SoilFER
Soil mapping
for resilient agrifood systems

Global Soil Nutrient and Nutrient Budgets maps (GSNmap) Phase I Technical Manual

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

The designations employed and the presentation of material in this information product do not imply the expression of any opinion whatsoever on the part of the Food and Agriculture Organization of the United Nations (FAO) concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The mention of specific companies or products of manufacturers, whether or not these have been patented, does not imply that these have been endorsed or recommended by FAO in preference to others of a similar nature that are not mentioned.

The views expressed in this information product are those of the author(s) and do not necessarily reflect the views or policies of FAO.

ISBN 978-92-5-135899-3
© FAO, 2022

Some rights reserved. This work is made available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; https://creativecommons.org/licenses/by-nc-sa/3.0/igo/legalcode).

Under the terms of this licence, this work may be copied, redistributed and adapted for non-commercial purposes, provided that the work is appropriately cited. In any use of this work, there should be no suggestion that FAO endorses any specific organization, products or services. The use of the FAO logo is not permitted. If the work is adapted, then it must be licensed under the same or equivalent Creative Commons licence. If a translation of this work is created, it must include the following disclaimer along with the required citation: "This translation was not created by the Food and Agriculture Organization of the United Nations (FAO). FAO is not responsible for the content or accuracy of this translation. The original [Language] edition shall be the authoritative edition."

Disputes arising under the licence that cannot be settled amicably will be resolved by mediation and arbitration as described in Article 8 of the licence except as otherwise provided herein. The applicable mediation rules will be the mediation rules of the World Intellectual Property Organization http://www.wipo.int/amc/en/mediation/rules and any arbitration will be conducted in accordance with the Arbitration Rules of the United Nations Commission on International Trade Law (UNCITRAL).

Third-party materials. Users wishing to reuse material from this work that is attributed to a third party, such as tables, figures or images, are responsible for determining whether permission is needed for that reuse and for obtaining permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

Sales, rights and licensing. FAO information products are available on the FAO website (www.fao.org/publications) and can be purchased through publications-sales@fao.org. Requests for commercial use should be submitted via: www.fao.org/contact-us/licence-request. Queries regarding rights and licensing should be submitted to: copyright@fao.org.

Cover design: ©FAO/Matteo Sala

Chapter 1

(PDF disabled)

Placeholder

Presentation and basics

Placeholder

What this Tutorial is not

The Training Dataset

Data Structure

How to Use This Manual

Manual Structure

Module 1: Introduction to R and Preparation of Soil Data for Digital Soil Mapping

Module 2: Soil Sampling Design

Module 3: GloSIS Data Preparation

Module 4: Digital Soil Mapping

Module 5: Soil Spectroscopy

Module 6: Soil Data Sharing

Required downloads

Step-by-step setup instructions

Step 1: Download and extract the GitHub repository

Step 2: Download the Google Drive data

Step 3: Inspect the project folder structure

Step 4: Place Google Drive files into the correct module folders

1.1 What is R?

1.1.1 Why R for Soil Science?

1.2 Installing R and RStudio

1.2.1 Installing R

1.2.2 Installing RStudio

1.2.3 Verifying Installation

1.3 Understanding the RStudio Interface

1.3.1 1. Console Pane (Bottom Left)

1.3.2 2. Code Editor Pane (Top Left)

1.3.3 3. Environment/History Pane (Top Right)

1.3.4 4. Files/Plots/Packages/Help Pane (Bottom Right)

1.4 R Packages: Extending R's Capabilities

1.4.1 What are Packages?

1.4.2 Installing Packages

1.4.2.1 Install Packages from GitHub or Other Sources

1.4.2.2 Manual Installation of R Packages

1.4.3 Loading Packages

1.4.4 Finding Help on Packages

1.5 R Basics: Objects and Data Types

1.5.1 Creating Objects with Assignment

1.5.2 Data Types in R

1.5.2.1 1. Numeric (Numbers)

1.5.2.2 2. Integer (Whole Numbers)

1.5.2.3 3. Character (Text Strings)

1.5.2.4 4. Logical (TRUE/FALSE)

1.5.2.5 5. Dates

1.5.3 Checking Data Types

1.6 Data Structures in R

1.6.1 Vectors

1.6.1.1 Creating Vectors

1.6.1.2 Vector Operations

1.6.1.3 Accessing Vector Elements

1.6.2 Factors

1.6.3 Matrices

1.6.4 Data Frames

1.6.4.1 Creating Data Frames

1.6.4.2 Accessing Data Frame Elements

1.6.4.3 Adding Columns

1.6.5 Lists

1.7 Operators in R

1.7.1 Arithmetic Operators

1.7.2 Comparison Operators

1.7.3 Logical Operators

1.8 Control Structures: Conditional Statements

1.8.1 If-Else Statements

1.8.2 Vectorized If-Else: `ifelse()`

1.8.3 Vectorized Cut: `cut()`

1.9 Control Structures: Loops

1.9.1 For Loops

1.9.2 While Loops

1.10 Functions in R

1.10.1 Using Built-in Functions

1.10.2 Creating Custom Functions

1.10.3 Function Arguments and Defaults

1.11 Data Manipulation with Base R

1.11.1 Subsetting and Filtering

1.11.2 Deleting Columns

1.11.3 Sorting Data

1.11.4 Aggregating Data

1.12 Data Manipulation with the Tidyverse

1.12.1 Loading the Tidyverse

1.12.2 Tibbles: Modern Data Frames

1.12.3 The Pipe Operator: %>%

1.12.4 Data Manipulation with {dplyr}

1.12.4.1 Selecting Columns

1.12.4.2 Filtering Rows

1.12.5 Renaming Columns with rename()

1.12.5.1 Creating/Modifying Columns with mutate()

1.12.5.2 Arranging (Sorting) Data

1.12.6 Grouping and Summarizing Data by columns

1.13 Combining Data Frames

1.13.1 Understanding Joins

1.13.1.1 Join types

1.13.1.2 Example datasets and join operations

1.13.2 Stacking Data with bind_rows()

1.14 Missing Data

1.14.1 Identifying Missing Data

1.14.2 Handling Missing Data

1.15 Data Reshaping with {tidyr} package: Pivoting

1.15.0.1 Wide vs Long Data Format

1.15.0.2 pivot_longer(): Wide to Long

1.15.0.3 pivot_wider(): Long to Wide

1.15.0.4 Advanced Pivoting: Multiple Measurements per Site (Replicates)

1.16 Handling Below Detection Limit Data

1.17 Working with Soil Data

1.17.1 Setting Working Directory

1.17.2 Load packages

1.17.3 Importing Data from Files

1.17.3.1 Reading CSV Files

1.17.3.2 Reading Excel Files

1.17.4 Exploring the Data

1.18 References

Chapter 2

Data preparation for Digital Soil Mapping in R

Placeholder

2.1 Introduction

2.2 Loading and exploring raw soil data

2.2.1 Basic data loading and structure

2.2.2 Data components and organization

2.3 Preparing site data for analysis

2.3.1 Adding a unique row identifier

2.3.2 Extracting and standardizing column names for sites

2.3.3 Creating unique profile identifiers

2.3.4 Remove exact duplicate site records

2.4 Coordinate validation and correction

2.4.1 Check 1: Missing coordinates

2.4.2 Check 2: Valid coordinate ranges

2.5 Soil depth validation and correction

2.5.1 Check 1: Missing depth boundaries

2.5.2 Check 2: Negative depth values

2.5.3 Check 3: Zero-thickness intervals

2.5.4 Check 4: Invalid depth logic

2.5.5 Check 5: Profiles without a surface horizon (top > 0)

2.6 Preparing lab data: harmonization and validation

2.6.1 Prepare and standardize laboratory columns

2.6.2 Check 1: Check each property against feasible analytical thresholds

2.6.2.1 Load property thresholds

2.6.2.2 Identify out-of-bounds values

2.6.2.3 Reporting out-of-bounds for properties and audit trail

2.6.3 Check 2: Texture validation

2.6.4 Check 3: Correction of out-of-bounds laboratory values

2.6.4.1 Option 1: Targeted corrections (when error mechanism is known)

2.6.4.2 Option 2: Replace out-of-bounds values with `NA`

2.7 Resolving duplicated data in soil profiles

2.7.1 Why duplicates occur

2.7.2 Types of duplicate situations

2.7.3 Check 1: Detect potential horizon duplicates within profiles

2.7.4 Resolving Profile Duplicates

2.7.4.1 Check 1: Average duplicated horizons (same depth intervals)

2.7.4.2 Check 2: Resolve ProfID for multiple depth sequences

2.7.5 Remove profiles not starting at the surface

2.7.5.1 Check 1: Remove profiles with incomplete surface coverage

2.7.6 Result

2.8 Harmonizing data for DSM

2.8.1 Depth standardization

2.8.2 Processing standardized data

2.8.3 Exporting standardized data for Digital Soil Mapping

2.9 Preparing data for spectroscopy analyses

2.10 Summary of exported files

References

2.11 Spatial Data Concepts

2.11.1 Vector vs Raster Data

2.11.2 Coordinate Reference Systems (CRS) and Projections

2.11.2.1 Geographic vs projected coordinate systems.

2.12 Spatial Packages in R: {sf} and {terra}

2.12.1 Installing and Loading Spatial Packages

2.12.2 Working with Vector Data using {sf}

2.12.3 Importing Vector Data

2.12.3.1 Creating spatial vector objects from data frames

2.12.3.2 Reading Shapefiles with st_read()

2.12.3.3 Inspecting and Setting CRS

2.12.3.4 Reprojecting Spatial Data

2.12.3.5 Inspecting and manipulating {sf} Object

2.12.3.6 Geometry Operations

2.12.3.6.1 Intersection and Union with Administrative Boundaries

2.12.4 Spatial Relationships and Joins

2.13 Working with Raster Data using {terra}

2.13.1 Importing Raster Data

2.14 Combining Vector and Raster Data

2.14.1 Extracting Raster Values to Points and Polygons

2.15 Exporting and Sharing Spatial Data

2.15.1 Exporting Vector Data

2.15.2 Exporting Raster Data

Chapter 3

Sampling design for soil surveys

Placeholder

3.1 Introduction

3.2 Sampling methodologies for soil spatial survey

3.2.1 Selection of the sampling methodology

3.2.2 Inference Methods

3.2.3 Purpose of Sampling

3.2.4 Sampling Design Types

3.2.4.1 Example: Soils4Africa Sampling Design

3.3 SoilFER Sampling Design

3.3.1 Soil Properties

3.3.2 Environmental Covariates

3.3.3 Understanding the Methodology

3.3.3.1 PSU Selection Process

3.3.3.2 SSU and TSU Selection Process

3.3.3.3 Site Identification System

3.4 Tutorial Using R

3.4.1 Setting Up the Environment

3.4.1.1 Install Required Libraries

3.4.2 Define Variables and Parameters

3.4.3 Country ISO Code

3.4.3.1 Land Use Type

3.4.3.2 File Paths

3.4.3.3 Coordinate Reference System

3.4.3.4 Sample Size Calculation

3.4.3.5 Sampling Unit Definitions

3.4.3.6 Algorithm Parameters

3.4.4 Custom Functions

3.4.4.1 Covariate Space Coverage Function (CSIS)

3.4.4.2 TSU Generation Function

3.4.5 Load Country Boundaries and Legacy Data

3.4.5.1 Visualize Boundaries and Legacy Data

3.4.6 Load Environmental Covariates for PSUs

3.4.6.1 Optional Auxiliary Data

3.4.6.2 Load and Process Auxiliary Data

3.4.7 Load Main Environmental Covariates

3.4.7.1 Load and Process Soil Climate Data (Newhall)

3.4.7.2 Convert Categorical Variables to Dummy Variables

3.4.7.3 Merge All Covariate Layers

3.4.7.4 Crop and Save Covariates

3.5 Principal Component Analysis (PCA)

3.6 Organize the Sampling Universe

3.6.1 Load and Process Land Use Data

3.6.2 Exclude Protected Areas

3.6.3 Restrict to Accessible Slopes

3.6.4 Aggregate to 100m Resolution

3.6.5 Save Processed Land Use

3.6.6 Filter Legacy Data

3.7 Create Primary Sampling Units (PSU Grid)

3.7.1 Generate 2km × 2km Grid

3.7.2 Trim to Country Boundaries

3.7.3 Save PSU Grid

3.8 Select PSUs with Sufficient Land Use Coverage

3.8.1 Calculate Land Use Percentage

3.8.2 Filter PSUs

3.9 Rasterize PSUs for Covariate Space Coverage

3.10 Compute Optimal Sample Size

3.10.1 Run Optimization

3.10.2 Extract Optimal Sample Size

3.11 Covariate Space Coverage - Computing PSUs

3.11.1 Prepare Function Parameters

3.11.2 Perform K-means Clustering

3.11.3 Assign Clusters and Calculate Distances

3.11.4 Extract Selected PSUs

3.11.5 Visualize in Covariate Space

3.12 Compute SSUs and TSUs

3.12.1 Load High-Resolution Covariates

3.12.2 Process Each PSU

3.12.3 Combine SSUs and TSUs

3.13 Export Sampling Units

3.13.1 Create Cluster Raster

3.13.2 Join Cluster Info to TSUs

3.13.3 Create Site IDs

3.13.4 Export Shapefiles

3.14 Compute Alternative PSUs

3.15 Assign Country-Wide Unique IDs

3.16 Export Unified Datasets

3.17 Summary Statistics

3.18 Create Distribution Maps

3.19 GPS Device Preparation

3.19.1 Export to GPX Format

3.19.2 Load onto GPS Device

3.20 Sampling Procedure

3.21 Data Recording

3.22 Quality Control

3.23 References

Appendix A: Software and Data Sources

3.23.1 Software Requirements

3.23.2 Data Sources

Appendix B: Troubleshooting Common Issues

3.23.3 Memory Issues

3.23.4 CRS Misalignment

3.23.5 Empty Geometry

Appendix C: Acronyms and Abbreviations

Chapter 4

TBD

Chapter 5

GloSIS Soil Data Preparation

Placeholder

5.1 Introduction

5.1.1 `Plot Data` Sheet

5.1.2 `Profile Data` Sheet

5.1.3 `Element Data` Sheet

5.1.4 `Specimen Data` Sheet

5.1.5 `Procedures` Sheet

5.1.6 `Metadata` Sheet

5.1.7 `Original Data` Sheet

5.2 GloSIS-ETL

5.2.1 GloSIS Data Harmonization

5.2.2 Step 1: Data Ingestion & Variable Mapping

5.2.2.1 Data Upload Interface

5.2.2.2 Parameter Mapping Configuration

5.2.3 Step 2: Quality Check and Quality Report Generation

5.2.4 Step 2a: Geographic Data Validation

5.2.4.1 Coordinate Reference System Requirements

5.2.5 Step 2b: Soil Depth Validation

5.2.6 Step 2c: Soil Properties Validation

5.2.7 Step 3: Data Correction and Resubmission

5.2.8 Step 4: Download Harmonized Data

5.2.9 GloSIS Data Standardization

5.2.10 GloSIS Data Viewer

5.3 Installation and Setup of the `GloSIS-ETL`

5.3.1 Install Docker Desktop

5.3.1.1 Windows

5.3.1.2 macOS

5.3.1.3 Linux

5.3.2 Install the GloSIS-ETL Services

5.3.3 Verify Installation

5.3.3.1 Using Docker Desktop

5.3.3.2 Using the Terminal

5.3.3.3 Stopping the Services

5.3.4 Optional: pgAdmin Web Interface

5.4 Accessing the Services

5.4.1 Landing Page

5.4.2 Shiny Application Modules

5.4.3 GloSIS Database

5.5 The `docker-compose.yml` File

5.6 The `.env` File

Chapter 6

Digital Soil Mapping and Modeling

Placeholder

Part I: Theoretical Foundations

6.1 Fundamentals of Digital Soil Mapping

6.1.1 Introduction to Digital Soil Mapping

6.1.2 The SCORPAN Framework

6.1.3 Statistical Theory for Predictive Soil Mapping

6.1.3.1 Mechanistic versus Empirical Approaches

6.1.3.2 Sources of Residual Variance

6.1.3.3 Covariates as Proxies of Soil-Forming Factors

6.1.3.4 The Universal Model of Soil Variation

6.1.3.5 Extending the Model: Space, Depth, and Time (3D+T)

6.1.4 Types of Soil Variables

6.1.5 Prediction Methods

6.1.6 Nonlinearity and Uncertainty

Part II: Practical Application

6.2 Workflow Overview

6.3 Continuous Modelling (pH)

6.3.1 Nature of the Variable

6.3.2 Data and Covariates

6.3.3 Feature Selection (Boruta)

6.3.4 Model Training

6.3.5 Variable Importance

6.3.6 Validation Metrics

6.3.7 Prediction Maps

6.3.7.1 Mean pH Map

6.3.7.2 Uncertainty Map (Standard Deviation)

6.3.7.3 Coefficient of Variation Map

6.4 Nominal Classification (Clay_pH_Class)

6.4.1 Nature of the Variable

6.4.2 Feature Selection

6.4.3 Model Training

6.4.4 Validation

6.4.5 Prediction Maps

6.4.5.1 Predicted Class Map

6.4.5.2 Maximum Probability Map

6.4.5.3 Uncertainty Map

6.4.5.4 Confident Predictions Map

6.5 Ordinal Modelling (pH_Class)

6.5.1 Nature of the Variable

6.5.2 Ordinal Encoding

6.5.3 Feature Selection

6.5.4 Model Training

6.5.5 Validation

6.5.6 Prediction Maps

6.5.6.1 Latent Values Map

6.5.6.2 Reclassified Ordinal Class Map

6.5.6.3 Standard Deviation Map

6.5.6.4 Class Instability Map

6.6 Area of Applicability (AOA)

6.6.1 Why AOA is Necessary

6.6.2 Dissimilarity Index (DI)

6.6.3 AOA Binary Mask

6.6.4 Local Point Density (LPD)

6.6.5 Predictions Masked to AOA

6.7 Summary

References

Chapter 7

Soil Data Sharing

7.1 Data sharing and export formats

7.2 Metadata

7.3 Web Services

References

GLOBAL SOIL PARTNERSHIP

The Global Soil Partnership (GSP) is a globally recognized mechanism established in 2012. Our mission is to position soils in the Global Agenda through collective action. Our key objectives are to promote Sustainable Soil Management (SSM) and improve soil governance to guarantee healthy and productive soils, and support the provision of essential ecosystem services towards food security and improved nutrition, climate change adaptation and mitigation, and sustainable development.

Australian Government
Department of Agriculture, Water and the Environment

AFACI
Asian Food & Agriculture Cooperation Initiative

Ministry of Finance of the Russian Federation

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Rural Development Administration

© FAO, 2022
CC1717EN/1/09.22