First Section

MACHINE LEARNING AS A FRAMEWORK FOR PREDICTIVE SOIL MAPPING: incorporating distances and spatial connectivity into machine learning-based modeling

Machine Learning (ML) and data science in general are of increasing interest in to soil modeling and mapping. ML methods such as random forest, neural nets, deep learning and similar are now regularly used to generate soil maps. But unlike with geostatistical techniques, such as various versions of kriging, spatial dependence structure in the data is often ignored in ML methods. If there is spatial structure in the cross-validation residuals of ML predictions, this indicates that the predictions are suboptimal and could be improved by taking spatial structure into account. To address this spatial autocorrelation issue, this workshop introduces the use of ML algorithms (e.g. random forest, gradient boosting and support vector machines, neural networks) in combination with geographical distances to sampling locations, as additional covariates, to fit models and predict soil properties. Our recent experience shows that this produces more accurate predictions. This workshop aims to introduce both basic and new concepts in using ML in R for predictive mapping of soil classes and numeric soil properties. The lecturers are experienced digital soil mappers, have developed a number of R packages and functions and have extensive backgrounds in Predictive Soil Mapping with R. Some of the lectures connected with this topic are available via: https://www.youtube.com/c/OpenGeoHubFoundation and https://www.youtube.com/c/IsricOrg

Specific tutorials of this workshop will cover:

- Brief introduction to R and to R packages commonly used for predictive soil mapping;
- Importing soil data into R; conversion and harmonization of soil data; soil data classes (sp, sf, rgdal, raster, aqp packages);
- Spatial overlay and derivation of simple and complex measures of proximity, spatial position and spatial context (raster, SAGA GIS);
- Using Machine Learning algorithms for predictive soil mapping (caret, ranger, randomForestSRC, xgboost, h2o, mlr and SuperLearner packages);
- Using R to produce web maps and visualisations (ggplot, mapview, shiny, leaflet packages);

How to get to the workshop: THRN 1307

Literature:

- "Predictive Soil Mapping with R" (https://envirometrix.github.io/PredictiveSoilMapping/)
- Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518 https://doi.org/10.7717/peerj.5518
- Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A. 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL, 4, 1-22, https://doi.org/10.5194/soil-4-1-2018
- "Geocomputation with R" (https://geocompr.robinlovelace.net/)
- "Introduction to Data Science: Data Analysis and Prediction Algorithms with R" (https://rafalab.github.io/dsbook/)

All participants of the workshop will receive a free copy of the "Predictive Soil Mapping with R" book.

Workshop programme:

Sunday, 2nd June 2019

- 8:30–9:00 Arrival and setting up of computers
- 9:00–10:30 Programme overview, software installation and first steps (Tom Hengl & Gerard Heuvelink)
- 10:30–11:00 Coffee break
- 11:00–12:30 R tutorial: Building spatial Machine Learning models (Tom Hengl)
- 12:30–13:30 Lunch break
- 13:30–15:30: R tutorial: spatial cross-validation and prediction error (Gerard Heuvelink)
- 15:30–16:00 Coffee break

- 16:00–17:00 R tutorial: Ensemble methods for predictive soil mapping using SuperLearner package (Tom Hengl)

Requirements

- **Basic knowledge of R, experience with fitting geostatistical models to generate soil maps;**
- **Laptop computer (preferably with Linux OS and/or Windows 7+ OS) with at least 4GB (8GM recommended) RAM and wifi;**
- **Pre-installed software following the installation instructions;**
- **Bringing your own data sets is highly recommended but not required.**

Registration and participation:

- Date: Sunday, 2nd June 2019
- Registration fee: **CAD 100**
- A minimum of 20 participants is required

Lecturers

	T. (Tom) Hengl is a senior researcher at OpenGeoHub Foundation / EnvirometriX Ltd with core speciality in big data analytics and automated soil mapping based on Machine Learning. Tom has a background in soil mapping and geo-information science. He continuously runs hands-on-R training courses to promote use of Open Source software for spatial analysis / spatial modeling purposes. He is currently the project leader for LandGIS — system for automated global soil and vegetation mapping at fine spatial resolutions (100 m, 250 m to 1 km) and which aspires to be recognized as an "OpenStreetMap-type" system for environmental data. ORCID ID: http://orcid.org/0000-0002-9921-5129
	Gerard B.M. Heuvelink is professor in Pedometrics and Digital Soil Mapping at Wageningen University and a senior researcher with ISRIC — World Soil Information. He holds an MSc in Applied Mathematics and a PhD in Environmental Sciences. His PhD research on error propagation in spatial environmental modelling marks the start of a scientific career in spatial uncertainty analysis, with applications in the soil domain as a main focus. Gerard has also contributed substantially to the development of pedometrics. He chaired the Pedometrics Commission of the IUSS from 2003 to 2006. He has published over 125 articles on geostatistics, spatial uncertainty analysis and pedometrics in peer-reviewed international scientific journals and is the Richard Webster medal 2014 receiver.