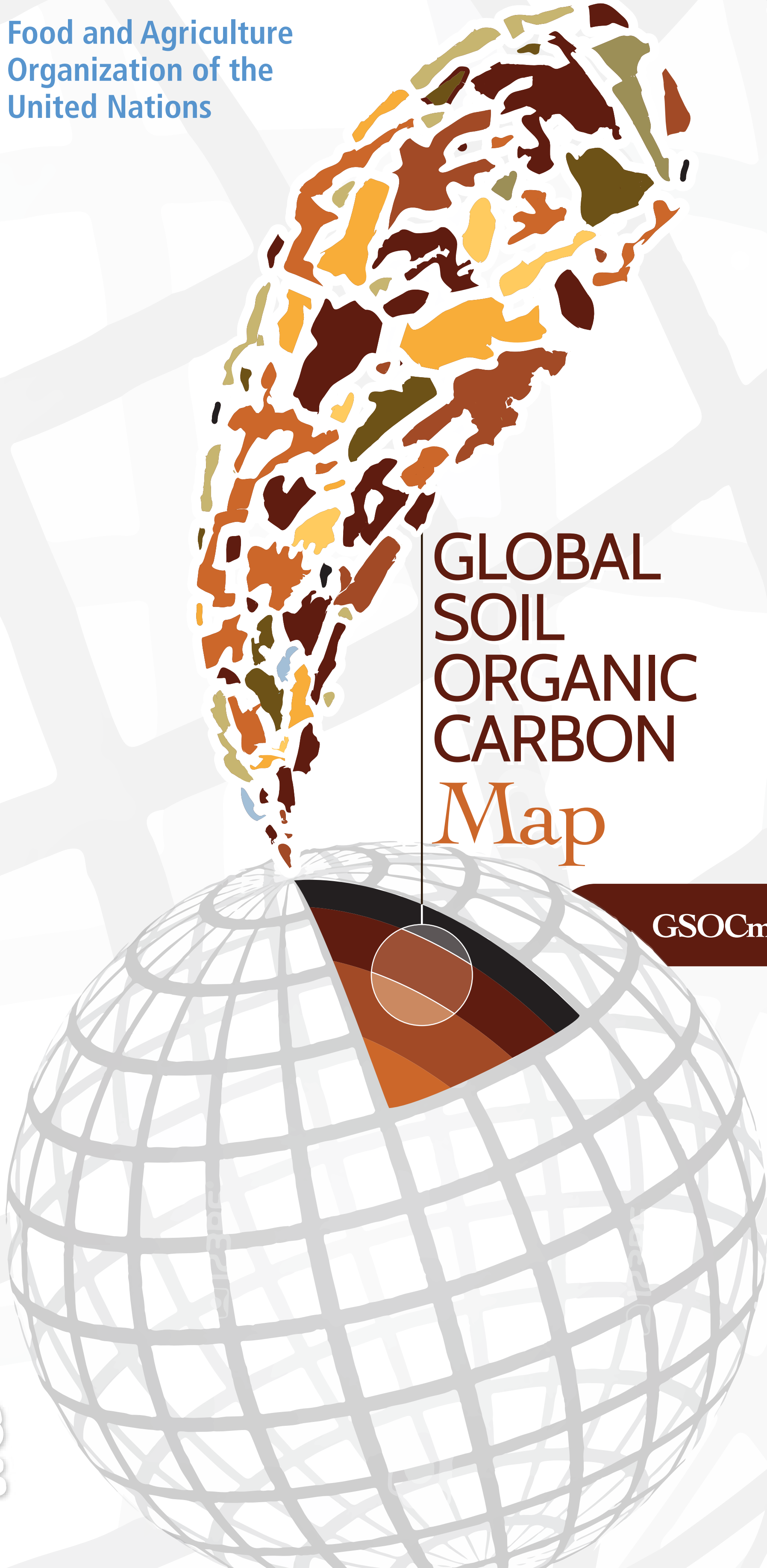
Food and Agriculture Organization of the United Nations

GLOBAL SOIL ORGANIC CARBON Map

TECHNICAL REPORT

GSOCmap v.1.6



itps
INTERGOVERNMENTAL TECHNICAL PANEL ON SOILS

GLOBAL SOIL PARTNERSHIP

Country Guidelines on Digital Soil Mapping

Food and Agriculture Organization of the United Nations
Rome, 2022

Required citation:
FAO. 2022. *Global Soil Organic Carbon Map* – GSOCmap v.1.6. *Technical report.* Rome. https://doi.org/10.4060/cb9015en

ISBN 978-92-5-135899-3

Cover design: ©FAO/Matteo Sala

Contents

Figures

Tables

Foreword

Abbreviations and acronyms

BD Bulk Density

CO_2 Carbon dioxide

CRF Coarse fragments

DM Dry matter

DSM Digital soil mapping

GAUL Global Administrative Unit Layers

GHG Greenhouse gas

GSOCmap Global Soil Organic Carbon Map

GSOCseq Global Soil Organic Carbon Sequestration Potential Map

GSP Global Soil Partnership

HWSD Harmonized World Soil Database

ISCN International Soil Carbon Network

INSII International Network of Soil Information Institutions

IPBES Intergovernmental Platform on Biodiversity and Ecosystem Services

IPCC Intergovernmental Panel on Climate Change

IPR Intellectual Property Rights

ITPS Intergovernmental Technical Panel on Soils

LDN Land Degradation Neutrality

NDVI Normalized difference in vegetation index

NPP Net Primary Production

P4WG Pillar 4 Working Group

QA/QC Quality Assurance/Quality Check

RMSE Root mean square error

SDF Soil Data Facility

SDG Sustainable Development Goals

SISLAC Latin America and the Caribbean's Soil Information System

SOC Soil organic carbon

SOM Soil organic matter

SPADE/M Soil Profile Analytical Database of Europe of Measured Parameters

SWRS Status of World's Soil Resources

UNCCD United Nations Convention to Combat Desertification

WFS Web Feature Service

WoSIS World Soil Information Service

Contributors

Prepared by:
Global Soil Partnership Secretariat
Ronald Vargas

Second Intergovernmental Technical Panel on Soils
Luca Montanarella - European Commission, Joint Research Centre (*Chair*); Saéb AbdelHaleem Khresat

Third Intergovernmental Technical Panel on Soils
Rosa Poch - Spain (*Chair*); Nsalambi V. Nkongolo - Democratic Republic of the Congo;

Chapter 1

Presentation

1.1 Background and objectives

To date, a total number of around 2.3 billion people are affected by moderate and severe food insecurity (FAO et al., 2022). In 2020, within the first year of the COVID-19 pandemic, an additional 320 million people became affected by food insecurity (FAO et al., 2021). The current conflicts and aggravating climate change further jeopardise achieving sustainable development goal (SDG) 2 (Zero Hunger) by 2030. The situation is alarming and urgent action is needed to revert the trends and increase food security.

The current global situation requires an increase of food production while preserving natural (soil) resources, lowering greenhouse gas emissions and optimising the use of goods such as fertilisers on agricultural sites (Eisenstein, 2020). Fertiliser prices more than doubled within one year and grain prices increased by around 25 percent (Jan. 2021 - Jan. 2022) (Hebebrand and Laborde, 2022). With the start of the armed conflict in Ukraine in February 2022, this trend became more pronounced.

Growing food insecurity and rapidly increasing fertiliser prices underscore the urgent need for informed decision-making and optimised soil nutrient management. However, a large data gap exists in regards to soil nutrient stocks and soil properties that govern nutrient availability. Therefore, FAO's Global Soil Partnership (GSP) has launched the Global Soil Nutrient and Nutrient Budget map

(GSNmap) initiative in an endeavour to provide harmonised and finely resolved soil nutrient data and information to stakeholders following a country-driven approach.

Up-to-date soil data on the status and spatial trends of soil nutrients and related soil attributes is key to guide policy-making to close yield gaps, and protect local natural resources. Therefore, locally-specific optimisation of soil nutrient and agricultural management are needed (Cunningham et al., 2013). The soil information collected in the GSNmap thereby serves as a cornerstone in delineating priority areas for action and thereby seizes the opportunity to reduce food insecurity, close yield gaps, and reduce environmental costs arising from mismanagement of soil nutrients and especially overfertilisation.

1.2 Global Soil Partnership

The Global Soil Partnership (GSP) was established in December 2012 as a mechanism to develop a strong interactive partnership and to enhance collaboration and generate synergies between all stakeholders to raise awareness and protect the world's soil resources. From land users to policymakers, one of the main objectives of GSP is to improve governance and promote sustainable management of soils. Since its creation, GSP has become an important partnership platform where global soil issues are discussed and addressed by multiple stakeholders at different levels.

The mandate of GSP is to improve governance of the planet's limited soil resources in order to guarantee productive agricultural soils for a food-secure world. In addition, it supports other essential soil ecosystem services in accordance with the sovereign right of each Member State over its natural resources. In order to achieve its mandate, GSP addresses six thematic action areas to be implemented in collaboration with its regional soil partnerships (Figure 1).

The area of work on Soil Information and Data (SID) of the GSP builds an enduring and authoritative global system (GloSIS) to monitor and forecast the condition of the Earth's soil resources and produce map products at the global level. The secretariat is working with the international network of soil data providers (INSII - International Network of Soil Information Institutions) to implement data related activities.

1.3 Country-driven approach and tasks

The GSNmap initiative will be jointly implemented by the International Network of Soil Information Institutions (INSII) and the GSP Secretariat. The process will be country-driven, involving and supporting all Member States in developing their national GSNmap data products. The GSNmap products will be developed following a two phase approach:

- Phase I: development of soil nutrient and associated soil property maps;
- Phase II: quantification, analysis, projections of nutrient budgets for agricultural land use systems at national, regional and global scale.

These guidelines only concern GSNmap Phase I, while the guidelines for the GSNmap Phase II will be published in the fourth quarter of 2022. Depending on national data availability and technical capacities, ad-hoc solutions will be developed by the GSNmap WG to support countries during the national GSNmap production and/or harmonisation phase. Where possible, GSP Secretariat will use publicly available data to gap-fill the areas which are not covered by the national submissions unless the country requests to be left blank on the GSNmap products.

Chapter 2

Setting-up the software environment

Y. Yigini

This cookbook focuses on SOC modeling using open source digital mapping tools. The instructions in this chapter will guide the user through installing and manually configuring the software to be used for DSM procedures for Microsoft Windows desktop platform. Instructions for other platforms (e.g. Linux Flavours, MacOS) can be found through free online resources.

2.1 Use of R, RStudio and R Packages

R is a language and environment for statistical computing. It provides a wide variety of statistical (e.g. linear modeling, statistical tests, time-series, classification, clustering, etc.) and graphical methods, and is highly extensible.

2.1.1 Obtaining and installing R

Installation files and instructions can be downloaded from the Comprehensive R Archive Network (CRAN).

1. Go to the following link https://cran.r-project.org/ to download and install **R**.
2. Pick an installation file for your platform.

2.1.2 Obtaining and installing RStudio

Beginners will find it very hard to start using **R** because it has no Graphical User Interface (GUI). There are some GUIs which offer some of the functionality of **R**. **RStudio** makes **R** easier to use. It includes a code editor, debugging and visualization tools. Similar steps need to be followed to install **RStudio**.

1. Go to https://www.rstudio.com/products/rstudio/download/ to download and install **RStudio**'s open source edition.
2. On the download page, *RStudio Desktop, Open Source License* option should be selected.
3. Pick an installation file for your platform.

2.1.3 Getting started with R

- **R** manuals: http://cran.r-project.org/manuals.html
- Contributed documentation: http://cran.r-project.org/other-docs.html
- Quick-**R**: http://www.statmethods.net/index.html
- Stackoverflow **R** community: https://stackoverflow.com/questions/tagged/r

2.2 R packages

When you download **R**, you get the basic **R** system which implements the **R** language. **R** becomes more useful with the large collection of packages that extend the basic functionality of it. **R** packages are developed by the **R** community.

refer to: - tidyverse book (R for data science) - caret (cookbook) - https://rspatial.org/

2.2.1 Finding R packages

The primary source for **R** packages is CRAN's official website, where currently about 12,000 available packages are listed. For spatial applications, various packages are available. You can obtain information about the available packages directly on CRAN with the `available.packages()` function. The function returns a matrix of details corresponding to packages currently available at one or more repositories. An easier way to browse the list of packages is using the *Task Views* link, which groups together packages related to a given topic.

2.3 GEE - google earth engine

- general info

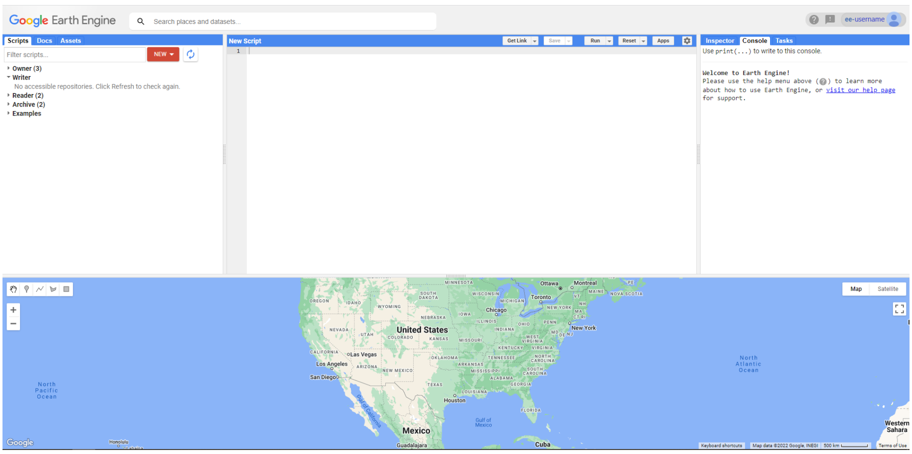


Figure 2.1: Google Earth Engine code editor.

- upload assets to GEE
- Explain how to import uploaded assets (?) ...

2.4 rgee - Extension to use google earth engine in R

The rgee package enables users to interact with the GEE servers using the R language. The package makes use of the Python language to interact with GEE. The package can be downloaded easily either directly from the GitHub repository or via CRAN.

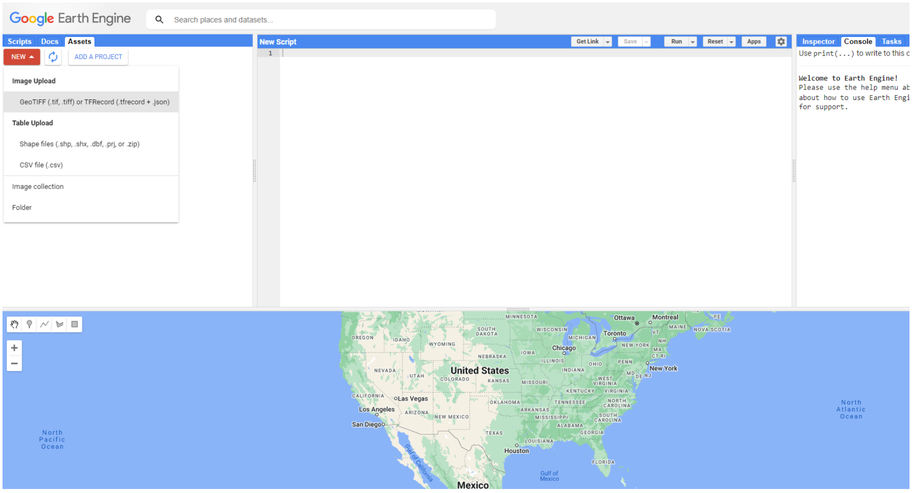


Figure 2.2: Select files and filetype to be uploaded as GEE assets.

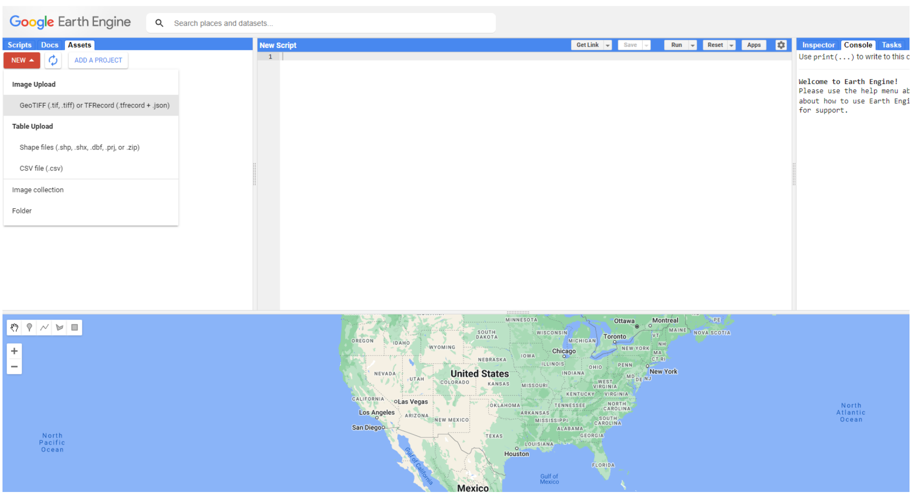


Figure 2.3: Upload interface.

```r
# Source: https://yabellini.github.io/curso_rgee/index.html
# Yanina Bellini Saibene

#install.packages('remotes')
# remotes::install_github("r-spatial/rgee")
```

To be able to interact with the GEE via Python, it is necessary to install certain R packages but also the so-called "Miniconda" commmand prompt which acts as Python interpreter mediating between R and GEE. The 'ee_install()' function automatically downloads and install all the software that is needed.

```r
# load rgee package and install dependencies
library(rgee)
```

```
## Registered S3 method overwritten by 'htmlwidgets':
##   method           from
##   print.htmlwidget tools:rstudio
```

```r
# ee_install() # installs miniconda
```

Once the dependencies are installed, it is necessary to initialize rgee by providing the user credentials of our GEE account. The ee_Initialize command must be run every time we want to use rgee.

```r
# Initialize Google Earth Engine! (you need to create a user account)
# ee_Initialize()



# Useful functions

#ee_check() # check the dependencies that do not belong to R
#ee_clean_credentials() # to remove the user credentials
#ee_clean_pyenv() # Delete variables of the system
```

Chapter 3

Introduction to Digital Soil Mapping of soil nutrients and associated soil attributes

Digital soil mapping (DSM) is a methodological framework to create soil attribute maps on the basis of the quantitative relationships between spatial soil databases and environmental covariates. The quantitative relations can be modelled by different statistical approaches, most of them considered machine learning techniques. Environmental covariates are spatially explicit proxies of soil-forming factors that are employed as predictors of the geographical distribution of soil properties. The methodology has evolved from the theories of soil genesis developed by Vasil Dokuchaev in his work the Russian Chernozems (1883), which later were formalised by Jenny (1941) with the equation of the soil-forming factors. The conceptual equation of soil-forming factors has been updated by McBratney, Santos and Minasny (2003) as follow:

$$S = f\left(s, c, o, r, p, a, n\right) \tag{3.1}$$

Where S is the soil classes or attributes (to be modelled) as a function of "s" as other soil properties, "c" as climatic properties, "o" as organisms, including land cover and human activity, "r" as terrain attributes, "p" as parent material, "a" as soil age, and "n" as the geographic position.

Digital soil mapping has been used to produce maps of soil nutrients. For instance, Hengl *et al.* (2017) predicted 15 soil nutrients at a 250 m resolution in Africa, using a random forest model (Wright and Ziegler, 2016), topsoil nutrient observations at point locations and a set of spatially-explicit environmental covariates. In 2021, Hengl et al. applied the same modelling approach to estimate total phosphorus in semi-natural soils at the global scale, as well.

In this technical manual, we present a DSM frameworks to map soil properties, including soil nutrients. One approach for soil observations with latitude and longitude data (point-support) (Figure 3.1).

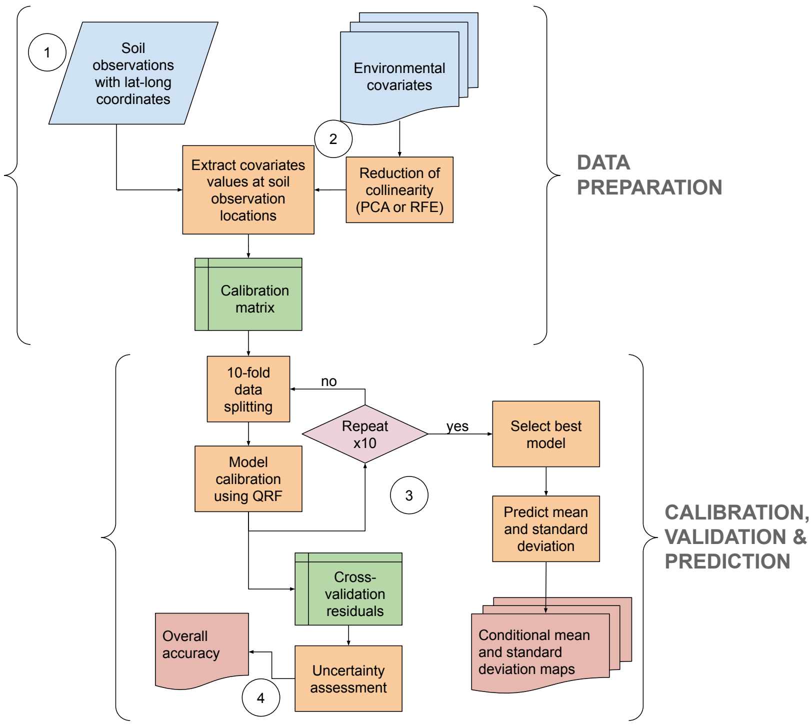


Figure 3.1: Digital soil mapping approach for point-support data. Circles are the steps.

Chapter 4

Step 1: soil data preparation

Soil data consist of measurement at a specific geographical location, time and soil depth. Therefore, it is necessary to arrange the data following the format shown in Table 4.1.

Table 4.1: Format example of a soil dataset

Profile ID	Horizon ID	Lat	Long	Year	Top	Bottom	Soil property	Value	Lab method
1	1_1	12.123456	1.123456	2018	0	20	SOC	3.4	W&B
1	1_2	12.123456	1.123456	2018	20	40	SOC	2.1	W&B
2	2_1	23.123456	2.123456	2019	0	30	SOC	2.9	W&B

Profile ID = unique profile identifier; Horizon ID = unique layer identifier; Lat = latitude in decimal degrees; Long = longitude in decimal degrees; Year = sampling year; Top = upper limit of the layer in cm; Bottom = lower limit of the layer in cm; Soil property = name of the soil property; Value = numerical value of the measure; Lab method = name of the laboratory protocol used for measuring the soil property.

Soil data usually require a preprocessing step to solve common issues such as, arranging the data format, fixing soil horizon depth consistency, detecting unusual soil property measurements, among others issues. Once the original dataset is clean and consistent, data harmonisation is needed to produce synthetic horizons (such as 0-30 cm layer), as well as to make compatible measurements from different lab methods. Horizon harmonisation will be done with the mass preserving spline function (Bishop, McBratney and Laslett, 1999; Malone *et al.*, 2009) fitted to each individual soil profile, which requires more than a layer per profile. In the cases of single-layer samples, which is common in sampling for nutrient determination, a pedotransfer function locally calibrated should be applied. Pedotransfer functions will be also required to harmonise the laboratory methods. Experts from GLOSOLAN will provide advice in this regard.

Chapter 5

Step 2: environmental covariates

5.1 Environmental covariates

The SCORPAN equation (Eq. 3.1) refers to the soil-forming factors that determine the spatial variation of soils. However, these factors cannot be measured directly. Instead, proxies of these soil forming factors are used. One essential characteristic of the environmental covariates is that they are spatially explicit, covering the whole study area. Table 2 shows a summary of the environmental covariates that can be implemented under the DSM framework.

Table 5.1: Environmental covariates

Factor	Description	Code	Resolution
Temperature	Mean air temperature (annual)	bio1	1000
	Mean daily temperature of warmest month	bio5	1000
	Mean daily temperature of coldest month	bio6	1000
Precipitation	Total precipitation (annual)	bio12	1000

	Mean precipitation of wettest month	bio13	1000
	Mean precipitation of driest month	bio14	1000
	Mean monthly precipitation of wettest quarter	bio16	1000
	Mean monthly precipitation of driest quarter	bio17	1000
Evapotranspiration	Mean monthly PET	pet_penman_mean	1000
	Minimum monthly PET	pet_penman_min	1000
	Range monthly PET	pet_penman_range	1000
	Maximum monthly PET	pet_penman_max	1000
Wind	Minimum monthly wind speed	sfcWind_min	1000
	Maximum monthly wind speed	sfcWind_max	1000
	Range monthly wind speed	sfcWind_range	1000
Growing season	Number of days with mean daily air temperature 10 °C	ngd10	1000
Vegetation Indices	NDVI (MOD13Q1), mean March-May from 2000-2022	ndvi_030405_mean	250
	NDVI (MOD13Q1), mean June-August from 2000-2022	ndvi_060708_mean	250
	NDVI (MOD13Q1), mean September-November from 2000-2022	ndvi_091011_mean	250
	NDVI (MOD13Q1), mean December-February from 2000-2022	ndvi_120102_mean	250
	NDVI (MOD13Q1), standard deviation March-May from 2000-2022	ndvi_030405_sd	250
	NDVI (MOD13Q1), standard deviation June-August from 2000-2022	ndvi_060708_sd	250
	NDVI (MOD13Q1), standard deviation September-November from 2000-2022	ndvi_091011_sd	250
	NDVI (MOD13Q1), standard deviation December-February from 2000-2022	ndvi_120102_sd	250
FPAR	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), mean March-May from 2000-2022	fpar_030405_mean	500

	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), mean June-August from 2000-2022	fpar_060708_mean	500
	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), mean September-November from 2000-2022	fpar_091011_mean	500
	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), mean December-February from 2000-2022	fpar_120102_mean	500
	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), standard deviation March-May from 2000-2022	fpar_030405_sd	500
	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), standard deviation June-August from 2000-2022	fpar_060708_sd	500
	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), standard deviation September-November from 2000-2022	fpar_091011_sd	500
	Fraction of photosynthetically active radiation (FPAR) (MOD15A2H), standard deviation December-February from 2000-2022	fpar_120102_sd	500
LST	Land Surface Temperature Day (MOD11A2), mean March-May from 2000-2022	lstd_030405_mean	1000
	Land Surface Temperature Day (MOD11A2), mean June-August from 2000-2022	lstd_060708_mean	1000
	Land Surface Temperature Day (MOD11A2), mean September-November from 2000-2022	lstd_091011_mean	1000
	Land Surface Temperature Day (MOD11A2), mean December-February from 2000-2022	lstd_120102_mean	1000

	Land Surface Temperature Day (MOD11A2), standard deviation March-May from 2000-2022	lstd_030405_sd	1000
	Land Surface Temperature Day (MOD11A2), standard deviation June-August from 2000-2022	lstd_060708_sd	1000
	Land Surface Temperature Day (MOD11A2), standard deviation September-November from 2000-2022	lstd_091011_sd	1000
	Land Surface Temperature Day (MOD11A2), standard deviation December-February from 2000-2022	lstd_120102_sd	1000
NDLST	Normalised Difference between LST Day and LST Night (MOD11A2), mean March-May from 2000-2022	ndlst_030405_mean	1000
	Normalised Difference between LST Day and LST Night (MOD11A2), mean June-August from 2000-2022	ndlst_060708_mean	1000
	Normalised Difference between LST Day and LST Night (MOD11A2), mean September-November from 2000-2022	ndlst_091011_mean	1000
	Normalised Difference between LST Day and LST Night (MOD11A2), mean December-February from 2000-2022	ndlst_120102_mean	1000
	Normalised Difference between LST Day and LST Night (MOD11A2), standard deviation March-May from 2000-2022	ndlst_030405_sd	1000
	Normalised Difference between LST Day and LST Night (MOD11A2), standard deviation June-August from 2000-2022	ndlst_060708_sd	1000
	Normalised Difference between LST Day and LST Night (MOD11A2), standard deviation September-November from 2000-2022	ndlst_091011_sd	1000

	Normalised Difference between LST Day and LST Night (MOD11A2), standard deviation December-February from 2000-2022	ndlst_120102_sd	1000
SWIR	Black-sky albedo for shortwave broadband (MCD43A3), mean June-August from 2000-2022	swir_060708_mean	500
Snow cover	MODIS Snow Cover (MOD10A1) mean	snow_cover	500
Land cover	Dynamic World 10m near-real-time (NRT) Land Use/Land Cover (LULC) dataset. Mean estimated probability of complete coverage by trees	trees	250
	Dynamic World 10m near-real-time (NRT) Land Use/Land Cover (LULC) dataset. Mean estimated probability ofcomplete coverage by shrub and scrub	shrub_and_scrub	250
	Dynamic World 10m near-real-time (NRT) Land Use/Land Cover (LULC) dataset. Mean estimated probability of complete coverage by flooded vegetation	flooded_vegetation	250
	Dynamic World 10m near-real-time (NRT) Land Use/Land Cover (LULC) dataset. Mean estimated probability of complete coverage by grass	grass	250
	Dynamic World 10m near-real-time (NRT) Land Use/Land Cover (LULC) dataset. Mean estimated probability of complete coverage by bare	crop	250
Terrain	Profile curvature	curvature	250
	Downslope curvature	downslopecurvature	250
	Uplslope curvature	upslopecurvature	250
	Deviation from Mean Value	dvm	250
	Deviation from Mean Value	dvm2	250
	Elevation	elevation	250
	Melton Ruggedness Number	mrn	250

Negative openness	neg-openness	250
Possitive openness	por-openness	250
Slope	slope	250
Topographic position index	tpi	250
Terrain wetness index	twi	250
Multirresolution of valley bottom flatness	vbf	250

Apart from the environmental covariates mentioned in Table 5.1, other types of maps could also be included, such as Global Surface Water Mapping Layers and Water Soil Erosion from the Joint Research Centre (JRC). At national level there may be very significant covariates that could complement or replace the covariates of Table 5.1.

Since environmental covariates are available at different resolutions and coordinate reference systems (CRS), they have to be harmonised at a common resolution and CRS. The target resolution in GSNmap is 250 m x 250 m, therefore, all covariates were aggregated (from higher to lower resolution) or disaggregated (from lower to higher resolution) to 250 m. This process involved a raster resampling method, which is usually implemented by a bilinear approach for continuous covariates, and by the nearest-neighbour approach for categorical covariates (not included in the current list).

Note that the target resolution of GSNmap has been set at 250 m, which can be considered a moderate resolution for a global layer. However, those countries that require a higher resolution are free to develop higher resolution maps and aggregate the resulting maps to the target resolution of GSNmap for submission.

5.2 Reducing collinearity in environmental covariates

Multicollinearity is usually present in remote sensing data and terrain attributes. While this was an issue for multiple linear regression models, current models such as random forest can deal with high dimensionality. However, the main reasons to reduce the number of environmental covariates are that a model with fewer predictors can be interpreted more easily, thus extracting new knowledge, redundant information increasing the computational demand, and improve prediction results (Behrens *et al.*, 2014).

Covariate selection can be done by supervised or unsupervised methods (Behrens *et al.*, 2010). Supervised methods work on the basis of prediction results, hence they are based on a given dataset. For instance, recursive feature elimination (RFE) in caret R package (Kuhn, 2022) provides a tool for selecting covariates according to their predicting contribution. Instead, unsupervised methods are used to reduce the dimensionality of the dataset by removing redundant information without taking into account a particular target variable. Principal component analysis is one of the most widely used for this purpose, however, it does not ensure that specific discriminant features are kept within the main factors (Behrens *et al.*, 2014). Another drawback of this technique is that model interpretation can be reduced when using factors instead of the original covariates.

5.3 Merging soil data and environmental covariates

A calibration dataset consists of soil observations and a matrix of predictors, where each row is a soil observation paired with the values of the corresponding covariates for the given spatial location. Some common issues and solution when merging soil observations and covariates are:

- Mismatch of coordinate reference system (CRS): it requires to convert the CRS of point data to the raster or polygon covariate CRS.
- Categorical covariates: some covariates may be categorical, such as land use/cover, legacy soil maps or geological maps. A common problem in this case is that some classes may not be sampled with any soil observation, causing an error when using the layer for prediction, since the model cannot predict over a class that was not part of the model calibration step. Also, because of the cross-validation procedure, it is advised to have, at least, three soil samples per class for the same reason.

Chapter 6

Step 3: Mapping continuous soil properties

6.1 Setting up repeated k-fold cross validation

Cross validation is one of the most used methods in DSM for assessing the overall accuracy of the resulting maps (Step 8, Figure 3). Since this is implemented along with the model calibration step, we explain the process at this stage.

Cross validation consists of randomly splitting the input data into a training set and a testing set. However, a unique testing dataset can bias the overall accuracy. Therefore, k-fold cross validation randomly splits the data into k parts, using 1/k part of it for testing and k-1/k part for training the model. In order to make the final model more robust in terms of parameter estimations, we include repetitions of this process. The final approach is called repeated k-fold cross-validation, where k will be equal to ten in this process. A graphical representation of the 10-fold cross validation is shown in Figure 6.1. Note that green balls represent the samples belonging to the testing set and yellow balls are samples of the training set. Each row is a splitting step of the 10-folds, while each block (repetitions) represent the repetition step.

Step 5 in Figure 3 represents the repeated cross-validation, but note that after each single splitting step (the rows in Figure 4) the training data go to model

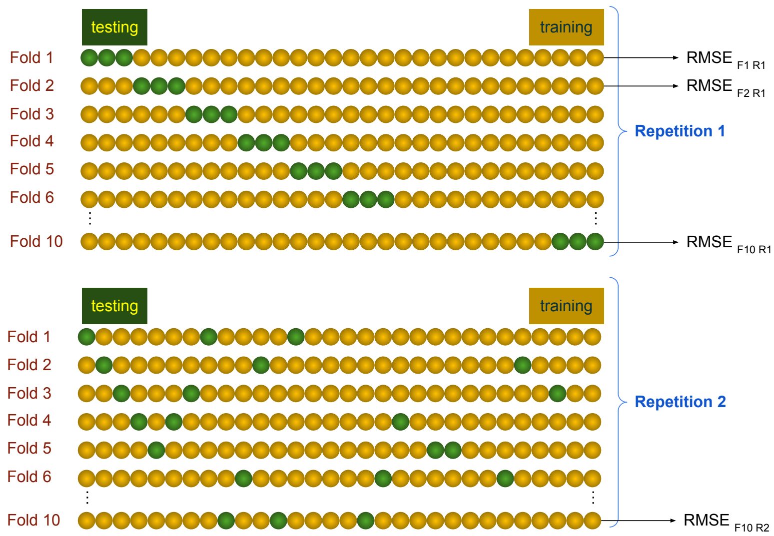


Figure 6.1: Schematic representation of the repeated cross-validation process.

calibration, which will be explained in Step 6 (next Section), and the testing data will be used with the calibrated model to produce the residuals (Step 8, Section 2.2.8). Repeated cross validation has been nicely implemented in the caret R package (Kuhn, 2022), along with several calibration methods.

6.2 Model calibration

The model calibration step involves the use of a statistical model to find the relations between soil observations and environmental covariates. One of the most widely used models in DSM is random forest (Breiman, 2001). Random forest is considered a machine learning method which belongs to the decision-tree type of model. Random forest creates an ensemble of trees using a random selection of covariate. The prediction of a single tree is made based on the observed samples mean in the leaf. The random forest prediction is made by taking the average of the predictions of the single trees. The size of the number of covariates at each tree (mtry) can be fine-tuned before calibrating the model.

Quantile regression forests (QRF, Meinshausen (2006)) are a generalisation of the random forest models, capable of not only predicting the conditional mean, but also the conditional probability density function. This feature allows one to estimate the standard deviation of the prediction, as well as the likelihood of the target variable falling below a given threshold. In a context where a minimum level of a soil nutrient concentration may be decisive for improving the crop yield, this feature can play an important role for the GSNmap initiative.

Model calibration will be implemented using the caret package (Kuhn, 2022). While we suggest to use QRF, caret provides a large set of models https://topepo.github.io/caret/available- models.html#) that might perform better in specific cases. In this regard, it is up to the user to implement a different model, ensuring the product specifications (Section Product Specifications).

6.3 Predicting soil attributes

After calibrating the model, caret will select the best set of parameters and will fit the model using the whole dataset. Then, the final model can be used to predict the target soil properties. The process uses the model and the values of the covariates at target locations. This is generally done by using the same input covariates as a multilayer raster format, ensuring that the names of the layers are the same as the covariates in the calibration dataset. In this step we will predict the conditional mean and conditional standard deviation at each raster cell.

Chapter 7

Step 4: uncertainty assessment

7.1 Introduction

Accuracy assessment is an essential step in digital soil mapping. One aspect of the accuracy assessment has been done in Step 7 by predicting the standard deviation of the prediction, which shows the spatial pattern of the uncertainty. Another aspect of the uncertainty is the estimation of the overall accuracy to measure the model performance. This will be measured using the model residuals generated by caret during the repeated cross validation step.

The residuals produced by caret consist of tabular data with observed and predicted values of the target soil property. They can be used to estimate different accuracy statistics. Wadoux, Walvoort and Brus (2022) have reviewed and evaluated many of them. While they concluded that there is not a single accuracy statistic that can explain all aspect of map quality, they recommended the following: mean prediction error (ME), that estimates the prediction bias; mean absolute prediction error (MAE) and root mean squared prediction error (RMSE) to estimate the magnitude of the errors; and model efficiency coefficient (MEC) (Janssen and Heuberger, 1995) as an estimator of the proportion of variance explained by the model.

While solar diagrams (Wadoux, Walvoort and Brus, 2022) are desired, we propose to produce a scatterplot of the observed vs predicted values maintaining the same range and scale for the X and Y axes.

Finally, note that accuracy assessment has been discussed in Wadoux *et al.* (2021), since the spatial distribution of soil samples might constrain the validity of the accuracy statistics. This is especially true in cases where the spatial distribution of observations is clustered. The authors recommended creating a kriging map of residuals before using them for assessing the map quality.

Chapter 8

Reporting results

Chapter 9

Compendium of R scripts

References

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X. & Scholten, T. 2014. Hyper-scale digital soil mapping and soil formation analysis, 213: 578–588. https://doi.org/10.1016/j.geoderma.2013.07.031

Behrens, T., Zhu, A.-X., Schmidt, K. & Scholten, T. 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping, 155: 175–185. https://doi.org/10.1016/j.geoderma.2009.07.010

Bishop, T.F.A., McBratney, A.B. & Laslett, G.M. 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines, 91: 27–45. https://doi.org/10.1016/s0016-7061(99)00003-8

Breiman, L. 2001. Random forests. *Machine Learning*, 45(1): 5–32. https://doi.org/10.1023/A:1010933404324

Hengl, T., Leenaars, J.G.B., Shepherd, K.D., Walsh, M.G., Heuvelink, G.B.M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I. & Kwabena, N.A. 2017. Soil nutrient maps of sub-saharan africa: Assessment of soil nutrient content at 250 m spatial resolution using machine learning, 109: 77–102. https://doi.org/10.1007/s10705-017-9870-x

Janssen, P.H.M. & Heuberger, P.S.C. 1995. Calibration of process-oriented models, 83: 55–66. https://doi.org/10.1016/0304-3800(95)00084-9

Kuhn, M. 2022. *Caret: Classification and regression training.* (also available at https://CRAN.R-project.org/package=caret).

Malone, B.P., McBratney, A.B., Minasny, B. & Laslett, G.M. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity, 154: 138–152. https://doi.org/10.1016/j.geoderma.2009.10.007

McBratney, A.B., Santos, M.L.M. & Minasny, B. 2003. On digital soil mapping, 117: 3–52. https://doi.org/10.1016/s0016-7061(03)00223-4

Meinshausen, Ni. 2006. Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Wadoux, A.M.J.-C., Heuvelink, G.B.M., Bruin, S. de & Brus, D.J. 2021. Spatial cross-validation is not the right way to evaluate map accuracy, 457: 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692
Wadoux, A.M.J.-C., Walvoort, D.J.J. & Brus, D.J. 2022. An integrated approach for the evaluation of quantitative soil maps through taylor and solar diagrams, 405: 115332. https://doi.org/10.1016/j.geoderma.2021.115332

Thanks to the financial support of

European Commission

Ministry of Economic Affairs of the Netherlands

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Ministry of Finance of the Russian Federation

ifa
International Fertilizer Industry Association

ISBN 978-92-5-135899-3

9 789251 358993

CB9015EN/1/03.22