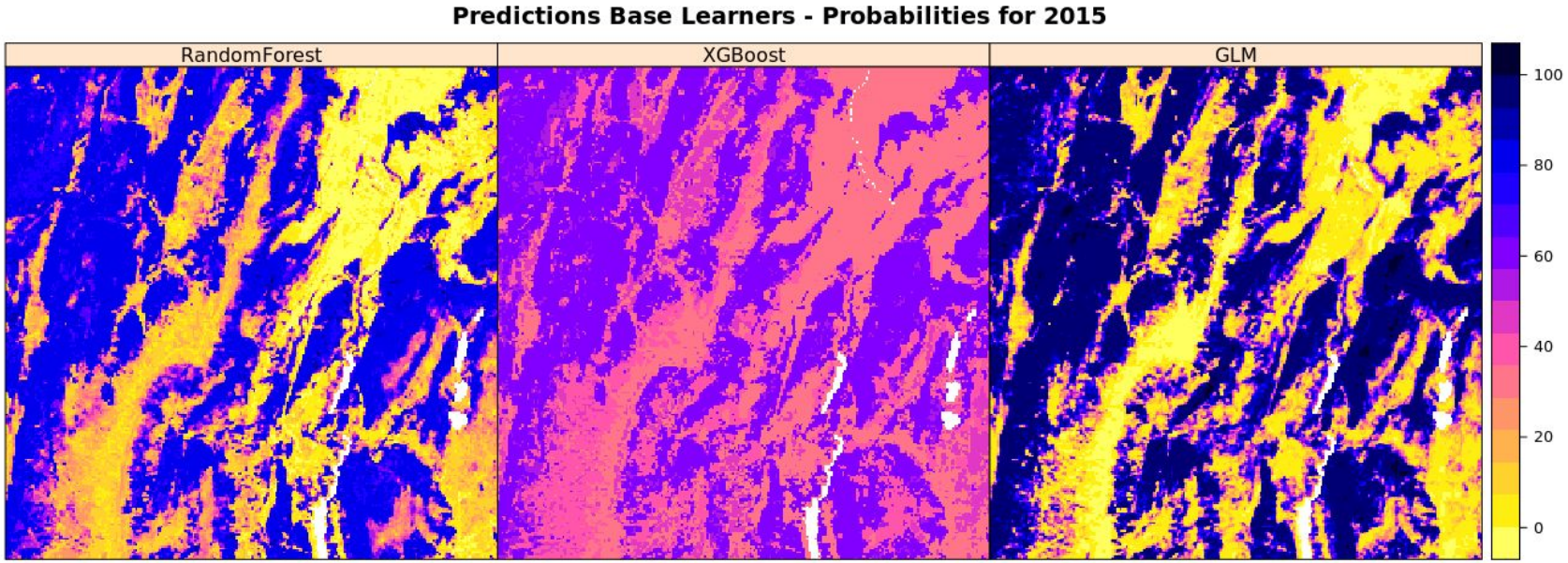
7 spectral indices computed on p50

Total amount of Landsat covariates: 1820



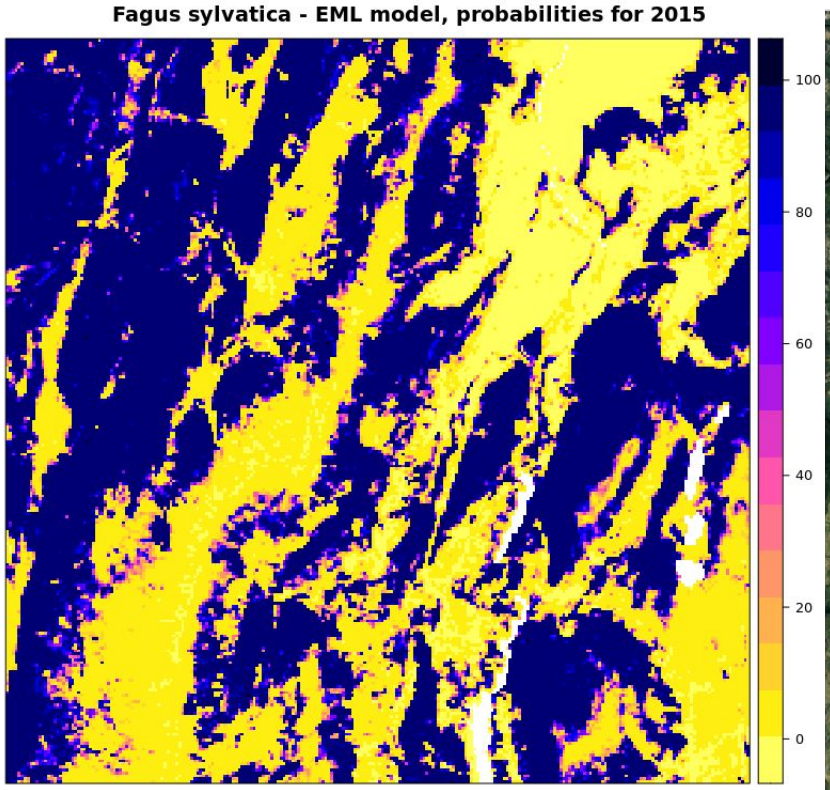
OpenDataScience

Combined predictions



OpenDataScience

Final predictions



OpenDataScience

Conclusions

- Ensemble Machine Learning is applicable to modeling spatiotemporal data for the purpose of predictive mapping;
- As an output of predictions one can produce a time-series of images / maps, then run trend / time-series analysis;
- Main advantages of using spatiotemporal Ensemble ML are:
  - With 1 model one can predict land cover, tree species, soil properties for the whole spacetime cube (for new years also);
  - If the model is unbiased, then the predictions are suited for analyzing trends;