Automated global soil mapping: discovering spatial soil patterns using machine learning

ISRIC World Soil Information

T. (Tom) Hengl <tom.hengl@isric.org>

ISRIC World Soil Information


SEERISK

Spatial Statistics

Spatial Statistics
Volume 14, Part A, Pages 1-114 (November 2015)
Spatial and Spatio-Temporal Models for Interpolating Climatic and Meteorological Data
Edited by Tomislav Hengl, Edzer Pebesma and Robert J. Hijmans

< Previous vol/iss | Next vol/iss >

Articles in Press

Open Access articles

Volumes 11 - 14 (2015)

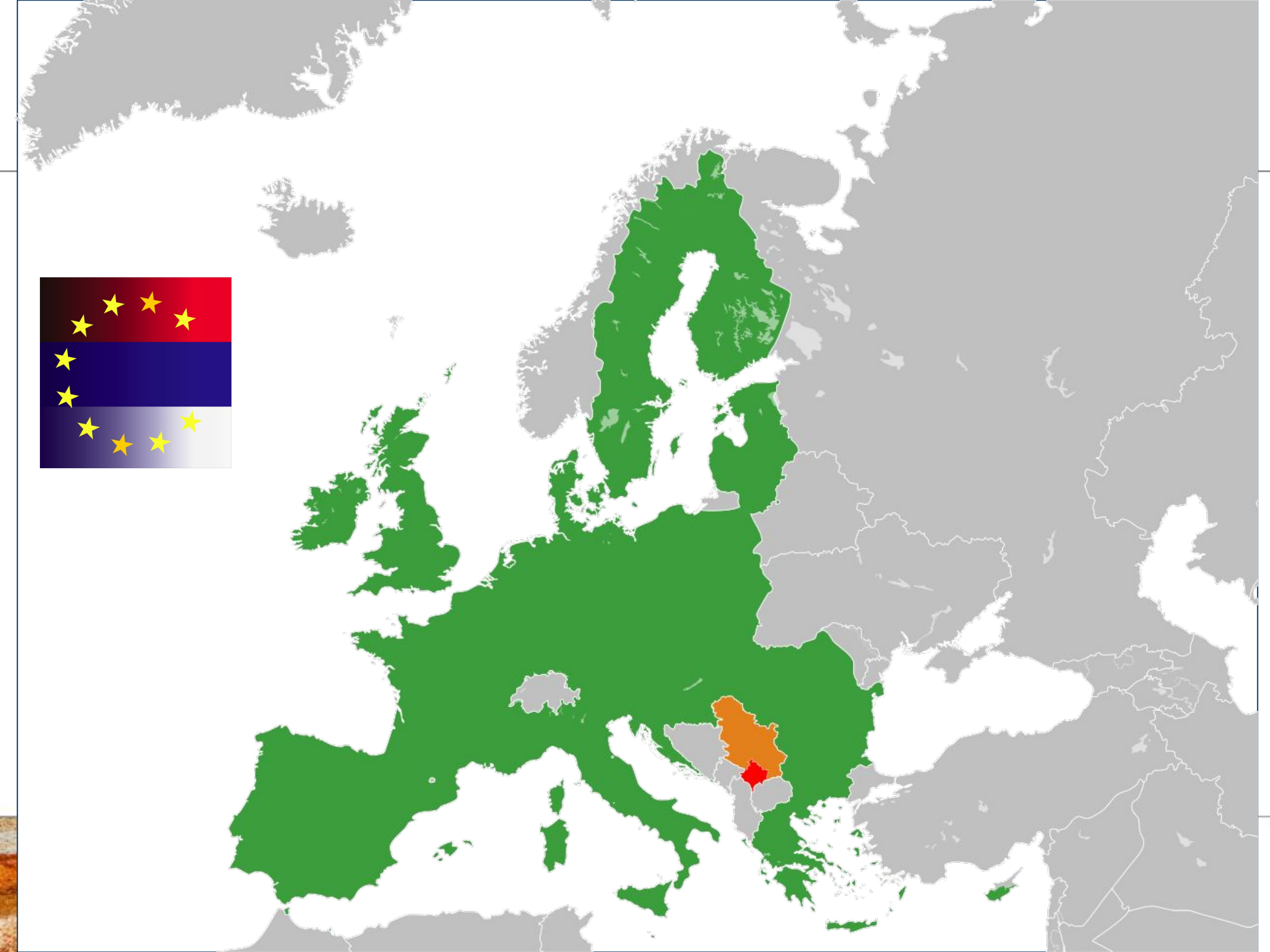
Volume 14, Part C
pp. 223-580 (November 2015)

Volume 14, Part B
pp. 115-222 (November 2015)

Download PDFs | Export

Editorial Board
Page IFC
PDF (45 K)

Spatial and spatio-temporal modeling of meteorological and climatic variables using Open Source software
Pages 1-3



Software outsourcing to Eastern Europe: Which countries work best?

Eastern Europe is building up a head of steam as a software outsourcing powerhouse, as companies focus more on cost-effectiveness and proximity, rather than just low cost.

By Andrada Fiscutean for Central European Processing | April 27, 2016 -- 11:13 GMT (04:13 PDT) | Topic: Developer

SoilGrids

A system for automated global soil mapping

www.soilgrids.org

ISRIC

World Soil Information

Automated (predictive) global soil mapping

We aims at implementing (all) knowledge of soil science through robust algorithms (BUP), and putting the system in operational use…

so that you do not need any soil mappers any more :)

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

SoilGrids1km — Global Soil Information Based on Automated Mapping

Tomislav Hengl, Jorge Mendes de Jesus, Robert A. MacMillan, Niels H. Batjes, Gerard B. M. Heuvelink, Eloi Ribeiro, Alessandro Samuel-Rosa, Bas Kempen, Johan G. B. Leenaars, Markus G. Walsh, Maria Ruiperez Gonzalez

Published: August 29, 2014 • DOI: 10.1371/journal.pone.0105992

Abstract

Background

Soils are widely recognized as a non-renewable natural resource and as biophysical carbon sinks. As such, there is a growing requirement for global soil information. Although several global soil information systems already exist, these tend to suffer from inconsistencies and limited spatial detail.

Methodology/Principal Findings

We present SoilGrids1km — a global 3D soil information system at 1 km resolution — containing spatial predictions for a selection of soil properties (at six standard depths): soil organic carbon (g kg−1), soil pH, sand, silt and clay fractions (%), bulk density (kg m−3), cation-exchange capacity (cmol+/kg), coarse fragments (%), soil organic carbon stock (t ha−1), depth to bedrock (cm), World Reference Base soil groups, and USDA Soil Taxonomy suborders. Our predictions are based on global spatial prediction models which we fitted, per soil variable,

Subject Areas

- Agricultural soil scie...
- Clay mineralogy
- Deserts
- Forecasting
- Remote sensing
- Surface temperature
- Taxonomy
- Urban areas

SoilGrids250m: global gridded soil information based on Machine Learning

Tomislav Hengl[1], Jorge Mendes de Jesus[1], Gerard B.M. Heuvelink[1], Maria Ruiperez Gonzalez[1], Milan Kilibarda[2], Aleksandar Blagotić[3], Wei Shangguan[4], Marvin N. Wright[5], Xiaoyuan Geng[6], Bernhard Bauer-Marschallinger[7], Mario Antonio Guevara[8], Rodrigo Vargas[8], Robert A. MacMillan[9], Niels H. Batjes[1], Johan G.B. Leenaars[1], Eloi Ribeiro[1], Ichsani Wheeler[10], Stephan Mantel[1], and Bas Kempen[1]

[1]ISRIC — World Soil Information, Wageningen, the Netherlands
[2]Faculty of Civil Engineering, University of Belgrade, Serbia
[3]GILab Ltd, Belgrade, Serbia
[4]College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
[5]Institut für Medizinische Biometrie und Statistik, Lübeck, Germany
[6]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada
[7]Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria
[8]University of Delaware, Newark DE, USA
[9]LandMapper Environmental Solutions Inc., Edmonton, Canada
[10]Envirometrix Inc., Wageningen, the Netherlands

Correspondence to: T. Hengl (tom.hengl@isric.org)

Abstract. This paper describes the technical development and accuracy assessment of the most recent and improved version of the SoilGrids system at 250 m resolution (June 2016 update). SoilGrids provides global predictions for standard numeric soil properties (organic carbon, bulk density, Cation Exchange Capacity (CEC), pH, soil texture fractions and coarse fragments) at seven standard depths (0, 5, 15, 30, 60, 100 and 200 cm), in addition to predictions of depth to bedrock and distribution of soil classes based on the World Reference Base (WRB) and USDA classification systems (ca. 280 raster layers in total). Predictions were based on ca. 150,000 soil profiles used for training and a stack of 158 remote sensing-based soil covariates (primarily derived from MODIS land products, SRTM DEM derivatives, climatic images and global landform and lithology maps), which were used to fit an ensemble of machine learning methods — random forest and gradient boosting and/or multinomial logistic regression — as implemented in the R packages ranger, xgboost, nnet and caret. The results of 10–fold

NATURE COMMUNICATIONS | ARTICLE OPEN

Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations

Christian Folberth, Rastislav Skalský, Elena Moltchanova, Juraj Balkovič, Ligia B. Azevedo, Michael Obersteiner & Marijn van der Velde

Affiliations | Contributions | Corresponding author

Nature Communications **7**, Article number: 11872 | doi:10.1038/ncomms11872

Received 29 July 2015 | Accepted 04 May 2016 | Published 21 June 2016

Abstract

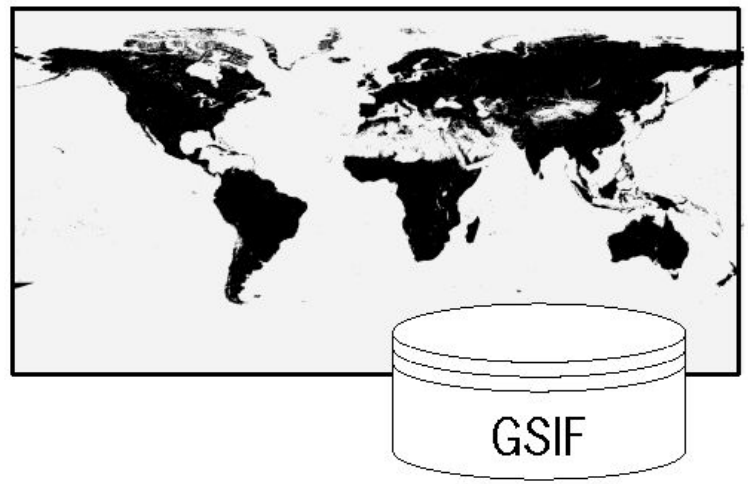
Global gridded crop models (GGCMs) are increasingly used for agro-environmental assessments and estimates of climate change impacts on food production. Recently, the influence of climate data and weather variability on GGCM outcomes has come under detailed scrutiny, unlike the influence of soil data. Here we compare yield variability caused by the soil type selected for GGCM simulations to weather-induced yield variability. Without fertilizer application, soil-type-related yield variability generally outweighs the simulated inter-annual variability in yield due to weather. Increasing applications of fertilizer and irrigation reduce this variability until it is practically negligible. Importantly, estimated climate change effects on yield can be either negative or positive depending on the chosen soil type. Soils thus have the capacity to either buffer or amplify these impacts. Our findings call for improvements in soil data available for crop modelling and more explicit accounting for soil variability in GGCM

Global Soil Information Facilities

Global Soil Information Facilities

ISRIC

SoilGrids



Query soil data

SoilInfo App

Spatial location (site)

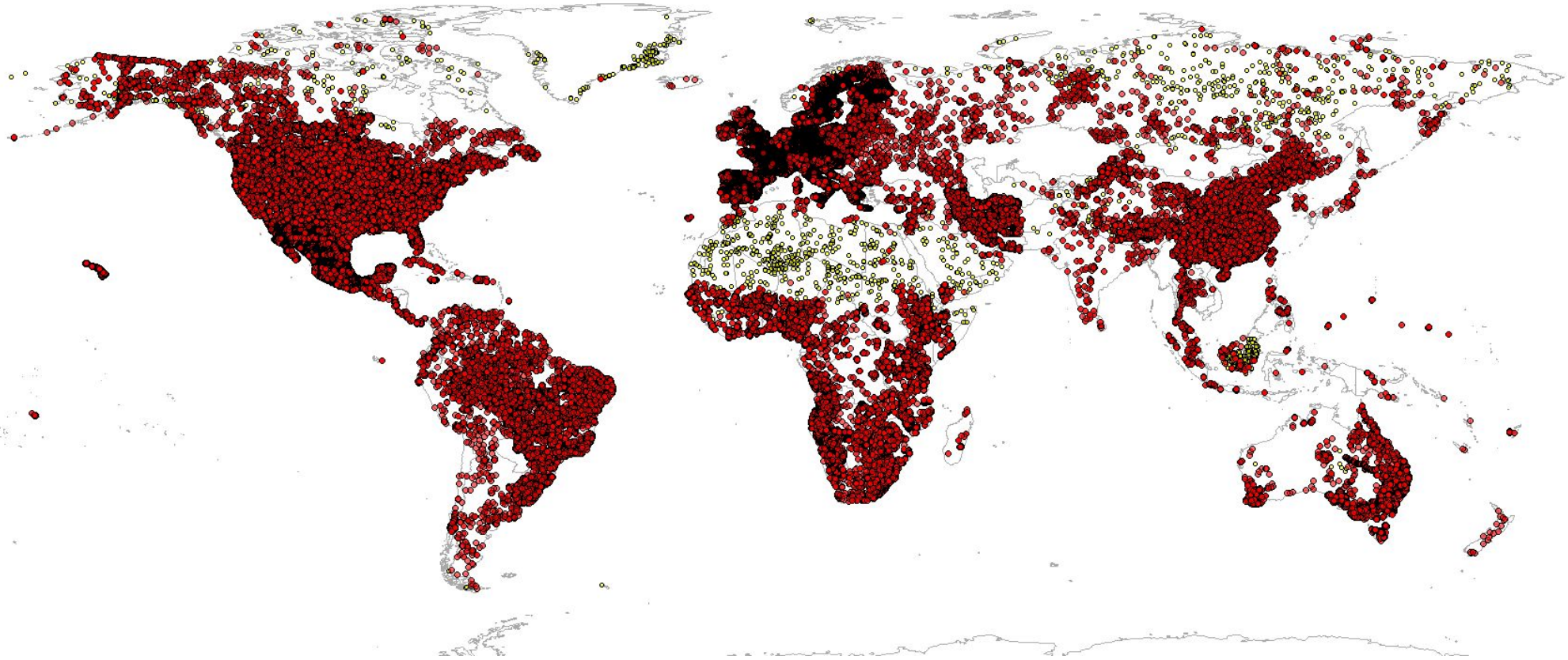
GPS-enabled mobile phone

- Soil property-depth curves (with confidence intervals)
- Dominant, 2nd and 3rd soil type
- Nutrient status (e.g. soil organic carbon stock)
- Available water capacity
- Current and past land use
- Closest soil data provider / agricultural extension service

SoilGrids inputs

- ➔ **"World's largest" compilation of soil profile / soil sample data sets** (national and international datasets from over 45 countries) → these contain both soil classification data + soil analytical/laboratory data.
- ➔ A 40TB repository of MODIS land products, climatic images, DEM derivatives, geological and geomorphological data (all at 250 m resolution)
- ➔ ISRIC's international network that can cross-check and validate spatial prediction patterns / values.

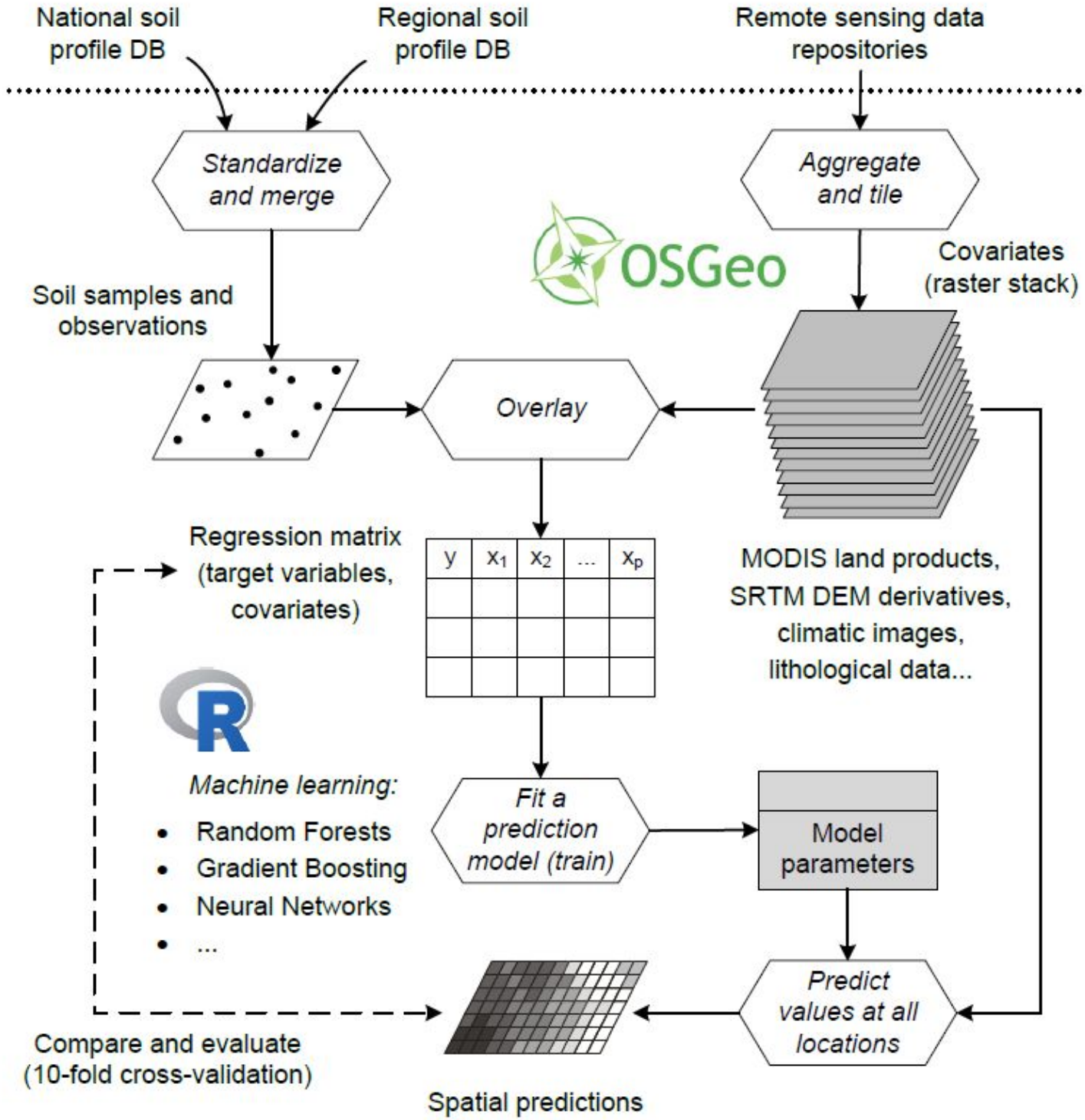
ca 150,000 points shown on this map



ISRIC World Soil Information

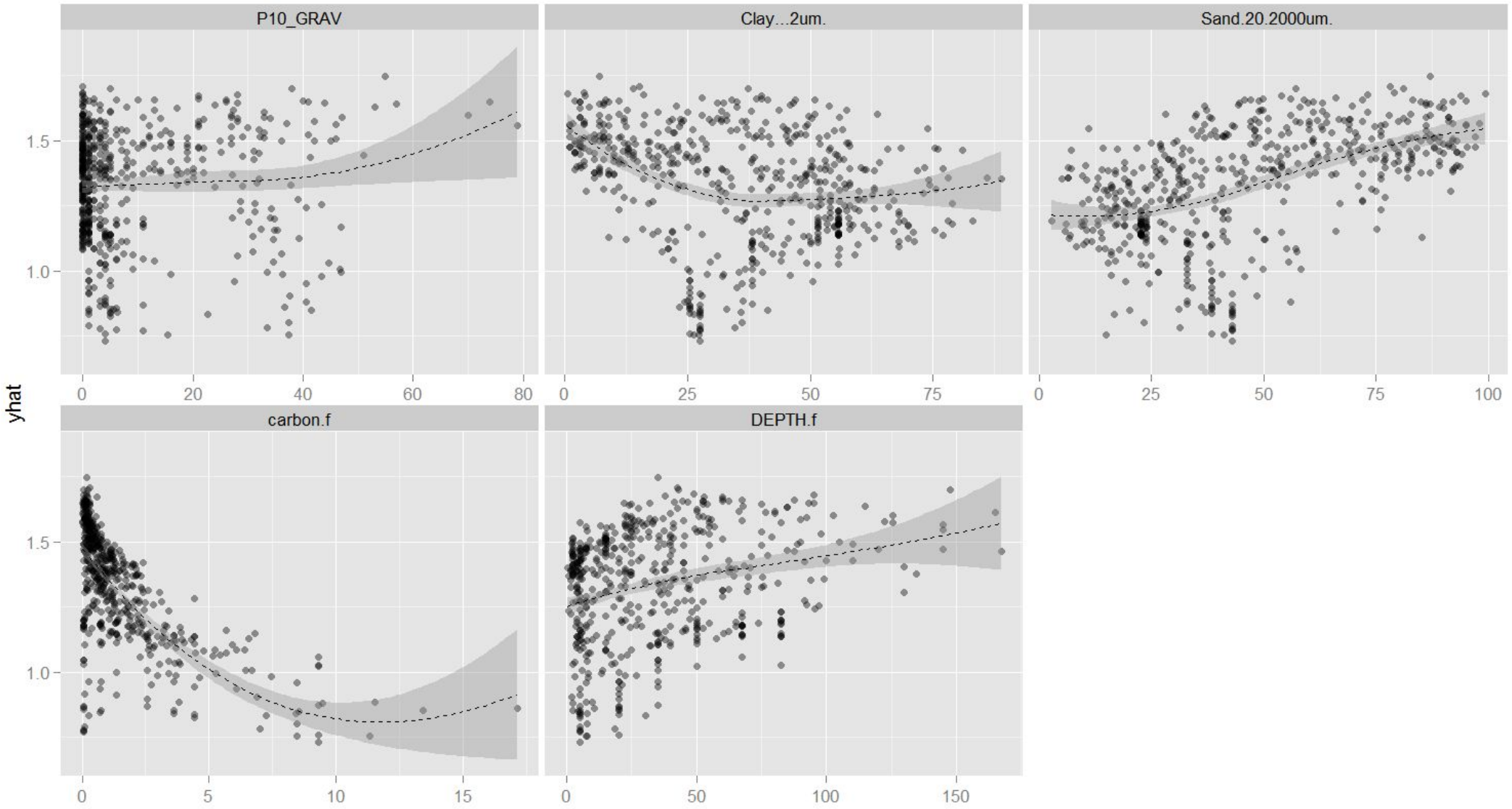
>850,000 measurements of soil organic carbon

Machine learning


National soil profile DB
Regional soil profile DB
Remote sensing data repositories
Standardize and merge
Aggregate and tile
OSGeo
Soil samples and observations
Covariates (raster stack)
Overlay
Regression matrix (target variables, covariates)
y
x1
x2
...
xp
MODIS land products, SRTM DEM derivatives, climatic images, lithological data...
Machine learning:
Random Forests
Gradient Boosting
Neural Networks
...
Fit a prediction model (train)
Model parameters
Predict values at all locations
Compare and evaluate (10-fold cross-validation)
Spatial predictions

Some common misconceptions

- ➔ MLA is a black box — it is not!
- ➔ Most of new implementations of random forest and similar will penalize for overfitting.
- ➔ MLA is non-statistical / there are no probabilities and the output (prediction) uncertainty is unknown — it is not true. **"MLA is a marriage between stats and knowledge representation"**.


P10_GRAV
Clay...2um.
Sand.20.2000um.
carbon.f
DEPTH.f
yhat
1.5
1.0
0
20
40
60
80
25
50
75
100
5
10
15
150

Watch out from overfitting

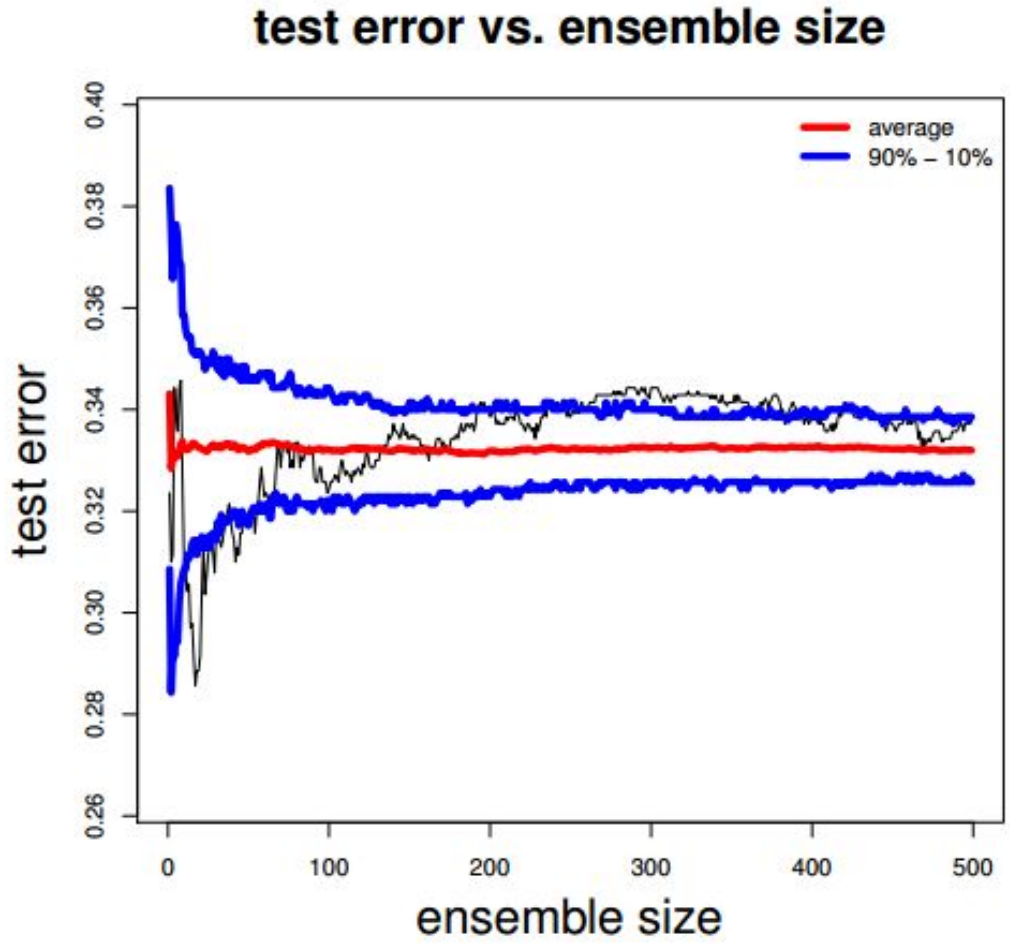


Figure 1: The black curve shows how Err_t varies as a function of t for a particular realization of $Q_1, \ldots, Q_t$. The red curve was obtained by generating many ensembles, and computing the average of the respective Err_t values at each t. Similarly, the blue curves depict the 10th and 90th percentiles of Err_t at each t.

Based on Lopez (2015)

Statistical Science
2001, Vol. 16, No. 3, 199–231

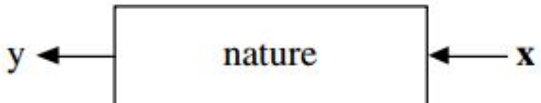
Statistical Modeling: The Two Cultures

Leo Breiman

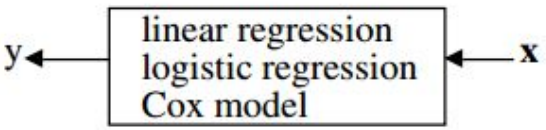
Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

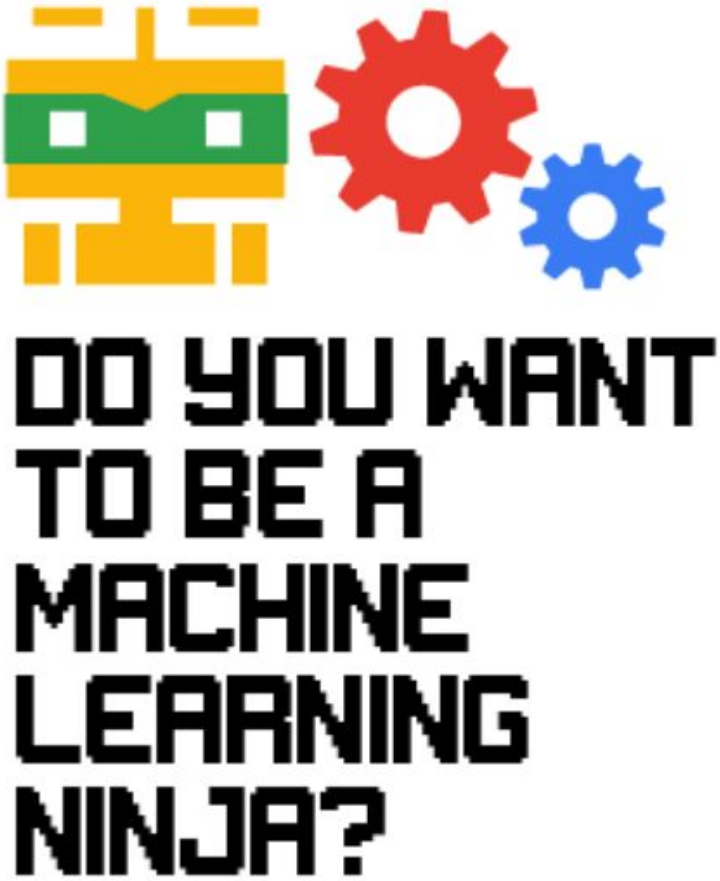


The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.

Estimated culture population. 98% of all statisti-



"The tagline is, *Do you want to be a machine learning ninja?*" says Christine Robson, a product manager for Google's internal machine learning efforts, who helps administer the program. "So we invite folks from around Google to come and spend six months embedded with the machine learning team, sitting right next to a mentor, working on machine learning for six months, doing some project, getting it launched and learning a lot."

MLAs of interest to global soil mapping

Some logical choices

- ➔ **Soils vary locally** (there is a strong local component) — tree-based models probably more suited.
- ➔ **Soils are shaped as a result of complex processes running over large periods of time**... but there are also sudden disturbances (soil slides, erosion)... in short: soil genesis is complex.
- ➔ Remote sensing (VISNIR) can not be used to directly map soils, BUT it can be used to represent soil forming factors.

Proposed solution:

- ➔ An extensive stack of carefully selected covariates (158)
- ➔ **Random forest + Gradient boosting**
- ➔ Put more effort to optimize the system so that it can be updated relatively fast

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE BIOMETRIE UND STATISTIK

IMBS

Home

Main Menu
- Home
- Staff
- How to find us
- Weblinks

Research
- Awards
- Projects
- Publications
- Software
- Employment Ad

Courses
- Biometrical Consulting Service
- Lectures
- Leselunch
- Seminar
- Student Research Project
- Veranstaltungen

Search

Search this site:

Search

Marvin N. Wright, M.Sc.

Institut für Medizinische Biometrie und Statistik
Universität zu Lübeck
Universitätsklinikum Schleswig-Holstein, Campus Lübeck
Ratzeburger Allee 160
23562 Lübeck
fon: 0451-500 2782
fax: 0451-500 2999
wright@imbs.uni-luebeck.de

Publications

- Wright, M. N., Ziegler, A. & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics* 17:145. http://dx.doi.org/10.1186/s12859-016-0995-8.
- Wright, M. N. & Ziegler, A. (2016). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, in press. http://arxiv.org/abs/1508.04409.
- Schirmer, J. H., Wright, M. N., Vonthein, R., Herrmann, K., Nölle. B., Both, M., Henes, F., Arlt, A., Gross, W. L., Schinke, S., Reinhold-Keller, E., Moosig, F. & Holle, J. U. (2016). Clinical presentation and long-term outcome of 144 patients with microscopic polyangiitis in a monocentric German cohort. *Rheumatology (Oxford)* 55:71-79. http://dx.doi.org/10.1093/rheumatology/kev286.
- Wright, M. N. & Ziegler, A. (2015). Multiple censored data in dentistry: A new statistical model for analyzing lesion size in randomized controlled trials. *Biometrical Journal* 57:384–394. http://dx.doi.org/10.1002/bimj.201400118.
- Paulick, C., Wright, M. N., Verleger, R. & Keller, K. (2014). Decomposition of 3-way arrays: A comparison of different PARAFAC algorithms. *Chemometrics and Intelligent Laboratory Systems* 137:97-109. http://dx.doi.org/10.1016/j.chemolab.2014.06.009.

Technical Reports

- Schmid, M., Wright, M. N. & Ziegler, A. (2015). On the Use of Harrell's C for Node Splitting in Random Survival Forests. *Technical Report*. http://arxiv.org/abs/1507.03092.

ISRIC World Soil Information

>850,000 measurements of soil organic carbon

ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R

Marvin N. Wright
Universität zu Lübeck

Andreas Ziegler
Universität zu Lübeck,
University of KwaZulu-Natal

Abstract

We introduce the C++ application and R package **ranger**. The software is a fast implementation of random forests for high dimensional data. Ensembles of classification, regression and survival trees are supported. We describe the implementation, provide examples, validate the package with a reference implementation, and compare runtime and memory usage with other implementations. The new software proves to scale best with the number of features, samples, trees, and features tried for splitting. Finally, we show that **ranger** is the fastest and most memory efficient implementation of random forests to analyze data on the scale of a genome-wide association study.

Keywords: C++, classification, machine learning, R, random forests, **Rcpp**, recursive partitioning, survival analysis.

otto group

Understanding XGBoost Model on Otto Da...

by Tianqi Chen · last run 11 months ago · R notebook · 44107 views
using data from Otto Group Product Classification Challenge

100 voters

Report

Understanding XGBoost Model on Otto Dataset

Michaël Benesty

1 Introduction

XGBoost is an implementation of the famous gradient boosting algorithm. This model is often described as a *blackbox*, meaning it works well but it is not trivial to understand how. Indeed, the model is made of hundreds (thousands?) of decision trees. You may wonder how possible a human would be able to have a general view of the model?

XGBoost: A Scalable Tree Boosting System

Tianqi Chen
University of Washington
tqchen@cs.washington.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

ABSTRACT

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

CCS Concepts

•**Methodologies** → **Machine learning;** •**Information systems** → **Data mining;**

Keywords

many applications. Tree boosting has been shown to give state-of-the-art results on many standard classification benchmarks [14]. LambdaMART [4], a variant of tree boosting for ranking, achieves state-of-the-art result for ranking problems. Besides being used as a stand-alone predictor, it is also incorporated into real-world production pipelines for ad click through rate prediction [13]. Finally, it is the de-facto choice of ensemble method and is used in challenges such as the Netflix prize [2].

In this paper, we describe XGBoost, a scalable machine learning system for tree boosting. The system is available as an open source package[2]. The impact of the system has been widely recognized in a number of machine learning and data mining challenges. Take the challenges hosted by the machine learning competition site Kaggle for example. Among the 29 challenge winning solutions [3] published at Kaggle's blog during 2015, 17 solutions used XGBoost. Among these solutions, eight solely used XGBoost to train the model, while most others combined XGBoost with neural nets in en-

Journal of Statistical Software

November 2008, Volume 28, Issue 5. *http://www.jstatsoft.org/*

Building Predictive Models in R Using the caret Package

Max Kuhn
Pfizer Global R&D

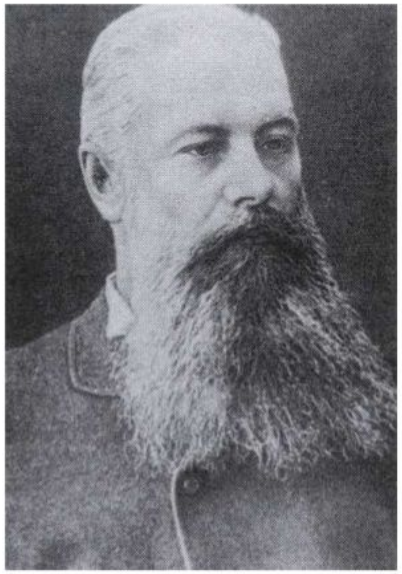
Abstract

The **caret** package, short for classification and regression training, contains numerous tools for developing predictive models using the rich set of models available in R. The package focuses on simplifying model training and tuning across a wide variety of modeling techniques. It also includes methods for pre-processing training data, calculating variable importance, and model visualizations. An example from computational chemistry is used to illustrate the functionality on a real data set and to benchmark the benefits of parallel processing with several types of models.

Keywords: model building, tuning parameters, parallel processing, R, **NetWorkSpaces**.

Results

They would have been interested in this…



Vasili Dokuchaev

The Russian School

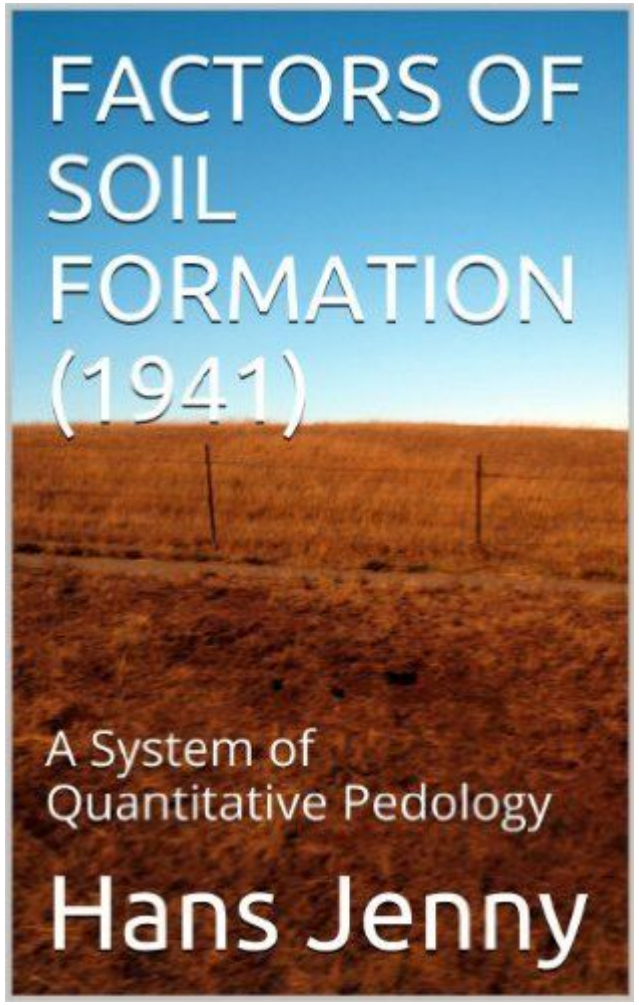
Soil forming factors

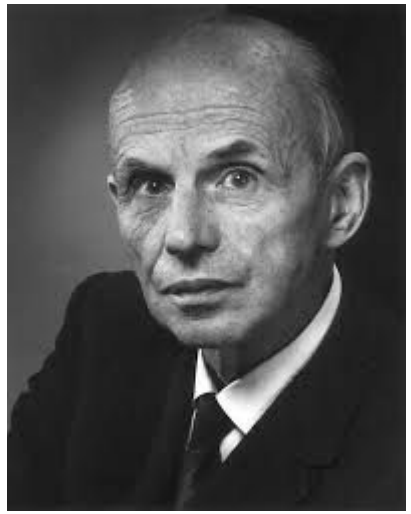
⇩

Soil forming processes

⇩

Different Soils





Bulk density

Soil pH

Soil organic carbon

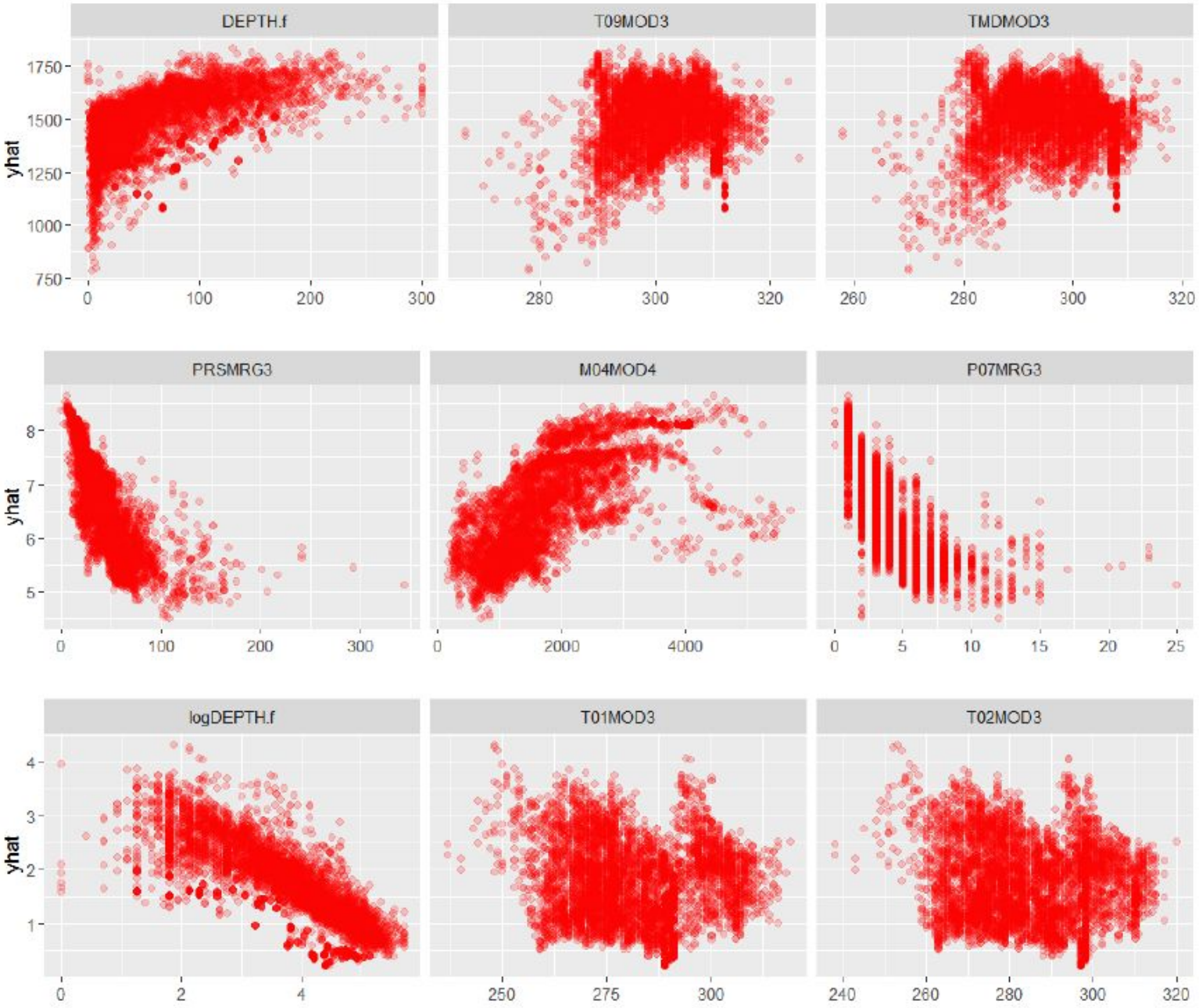


Figure 6. Examples of fitted relationships for bulk density (above), pH (middle) and soil organic carbon (below). Plots show target variables and top three most important covariates as reported by the random forest model. `DEPTH.f` is the depth from soil surface, `T09MOD3` is mean monthly temperature for September, `TMDMOD3` is mean annual temperature, `PRSMRG3` is total annual precipitation, `M04MOD4` is mean monthly MODIS NIR band reflectance, `P07MRG3` is mean monthly precipitation for July, `T01MOD3` is mean monthly temperature for January, and

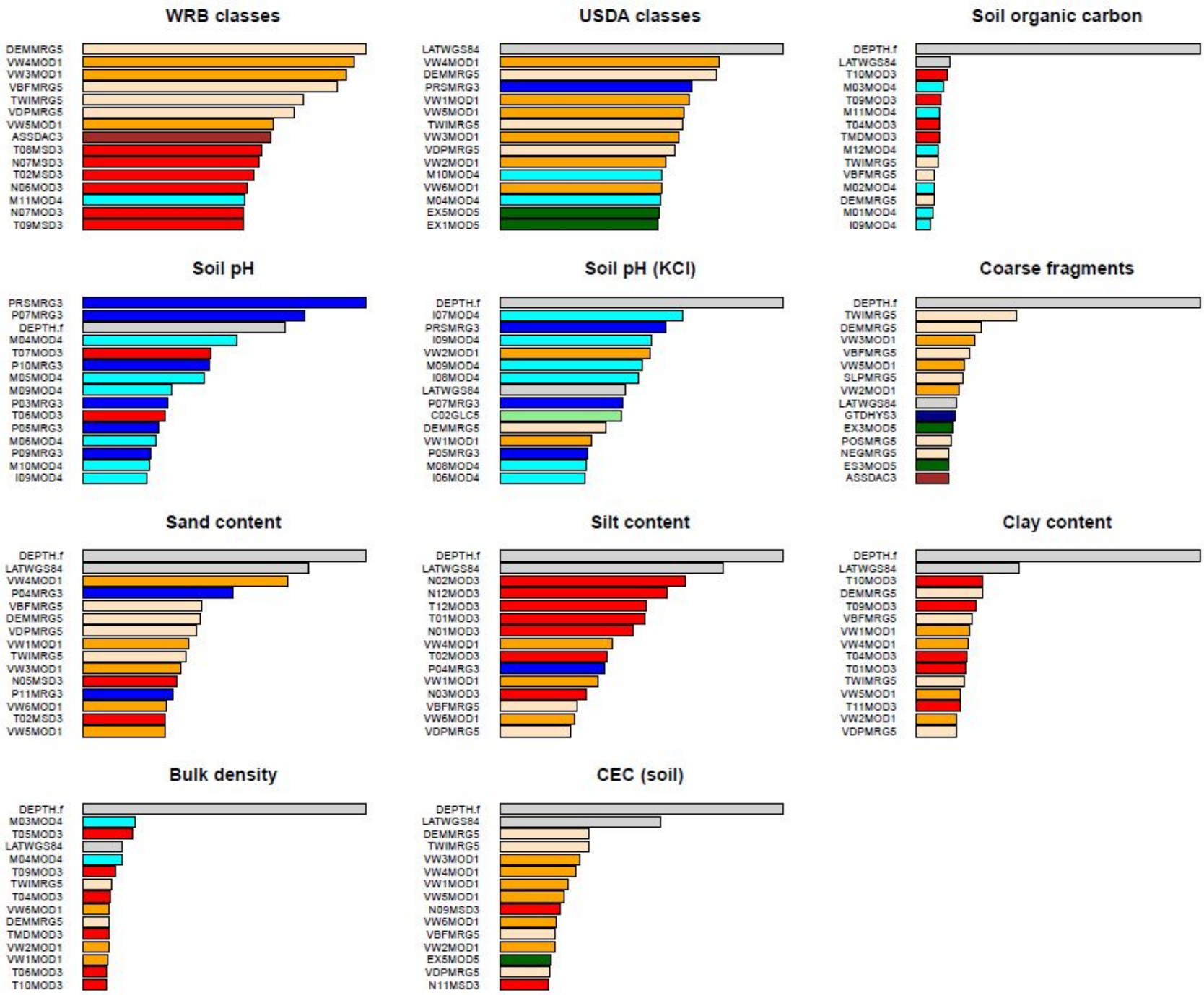
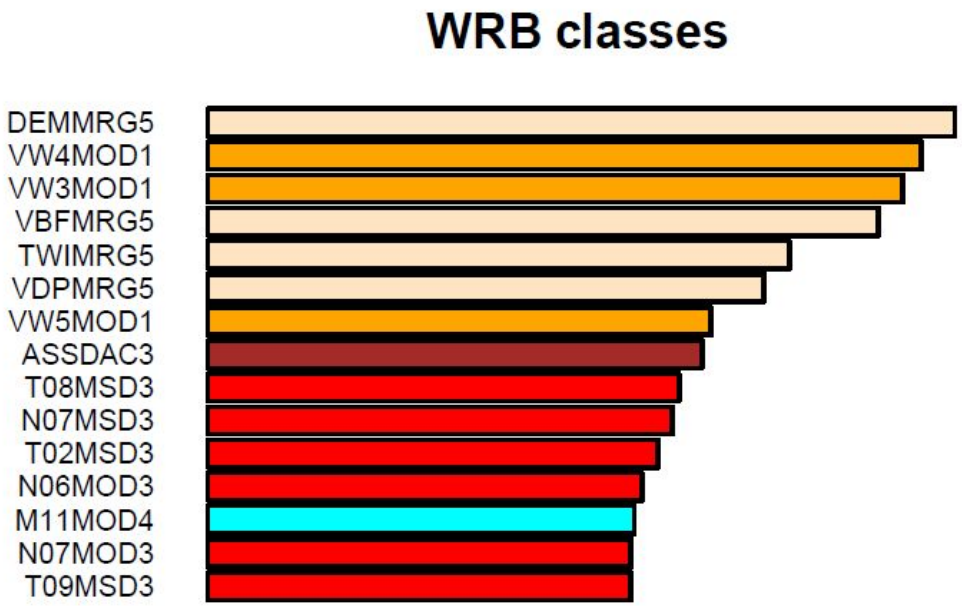
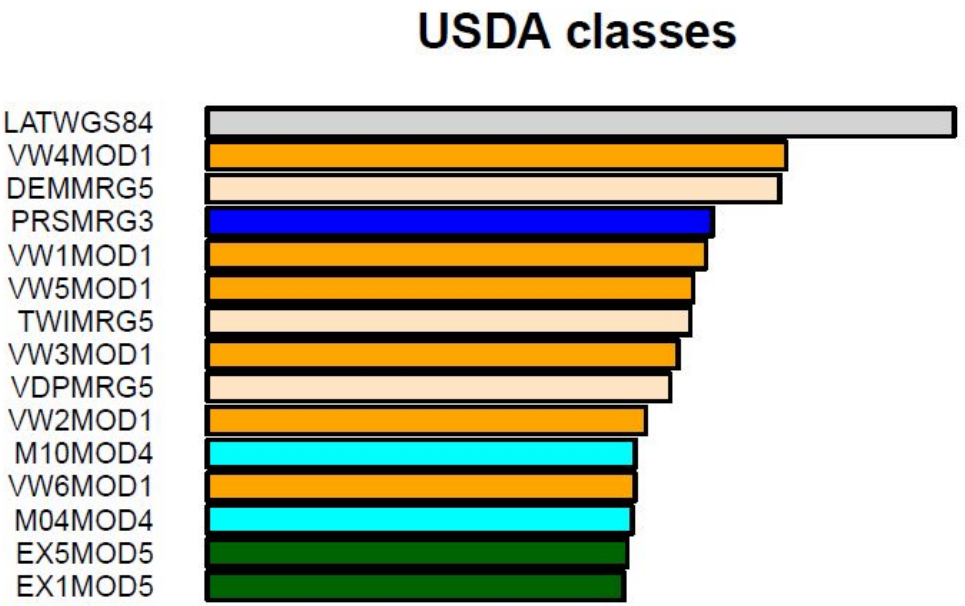


Figure 5. Fitted variable importance plots for target variables. Generated as an average between using the `ranger` and `xgboost` packages, (for soil types results are based on the `ranger` model only). `DEPTH.f` is the depth from soil surface, `T**MOD3` and `N**MOD3` are mean monthly temperatures daytime and nighttime (red color), `TWI`, `DEM`, `VBF` and `VDP` are DEM-parameters (bisque color), `M**MOD4` are mean

Variable importance soil types





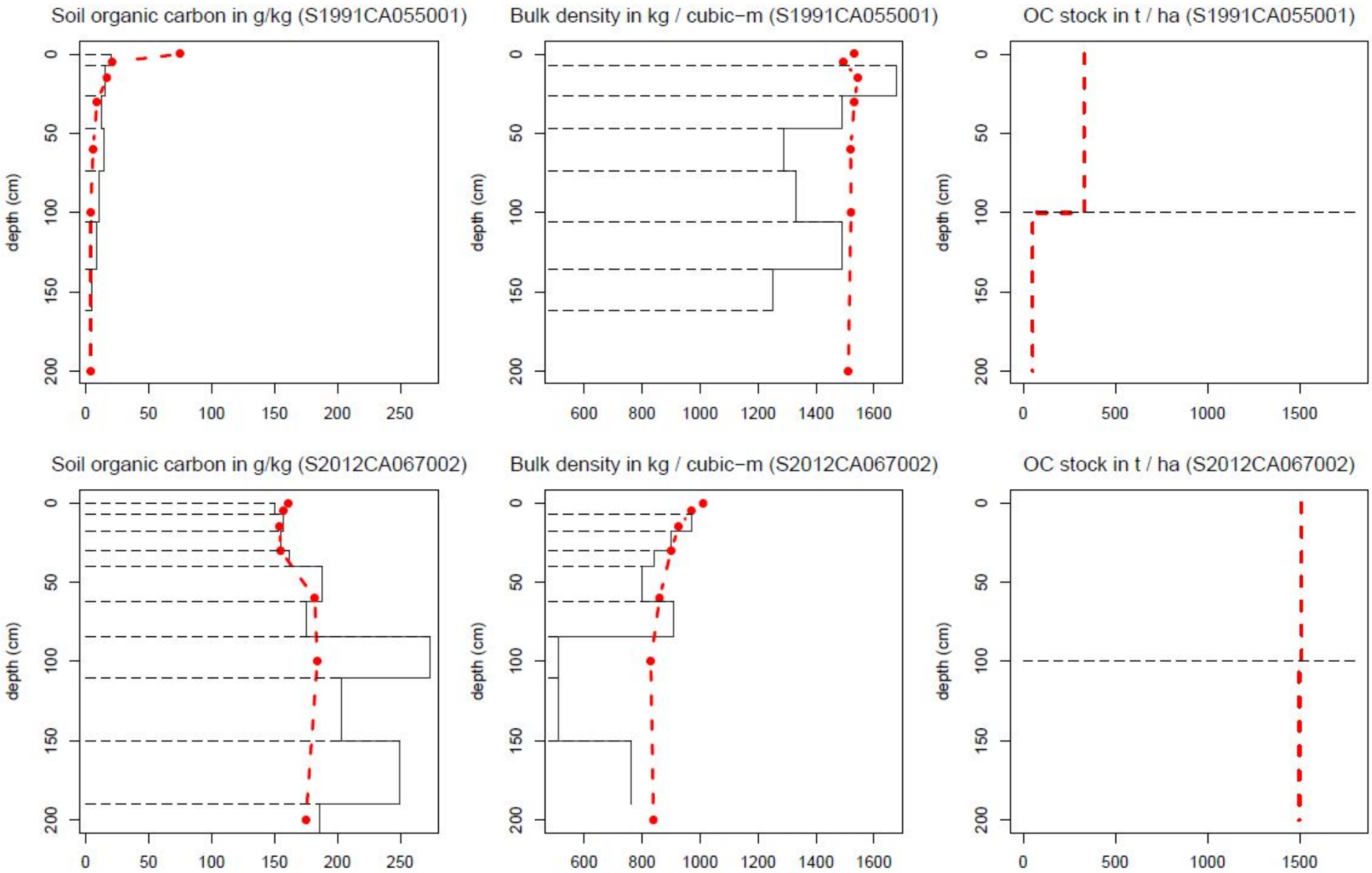
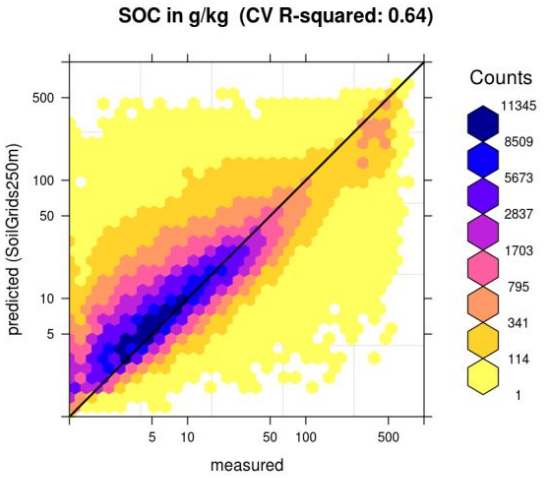
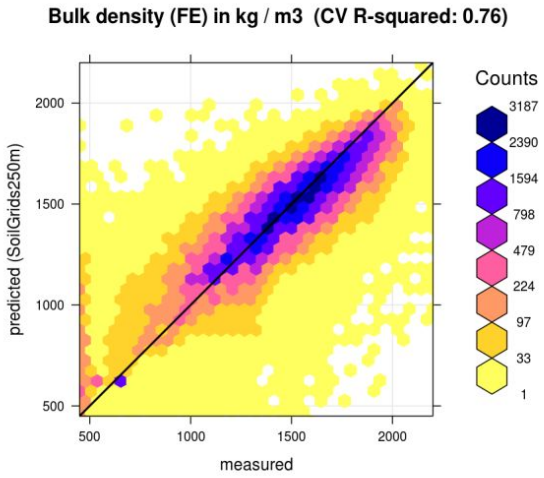
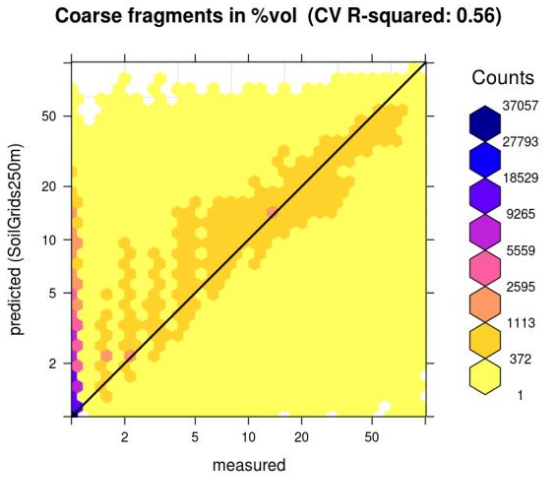
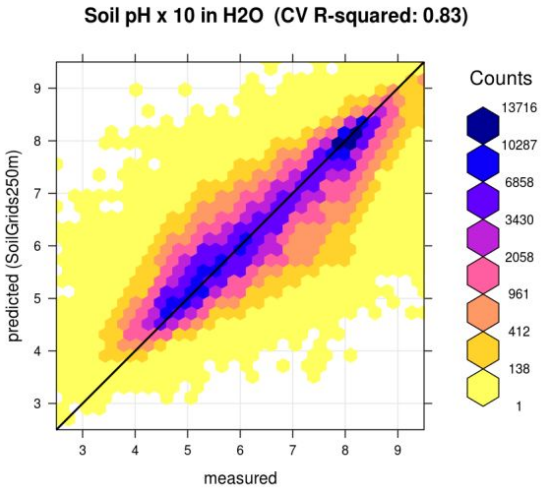


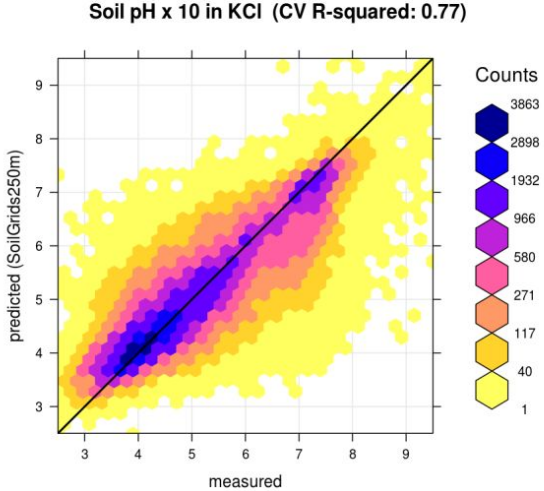
Figure 2. Example of soil variable-depth curves: original sampled soil profiles vs predicted values (SoilGrids) at seven standard depths (broken red line) and estimated soil organic carbon stock for depths 0–100 and 100–200 cm. Locations of points: mineral soil `S1991CA055001` (-122.37°W, 38.25°N), and an organic soil profile `S2012CA067002` (-121.62°W, 38.13°N).

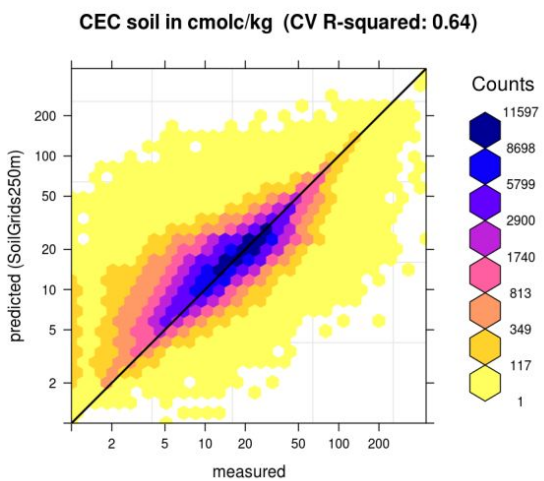

SOC in g/kg (CV R-squared: 0.64)
predicted (SoilGrids250m)
measured
Counts
11345
8509
5673
2837
1703
795
341
114
1

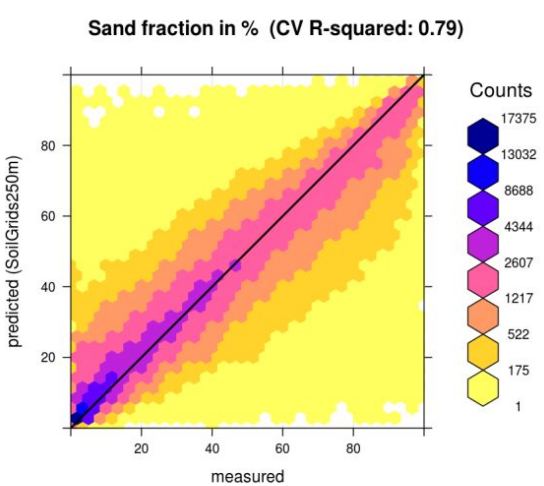

Bulk density (FE) in kg / m3 (CV R-squared: 0.76)
predicted (SoilGrids250m)
measured
Counts
3187
2390
1594
798
479
224
97
33
1

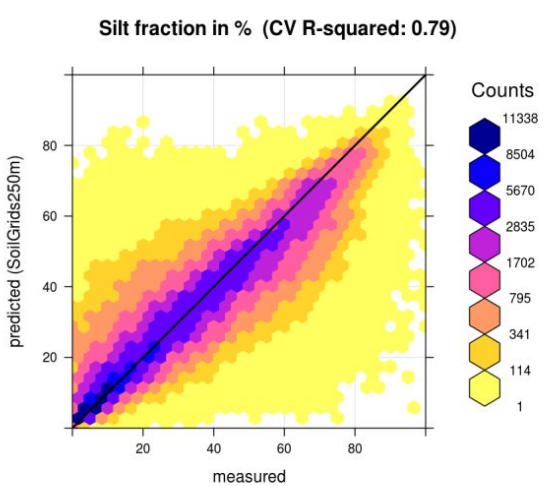

Coarse fragments in %vol (CV R-squared: 0.56)
predicted (SoilGrids250m)
measured
Counts
37057
27793
18529
9265
5559
2595
1113
372
1

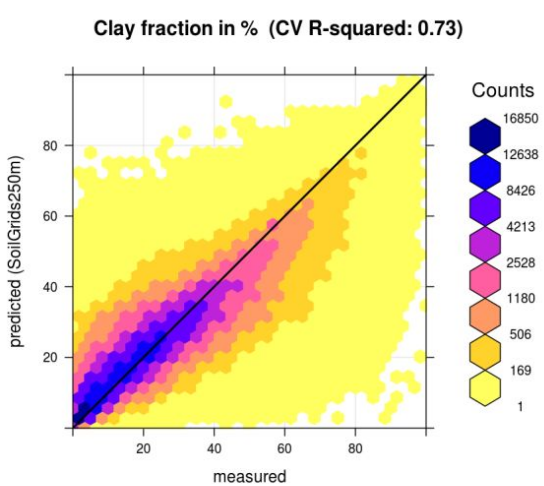

Soil pH x 10 in H2O (CV R-squared: 0.83)
predicted (SoilGrids250m)
measured
Counts
13716
10287
6858
3430
2058
961
412
138
1


Soil pH x 10 in KCl (CV R-squared: 0.77)
predicted (SoilGrids250m)
measured
Counts
3863
2898
1932
966
580
271
117
40
1

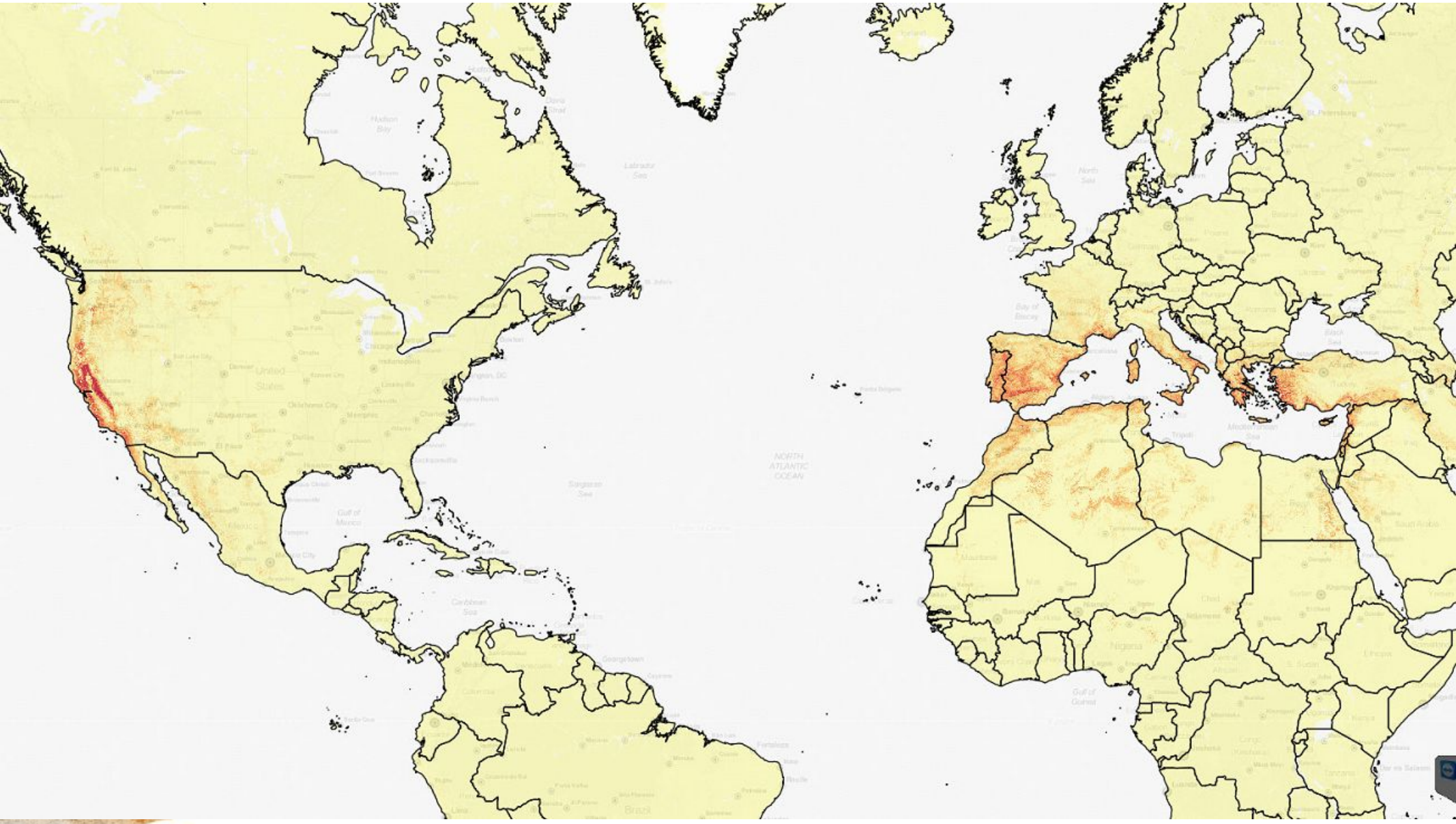

CEC soil in cmolc/kg (CV R-squared: 0.64)
predicted (SoilGrids250m)
measured
Counts
11597
8698
5799
2900
1740
813
349
117
1


Sand fraction in % (CV R-squared: 0.79)
predicted (SoilGrids250m)
measured
Counts
17375
13032
8688
4344
2607
1217
522
175
1

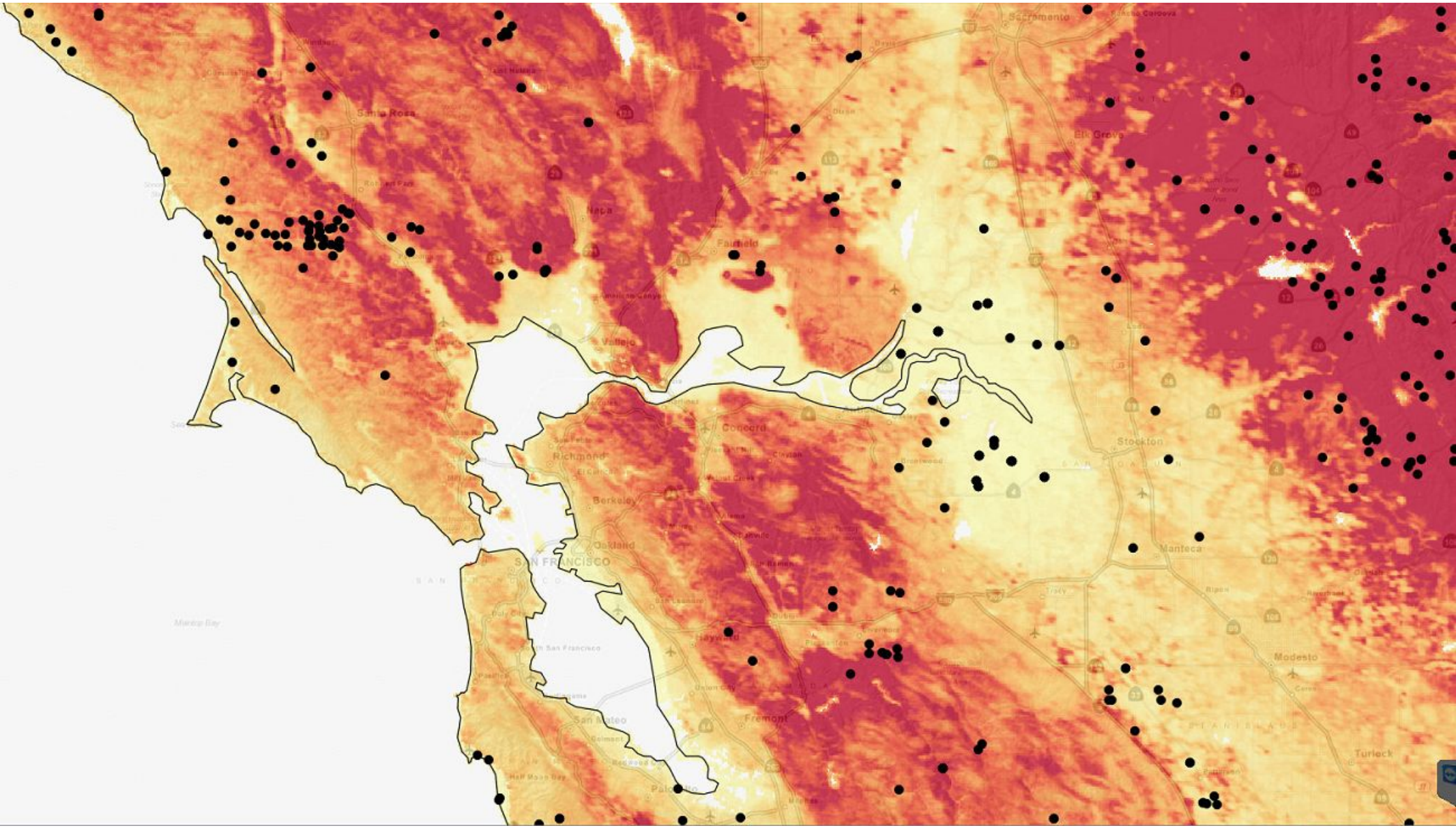

Silt fraction in % (CV R-squared: 0.79)
predicted (SoilGrids250m)
measured
Counts
11338
8504
5670
2835
1702
795
341
114
1

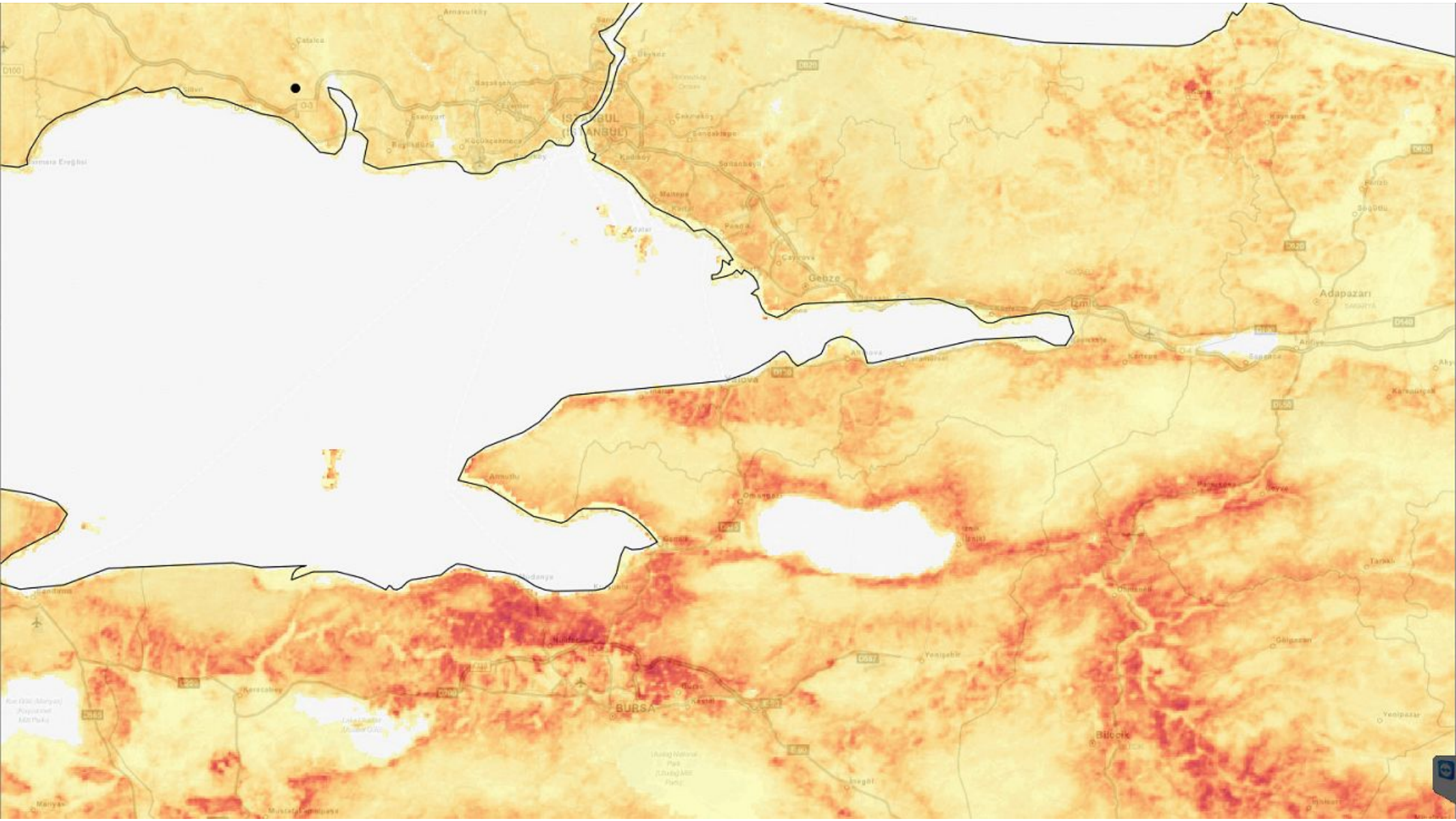

Clay fraction in % (CV R-squared: 0.73)
predicted (SoilGrids250m)
measured
Counts
16850
12638
8426
4213
2528
1180
506
169
1

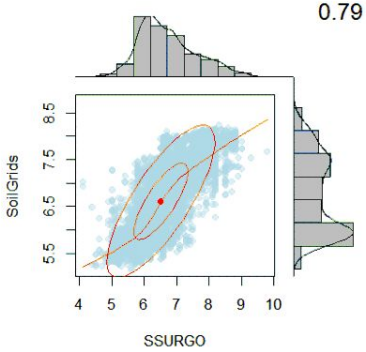
Xeralfs



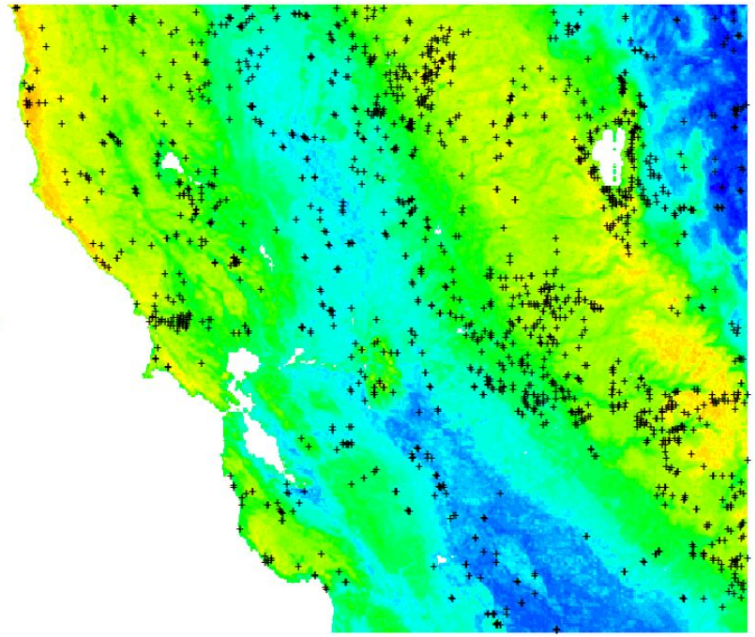
ISRIC World Soil Information


Sacramento
Santa Rosa
Napa
Fairfield
Vallejo
Concord
Richmond
Berkeley
Oakland
SAN FRANCISCO
Stockton
Manteca
Modesto
Turlock
Fremont
San Mateo
Half Moon Bay

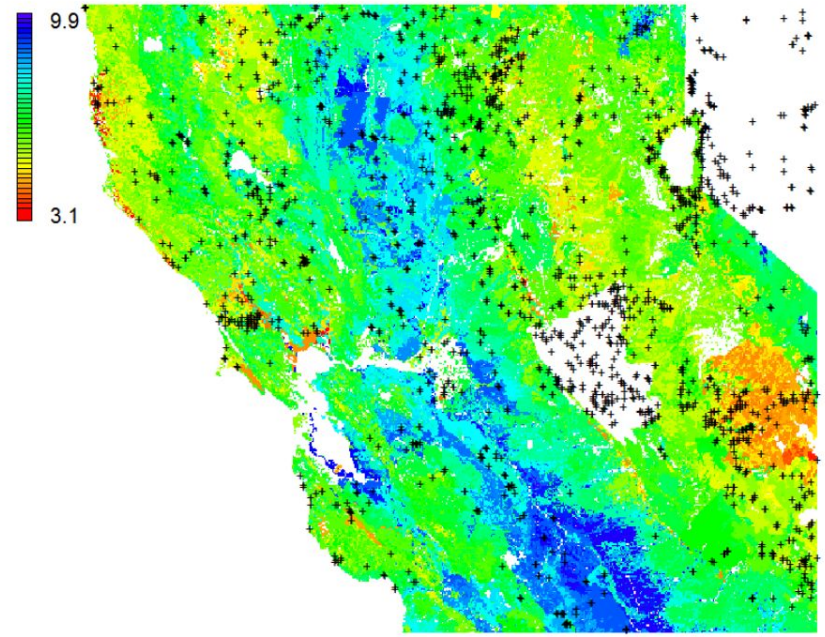

ISTANBUL
İSTANBUL
Gebze
İzmit
Adapazarı
BURSA
Bilecik

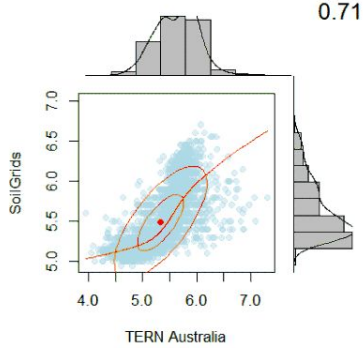


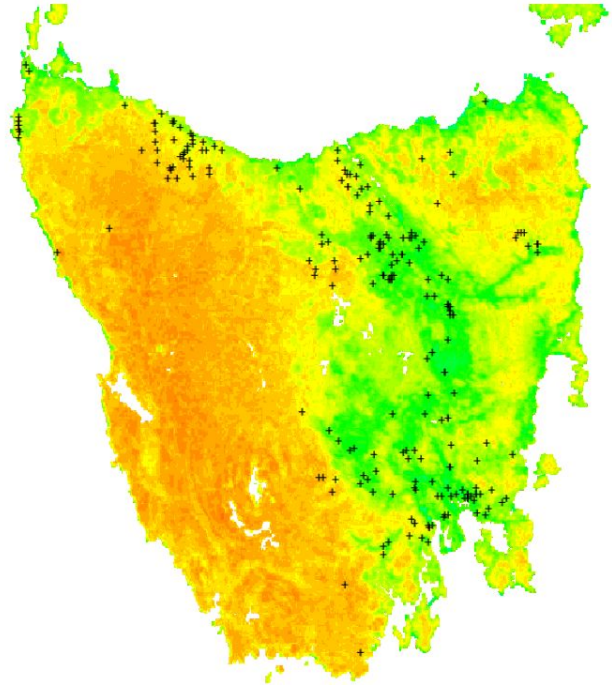
SoilGrids

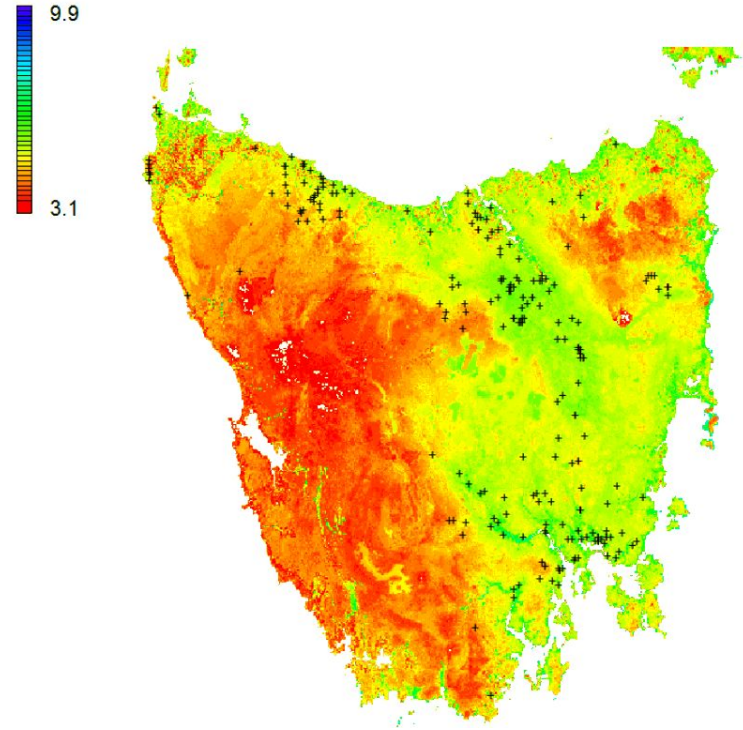


SSURGO





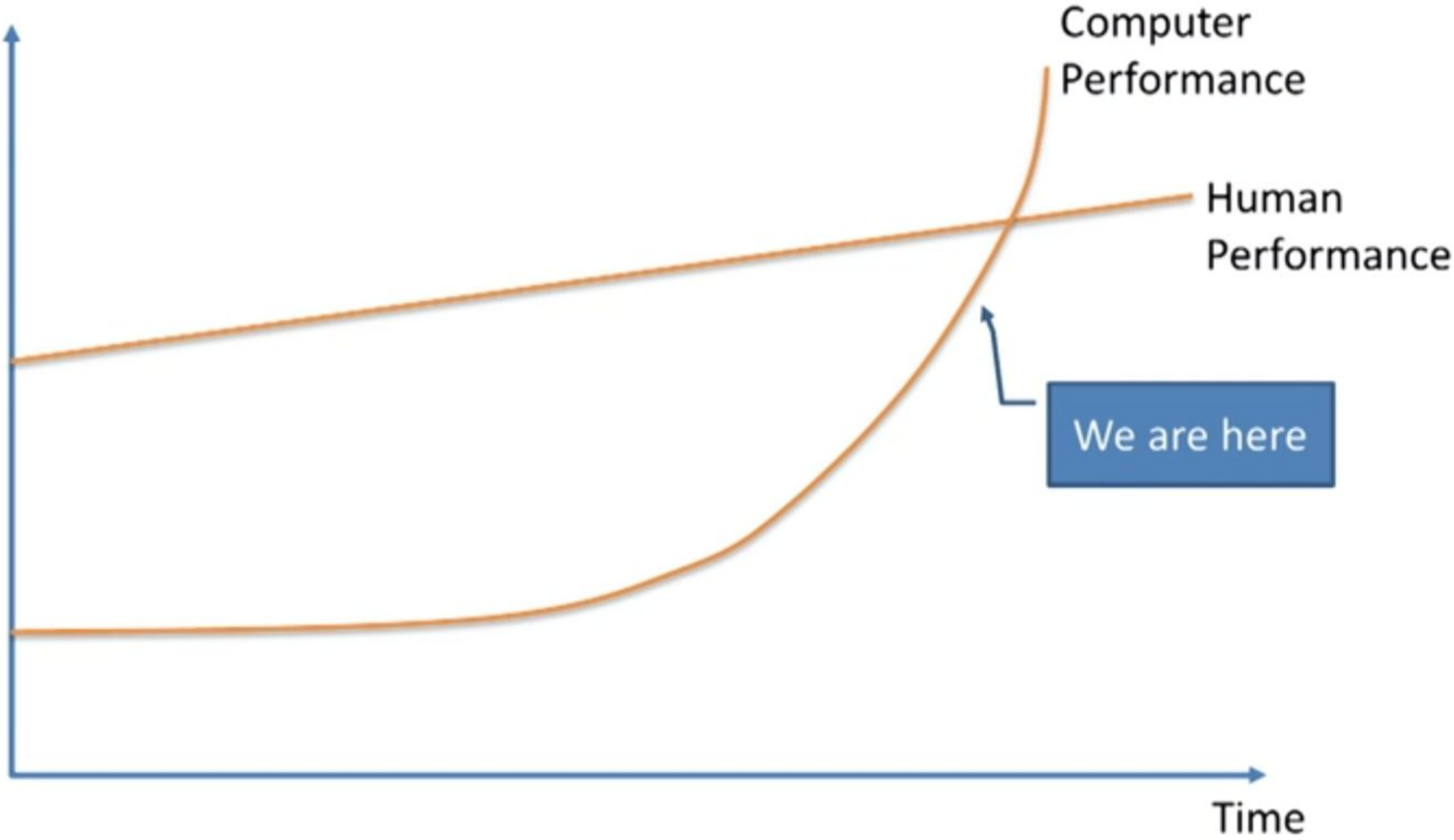




Conclusions

Conclusions

- ➔ Traditional soil surveyors got it right! — distribution of soil classes is mainly controlled by DEM morphometry (especially hydrological parameters).
- ➔ Soil classification and polygon models of soils seem to make sense — in many parts of the world we see ”soil groupings i.e. soil bodies”... but there are also transition zones and small individual patches... so it is really a hybrid that we need.
- ➔ In the machine learning framework, much more time needs to be spent on preparing data


Computer
Performance
Human
Performance
We are here
Time

Viewpoint: How creativity is helped by failure

14 November 2015 | Magazine

When it comes to creating a great work of art, practice makes perfect, writes Matthew Syed.

In today's Magazine

Top Stories

First Paris attacker identified

22 minutes ago

Democrats debate IS after Paris attacks

5 hours ago

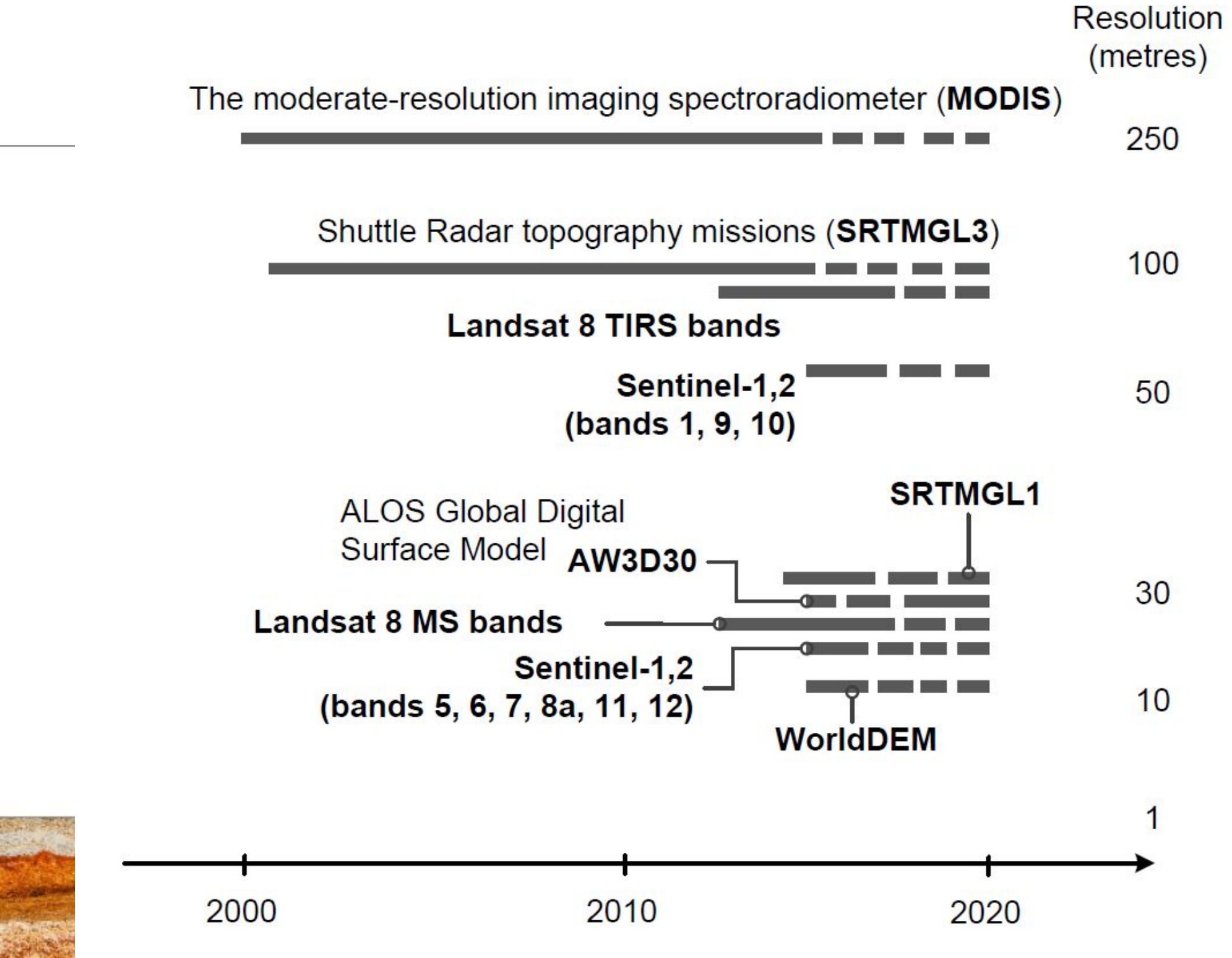
Phones 'need bed mode' to help sleep

15 November 2015

Features & Analysis

Eau de Belgrade

How a perfumer survived two wars and communism


Resolution
(metres)
The moderate-resolution imaging spectroradiometer (MODIS)
250
Shuttle Radar topography missions (SRTMGL3)
100
Landsat 8 TIRS bands
Sentinel-1,2
(bands 1, 9, 10)
50
SRTMGL1
ALOS Global Digital
Surface Model
AW3D30
30
Landsat 8 MS bands
Sentinel-1,2
(bands 5, 6, 7, 8a, 11, 12)
10
WorldDEM
1
2000
2010
2020