Automated soil mapping based on Machine Learning: towards a soil data revolution

T. (Tom) Hengl <tom.hengl@isric.org>
Since the last time you saw me...

300GB

SoilGrids1km

30GB

SoilGrids250m
Since the last time you saw me...

- SoilGrids1km
- SoilGrids250m

1. More points
2. More covariates
3. 16x more pixels
4. MLA (ensemble)
5. HPC
6. Higher accuracy
SOIL GRIDS
A system for automated global soil mapping
www.soilgrids.org
SoilGrids250m: global gridded soil information based on Machine Learning

Tomislav Hengl¹, Jorge Mendes de Jesus¹, Gerard B.M. Heuvelink¹, Maria Ruiperez Gonzalez¹, Milan Kilibarda², Aleksandar Blagotić³, Wei Shangguan⁴, Marvin N. Wright⁵, Xiaoyuan Geng⁶, Bernhard Bauer-Marschallinger⁷, Mario Antonio Guevara⁸, Rodrigo Vargas⁸, Robert A. MacMillan⁹, Niels H. Batjes¹, Johan G.B. Leenaars¹, Eloi Ribeiro¹, Ichsani Wheeler¹⁰, Stephan Mantel¹, and Bas Kempen¹

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

➔ OpenStreetMap
➔ OpenWeatherMap
➔ Wikipedia
➔ Global Biodiversity Information Facilities
1. Open Data license
ODC Open Database License (ODbL) Summary

This is a human-readable summary of the ODbL 1.0 license. Please see the disclaimer below.

You are free:

To Share: To copy, distribute and use the database.

To Create: To produce works from the database.

To Adapt: To modify, transform and build upon the database.

As long as you:

Attribute: You must attribute any public use of the database, or works produced from the database, in the manner specified in the ODbL. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database.

Share-Alike: If you publicly use any adapted version of this database, or works produced from an adapted database, you must also offer that adapted database under the ODbL.

Keep open: If you redistribute the database, or an adapted version of it, then you may use technological measures that restrict the work (such as DRM) as long as you also redistribute a version without such measures.

Disclaimer

This is not a license. It is simply a handy reference for understanding the ODbL 1.0 — it

World Soil Information

http://opendatacommons.org/licenses/odbl/summary/
2. Versioning
(automated mapping system)
3. Reproducibility / open code
https://github.com/ISRICWorldSoil/

SoilGrids250m

Global spatial predictions of soil properties and classes at 250 m resolution

What can you find on this github repository:
- R scripts documenting processing steps,
- Sample code explaining the modelling framework,
- Functions for Cross-validation of ensemble models with examples,
- Examples of predictions, outputs and visualizations.
SoilGrids inputs

➔ ca 150,000 points ("World's largest" compilation of soil profile / soil sample datasets) based on national and international datasets from over 45 countries.

➔ 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.
Data holdings in WoSIS 2
(December 2015)

- About 98,000 unique profiles
- Some 76,000 profiles are georeferenced within defined limits
- Number of measured data for each property varies between profiles with depth, generally depending on the purpose of the initial studies
- Source data based on diverse (inter)national standards
- Generally, limited quality information provided with the source (analytical) data

Lineage:
- Datasets, reports & maps

Soil observations and measurements:
- Feature (georeferenced profiles & layers)
- Attribute (x-y-z-t, map, class, site, layer-field, layer-lab)
- Method
- Value, including units of expression
Big thanks to:
AfSIS project

Bill & Melinda Gates Foundation

Business Operation

Bill & Melinda Gates Foundation is one of the largest private foundations in the world, founded by Bill and Melinda Gates. It was launched in 2000 and is said to be the largest transparently operated private foundation in the world.

Wikipedia

Nonprofit category: Private Grantmaking Foundations

Founded: 2000

Assets: 36.79 billion USD (2010)

Income: 53 billion USD (2010)

Founders: Melinda Gates, Bill Gates
Also thanks to:

Center for International Forestry Research

UN-REDD Programme

World Soil Information
Machine learning as a framework for automated soil mapping
Methods

➔ 2D and 3D soil properties: **ensemble random forest and gradient boosting** (ranger, xgboost)

➔ soil types: **ensemble random forest and nnet::multinom**

➔ Cross-validation, post-processing, pseudo-observations
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, Repp, recursive partitioning, survival analysis.
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\textsuperscript{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\textsuperscript{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-
Building Predictive Models in \texttt{R} Using the \texttt{caret} Package

Max Kuhn  
Pfizer Global R&D

Abstract

The \texttt{caret} package, short for classification and regression training, contains numerous tools for developing predictive models using the rich set of models available in \texttt{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.

\textit{Keywords}: model building, tuning parameters, parallel processing, \texttt{R}, \texttt{NetWorkSpaces}. 
Results
They would have been interested in this…

The Russian School

Soil forming factors

Soil forming processes

Different Soils

Vasili Dokuchaev

FACTORS OF SOIL FORMATION (1941)

A System of Quantitative Pedology

Hans Jenny

World Soil Information
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, T00MOD3 is mean annual temperature, PRSMR3G3 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 T02MOD3 is mean monthly temperature for February.
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
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).
Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations

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

Abstract

Global gridded crop models (GGMs) 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 GGM outcomes has come under detailed scrutiny, unlike the influence of soil data. Here we compare yield variability caused by the soil type selected for GGM 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 GGM
Maps
USDA suborders -> in Europe!
WRB 2nd level -> in USA!
Predicted USDA Soil Taxonomy class (Twelfth Edition; 2014)

**Fibrists (26%)**
(TAXOUSDA)
Histosols that are primarily made up of only slightly decomposed organic materials, often called peat.

Hemists (14%) | Saprist (12%)

Predicted World Reference Base (2006) soil class

**Fibric Histosols (18%)**
(TAXNWRB)
Histosols = Soils consisting primarily of organic materials. They are defined as having 40 centimetres or more of organic soil material in the upper 80 centimetres. Having, after rubbing, two-thirds or more (by volume) of the organic material consisting of recognizable plant tissue within 100 cm of the soil surface (in Histosols only).

Haplic Acrisols (10%) | Hemic Histosols (10%)
Autumn tints line a road pointing to Mount Barrow. 

By Jim van Oord

Fly to the church's horizon

Add text:

Comment is

Pax@amio

Upload your photos
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...
God’s Word is like an iceberg—there is more truth unseen than seen
The moderate-resolution imaging spectroradiometer (MODIS)

Shuttle Radar topography missions (SRTMGL3)

Landsat 8 TIRS bands
  Sentinel-1,2
  (bands 1, 9, 10)

ALOS Global Digital Surface Model
  AW3D30

Landsat 8 MS bands
  Sentinel-1,2
  (bands 5, 6, 7, 8a, 11, 12)

SRTMGL1

WorldDEM

Resolution (metres)
250
100
50
30
10
1

2000 2010 2020
Towards 100 m, 30 m resolution...
Get ready for the Soil Data Revolution!