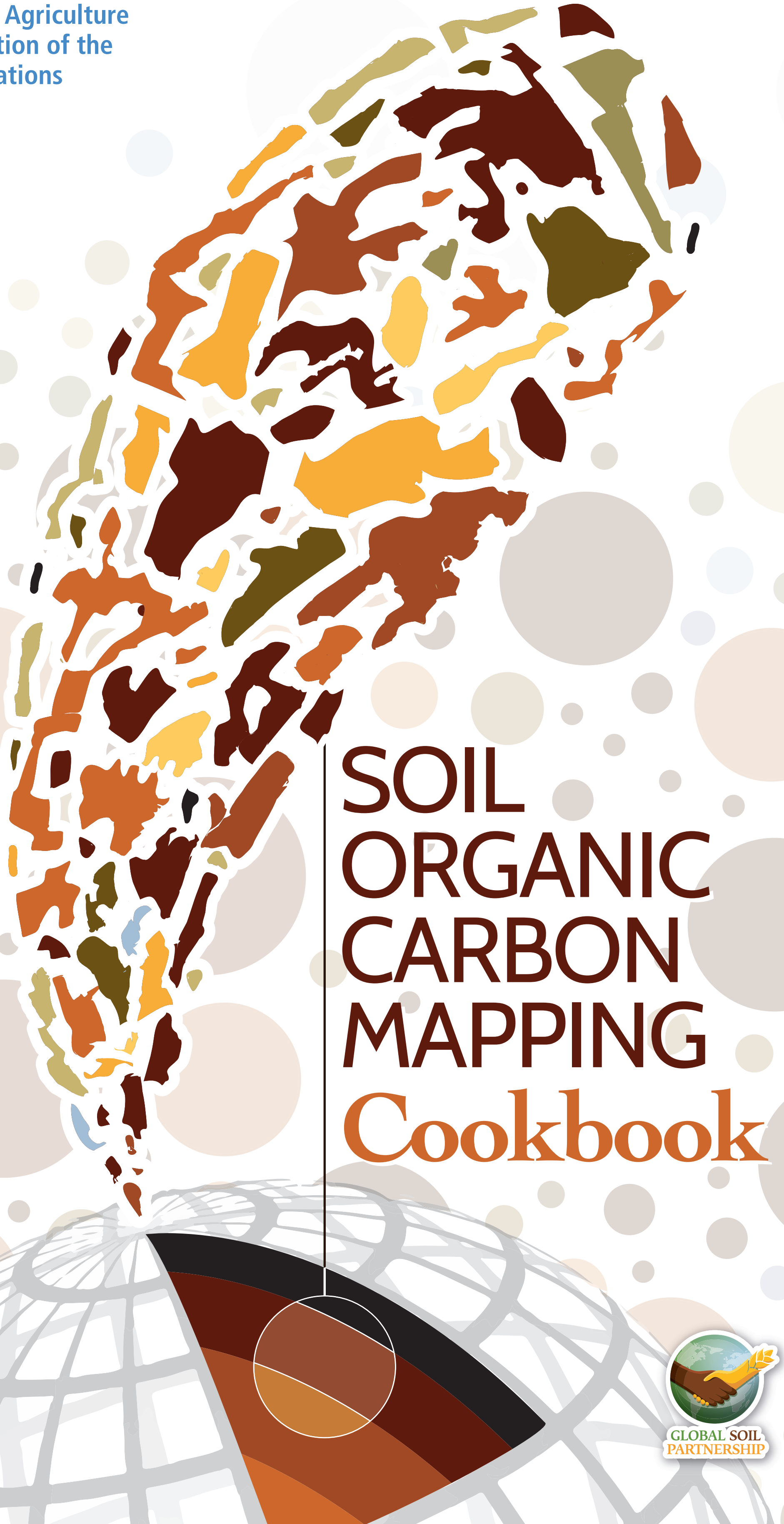
Food and Agriculture
Organization of the
United Nations



SOIL ORGANIC CARBON MAPPING Cookbook

GLOBAL SOIL
PARTNERSHIP

SOIL ORGANIC CARBON MAPPING

GSOC Map
Cookbook Manual

April 2017 - First Edition

EDITORS

Yusuf Yigini, **Rainer Baritz**, **Ronald R. Vargas**

Global Soil Partnership, Food and Agriculture Organization of the United Nations

AUTHORS

Dick Brus - Department of Plant Sciences, Wageningen University, the Netherlands
Tomislav Hengl - ISRIC - World Soil Information, Wageningen, the Netherlands
Gerard Heuvelink - World Soil Information, Wageningen, the Netherlands
Bas Kempen - ISRIC, World Soil Information, Wageningen, the Netherlands
Titia VL Mulder - Wageningen University, Department of Environmental Sciences, The Netherlands
Guillermo Federico Olmedo - INTA, Instituto Nacional de Tecnología Agropecuaria, Argentina
Laura Poggio - The James Hutton Institute, Craigiebuckler Aberdeen, Scotland UK
Eloi Ribeiro - ISRIC , World Soil Information, Wageningen, the Netherlands
Christian Thine Omuto - Department of Environmental and Biosystems Engineering, University of Nairobi, Kenya

COVER DESIGN AND CONTENT FORMATTING

Matteo Sala, FAO

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
ROME, 2017

COPYRIGHT AND DISCLAIMER

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 [insert number]

FAO information products are available on the FAO website (www.fao.org/publications) and can be purchased through publications-sales@fao.org.

CONTENTS

PRESENTATION

The Global Soil Partnership (GSP) aims to promote sustainable soil management at all levels and in all land uses through normative tools that rely on evidence-based science. Understanding the status of a given soil, including its properties and functions, and relating this information to the ecosystem services that soil provides becomes a mandatory action before making decisions on how to manage a soil sustainably. To achieve this, the availability of and use of soil data and information is fundamental to underpin soil management decisions. For this reason, members of the GSP have decided to establish a Global Soil Information System (GLOSIS) that relies on national soil information systems.

In the process of establishing GLOSIS, a number of tools and networks are being created, including the International Network of Soil Information Institutions (INSII), a soil data policy and more. Taking advantage of this process and responding to a request for support in developing the Sustainable Development Goal Indicators, especially Indicator 15.3, the GSP Plenary Assembly instructed the Intergovernmental Technical Panel on Soils and the GSP Secretariat to develop a Global Soil Organic Carbon Map (GSOCMap) following the same bottom-up approach as GLOSIS. To this end, members under the INSII umbrella developed guidelines and technical specifications for the preparation of the GSOCMap (*http://www.fao.org/3/a-bp164e.pdf*) and countries were invited to prepare their national soil organic carbon maps according to these specifications.

Given the scientific advances in tools for mapping soil organic carbon (SOC), many countries requested the GSP Secretariat to support them in the process of preparing these national maps. An intensive capacity development programme on SOC carbon mapping was the answer to support countries in this process. Various regional and national training sessions were organized using an on-the-job-training modality to ensure that national experts were trained using their own datasets. To support this capacity development process, a reference knowledge source was needed, hence the GSP Secretariat invited a group of top experts to prepare a Soil Organic Carbon Mapping Cookbook.

This cookbook provides generic methodologies and the technical steps to produce a SOC map. This includes step-by-step guidance for developing 1 km grids for SOC stocks, as well as for the preparation of local soil data, the compilation and pre-processing of ancillary spatial data sets, upscaling methodologies, and uncertainty assessments. Guidance is mainly specific to soil carbon data, but also contains many generic sections on soil grid development due to its relevance for other soil properties.

The main focus of the guidance is on the upscaling of SOC stocks in the GSOCMap and as such the cookbook supplements the "GSP Guidelines for sharing national data/information to compile a Global Soil Organic Carbon (GSOC) map". It provides technical guidelines to prepare and evaluate spatial soil data sets to:

- Determine SOC stocks from local samples to a target depth of 30 cm;
- Prepare spatial covariates for upscaling; and
- Select and apply the best suitable upscaling methodology.

In terms of statistical upscaling methods, the use of conventional upscaling methods using soil maps and soil profiles is still very common, although this approach is mostly considered empirical by soil mappers. Even though evaluations are based on polygon soil maps, the resulting SOC maps can be rasterized to any target grid. However, a spatially-explicit assessment of uncertainties is impossible. The use of digital soil mapping to upscale local soil information is increasingly applied and recommended. This cookbook presents two approaches in detail, namely spatial modelling using either regression or data mining analysis, combined with geostatistics as regression kriging.

This first edition of the cookbook will be followed by a series of updates and extensions that would be necessary to cover a larger variety of upscaling approaches. The experiences gained throughout 2017 during the implementation of the GSOCMap capacity development programme will be considered in the next editions. This will especially include updates in the section on uncertainties which will be adjusted to provide more practical implementation steps.

It is our hope that this cookbook will fulfil its mandate of easily enabling any user to produce a SOC or other soil property map using soil legacy data and modern methods of digital soil mapping in contribution to improved decision making on soil management.

1. SOIL PROPERTY MAPS

1.1. DEFINITIONS, OBJECTIVES

Soil property maps represent spatial information about soil properties to a certain depth or for soil horizons. Conventionally, soil property maps are generated as polygon maps, with properties from typical soil profiles representing soil mapping units.

Digital Soil Mapping (DSM) allows more accurate spatial mapping of soil properties, including the spatial quantification of the prediction error. The quality of such predictions improves with increasing number of local observations (e.g. soil profiles) available to build prediction model. Whenever possible, DSM is recommended.

The development of soil property maps via digital soil mapping is spatially flexible. For different soil properties (e.g. concentration and stocks of nutrients in the soil, carbon, heavy metals, pH, cation exchange capacity, physical soil properties such as particle sizes and bulk density, etc.), various depth classes and spatial resolution can be modelled depending on project and mapping objectives and available input data. For GSOCmap, a 1 km grid is pursued. The same methodology and input data can also be used to produce higher resolution soil grids.

The mapping of global soil organic carbon stocks (GSOC) will be the first implementation of a series of other soil property grids to be developed for GLOSIS, based on the typical GSP country-driven system. GSOCmap will demonstrate the capacity of countries all around the globe to compile and manage national soil information system and to utilize and evaluate these data following agreed international specifications. The GSP Secretariat, FAO and its regional offices, as well as the Regional Soil Partnerships, are all challenged together with the GSP members, especially the members of the International Network of Soil Information Institutions INSII), to establish national capacity and soil data infrastructures to enable soil property mapping.

1.2. GENERIC MAPPING OF SOIL GRIDS: UPSCALING OF PLOT-LEVEL MEASUREMENTS AND ESTIMATES

The following table presents an overview of different geographic upscaling approaches, recommended to produce soil property maps, in particular GSOCmap.

Table 1.1 An overview of common upscaling methods

Conventional upscaling[1]	**Class-matching**	Derive average SOC stocks per "class": soil type for which a national map exists, or combination with other spatial covariates, e.g. land use category, climate type, biome, etc. This approach is used in the absence of spatial coordinates of the source data.
	Geomatching	Point locations with spatial referencing are overlaid with GIS layers of important covariates (such as a soil map). Upscaling is based on averaged SOC values per mapping unit.
Digital soil mapping[2]	**Data Mining and Geostatistics**	Multiple regression, classification tree, random forests, Regression kriging, kriging with external drift

[1] Lettens, S., J. Van Orshoven, B. Van Wesemael and B. Muys (2004). Soil organic and inorganic carbon content of landscape units in Belgium for 1950 – 1970. Soil Use and Management 20: 40-47.

[2] Dobos, E., F. Carré, T. Hengl, H.I. Reuter and G. Tóth (2006). Digital Soil Mapping as a support to production of functional maps. EUR 22123 EN, 68 pp. Office for Official Publications of the European Communities, Luxemburg.

Digital soil mapping is based on the development of functions for upscaling point data (with soil measurements) to a full spatial extent using correlated environmental covariates, for which spatial data are available.

DEFINITION

DSM: CONCEPT OF ENVIRONMENTAL CORRELATION THAT EXPLORES THE QUANTITATIVE RELATIONSHIP AMONG ENVIRONMENTAL VARIABLES AND SOIL PROPERTIES AND COULD BE USED TO PREDICT THE LATTER; MULTIVARIATE PREDICTION TECHNIQUES

2. PREPARATION OF LOCAL SOIL PROPERTY DATA

2.1. SOIL PROFILES AND SOIL AUGERS

Soil profiles are complex real world entities. Soil profiles are composed of soil layers which form soil horizons; the soil layers have different properties and these properties are evaluated with different methods. As we know, soil and vertical soil properties are landscape elements and part of matter dynamics (water, nutrients, gases, habitat). Local soil samples or soil profiles add a third dimension into the spatial assessment of soil properties in the landscape.

Most commonly, soil are described as vertical profiles using soil pits (sometimes also augerings, but this is less accurate). Soil profiles are described using macro-morphological properties. These properties can be assessed in the field without analysis by making a field inventory or land evaluation. For additional quantitative analysis, soils are then sampled by genetic horizon or by depth class.

Sampling of soils is the basis to obtain quantitative information. Depending on the goal of a project, sampling can be quite diverse. Sampling can follow the description of the soil, or can be conducted without, for example using a spade or auger to generate a composite sample (for a certain depth independent of the morphological features such as soil horizons).

Sampling locations can be representative for a certain location, project, field, or mapped object, such as a soil type.

STEP

STEP 1: PREPARE METADATA ABOUT FOR THE SOIL SOURCE DATA (SEE THE METADATA SECTION)

2.2. SOIL DATABASE

In order to process and evaluate soil information from field assessments, soil profile and analytical information needs to be stored in a data base. This can be a set of simple Excel Spreadsheets, or a relational or object-oriented data base management system (Baritz *et al.* 2009). When working in R, *SoilProfileCollections* from the R ***'aqp'*** package could be a useful tool. Tables 2.1 – 2.3 are examples of how soil information can be stored. The advantage of such organization is the possibility to develop relational databases which can be easily queried. Such a systematic approach will support the organization of national soil information and will reduce errors in future modelling exercises (Baritz *et al.* 2009).

Table 2.1 stores site-level data, which describe the location of the soil description and/or sampling site: spatial coordinates, landscape attributes such as slope gradient and slope form, soil class, land cover type, rock type etc. In this table every row should hold a single soil profile. One column, usually the first one, should be the soil profile's unique identifier. Using the latter, soil information can be easily linked from one table to another.

Table 2.2 stores information from the soil description, such as horizon name, horizon thickness, organic matter content, carbonate content, soil color, etc. The first column contains the soil profile's unique identifier. It is important to include the upper and lower limits for each soil layer; in case the sampling strategy deviates from soil layers/soil horizons, the upper and lower depth of the sampling locations should be specified if possible. This information is needed for modelling soil properties over the soil profile.

Table 2.3 contains the results from the laboratory soil analysis and again lists the soil profile's unique identifier. Both tables 2.2 and 2.3 could also contain data for O horizons of forests, and H horizons for peat soils.

STEP 2: PREPARE TABLE WITH SOIL PROPERTIES FOR SOC MAPPING

Table 2.1 Example for site-level data table

Profile_ID	X coord	Y coord	Soil Type	Land Cover	Parent material	Solum depth [cm][1]
AB1	- 33.0109	- 69.9668	Luvic Calcisol	Shrubland	Limestone	45
BJ12	- 33.5727	- 69.8331	Eutric Cambisol	Crops	Basalt	110
...	...	...	...	...		

[1] SOLUM DEPTH DESCRIBES THE TOTAL DEPTH OF THE DEVELOPED SOIL. THIS IS IMPORTANT TO KNOW IF SOC STOCKS IN THE TARGET DEPTH 0-30 CM REFER TO A SOIL WHICH IS LESS DEEPLY DEVELOPED (E.G. REACHES BEDROCK OR GROUNDWATER WITHIN 30 CM DEPTH).

Table 2.2 Example for profile-description table

Profile_ ID	Horizon name	Upper Limit	Lower Limit	Organic matter content [%][2]	Carbonate content	Texture class	Stone content
AB1	A	0	18	2-4	SL	SiL	F
AB1	BC	18	53	1-1.5	MO	SiC	C
AB1	C	53	100	1-1.5	MO	SiC	C
BJ12	A	0	27	2-4	N	SiCL	V
...	...	...	...	...			

[2] CLASSES BASED ON FAO (2006)

Table 2.3 Example for soil analytical table

Profile_ID	Organic carbon [%]	Bulk density[3]	Sand	Clay	Silt	...
AB1	1.35					
AB1	0.62					
AB1	0.35					
BJ12	1.86					
...	...					

[3] BULK DENSITY CAN ALSO BE ESTIMATED USING ORGANIC CARBON AND TEXTURAL DATA.

RULE 1: CONVENTIONS FOR FILLING TABLES: 0 = 0, NO DATA = NA, ALL NUMERIC

2.3. COMPLETENESS OF MEASUREMENTS/ESTIMATES

The GSOC mapping guideline specifies which soil parameters are needed to produce a GSOCmap. Of course, other soil properties can be evaluated and modelled using this cookbook as well.

In order to calculate stocks we need soil properties A:Z...Explain pedotransfer functions, explain that you go from SOM to SOC, which soil properties you need and how you come to an estimate up to 30cm. If one needs to add a column to their table with calculated stocks than you need to explain them how to do that!

a) Stones

The estimation of stoniness is difficult and time consuming, and therefore not carried out in many national soil inventories, or only estimated visually in the profile. Unfortunately, if soil inventories and sampling are done with simple pits or augers rather than standard soil pits, stones are very often not assessed.

As a proxy, it is recommended to derive national default values from well described soil profile pits by soil type.

b) Bulk density

The amount of fine earth is one of the basic estimation parameters to estimate SOC stocks in the mineral soil as well as in peat layers. It depends on the volume of soil considered (depth × reference area) and the bulk density (BD). BD expresses the soil weight per unit volume. When determining BD, it is important to subtract stones, if any, from the cylinder samples; if this is not done, BD is underestimated, and the resulting SOC stocks are overestimated. Stones in the cylinders are added to the total stone content in order to correct for the total amount of fine earth per volume of soil in a given area.

Most of the soil profiles in national databases come from agricultural land. Very often, BD estimates do not consider fine stones because top soils (e.g. plough layers) seem to be free of visible stones.

Table 2.4 Bulk density values for mineral, forest and peat soils

Mineral soil	Default values: [General Guide for Estimating Moist Bulk Density]()	If analytical BD is missing, BD can be estimated using pedotransfer functions (examples are listed below)
Forest floor	Default bulk densities: Ottmar and Andreu (2007) Pine: L 0.018 g/cm^3 ,F/H 0.057 g/cm^3 Hardwood L 0.012 g/cm^3, F/H 0.043 g/cm^3 Barney *et al.* (1981)[1]: Birch: L/H 0.17 g/cm^3 Spruce: L 0.051 g/cm^3, H 0.13 g/cm^3	L (litter) layer (or Oi horizon, U.S. soil taxonomy) Organic layer, or duff layer: partially decomposed material above the mineral soil and beneath the litter layer; F (fermentation) and H (humus) horizons (Oe and Oa, U.S. soil taxonomy)
	Alternative: calculation of litter C stocks based on the weight per area of O layer horizons (e.g. if sampled with metal frames)	0.057 0.018 0.043 0.012 F/H L
Peat	Default value: 0.31 g/m^3 (Batjes 1996) Agus *et al.* (2011) distinguish different peat decomposition types (with different C content): Sapric 0.174 (48.90 % C) Hemic 0.117 (52.27% C) Fibric 0.089 (53.56 % C)	The range of peat BD is generally about 0.02–0.3 t/m3 depending on maturity and compaction, as well as the ash content (Agus *et al.* 2011)

[1] DIFFICULTY WITH THE DERIVATION OF DEFAULT VALUES: THE OLDER STUDY BY BARNEY *ET AL.* (1981) SEEMS TO OVERESTIMATE BD: THE SAMPLING AND ANALYSIS METHOD NEEDS TO BE CAREFULLY REVIEWED. HOWEVER, THE AUTHORS ALSO LIST CITATIONS AND REFERENCE VALUES OF OTHER AUTHORS (SPRUCE, PINE, MIXED POPULAR/SPRUCE, MIXED SPRUCE/FIR). EXAMPLE FOR PEDOTRANSFER FUNCTIONS TO ESTIMATE BD, BASED ON THE SOIL ORGANIC MATTER CONTENT (SOC × 1.724)

Example for Pedotransfer functions to estimate BD, based on the soil organic matter content (SOC × 1.724)

Saini (1996)	BD = 1,62-0,06 * OM
Drew (1973)	BD = 1/(0,6268 + 0,0361 * OM)
Jeffrey (1979)	BD = 1.482 - 0,6786 * (log OM)
Grigal et. al (1989)	BD = 0,669 + 0,941* e ^(-0,06 * OM)
Adams (1973)	BD = 100/(OM/0,244 + (100-OM))/MBD
Honeysett & Ratkowsky (1989)	BD = 1/(0,564 + 0,0556*OM)

(MDB: Mineral particle density, assumed to be the specific gravity of quartz, 2.65 Mg m-3)

Each method is derived from a specific set of regional soils that is regionally adapted. Selection of the proper method for a given country shall be based on existing reviews and comparisons.

c) Soil carbon analysis

Rosell *et al.* (2001) have closely reviewed the different SOC and SOM estimation procedures, and have also drawn some conclusions about the sources of errors. Determination of SOC from dry combustion methods is least susceptible to errors.

Dry combustion by Loss on Ignition, LoI: SOC is re-calculated applying a *conversion factor:* It is commonly assumed, that organic matter contains an average of 58% organic carbon (so-called Van Bemmelen factor 1.724; for non-organic horizons: SOC = SOM / 1.724). For organic horizons, conversion factor ranges from 1.9 to 2.5 (Nelson and Sommers 1982). The inorganic carbon is not resolved, since typically, temperatures between 400 and 550°C are used.

Wet oxidation: Since wet oxidation is applied without additional (external) heating, low temperatures of around 120° (internal heat) are typical. Thus, the oxidation of carbon is incomplete, and a so-called oxidation factor needs to be applied. With external heating, the C-recovery of the method becomes improved, up to complete recovery. No correction against the mineral carbon is needed. Wet oxidation should typically only be applied to samples with < 5% organic matter.

Usually, an average of 76% organic carbon is recovered, leading to a standard oxidation factor or 1.33 (Lettens *et al.* 2005).

d) Carbonates

In case the total organic carbon is determined with temperatures &gt; 600-800°C, the proportion of mineral soil in CaCO3 has to be subtracted in order to derive the amount of organic carbon (inorganic carbon is also oxidized). The pH value gives the first indication whether the sample has to be analyzed for inorganic carbon or not.

It is crucial to report in the metadata whether national SOC values refer to total C or if the inorganic component has been considered.

e) Depth

The standard depth for GSOCmap is 0-30 cm. Subdivisions are possible depending on the available data, by genetic horizon or depth classes. The following depths are additionally considered for GSOC map (optional):

Forest floor: thickness [cm] subdivision in horizons depending on national soil inventory method (e.g. L, F, H)

Peat: > 30 , < 100 depending on national data

STEP 3: FILL DATA GAPS TO CALCULATE LOCAL SOC STOCKS

2.4. COMPLETENESS OF DEPTH ESTIMATE

Soil properties are commonly collected from the field inventories (see Table 2.2) or from sampling and analysing horizons and/or fixed depths. Since a fixed target depth of **30 cm** is required for GSOC (other depth classes will be recommended in the future, following the GlobalSoilMap specifications (reference)), data holders are confronted with the following options:

Option 1: Soil sampling has already considered this depth: data can be directly used for upscaling see "Upscaling Methods" section)

Option 2: Horizons or layers/depth classes are sampled; but aggregation is needed over the 0-30 cm.

Option 3: The target depth (0-30 cm) was not completely covered by sampling e.g. only the A horizon or a topsoil layer (e.g. 0-20 cm) has been sampled. For both options 2 and 3, additional processing is needed (e.g. equal-area splines).

For both options 2 and 3, transformation is needed using e.g. equal-area splines.

In the case of option 3, the use of equal-area splines was first proposed by Ponce-Hernandez *et al.* (1986), and later tested against real data (Bishop *et al.* 1999). This technique is based on fitting continuous depth functions for modelling the variability of soil properties with depth. Thus, it is possible to convert soil profiles to standard depths, but also to fill gaps. The equal-area spline function consists of a series of local quadratic polynomials that join at 'knots' located at the horizon boundaries thereby the mean value of each horizon is maintained by the spline fit. They are called equal-area splines because the area to the left of the fitted spline curve is equal to the area to the right of the curve.

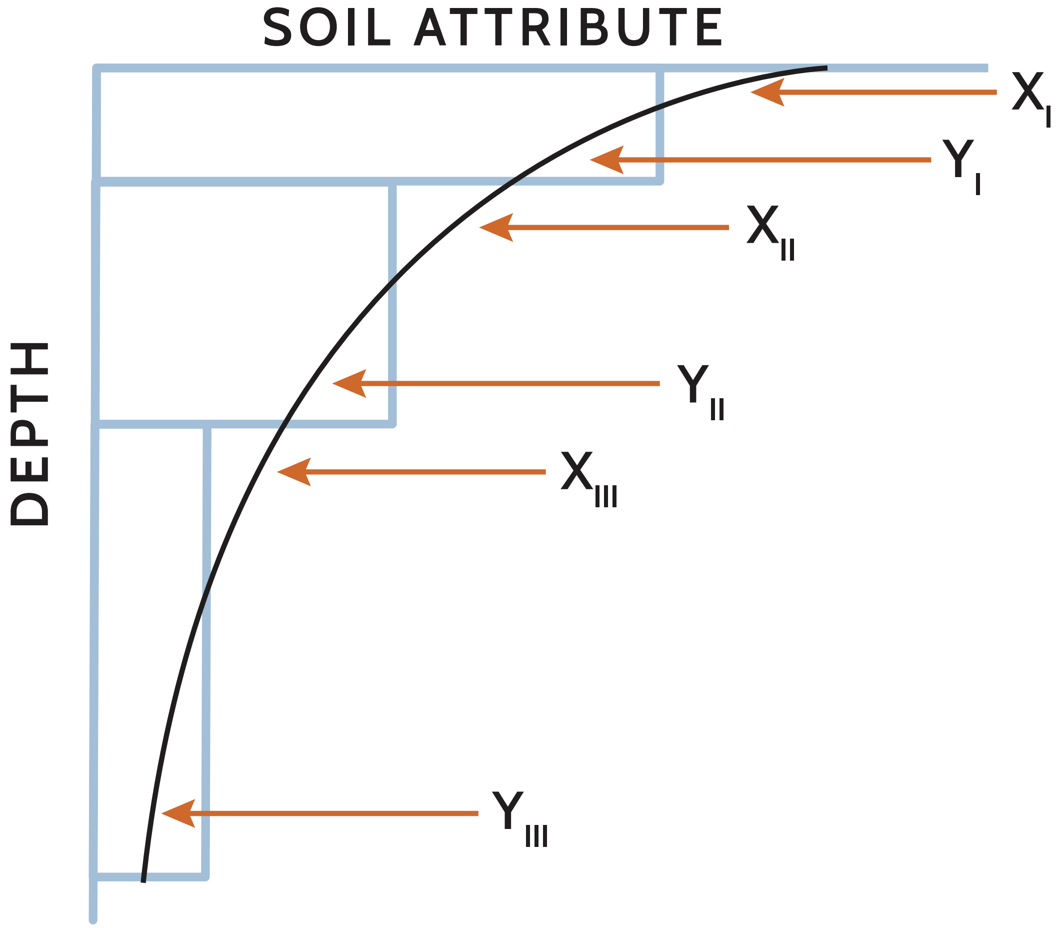


FIGURE 2.1 AN EQUAL-AREA QUADRATIC SPLINE FROM PONCE-HERNANDEZ *ET AL.* (1986). (CITED BY BISHOP *ET AL.*, 1999).

The equal-area spline function is composed of two terms. The first term represents fidelity to the data. The second term measures roughness of the function. The parameter lambda controls the trade-off between the fidelity term and the roughness penalty. The choice of lambda is itself a non-trivial problem. When non prior information is available, many authors recommend using a lambda value between 0.01 and 0.1.

Another set of parameters for these functions in many different softwares is the target standard depths. By defining a set of standard depths we can obtain the value of the variable for a synthetic horizon of for example, 0 – 20 cm, regardless of how the profiles were originally sampled in the field.

In this cookbook manual we proposed two solutions to fill depth related gaps. One is based on 'R' and the other is CSIRO Spline Tool. While 'R' based approach requires background knowledge on 'R',CSIRO Spline Tool requires less.

2.4.1 TECHNICAL STEPS (EQUAL AREA SPLINES USING R)

In R environment, the easiest way to apply equal-area splines is using the function GSIF::mpspline from the R package GSIF (Hengl 2016, see section 4.3.2). For illustration, a sample dataset has been used (see Chapter 5.). This function requires data stored as SoilProfileCollection (SPC) using package aqp. Nevertheless, data in any local soil database or in tables like the ones proposed before (Tables 2.1, 2.2 and 2.3) can be transformed to a SPC.

The function GSIF::mpspline has several arguments. One of the arguments is the lambda value mentioned before. The proposed default value is 0.1. Another argument for this function is the target standard depths. The function produces spline-estimated values at these depths. However, this function also produces spline-estimated values at 1 cm increments. The following technical steps require 'R' and certain packages.

STEP 1: LOAD NEEDED PACKAGES AND SET WORKING FOLDER

```
# load aqp package
library(aqp)
# load gsif package
library(gsif)
# set working folder, please change accordingly
setwd("c://..../cookbook/")
```

STEP 2: LOAD DATA AND PROMOTE TO SOILPROFILECOLLECTION

```
data <- read.csv(“profile-data.csv”, na.strings = “-9999”)
# We convert our table to a SoilProfileCollection (aqp) using function depths(
depths(data) <- ProfID ~ DepthFrom + DepthTo
# inspect the new object
str())
```

STEP 3 - APPLYING THE SPLINE FUNCTION

```
Step 3: Apply mpsspline function to estimate values at fixed depths
```

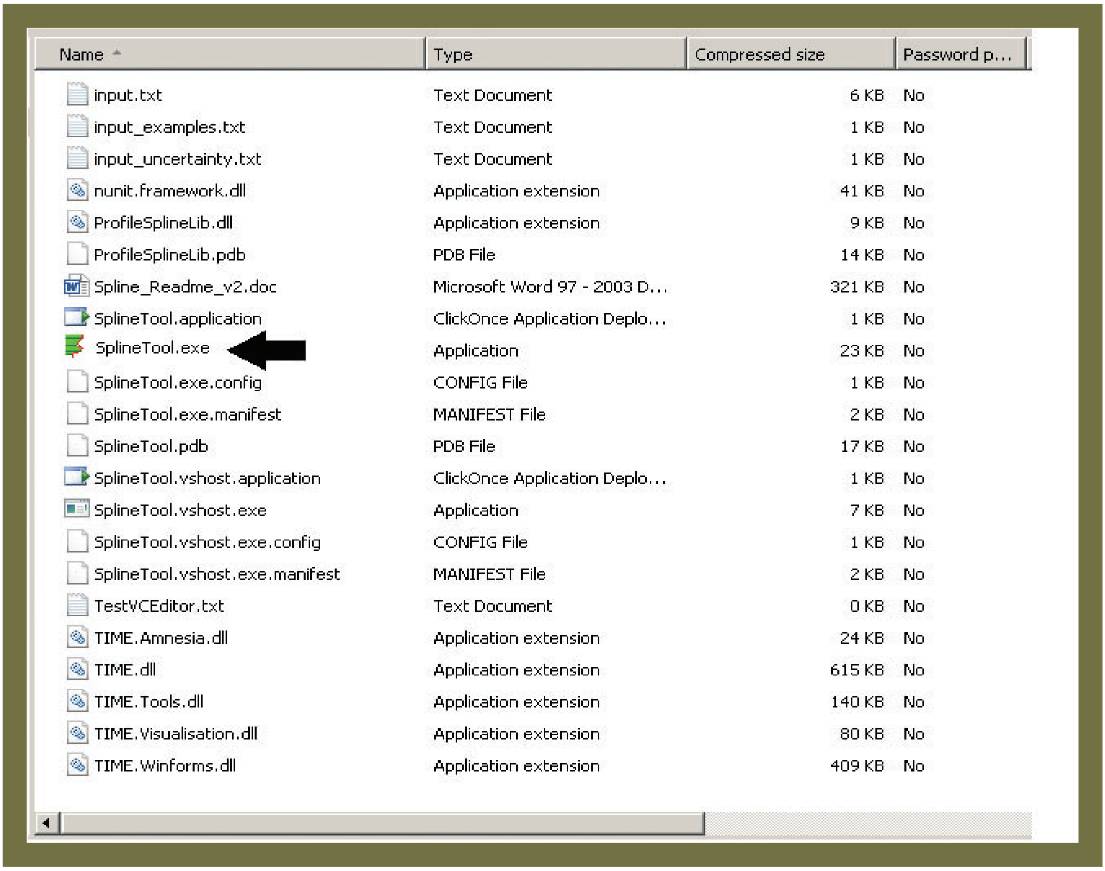
2.4.2 TECHNICAL STEPS (SPLINE TOOL V2.0 -ACLEP/CSIRO - AUSTRALIA-)

The Spline Tool is developed by CSIRO Land and Water. This standalone tool has an easy graphical user interface and allows user to estimate soil properties for standard depth intervals using mass preserving splines from input profiles with irregular or non-contiguous depth intervals.

The tool is available at: *http://www.asris.csiro.au/downloads/GSM/SplineTool_v2.zip*

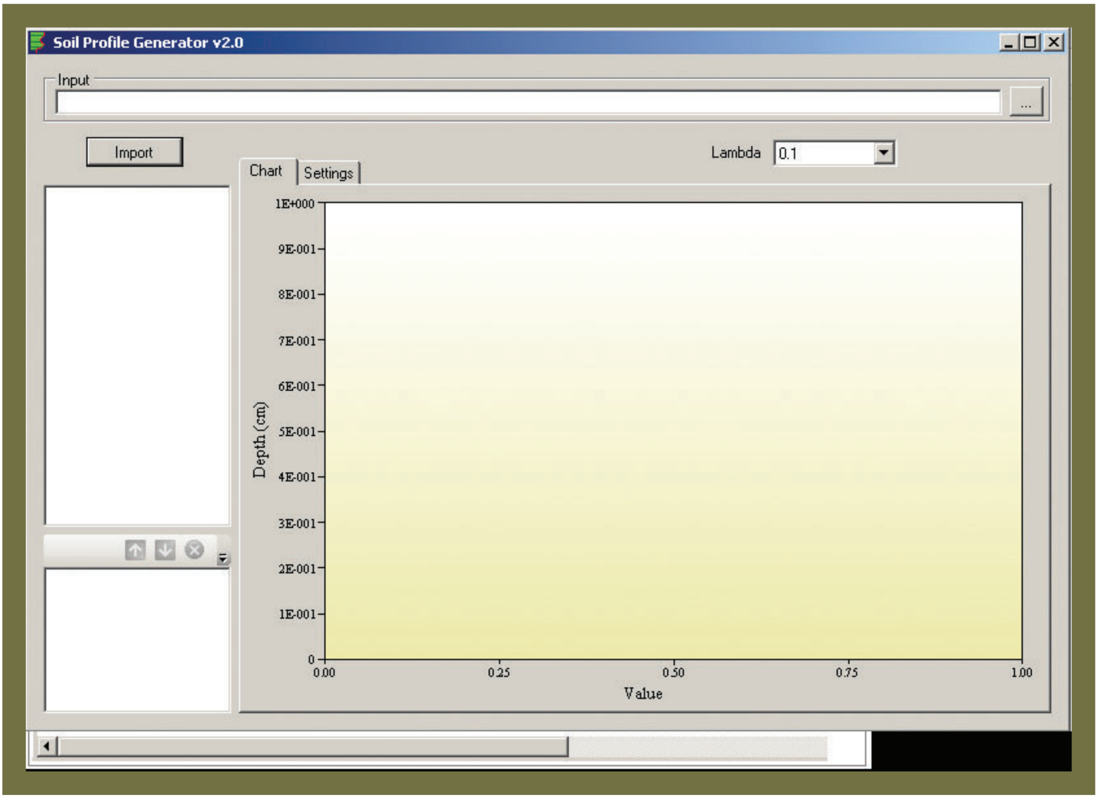
To install the Spline Tool simply extract the .zip content into a new folder.

STEP 1: DOWNLOAD THE SPLINE TOOL AND EXTRACT THE .ZIP FILE CONTENT IN A NEW FOLDER

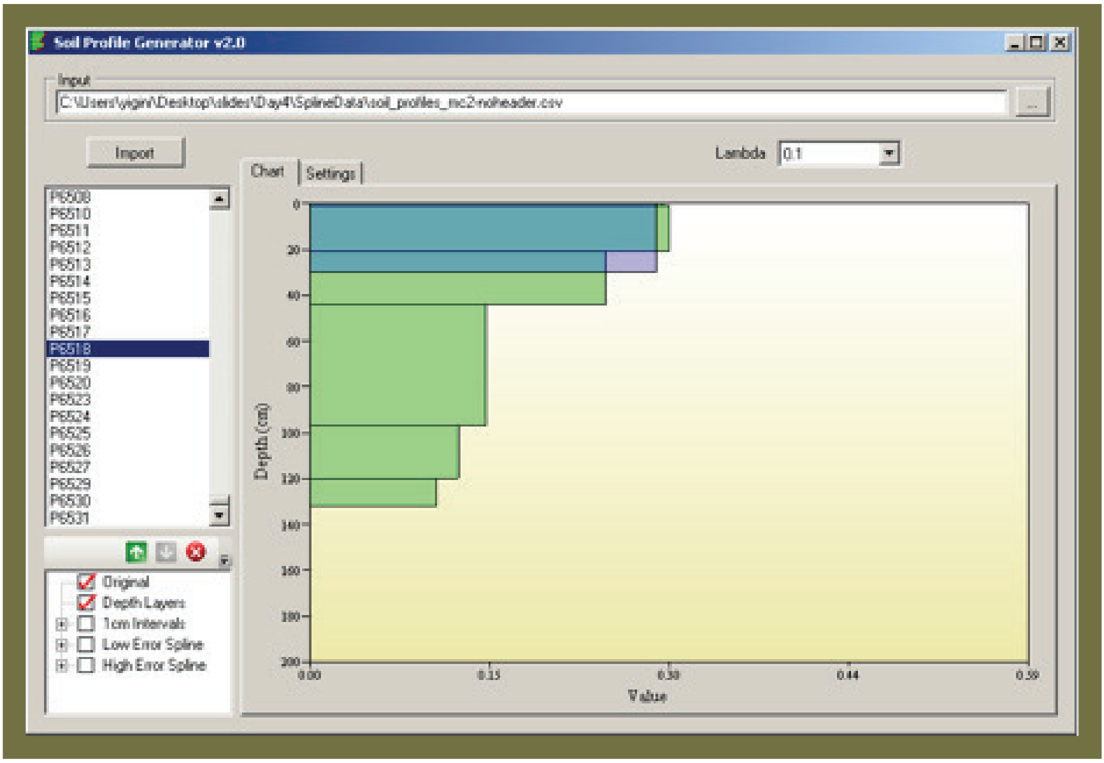


Name	Type	Compressed size	Password p...
input.txt	Text Document	6 KB	No
input_examples.txt	Text Document	1 KB	No
input_uncertainty.txt	Text Document	1 KB	No
nunit.framework.dll	Application extension	41 KB	No
ProfileSplineLib.dll	Application extension	9 KB	No
ProfileSplineLib.pdb	PDB File	14 KB	No
Spline_Readme_v2.doc	Microsoft Word 97 - 2003 D...	321 KB	No
SplineTool.application	ClickOnce Application Deplo...	1 KB	No
SplineTool.exe	Application	23 KB	No
SplineTool.exe.config	CONFIG File	1 KB	No
SplineTool.exe.manifest	MANIFEST File	2 KB	No
SplineTool.pdb	PDB File	17 KB	No
SplineTool.vshost.application	ClickOnce Application Deplo...	1 KB	No
SplineTool.vshost.exe	Application	7 KB	No
SplineTool.vshost.exe.config	CONFIG File	1 KB	No
SplineTool.vshost.exe.manifest	MANIFEST File	2 KB	No
TestVCEditor.txt	Text Document	0 KB	No
TIME.Amnesia.dll	Application extension	24 KB	No
TIME.dll	Application extension	615 KB	No
TIME.Tools.dll	Application extension	140 KB	No
TIME.Visualisation.dll	Application extension	80 KB	No
TIME.Winforms.dll	Application extension	409 KB	No

STEP 2: DOUBLE CLICK ON SPLINETOOL.EXE AND RUN THE APPLICATION. THE TOOL REQUIRES THE MICROSOFT .NET FRAMEWORK, WHICH IS DEVELOPED, SERVICED AND SUPPORTED BY MICROSOFT AND AVAILABLE AT *https://www.microsoft.com/net/download/framework*



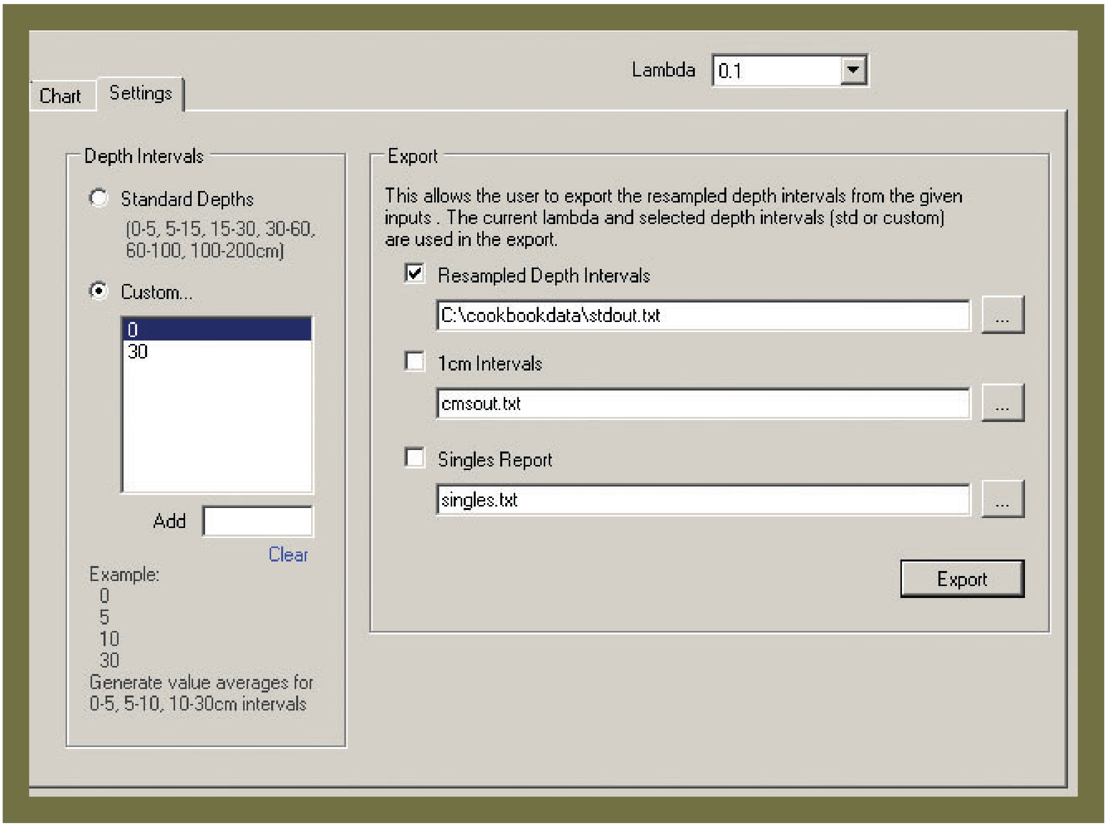
STEP 3: USE BROWSE BUTTON TO BROWSER YOUR .CSV FILE AND CLICK IMPORT BUTTON TO LOAD YOUR DATA. THE SPLINE TOOL USES COMMA DELIMITED INPUT TEXT FILES AND GENERATES **CMSOUT.TXT**, **STDOUT.TXT** AND **SINGLE.TXT**. THE SPLINE TOOL CURRENTLY INPUTS AND OUTPUTS ONE SOIL PROPERTY AT EACH TIME. THE CMSOUT.TXT FILE CONTAINS SOIL ATTRIBUTE VALUES FOR DEFAULT DEPTH INTERVALS TO A DEPTH OF 200 CM. THE MAXIMUM DEPTH OF THE OUTPUT IS LIMITED BY THE MAXIMUM INPUT DEPTH. THESE STANDARD DEPTHS CAN BE EASILY CHANGED ON THE SETTINGS TAB. 0-30 CM WOULD BE USED FOR GSOC MAP MANDATORY DEPTH.


Soil Profile Generator v2.0

Input: C:\Users\yigini\Desktop\slides\Day4\SplineData\soil_profiles_mc2-noheader.csv

Lambda: 0.1

☑ Original
☑ Depth Layers
☐ 1cm Intervals
☐ Low Error Spline
☐ High Error Spline


Lambda: 0.1

Depth Intervals

○ Standard Depths (0-5, 5-15, 15-30, 30-60, 60-100, 100-200cm)

● Custom...

0
30

Add

Example:
0
5
10
30
Generate value averages for 0-5, 5-10, 10-30cm intervals

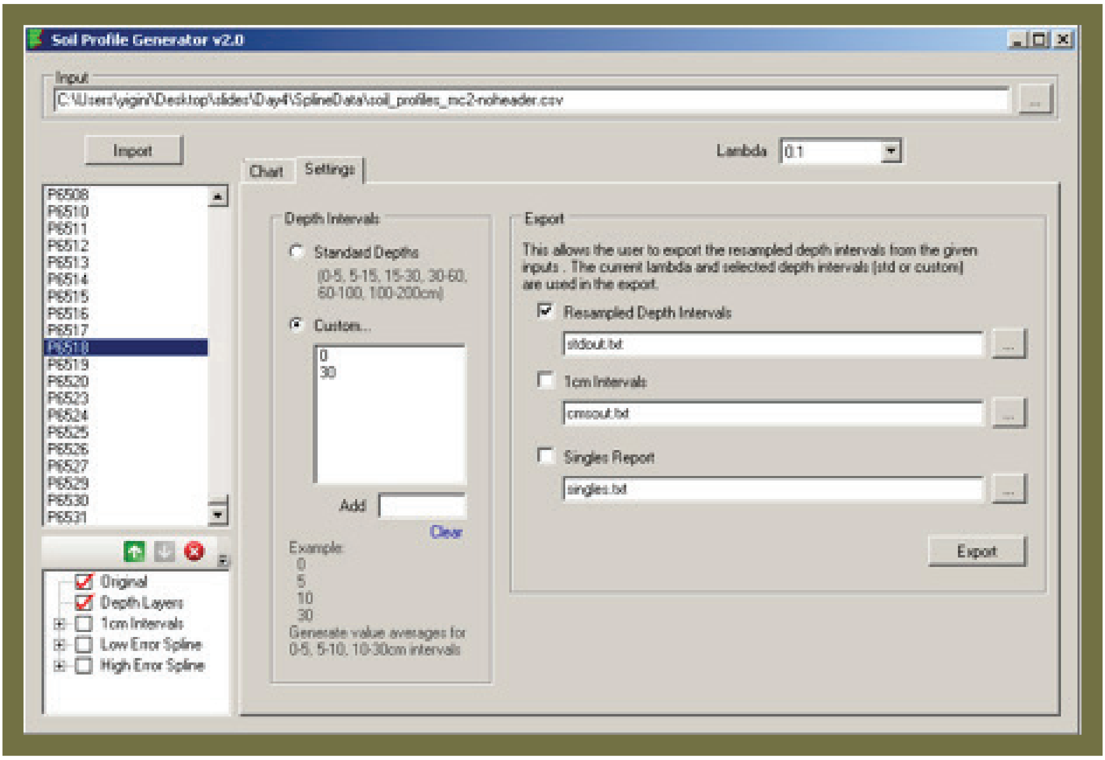
Export

This allows the user to export the resampled depth intervals from the given inputs . The current lambda and selected depth intervals (std or custom) are used in the export.

☑ Resampled Depth Intervals: C:\cookbookdata\stdout.txt

☐ 1cm Intervals: cmsout.txt

☐ Singles Report: singles.txt



Input: C:\Users\yigini\Desktop\slides\Day4\SplineData\soil_profiles_mc2-noheader.csv

Lambda: 0.1

P6508
P6510
P6511
P6512
P6513
P6514
P6515
P6516
P6517
P6518
P6519
P6520
P6523
P6524
P6525
P6526
P6527
P6529
P6530
P6531

Depth Intervals

○ Standard Depths
(0-5, 5-15, 15-30, 30-60, 60-100, 100-200cm)

● Custom...
0
30

Add

Example:
0
5
10
30
Generate value averages for 0-5, 5-10, 10-30cm intervals

Export

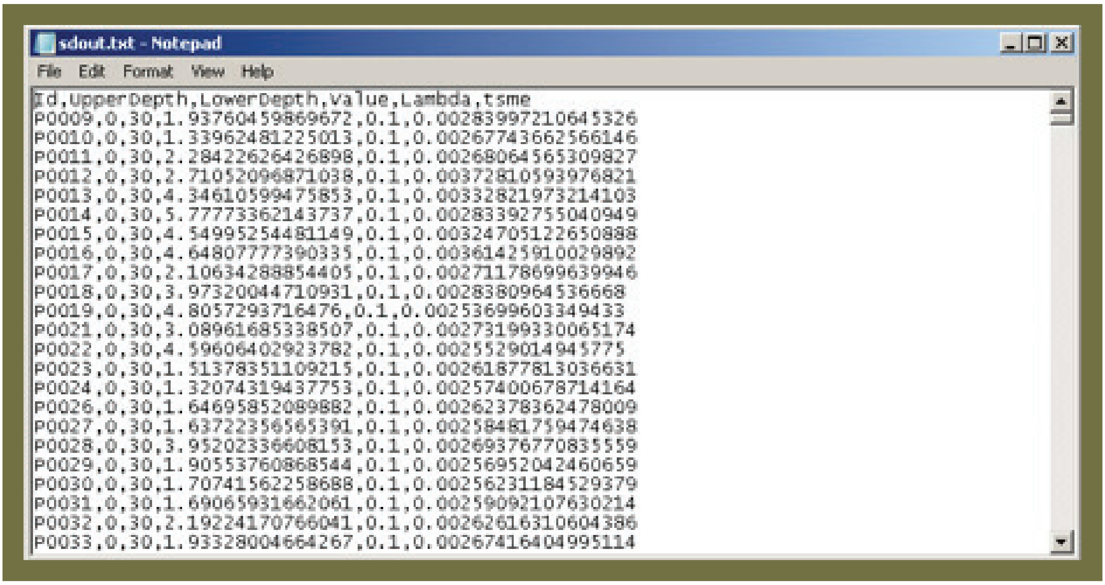
This allows the user to export the resampled depth intervals from the given inputs . The current lambda and selected depth intervals (std or custom) are used in the export.

☑ Resampled Depth Intervals
stdout.txt

□ 1cm Intervals
cmsout.txt

□ Singles Report
singles.txt

STEP 4: CLICK "EXPORT" BUTTON TO EXPORT THE DATA AS .TXT. THE RESULTS ARE NOW STORED IN THE STDOUT.TXT FILE. NOTE THAT THE OUTPUT FILES (STDOUT.TXT AND CMSOUT.TXT) CONTAIN VALUES FOR LAMBDA (THE SMOOTHNESS OF THE FUNCTION) AND TMSE (ESTIMATED MEAN SQUARED ERROR OF THE SPLINE).



```
Id,UpperDepth,LowerDepth,Value,Lambda,tsme
P0009,0,30,1.93760459869672,0.1,0.00283997210645326
P0010,0,30,1.33962481225013,0.1,0.00267743662566146
P0011,0,30,2.28422626426898,0.1,0.00268064565309827
P0012,0,30,2.71052096871038,0.1,0.00372810593976821
P0013,0,30,4.34610599475853,0.1,0.00332821973214103
P0014,0,30,5.77773362143737,0.1,0.00283392755040949
P0015,0,30,4.54995254481149,0.1,0.00324705122650888
P0016,0,30,4.64807777390335,0.1,0.00361425910029892
P0017,0,30,2.10634288854405,0.1,0.00271178699639946
P0018,0,30,3.97320044710931,0.1,0.00283809645366668
P0019,0,30,4.8057293716476,0.1,0.00253699603349433
P0021,0,30,3.08961685338507,0.1,0.00273199330065174
P0022,0,30,4.59606402923782,0.1,0.00255290149 45775
P0023,0,30,1.51378351109215,0.1,0.00261877813036631
P0024,0,30,1.32074319437753,0.1,0.00257400678714164
P0026,0,30,1.64695852089882,0.1,0.00262378362478009
P0027,0,30,1.63722356565391,0.1,0.00258481759474638
P0028,0,30,3.95202336608153,0.1,0.00269376770835559
P0029,0,30,1.90553760868544,0.1,0.00256952042460659
P0030,0,30,1.70741562258688,0.1,0.00256231184529379
P0031,0,30,1.69065931662061,0.1,0.00259092107630214
P0032,0,30,2.19224170766041,0.1,0.00262616310604386
P0033,0,30,1.93328004664267,0.1,0.00267416404995114
```

STEP 5: THE OUTPUT .TXT FILE WILL HAVE THE FOLLOWING COLUMNS

- PROFILE ID/SAMPLE ID
- UPPER DEPTH
- LOWER DEPTH
- OUTPUT VALUE
- LAMBDA (THE SMOOTHNESS OF THE SPLINE FUNCTION)
- TMSE (ESTIMATED MEAN SQUARED ERROR)

The tool exports the data in it's own folder as sdout.txt. Even though the input file has X, Y coordinates, they are missing from the exported file since the tool does not keep any other information than Id, UpperDepth, LowerDepth, Value, Lambda, tsme. We need to add X, Y columns back in the data table to be able to use the data in a DSM framework.

Further reading

Spline_Readme_v2.doc in the .zip archive and;

B.P. Malone, A.B. McBratney, B. Minasny, G.M. Laslett (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma, 154, 138-152

T.F.A. Bishop, A.B. McBratney & G.M. Laslett (1999). Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma, 91, 27-45.

2.5 References

Agus F, Hairiah K, Mulyani A. (2011). Measuring carbon stock in peat soils: practical guidelines. Bogor, Indonesia: World Agroforestry Centre (ICRAF) Southeast Asia Regional Program, Indonesian Centre for Agricultural Land Resources Research and Development. 60 p.

Baritz, R., E. Eberhardt, M. Van Liedekerke and P. Panagos (2009). Environmental Assessment of Soil for Monitoring Volume III: Database Design and Selection. EUR 23490 EN/3 – 2008. Office for Official Publications of the European Union, Luxembourg, 2009.

Bishop, T. F. A., McBratney, A. B., & Laslett, G. M. (1999). Modeling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma, 91(1–2), 27–45. https://doi.org/10.1016/S0016-7061(99)00003-8

Hiederer, R. (2009). Distribution of Organic Carbon in Soil Profile Data. European Commission JRC Scientific and Technical Research Reports EUR 23980 EN. Office for Official Publications of the European Union, Luxembourg, 2009. OPOCE LB-NA-23980-EN-C.

Hiederer, R. (2013). Mapping Soil Properties for Europe - Spatial Representation of Soil Database Attributes. JRC Technical Reports. EUR 26082 EN. Office for Official Publications of the European Union, Luxembourg, 2013. ISBN 978-92-79-32516-8 (pdf)

Lettens, S., B. De Vos, J. Van Orshoven, B. Muys and B. van Wesemael (2005). Variable carbon recovery of Walkley-Black analysis and implications for national soil organic carbon inventories. European Journal of Soil Science.

Malone, B. P., McBratney, A. B., & Minasny, B. (2011). Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma, 160(3–4), 614–626. *https://doi.org/10.1016/j.geoderma.2010.11.013*

Ottmar, R. and A. Andreu (2007). Litter and Duff Bulk Densities in the Southern United States. Joint Fire Science Program Project #04-2-1-49. Final Report. USDA Forest Service, Seattle. *https://www.firescience.gov/projects/04-2-1-49/project/04-2-1-49_final_report.pdf*

Ponce-Hernandez, R., Marriott, F.H.C., Beckett, P.H.T., 1986. An improved method for reconstructing a soil profile from analyses of a small number of samples. Journal of Soil Science 37, 455–467.

Nelson, D.W. and L.E. Sommers (1982). Total carbon, organic carbon and organic matter. p. 539-537. In: Page, A.L. *et al.* (eds.), Methods for soil analysis. Part 2. Chemical and microbial processes. American Society of Agronomy, Madison, Wisconsin, USA.

Rosell, R.A., J.C. Gasparoni and J.A. Galantini (2001). Soil Organic Matter Evaluation. In: Lal, R., J.M. Kimble, R.F. Follett and B.A. Stewart (eds.). Assessment Methods for Soil Carbon. CRC Press LLC P. Lewis Publishers, Boca Raton, USA. p. 311-322.

3. PREPARATION OF SPATIAL COVARIATES

3.1 DEM-DERIVED COVARIATES

3.1.1 DEM SOURCE DATA SETS

Currently, two global level 30 m DEMs are freely available: the Shuttle Radar Topographic Mission (SRTM) and the ASTER Global Digital Elevation Model (GDEM). They provide topographic data at the global scale, which are freely available for users. Both DEMs were compared by Wong *et al.* (2014). Comparison against high-resolution topographic data of Light Detection and Ranging (LiDAR) in a mountainous tropical montane landscape showed that the SRTM (90 m) produced better topographic data in comparison with ASTER GDEM.

DEFINITION

- RECOMMENDED FOR NATIONAL LEVEL APPLICATIONS: 30 M GDEM / SRTM
- RECOMMENDED FOR GLOBAL LEVEL APPLICATIONS: SRTM 90 M, RESAMPLED 1 KILOMETRE.

In both cases noise and artefacts need to be filtered out. ASTER seems to contain more large artefacts (e.g. peaks), particularly in flat terrain, which are very difficult to remove through filtering.

DEFINITION

GRASS GIS OR GDAL: USE "MDENOISE" MODULE/UTILITY TO REMOVE NOISE WHILE PRESERVING SHARP FEATURES LIKE RIDGES, LINES AND VALLEYS.

SRTM contains many gaps (pixels with no-data). These gaps could be filled using splines. SAGA GIS has a module called 'Close Gaps with Splines' and other similar tools for doing this.

3.2 PARENT MATERIAL

Parent material has a crucial impact on soil formation, soil geochemistry and soil physics. Parent material, if not specifically mapped by soil mappers and included in soil maps, is usually available from Geology maps. These maps focus on rock formation, mineral components and age, and often lack younger surface sediments (even in quaternary maps). Parent material/rock types classified by soil mappers considers more strongly geochemistry and rock structure. The most commonly available approximation to parent material is certainly a geology map. Its geochemistry has essential impact on the soil chemistry, e.g. cation exchange capacity, base saturation, and nutrient stock. The rock structure determines the ability to disintegrate, which has impact on soil physical properties, like texture, skeleton content, permeability, and soil thickness.

National parent material and geology maps may be used. Other available datasets and data portals are given on the ISRIC WorldGrids website (worldgrids.org).

- OneGeology: The world geological maps are now being integrated via the OneGeology project which aims at producing a consistent Geological map of the world in approximate scale 1:1M (Jackson, 2007) *www.onegeology.org*
- USGS has several data portals, e.g. that allow browsing of the International Surface Geology (split into South Asia, South America, Iran, Gulf of Mexico, Former Soviet Union, Europe, Caribbean, Bangladesh, Asia Pacific, Arctic, Arabian Peninsula, Africa and Afghanistan) *https://mrdata.usgs.gov/geology/world*
- Hartmann and Moosdorf (2012) have assembled a global, purely lithological database called GLiM (Global Lithological Map). GLiM consists of over 1.25 million digital polygons that are classified in three levels (a total of 42 rock-type classes). *https://www.geo.uni-hamburg.de/en/geologie/forschung/geochemie/glim.html*
- USGS jointly with ESRI has released in 2014 a Global Ecological Land Units map at 250 m resolution. This also includes world layer of rock types. This data can be downloaded from the USGS site *http://rmgsc.cr.usgs.gov/outgoing/ecosystems/Global*

3.3 SOIL MAPS

Soil maps play a crucial role for upscaling soil property data from point locations. They can be the spatial layer for conventional upscaling, they can also serve as a covariate in digital soil mapping. Predicted soil property maps have lower quality in areas where the covariates such as relief, geology and climate so not correlate well with the dependent variable, here soil carbon stocks. This is especially true for soils under groundwater or stagnic water influence. This information is well-represented in soil maps.

FAO, IIASA, ISRIC, ISS CAS and JRC produced a gridded 1 km soil class map (HWSD). Global HWSD-derived soil property maps can be downloaded as geotiffs at *http://worldgrids.org/doku.php/wiki:layers#harmonized_world_soil_database_images_5_km* (see also section 3.6).

3.4 LAND COVER/LAND USE

Besides soil, geology and climate, land use and/or land cover data are unarguably vital data for any statistical effort to map soil properties. There are many of various sources of data on land cover including global and continental products, such as GlobCover, GeoCover, Globeland30, CORINE Land Cover.

3.4.1 GLOBCOVER (GLOBAL)

GlobCover is a European Space Agency (ESA) initiative which began in 2005 in partnership with JRC, EEA, FAO, UNEP, GOFC-GOLD and IGBP. The aim of the project was to develop a service capable of delivering global composites and land cover maps using as input observations from the 300 m MERIS sensor onboard the ENVISAT satellite mission. ESA makes available the land cover maps, which cover 2 periods: December 2004 - June 2006 and January - December 2009. The classification module of the GlobCover processing chain consists in transforming the MERIS-FR multispectral mosaics produced by the pre-processing modules into a meaningful global land cover map. The global land cover map has been produced in an automatic and global way and is associated with a legend defined and documented using the UN LCCS. The GlobCover 2009 land cover map is delivered as one global land cover map covering the entire Earth. Its legend, which counts 22 land cover classes, has been designed to be consistent at the global scale and therefore, it is determined by the level of information that is available and that makes sense at this scale (Bontemps *et al.*, 2011).

The GlobCover data can be downloaded at: *http://due.esrin.esa.int/page_globcover.php*

3.4.2 LANDSAT GEOCOVER (GLOBAL)

The Landsat GeoCover collection of global imagery was merged into mosaics by the Earth Satellite Company (now MDA Federal). The result was a series of tiled imagery that is easier to wield than individual scenes, especially since they cover larger areas than the originals. The great detail in these mosaic scenes, however, makes them large in storage size, so the Mr.Sid file format, which includes compression operations, was chosen for output. While GeoCover itself is available in three epochs of 1975, 1990 and 2000, only the latter two epochs were made into mosaics.

Coverage: The GeoCover Landsat mosaics are delivered in a Universal Transverse Mercator (UTM) / World Geodetic System 1984 (WGS84) projection. The mosaics extend north-south over 5 degrees of latitude, and span east-west for the full width of the UTM zone. For mosaics below 60 degrees north latitude, the width of the mosaic is the standard UTM zone width of 6 degrees of longitude. For mosaics above 60 degrees of latitude, the UTM zone is widened to 12 degrees, centred on the standard even-numbered UTM meridians. To insure overlap between adjacent UTM zones, each mosaic extends for at least 50 kilometres to the east and west, and 1 kilometre to the north and south.

Pixel size: 14.25 meters (V 2000)

The data is available at: *ftp://ftp.glcf.umd.edu/glcf/Mosaic_Landsat* (FTP Access)

3.4.3 GLOBELAND30 (GLOBAL)

GlobeLand30, the world's first global land cover dataset at 30 m resolution for the years 2000 and 2010, was recently released and made publicly available by China.

The National Geomatics Center of China under the "Global Land Cover Mapping at Finer Resolution" project has recently generated a global land cover map named GlobeLand30. The dataset covers two timestamps of 2000 and 2010, primarily acquired from Landsat TM and ETM+ sensors, which were then coupled/checked with some local products.

The data is publicly available for non-commercial purposes at:
http://www.globallandcover.com/GLC30Download/index.aspx

Further reading and other global data sources:
http://worldgrids.org/doku.php/wiki:land_cover_and_land_use

3.4.4 CORINE LAND COVER (EUROPE ONLY)

The pan-European component is coordinated by the European Environment Agency (EEA) and produces satellite image mosaics, land cover / land use (LC/LU) information in the CORINE Land Cover data, and the High Resolution Layers.

The CORINE Land Cover is provided for 1990, 2000, 2006 and 2012. This vector-based dataset includes 44 land cover and land use classes. The time-series also includes a land-change layer, highlighting changes in land cover and land-use. The high-resolution layers (HRL) are raster-based datasets (100 m, 250 m) which provide information about different land cover characteristics and is complementary to land-cover mapping (e.g. CORINE) datasets.

The CORINE Land Cover Data are available at:
http://www.eea.europa.eu/data-and-maps/data

3.5 CLIMATE

3.5.1 WORLDCLIM V1.4 AND V2 (GLOBAL)

WorldClim is a set of global climate layers (gridded climate data) with a spatial resolution of about 1 km2 (10 minutes, 5 minutes, 2.5 minutes are also available). These data can be used for mapping and spatial modelling. The current version is Version 1.4. and a preview of Version 2 is available for testing at worldclim.org. The data can be downloaded as generic grids or in ESRI Grid format.

The WorldClim data layers were generated by interpolation of average monthly climate data from weather stations on a 30 arc-second resolution grid. In V1.4, variables included are monthly total precipitation, and monthly mean, minimum and maximum temperatures, and 19 derived bioclimatic variables. The WorldClim precipitation data were obtained from a network of 1,473 stations, mean temperature from 24,542 stations, and minimum and maximum temperatures from 14,835 stations (Hijmans *et al.* 2005).

The Bioclimatic parameters are: annual mean temperature, mean diurnal range, iso-thermality, temperature seasonality, max temperature of warmest month, minimum temperature of coldest month, temperature annual range , mean temperature of wettest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality (coefficient of variation), precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter, precipitation of coldest quarter.

WorldClim Climate Data are available at: *www.worldclim.org* (WorldClim 1.4 (current conditions) by *www.worldclim.org*; Hijmans *et al.*, 2005. Int. J. of Clim. 25: 1965-1978. Is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License).

3.5.2 GRIDDED AGRO-METEOROLOGICAL DATA IN EUROPE (EUROPE)

CGMS database contains meteorological parameters from weather stations interpolated on a 25×25 km grid. Meteorological data are available on a daily basis from 1975 to the last calendar year completed, covering the EU Member States, neighbouring European countries.

The following parameters are available at 1 day time resolution;

maximum air temperature (°C),
minimum air temperature (°C),
mean air temperature (°C),
mean daily wind speed at 10m (m/s),
mean daily vapour pressure (hPa),
sum of precipitation (mm/day),
potential evaporation from a free water surface (mm/day),
potential evapotranspiration from a crop canopy (mm/day),
potential evaporation from a moist bare soil surface (mm/day),
total global radiation (KJ/m2/day),Snow Depth

Data Access: http://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx

3.6 GSOC MAP - DATA REPOSITORY (ISRIC, 2017)

ISRIC World Soil Information has established a data repository contains raster layers of various biophysical earth surface properties for each territory in the world. These layers can be used as covariates in a digital soil mapping exercise.

3.6.1 COVARIATES AND EMPTY MASK

The territories and their boundaries are obtained from from the Global Administrative Unit Layers (GAUL)dataset:
each folder contains three subfolders;
* **covs:** GIS layers of various biophysical earth surface properties
* **mask:** an 'empty' grid file of the territory with territory boundary according to GAUL. This grid can for instance be used as a mapping mask.

* **soilgrids:** all SoilGrids250m soil class and property layers as available through www.soilgrids.org. Layers are aggregated to 1 km.

3.6.2 DATA SPECIFICATIONS

File format: GeoTiff
Coordinate system: WGS84, latitude-longitude in decimal degrees
Spatial resolution: 1km

3.6.3 DATA ACCESS

ftp://gsp.isric2.org/ (username: gsp, password: gspisric) or
ftp://85.214.253.67 (username: gsp, password: gspisric)

LICENCE and ACKNOWLEDGEMENT
The GIS layers can be freely used under the condition that proper credit should be given to the original data source in each publication or product derived from these layers. Licences, data sources, data citations are indicated the data description table.

3.7 PREPARATION OF A SOIL PROPERTY TABLE FOR SPATIAL STATISTICS

The upscaling procedures (Chapter 6) depend on the rationale that the accumulation of local soil carbon stocks (and also other properties) depend on parameters for which spatial data are available, such as climate, soil type, parent material, slope, management. This information (Covariates) must be collected first. Details are provided above. The properties contained in the covariates can be extracted to each georeferenced sample site and added to the soil property table (Table 3.1). This table is used for training and validation of the statistical model for predicting the SOC stocks which subsequently can be applied to the full spatial extent.

Table 3.1 Extended input table for spatial analysis (spatial SOC prediction table)

Profile_ID	SOC stock [t/ha]	Properties	Land form	Topographic wetness index	Avg annual temperature	Other covariates
AB1						
BJ12						
…	…					

ISRIC World Soil Information offers ca. 130 different national covariates for download.

This table is then used for the main upscaling procedures (See the Upscaling Methods Section)

3.8 PREPARATION OF A SOIL PROPERTY TABLE FOR SPATIAL STATISTICS

The upscaling procedures (section 4) depend on the rationale, that the accumulation of local soil carbon concentrations and stocks (and also other properties) depends on influential parameters for which spatial data are available, such as climate, soil type, parent material, slope, management. Any parameter in the table of local soil properties, for which a spatial layer is available, may be included in the final table. Other covariates will be added in section 3. An example is the clay content, which may be derived from a soil type or parent rock map.

Table 3.2 Input table for spatial analysis (spatial SOC prediction table)

Profile_ID	SOC stock [t/ha] 0-30	SOC stock [t/ha] litter	Soil type	Clay [%]	Other soil properties
AB1					
BJ12					
…	…				

3) BULK DENSITY CAN ALSO BE ESTIMATED USING ORGANIC CARBON AND TEXTURAL DATA.

IN CASE THIS TABLE IS PREPARED FOR DIFFERENT DEPTHS, 0-10 CM, 10-30 CM, AND IF THE HOST INSTITUTION INTENDS TO DEVELOP DIFFERENT SPATIAL MODELS FOR DIFFERENT DEPTHS (E.G. SEPARATE SPATIAL PREDICTION MODEL FOR LITTER AND MINERAL SOIL 0-30), THEN THE SEPARATE GRIDS HAVE TO BE ADDED.

3.9 References

Bontemps, S., Defourny, P., Van Bogaert, E., Arino, O., Kalogirou, V., Perez, J.R., 2011. GLOBCOVER 2009-Product Description and Validation Report.

Gupta, S. (2015). A Multiple Regression Technique in Data Mining. International Journal of Computer Applications, 126(5).

Hengl, T. (2009). A practical guide to geostatistical mapping (Vol. 52). Hengl.

Hijmans R J, Cameron S E, Parra J L, Jones P G, Jarvis A, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Hoffmann, J. P., & Shafer, K. (2005). Linear regression analysis: Assumptions and applications. Department of Sociology Brigham Young University.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. Applied Statistics, 300-303.

Vasques, G. M., Grunwald, S., & Sickman, J. O. (2009). Modeling of soil organic carbon fractions using visible–near-infrared spectroscopy. Soil Science Society of America Journal, 73(1), 176-184.

Wong, W.V.C., S. Tsuyuki, K. Ioki and M.-H. Phua (2014). Accuracy assessment of global topographic data (SRTM & ASTER GDEM) in comparison with lidar for tropical montane forest. Conference Paper, October 2014. The 35th Asian Conference on Remote Sensing 2014, At Nay Pyi Taw, Myanmar. *https://www.researchgate.net/publication/267811614_Accuracy_assessment_of_global_topographic_data_SRTM_ASTER_GDEM_in_comparison_with_lidar_for_tropical_montane_forest*

4. SETTING-UP THE SOFTWARE ENVIRONMENT

This cookbook focuses on soil organic carbon modelling using open source digital mapping tools. The instructions and screen captures in this section will guide you through installing and manually configuring the software to be used for digital soil mapping procedures for Microsoft Windows desktop platform. Instructions for the other platforms (Linux Flavours, MacOS) can be found through free online resources.

4.1 USE OF 'R' , RSTUDIO AND R PACKAGES

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

4.1.1 OBTAINING AND INSTALLING R

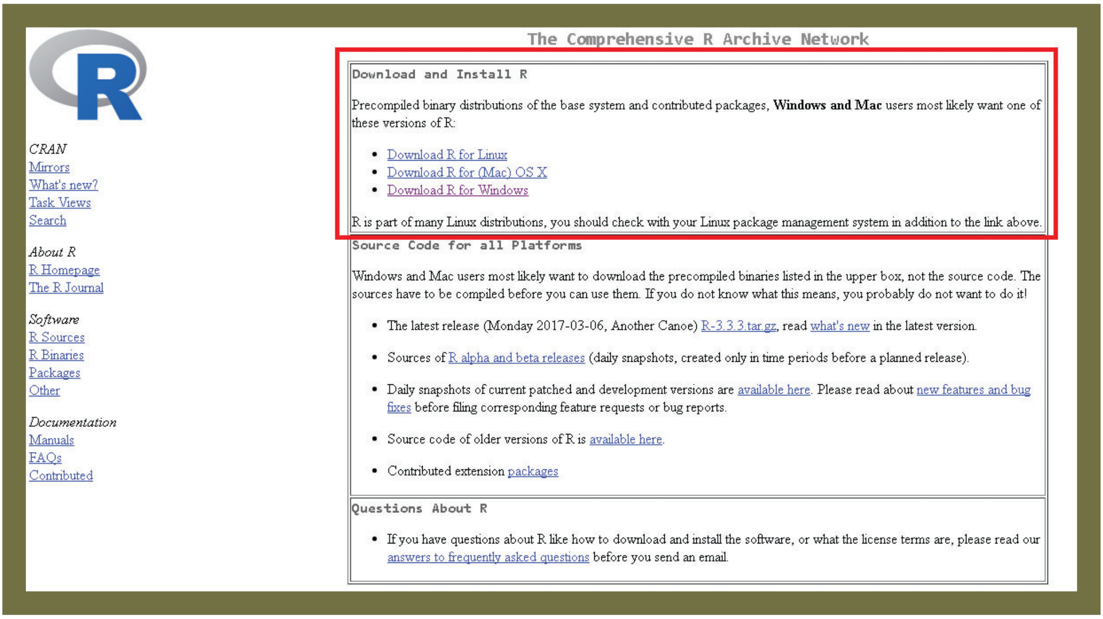
4.1.2 INSTALLATION

STEP

STEP 1: GO TO *HTTPS://CLOUD.R-PROJECT.ORG/INDEX.HTML*

STEP

STEP 2: PICK AN INSTALLATION FILE FOR YOUR PLATFORM



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

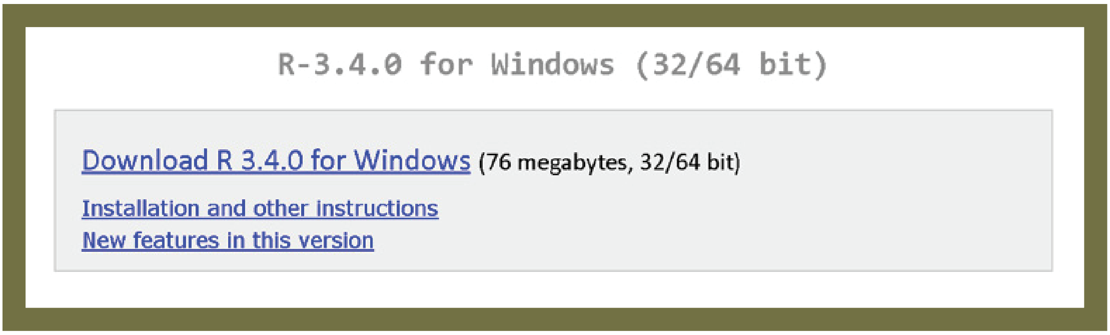
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Monday 2017-03-06, Another Canoe) R-3.3.3.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

STEP 3: FOR WINDOWS OS YOU CAN CLICK ON THE LINK "DOWNLOAD R FOR WINDOWS"

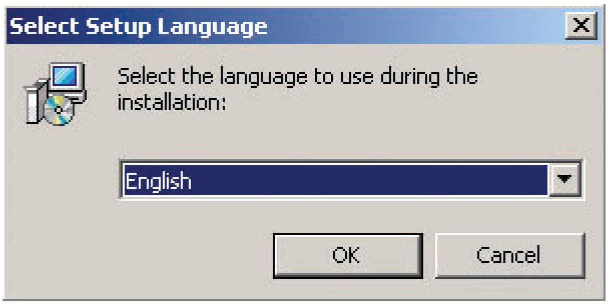


R-3.4.0 for Windows (32/64 bit)

Download R 3.4.0 for Windows (76 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

STEP 4: YOU CAN START THE INSTALLATION BY CLICKING ON THE DOWNLOADED .EXE FILE. R CAN BE INSTALLED IN MORE THAN 20 LANGUAGES.

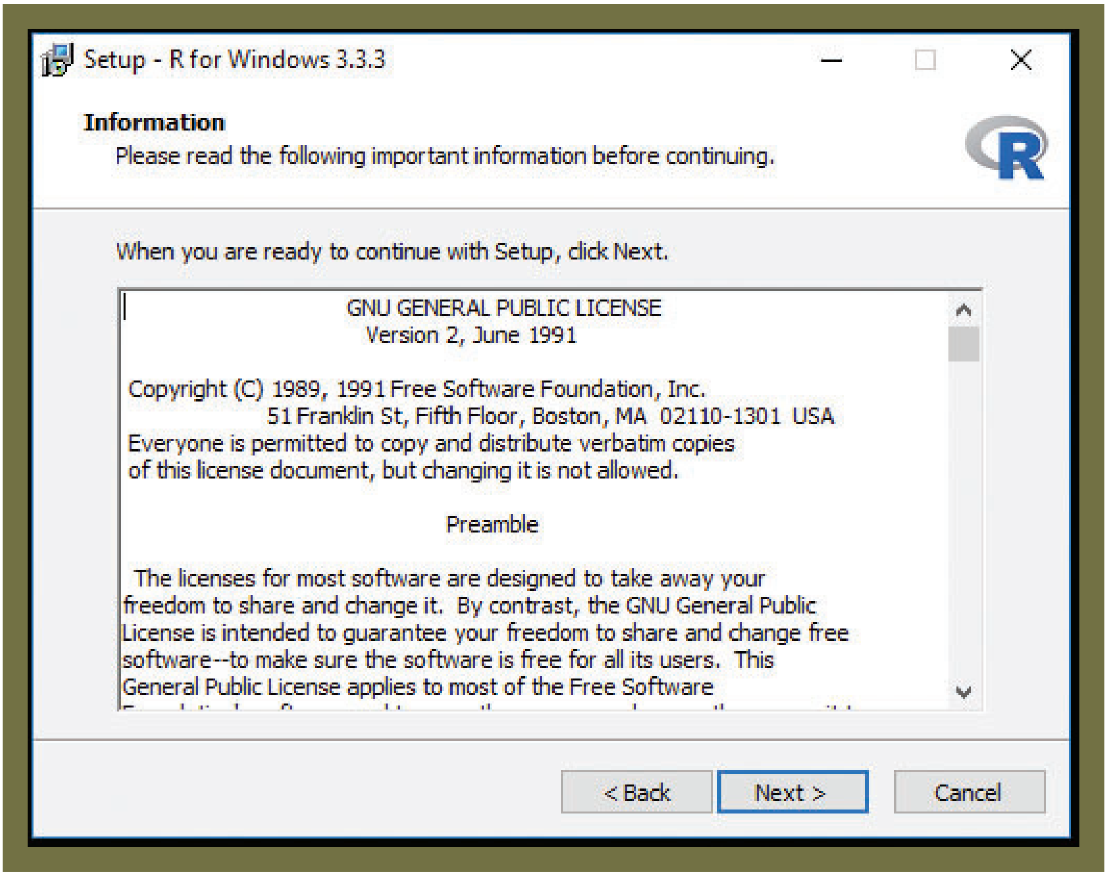


Select Setup Language

Select the language to use during the installation:

English

STEP 5: READ GNU-GPL LICENSE INFORMATION AND CLICK› NEXT



Setup - R for Windows 3.3.3

Information

Please read the following important information before continuing.

When you are ready to continue with Setup, click Next.

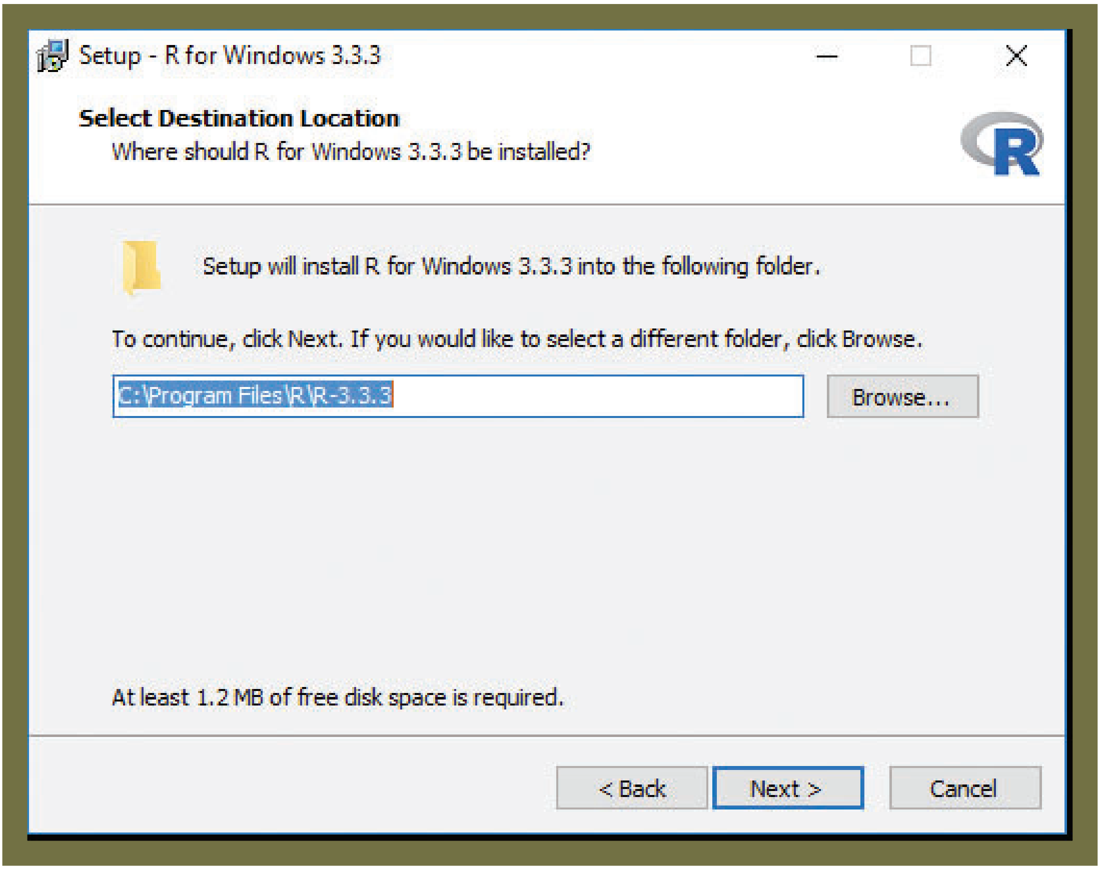
```
                    GNU GENERAL PUBLIC LICENSE
                       Version 2, June 1991

 Copyright (C) 1989, 1991 Free Software Foundation, Inc.
               51 Franklin St, Fifth Floor, Boston, MA  02110-1301  USA
 Everyone is permitted to copy and distribute verbatim copies
 of this license document, but changing it is not allowed.

                            Preamble

  The licenses for most software are designed to take away your
freedom to share and change it.  By contrast, the GNU General Public
License is intended to guarantee your freedom to share and change free
software--to make sure the software is free for all its users.  This
General Public License applies to most of the Free Software
```

STEP 6: SELECT DESTINATION LOCATION AND CLICK> NEXT



Setup - R for Windows 3.3.3

Select Destination Location

Where should R for Windows 3.3.3 be installed?

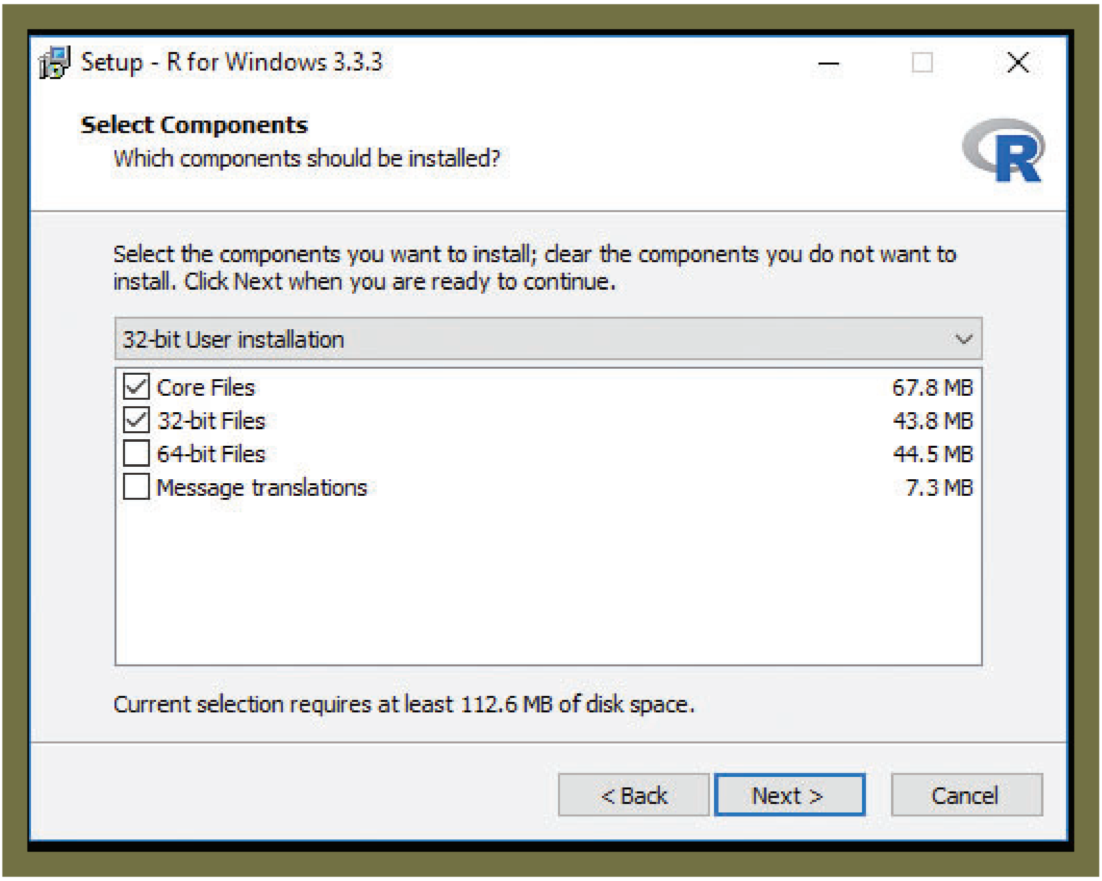
Setup will install R for Windows 3.3.3 into the following folder.

To continue, click Next. If you would like to select a different folder, click Browse.

C:\Program Files\R\R-3.3.3

At least 1.2 MB of free disk space is required.

STEP 7: SELECT COMPONENTS (LEAVE DEFAULTS) AND CLICK> NEXT



Setup - R for Windows 3.3.3

Select Components

Which components should be installed?

Select the components you want to install; clear the components you do not want to install. Click Next when you are ready to continue.

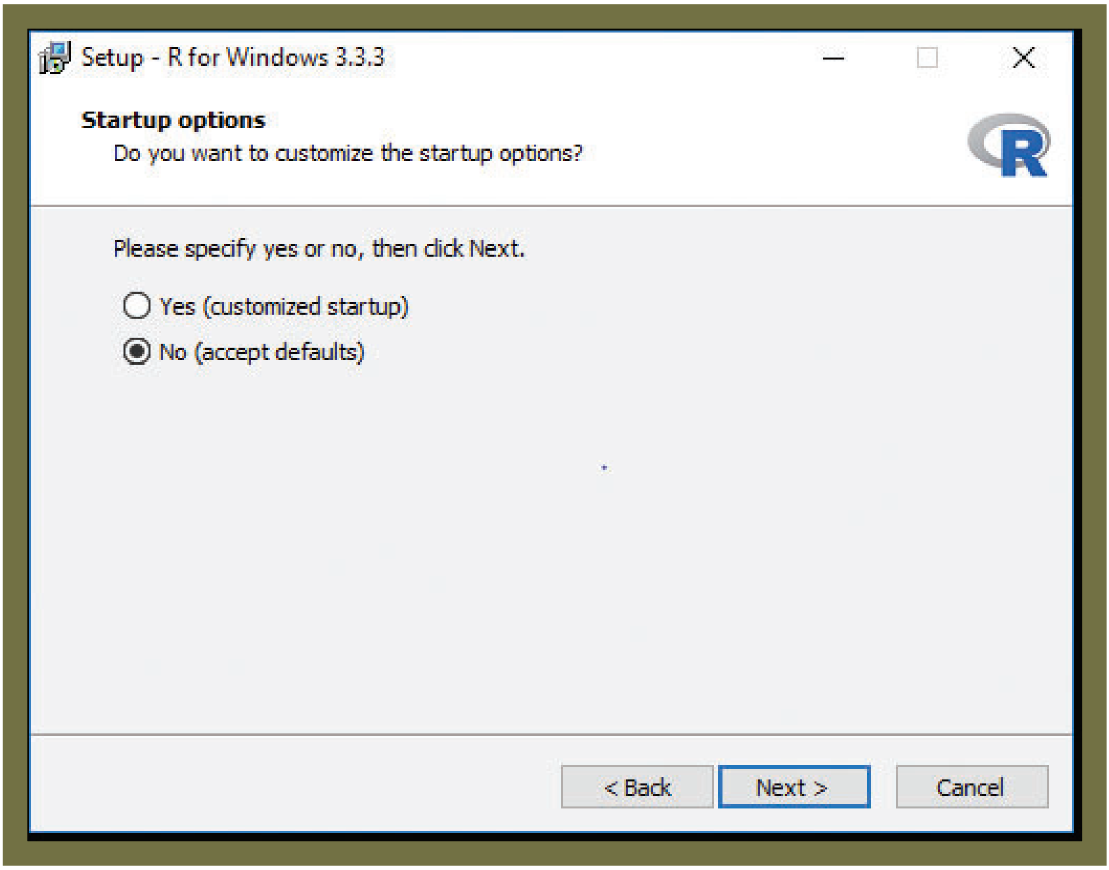
32-bit User installation

☑ Core Files	67.8 MB
☑ 32-bit Files	43.8 MB
□ 64-bit Files	44.5 MB
□ Message translations	7.3 MB

Current selection requires at least 112.6 MB of disk space.

STEP

STEP 8: STARTUP OPTIONS (LEAVE DEFAULTS) AND CLICK> NEXT



Setup - R for Windows 3.3.3

Startup options

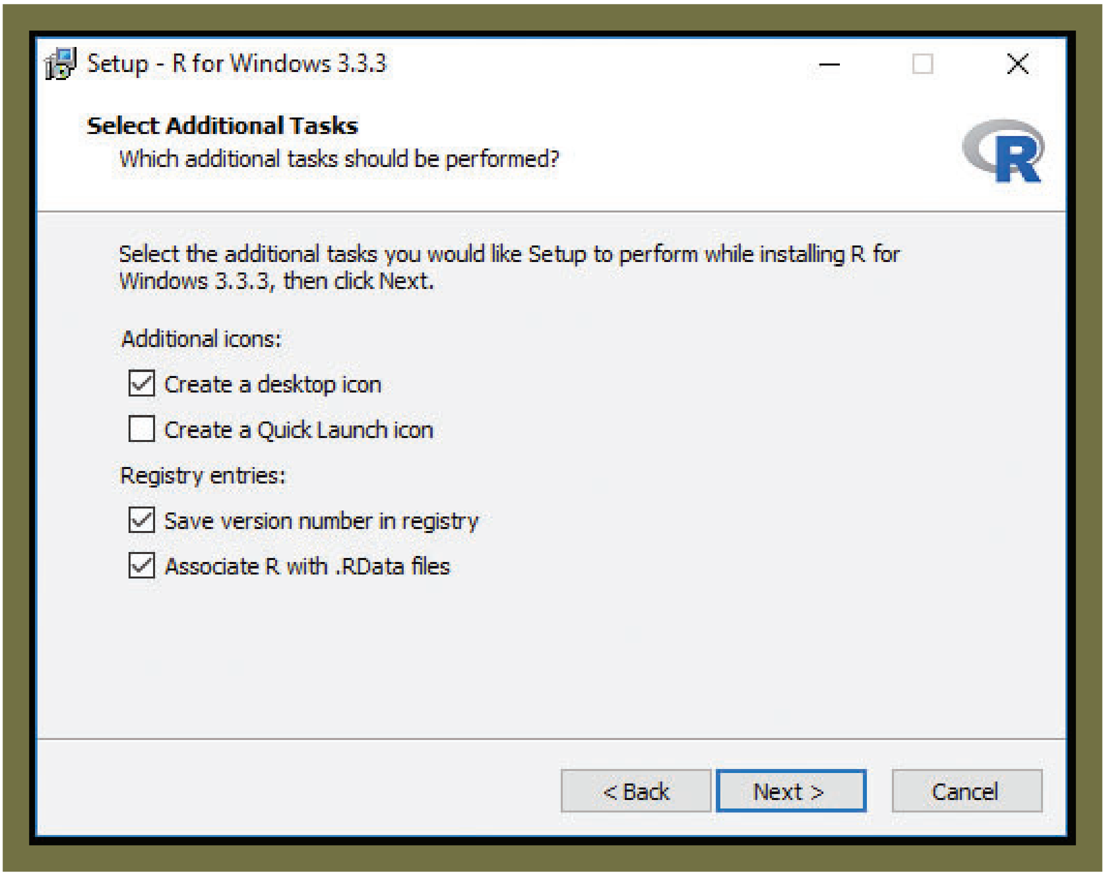
Do you want to customize the startup options?

Please specify yes or no, then click Next.

○ Yes (customized startup)

● No (accept defaults)

STEP 9: SELECT ADDITIONAL TASKS (LEAVE DEFAULTS) AND CLICK› NEXT



Setup - R for Windows 3.3.3

Select Additional Tasks

Which additional tasks should be performed?

Select the additional tasks you would like Setup to perform while installing R for Windows 3.3.3, then click Next.

Additional icons:

☑ Create a desktop icon

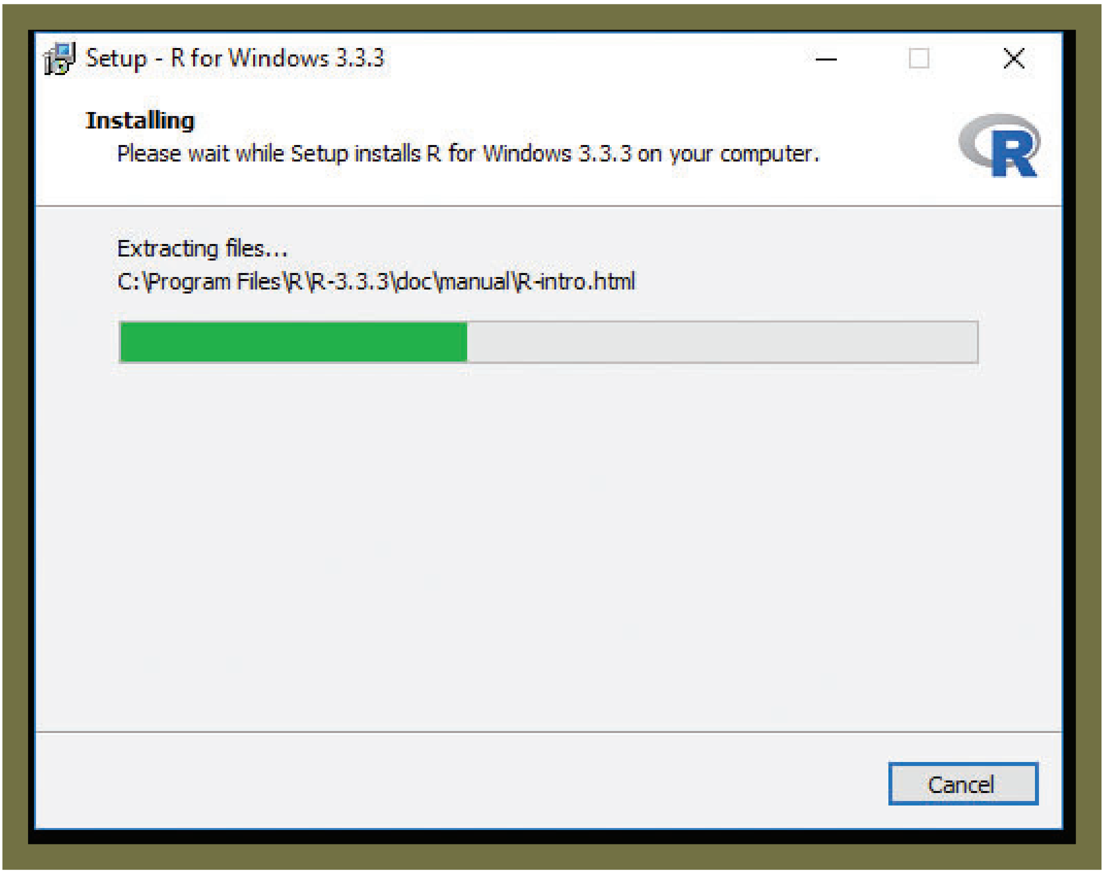
□ Create a Quick Launch icon

Registry entries:

☑ Save version number in registry

☑ Associate R with .RData files

STEP 10: INSTALLER WILL INSTALL R ON YOUR COMPUTER



Setup - R for Windows 3.3.3

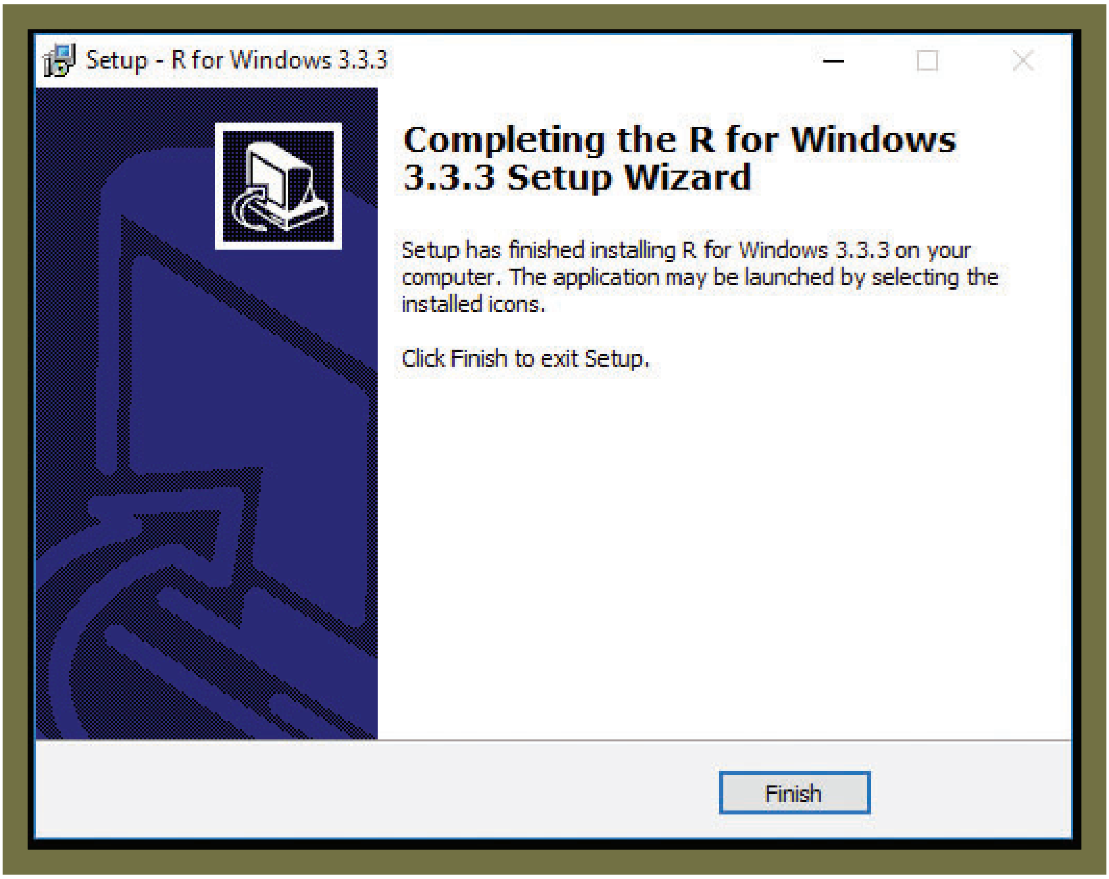
Installing

Please wait while Setup installs R for Windows 3.3.3 on your computer.

Extracting files...

C:\Program Files\R\R-3.3.3\doc\manual\R-intro.html

STEP 11: AFTER INSTALLATION FINISHED CLICK > FINISH



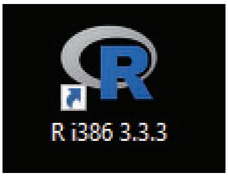
Setup - R for Windows 3.3.3

Completing the R for Windows 3.3.3 Setup Wizard

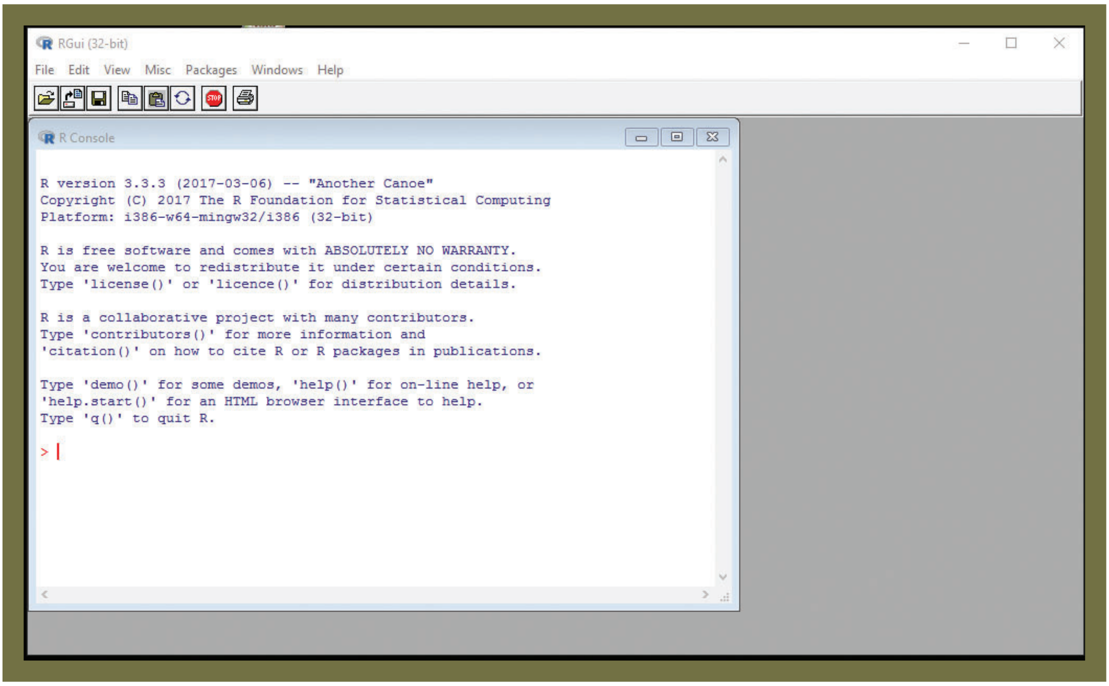
Setup has finished installing R for Windows 3.3.3 on your computer. The application may be launched by selecting the installed icons.

Click Finish to exit Setup.

STEP 12: CLICK ON THE "R" ICON ON THE DESKTOP OR IN THE START MENU



STEP 13: R NOW SHOULD START



```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

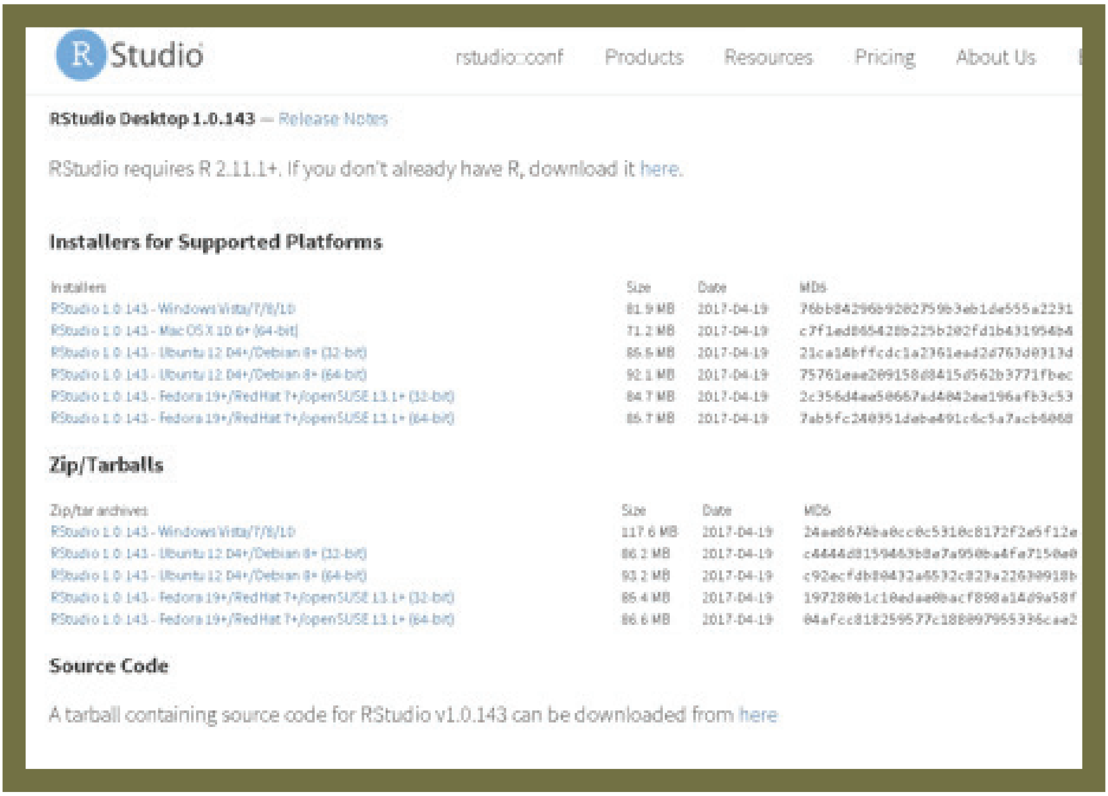
4.2 OBTAINING AND INSTALLING R STUDIO

Beginners will find 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. In this cookbook we would like focus on a GUI which makes R easier to use. R Studio's Open Source Edition can be downloaded at *https://www.rstudio.com/products/rstudio/download/* . On the download page, "RStudio Desktop, Open Source License" option should be selected.

4.2.1 INSTALLATION

You can follow very similar steps to install RStudio.

STEP 1: PICK THE VERSION THAT IS LISTED AS RECOMMENDED FOR YOUR SYSTEM. INSTALLING SHOULD BE STRAIGHTFORWARD.



RStudio Desktop 1.0.143 — Release Notes

RStudio requires R 2.11.1+. If you don't already have R, download it here.

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.0.143 - Windows Vista/7/8/10	81.9 MB	2017-04-19	76bb84296b9282759b3eb1de555a2231
RStudio 1.0.143 - Mac OS X 10.6+ (64-bit)	71.2 MB	2017-04-19	c7f1ed865428b225b202fd1b431954b4
RStudio 1.0.143 - Ubuntu 12.04+/Debian 8+ (32-bit)	85.6 MB	2017-04-19	21ca14bffcdc1a2361ead2d763d0313d
RStudio 1.0.143 - Ubuntu 12.04+/Debian 8+ (64-bit)	92.1 MB	2017-04-19	75761eae209158d8415d562b3771fbec
RStudio 1.0.143 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	84.7 MB	2017-04-19	2c356d4ae50667ad4042ae196afb3c53
RStudio 1.0.143 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	85.7 MB	2017-04-19	7ab5fc240351daba491c6c5a7acb6068

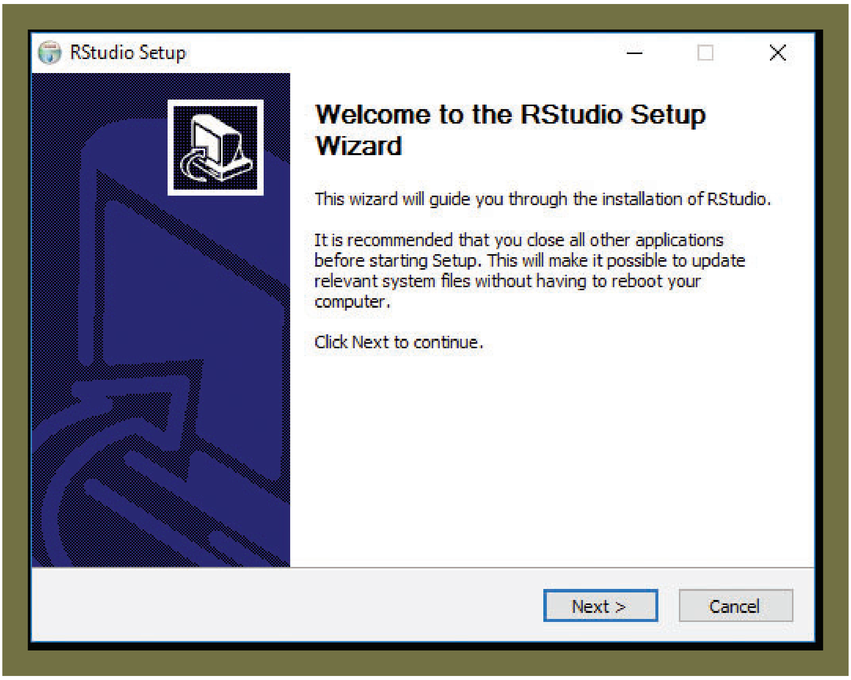
Zip/Tarballs

Zip/tar archives	Size	Date	MD5
RStudio 1.0.143 - Windows Vista/7/8/10	117.6 MB	2017-04-19	24ae8674ba0cc0c5310c8172f2e5f12e
RStudio 1.0.143 - Ubuntu 12.04+/Debian 8+ (32-bit)	86.2 MB	2017-04-19	c4444d8159463b8a7a950ba4fa7150a0
RStudio 1.0.143 - Ubuntu 12.04+/Debian 8+ (64-bit)	93.2 MB	2017-04-19	c92ecfdb80432a6532c823a22630918b
RStudio 1.0.143 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	85.4 MB	2017-04-19	197280b1c10edae0bacf898a14d9a58f
RStudio 1.0.143 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	86.6 MB	2017-04-19	04afcc8182595 77c188897955336cae2

Source Code

A tarball containing source code for RStudio v1.0.143 can be downloaded from here

STEP 2: RUN THE RSTUDIO INSTALLER BY CLICKING ON THE DOWNLOADED .EXE FILE.



RStudio Setup

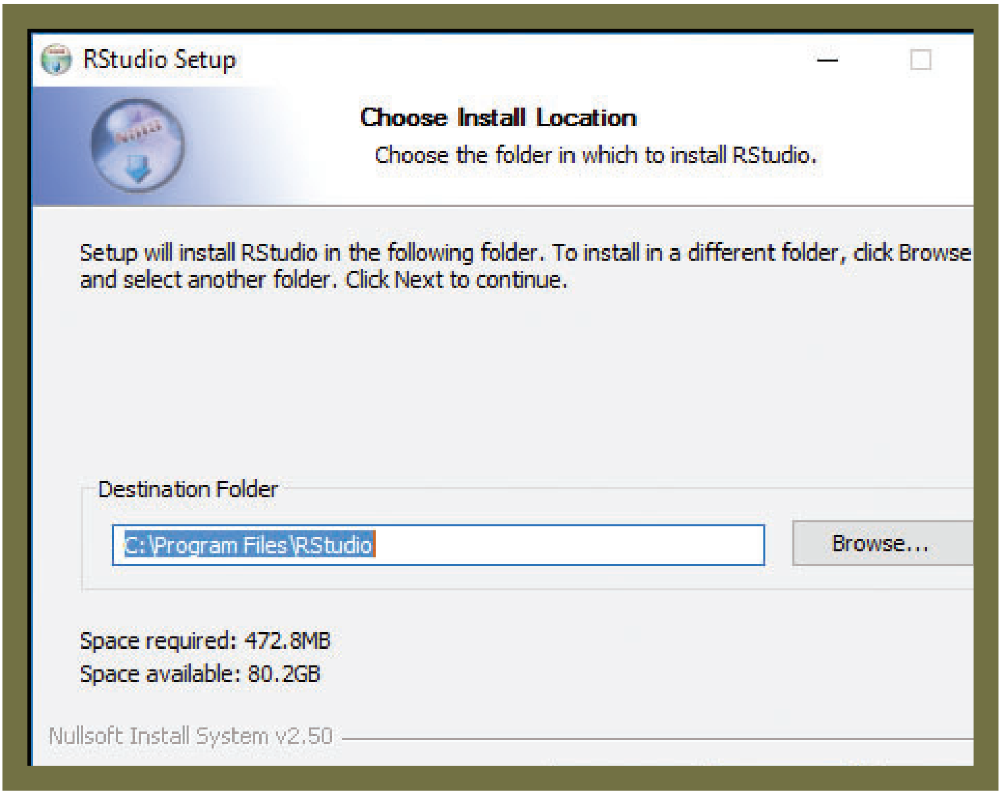
Welcome to the RStudio Setup Wizard

This wizard will guide you through the installation of RStudio.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

STEP 3: SELECT INSTALL LOCATION AND CLICK > NEXT.



RStudio Setup

Choose Install Location

Choose the folder in which to install RStudio.

Setup will install RStudio in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

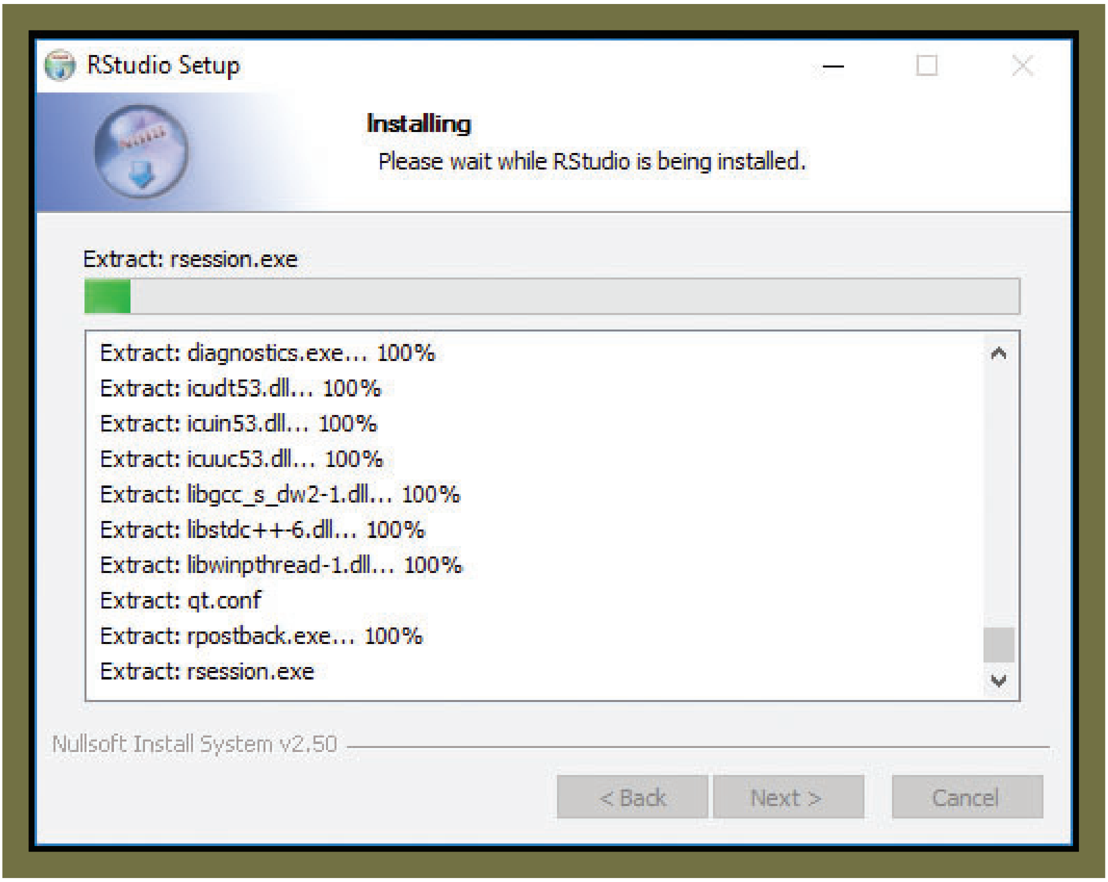
Destination Folder

C:\Program Files\RStudio

Space required: 472.8MB
Space available: 80.2GB

Nullsoft Install System v2.50

STEP 4: THE INSTALLER WILL INSTALL THE SOFTWARE.



RStudio Setup

Installing

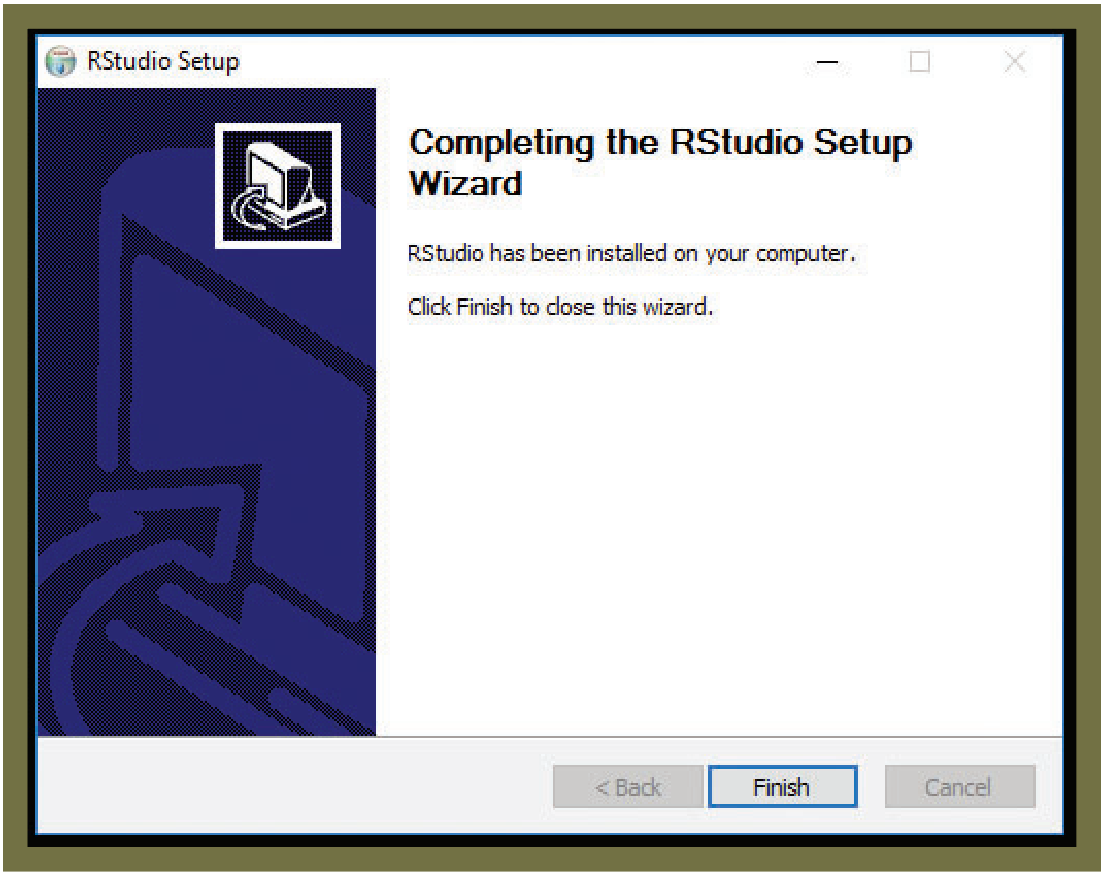
Please wait while RStudio is being installed.

Extract: rsession.exe

```
Extract: diagnostics.exe... 100%
Extract: icudt53.dll... 100%
Extract: icuin53.dll... 100%
Extract: icuuc53.dll... 100%
Extract: libgcc_s_dw2-1.dll... 100%
Extract: libstdc++-6.dll... 100%
Extract: libwinpthread-1.dll... 100%
Extract: qt.conf
Extract: rpostback.exe... 100%
Extract: rsession.exe
```

Nullsoft Install System v2.50

STEP 5: CLICK> FINISH WHEN THE WIZARD FINISHED THE INSTALLATION

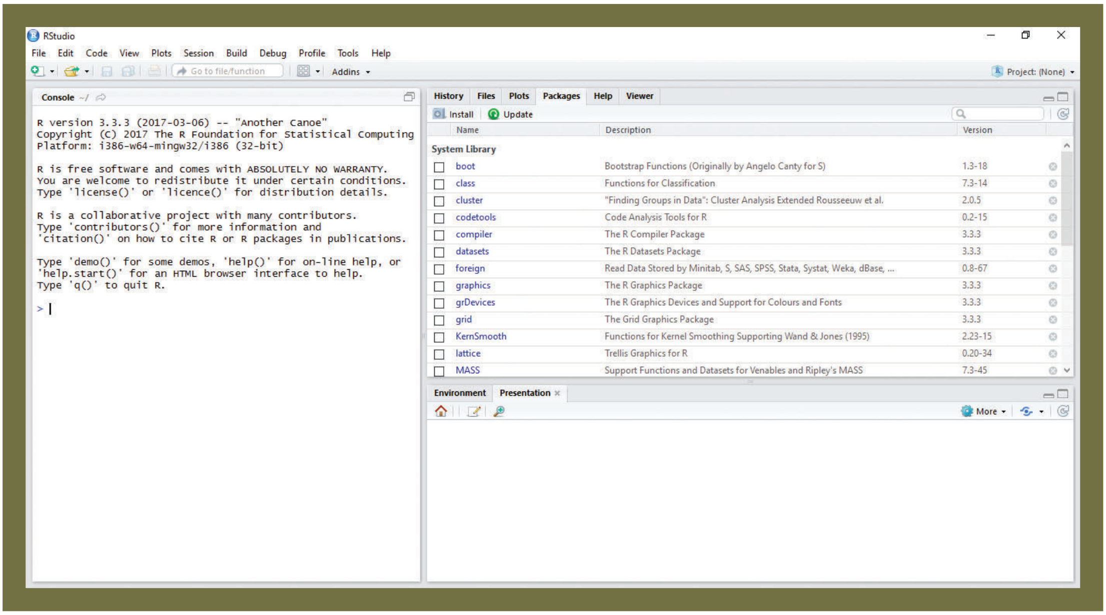


Completing the RStudio Setup Wizard

RStudio has been installed on your computer.

Click Finish to close this wizard.

STEP 7: AFTER THE INSTALLATION,YOU ONLY NEED TO OPEN RSTUDIO, BECAUSE IT WILL AUTOMATICALLY ALSO START R.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console ~/
History Files Plots Packages Help Viewer
Install Update
Environment Presentation
More

```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

	Name	Description	Version
System Library			
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18
□	class	Functions for Classification	7.3-14
□	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.5
□	codetools	Code Analysis Tools for R	0.2-15
□	compiler	The R Compiler Package	3.3.3
□	datasets	The R Datasets Package	3.3.3
□	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...	0.8-67
□	graphics	The R Graphics Package	3.3.3
□	grDevices	The R Graphics Devices and Support for Colours and Fonts	3.3.3
□	grid	The Grid Graphics Package	3.3.3
□	KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)	2.23-15
□	lattice	Trellis Graphics for R	0.20-34
□	MASS	Support Functions and Datasets for Venables and Ripley's MASS	7.3-45

4.2.2 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*

4.3. R PACKAGES

- When you download R, you get that ``base" R system
- The R system comes with basics; implements the R language
- R becomes so useful with the large collection of packages that extend the basic functionality of R
- R packages are developed by the R community

4.3.1 FINDING R PACKAGES

The primary source for the R packages is CRAN's official website. For spatial applications, many packages are available. You can obtain information about the available packages on CRAN with the available.packages() function. The function returns a matrix of details corresponding to packages currently available at one or more repositories. However, there are more than 10000 packages in the CRAN repository.

An easier way to browse the list of packages is using the Task Views link, which groups together many packages related to a given topic.

DEFINITION

HTTP://CRAN.R-PROJECT.ORG/WEB/VIEWS/

For example, the Task View for analysis of Spatial Data can be accessed at: *https://CRAN.R-project.org/view=Spatial.*

The following code installs the "ggplot2" package from CRAN

```
> install.packages("ggplot2")
```

The packages can be installed also using the graphical user interface.

CLICK "PACKAGES" IN THE TOP MENU THEN CLICK "INSTALL PACKAGE(S)".
OR USE THE APPROPRIATE TAB ON THE BOTTOM RIGHT, BY DEFAULT.

4.3.2 MOST USED R PACKAGES FOR DIGITAL SOIL MAPPING

As was previously mentioned, R is extensible trough packages. R packages are collections of R functions, data, documentation and compiled code easy to share with others. They are more than 10000 R packages available at the Comprehensive R Archive Network (CRAN) (cran.r-project.org). In the following subsections we are going to present the most used packages related with soil property mapping.

Soil science and Pedometrics

aqp: Algorithms for quantitative pedology. *http://cran.r-project.org/web/ packages/ aqp/index.html.* A collection of algorithms related to modeling of soil resources, soil classification, soil profile aggregation, and visualization.

GSIF: Global soil information facility. *http://cran.r-project.org/web/packages/GSIF/index.html*. Tools, functions and sample datasets for digital soil mapping. Global Soil Information Facilities - tools (standards and functions) and sample datasets for global soil mapping.

soiltexture: "The Soil Texture Wizard" is a set of R functions designed to produce texture triangles (also called texture plots, texture diagrams, texture ternary plots), classify and transform soil textures data. These functions virtually allows to plot any soil texture triangle (classification) into any triangle geometry (isosceles, right-angled triangles, etc.). This set of function is expected to be useful to people using soil textures data from different soil texture classification or different particle size systems. Many (&gt; 15) texture triangles from all around the world are predefined in the package. A simple text based graphical user interface is provided: soiltexture_gui().

Spatial Analysis

sp: *http://cran.r-project.org/web/packages/sp/index.html*. A package that provides classes and methods for spatial data. The classes document where the spatial location information resides, for 2D or 3D data.

raster: *http://cran.r-project.org/web/packages/raster/index.html*. Reading, writing, manipulating, analyzing and modeling of gridded spatial data. The package implements basic and high-level functions and processing of very large files is supported.

rgdal: *http://cran.r-project.org/web/packages/rgdal/index.html*. Provides bindings to Frank Warmerdam's Geospatial Data Abstraction Library (GDAL).

RSAGA: *http://cran.r-project.org/web/packages/RSAGA/index.html*. RSAGA provides access to geocomputing and terrain analysis functions of SAGA GIS *http://www.saga-gis.org/en/index.html* from within R by running the command line version of SAGA.

Modeling

caret: *http://cran.r-project.org/web/packages/caret/index.html*. Extensive range of functions for training and plotting classification and regression models.

Cubist: *http://cran.r-project.org/web/packages/Cubist/index.html*. Regression modeling using rules with added instance-based corrections. Cubist models were developed by Ross Quinlan.

C5.0: *http://cran.r-project.org/web/packages/C50/index.html*. C5.0 decision trees and rule-based models for pattern recognition. Another model structure developed by Ross Quinlan.

gam: *http://cran.r-project.org/web/packages/gam/index.html*. Functions for fitting and working with generalized additive models.

nnet: *http://cran.r-project.org/web/packages/nnet/index.html*. Software for feed-forward neural networks with a single hidden layer, and for multinomial log-linear models.

gstat: http://cran.r-project.org/web/packages/gstat/. Variogram modelling; simple, ordinary and universal point or block (co)kriging, sequential Gaussian or indicator (co)simulation; variogram and variogram map plotting utility functions.

ithir: A collection of functions and algorithms specific to pedometrics. The package was developed by Brendan Malone at the University of Sydney.

Mapping and plotting

Both raster and sp have handy functions for plotting spatial data. Besides using the base plotting functionality, another useful plotting package is ggplot2.

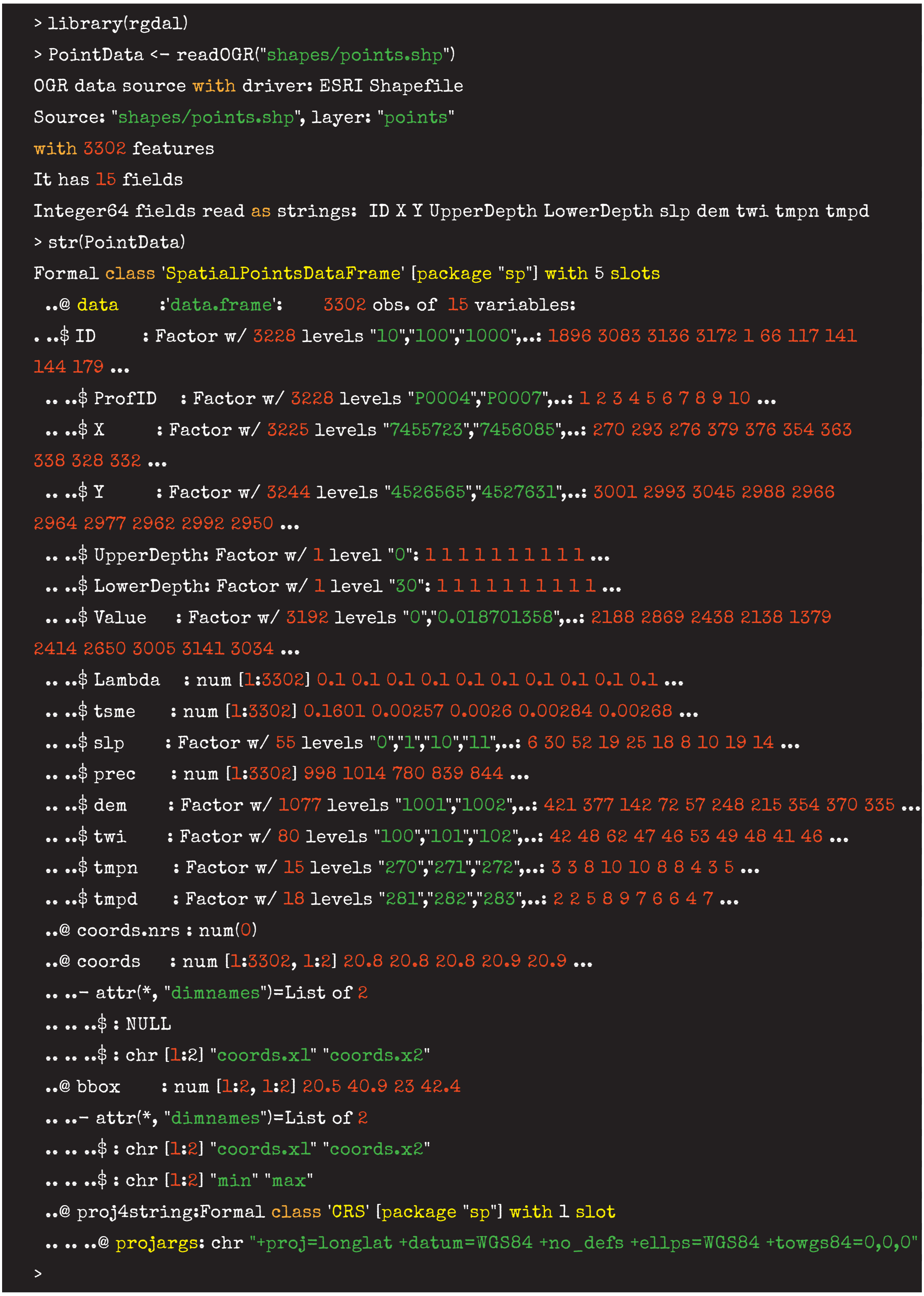
plotKML: Writes sp-class, spacetime-class, raster-class and similar spatial and spatio-temporal objects to KML following some basic cartographic rules.

4.4 R AND SPATIAL DATA

R has a large and growing number of spatial data packages. We recommend taking a quick browse on R's official website to see the spatial packages available: *http://cran.r-project.org/web/views/Spatial.html*

4.4.1 READING SHAPEFILES

The ESRI's Shapefile format is widely used for storing vector-based spatial data (i.e., points, lines, polygons). This example demonstrates use of rgdal package that provides functions for reading and/or writing shapefiles.



```
> library(rgdal)
> PointData <- readOGR("shapes/points.shp")
OGR data source with driver: ESRI Shapefile
Source: "shapes/points.shp", layer: "points"
with 3302 features
It has 15 fields
Integer64 fields read as strings:  ID X Y UpperDepth LowerDepth slp dem twi tmpn tmpd
> str(PointData)
Formal class 'SpatialPointsDataFrame' [package "sp"] with 5 slots
 ..@ data      :'data.frame':      3302 obs. of  15 variables:
. ..$ ID       : Factor w/ 3228 levels "10","100","1000",..: 1896 3083 3136 3172 1 66 117 141
144 179 ...
 .. ..$ ProfID    : Factor w/ 3228 levels "P0004","P0007",..: 1 2 3 4 5 6 7 8 9 10 ...
 .. ..$ X         : Factor w/ 3225 levels "7455723","7456085",..: 270 293 276 379 376 354 363
338 328 332 ...
 .. ..$ Y         : Factor w/ 3244 levels "4526565","4527631",..: 3001 2993 3045 2988 2966
2964 2977 2962 2992 2950 ...
 .. ..$ UpperDepth: Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 .. ..$ LowerDepth: Factor w/ 1 level "30": 1 1 1 1 1 1 1 1 1 1 ...
 .. ..$ Value     : Factor w/ 3192 levels "0","0.018701358",..: 2188 2869 2438 2138 1379
2414 2650 3005 3141 3034 ...
 .. ..$ Lambda    : num [1:3302] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 .. ..$ tsme      : num [1:3302] 0.1601 0.00257 0.0026 0.00284 0.00268 ...
 .. ..$ slp       : Factor w/ 55 levels "0","1","10","11",..: 6 30 52 19 25 18 8 10 19 14 ...
 .. ..$ prec      : num [1:3302] 998 1014 780 839 844 ...
 .. ..$ dem       : Factor w/ 1077 levels "1001","1002",..: 421 377 142 72 57 248 215 354 370 335 ...
 .. ..$ twi       : Factor w/ 80 levels "100","101","102",..: 42 48 62 47 46 53 49 48 41 46 ...
 .. ..$ tmpn      : Factor w/ 15 levels "270","271","272",..: 3 3 8 10 10 8 8 4 3 5 ...
 .. ..$ tmpd      : Factor w/ 18 levels "281","282","283",..: 2 2 5 8 9 7 6 6 4 7 ...
 ..@ coords.nrs : num(0)
 ..@ coords     : num [1:3302, 1:2] 20.8 20.8 20.8 20.9 20.9 ...
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : NULL
 .. .. ..$ : chr [1:2] "coords.x1" "coords.x2"
 ..@ bbox       : num [1:2, 1:2] 20.5 40.9 23 42.4
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:2] "coords.x1" "coords.x2"
 .. .. ..$ : chr [1:2] "min" "max"
 ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
 .. .. ..@ projargs: chr "+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"
>
```

We may want to use these data in other GIS environments such as ArcGIS, QGIS, SAGA GIS etc. This means we need to export the SpatialPointsDataFrame to an appropriate spatial data format such as a shapefile. “rgda”l is again used for this via the writeOGR() function. To export the data set as a shapefile:

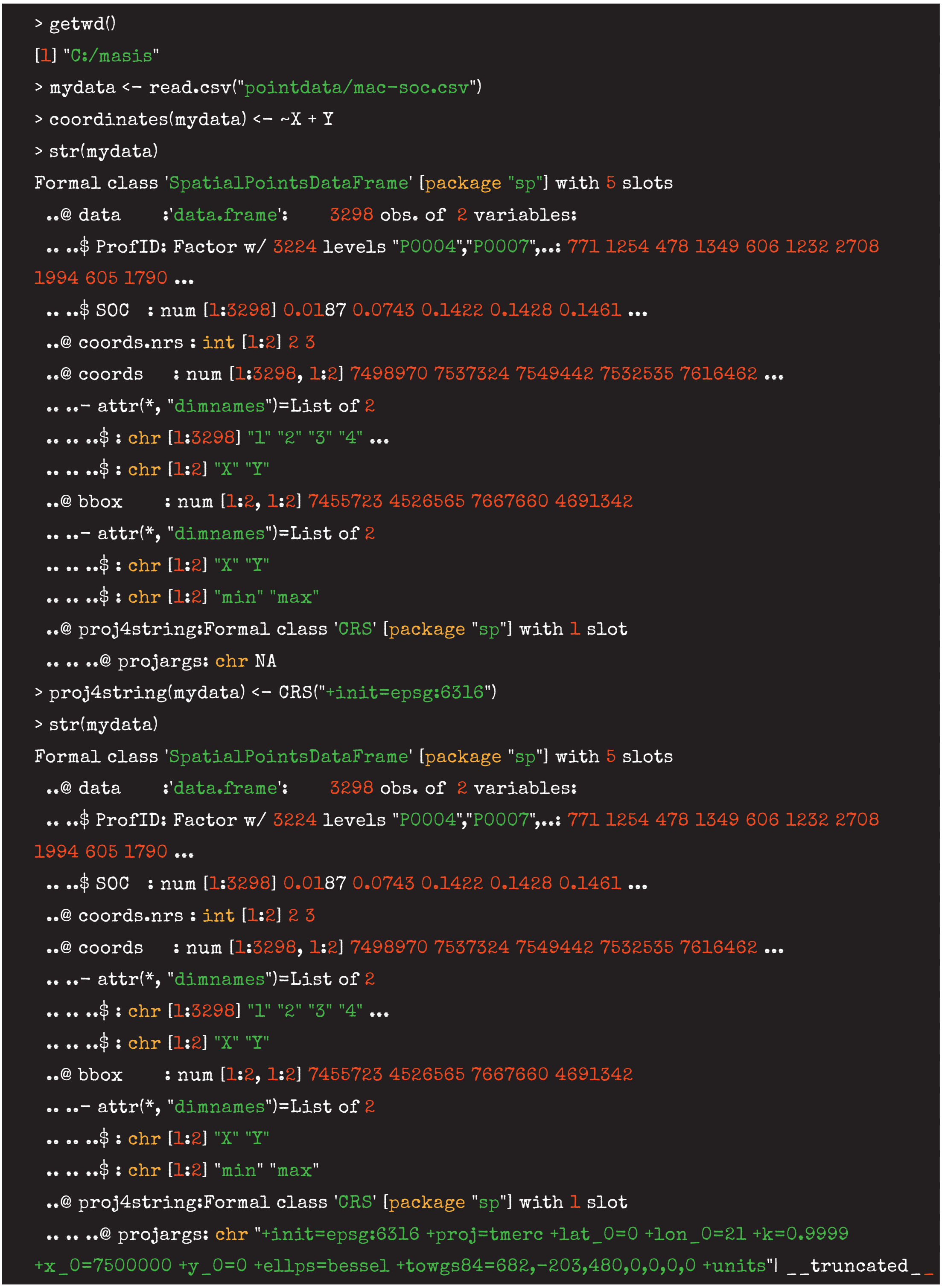
```
> writeOGR(PointData, ".", "pointdata-shape", "ESRI Shapefile")
# Check your working directory for presence of this file
```

4.4.2 COORDINATE REFERENCE SYSTEMS (CRS) IN R

We need to define the CRS (Coordinate Reference System) to be able to perform any sort of spatial analysis in R. To clearly tell R this information we define the CRS which describes a reference system in a way understood by the PROJ.4 projection library *http://trac.osgeo.org/proj*

An interface to the PROJ.4 library is available in the rgdal package. Alternative to using Proj4 character strings, we can use the corresponding yet simpler EPSG code (European Petroleum Survey Group). “rgdal” also recognizes these codes. If you are unsure of the Proj4 or EPSG code for the spatial data that you have, but know the CRS, you should consult *http://spatialreference.org* for assistance.

The following example shows how you can create a spatial object from a .csv file. We can use the coordinates() function from the sp package to define which columns in the data frame refer to actual spatial coordinates—here the coordinates are listed in columns X and Y.



```
> getwd()
[1] "C:/masis"
> mydata <- read.csv("pointdata/mac-soc.csv")
> coordinates(mydata) <- ~X + Y
> str(mydata)
Formal class 'SpatialPointsDataFrame' [package "sp"] with 5 slots
 ..@ data      :'data.frame':     3298 obs. of  2 variables:
 .. ..$ ProfID: Factor w/ 3224 levels "P0004","P0007",..: 771 1254 478 1349 606 1232 2708
1994 605 1790 ...
 .. ..$ SOC   : num [1:3298] 0.0187 0.0743 0.1422 0.1428 0.1461 ...
 ..@ coords.nrs : int [1:2] 2 3
 ..@ coords     : num [1:3298, 1:2] 7498970 7537324 7549442 7532535 7616462 ...
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:3298] "1" "2" "3" "4" ...
 .. .. ..$ : chr [1:2] "X" "Y"
 ..@ bbox      : num [1:2, 1:2] 7455723 4526565 7667660 4691342
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:2] "X" "Y"
 .. .. ..$ : chr [1:2] "min" "max"
 ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
 .. .. ..@ projargs: chr NA
> proj4string(mydata) <- CRS("+init=epsg:6316")
> str(mydata)
Formal class 'SpatialPointsDataFrame' [package "sp"] with 5 slots
 ..@ data      :'data.frame':     3298 obs. of  2 variables:
 .. ..$ ProfID: Factor w/ 3224 levels "P0004","P0007",..: 771 1254 478 1349 606 1232 2708
1994 605 1790 ...
 .. ..$ SOC   : num [1:3298] 0.0187 0.0743 0.1422 0.1428 0.1461 ...
 ..@ coords.nrs : int [1:2] 2 3
 ..@ coords     : num [1:3298, 1:2] 7498970 7537324 7549442 7532535 7616462 ...
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:3298] "1" "2" "3" "4" ...
 .. .. ..$ : chr [1:2] "X" "Y"
 ..@ bbox      : num [1:2, 1:2] 7455723 4526565 7667660 4691342
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:2] "X" "Y"
 .. .. ..$ : chr [1:2] "min" "max"
 ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
 .. .. ..@ projargs: chr "+init=epsg:6316 +proj=tmerc +lat_0=0 +lon_0=21 +k=0.9999
+x_0=7500000 +y_0=0 +ellps=bessel +towgs84=682,-203,480,0,0,0,0 +units"| __truncated__
```

4.4.3 WORKING WITH RASTERS

Most of the functions for handling raster data are available in the raster package. There are functions for reading and writing raster files from and to different formats. In digital soil mapping we mostly work with data in table format and then rasterise this data so that we can make a continuous map. For doing this in R environment, we will load raster data in a data frame. This data is a digital elevation model provided by ISRIC for FYROM (Chapter 3.6).

```
> DEM <- raster("cov/DEMENV5.tif")
> str(DEM)
Formal class 'RasterLayer' [package "raster"] with 12 slots
 ..@ file   :Formal class '.RasterFile' [package "raster"] with 13 slots
 .. .. ..@ name                 : chr "C:\\masis\\cov\\DEMENV5.tif"
 .. .. ..@ datanotation         : chr "INT2S"
 .. .. ..@ byteorder            : chr "little"
 .. .. ..@ nodatavalue          : num -Inf
 ... | __truncated__
```

We may want to export this raster to a suitable format to work in a standard GIS environment. See the help file for writeRaster (> ?writeRaster) to get information regarding the supported grid types that data can be exported. Here, we will export our raster to ESRI Ascii, as it is a common and universal raster format.

```
> writeRaster(DEM, filename = "mac-dem.asc",format = "ascii", overwrite = TRUE)
#Check your working space for presence of the ascii file!
```

We may also want to export our mac.dem to KML file using the KML function. Note that we need to re-project the data to WGS84 geographic. The raster re-projection is performed using the projectRaster function. Look at the help file for this (> ?projectRaster)KML is a handy function from raster for exporting grids to kml format.

```
> KML(DEM, "DEM.kml", col = rev(terrain.colors(255)),   overwrite = TRUE)
#Check your working space for presence of the kml file and try to open it in Google
EarthTM)
```

4.4 OTHER DSM RELATED SOFTWARE/TOOLS

QGIS: QGIS is available at: *http://www.qgis.org/en/site/forusers/download.html*

SAGA GIS: *https://sourceforge.net/projects/saga-gis/files/*

ArcGIS: 60 day trial can be downloaded at *http://www.esri.com/software/arcgis/* free-trial (needs registration)

4.5 References

Malone, B. P., Minasny, B., &amp; McBratney, A. B. (2016). Using R for Digital Soil Mapping, ISBN 978-3-319-44327-0

Walvoort, D.J.J., Brus, D.J., De Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Computers & Geosciences 36, 1261-1267.

R Development Core Team, 2016. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

Kuhn, M., 2015. A Short Introduction to the caret Package. Available at: https://cran.r-project.org/web/packages/caret/.

5. DEMONSTRATION DATASET

5.1. MASIS (CHAPTER 2 AND 6)

The sample dataset used for the spline function and the regression kriging has been extracted from the Macedonian Soil Information System Database (MASIS). The database contains around 4000 soil profiles with 11000 horizons. The data table overviews and the summary statistics are given below.

5.1.1 PROFILE TABLE

This table stores site level information (Profile no, Profile ID, X and Y Coordinates)

```
  # first 10 rows are shown here. head function can be used to display only the first few
  rows; #head(mydata, n=10) where head is the function, mydata is the object and n refers
  number of #rows to be shown

  > head(MASISProfiles, n=10)
     OBJECTID  ProfileNo     ProfID        X        Y
1        4318         43      P6517  7587329  4634774
2        4319         13      P5841  7528639  4582915
3        4320         35      P5842  7529064  4581678
4        4321         26      P5843  7529646  4584551
5        4322         20      P5844  7531852  4589188
6        4323         36      P5845  7530443  4589328
7        4324         18      P5846  7529855  4588422
8        4325         41      P5847  7530225  4587847
9        4326         21      P5848  7531111  4588093
10       4327         19      P5849  7532257  4589815
  ...
```

5.1.2 HORIZON DATA TABLE

This table stores information from the soil description, such as horizon name, horizon thickness, organic matter content, carbonate content, soil colour, etc.

```
# first 10 rows are shown here. head function can be used to display only the first few
rows; #head(mydata, n=10) where head is the function, mydata is the object and n refers
number of #rows to be shown

> MASISHorizons <- read.csv("HorizonData.csv")
> head(MASISHorizons, n=10)
    X OBJECTID  HorNO     HorID   DepthFrom     DepthTo    Code         SOC     ProfID
1  842    24154       1 P3234H01          0          29       0           0      P3234
2  843    24155       2 P3234H02         29          60       0           0      P3234
3  844    24156       3 P3234H03         60          80       0           0      P3234
4  845    27578       1 P1102H01          0          30       0           0      P1102
5  846    27579       2 P1102H02         30          60       0           0      P1102
6  847    27994       3 P4002H03         48          90       0           0      P4002
7  848    27995       4 P4002H04         90         115       0           0      P4002
8  849    28086       1 P4039H01          2          25       0           0      P4039
9  850    28087       2 P4039H02         25          57       0           0      P4039
10 851    28088       3 P4039H03         57          90       0           0      P4039
...
```

The summary statistics:

```
# R provides a wide range of functions for obtaining summary statistics. One method of
# obtaining descriptive statistics is to use the summary( ) function..
# usage: summary(object)

> summary(point.data$Value)
  Min.       1st Qu.  Median  Mean    3rd Qu.  Max.
 0.000       1.005    1.493   1.912   2.244    50.330
```

5.2 KENYA (RANDOM FOREST - CHAPTER 6.3)

The sample data used in the Random Forest Chapter was obtained from a study of SOC in north-eastern Kenya (Omuto, 2008). The data was collected using a Y-shape sampling frame for topsoil (0-30 cm).

5.3 DATA ACCESS

Please contact GSP Secretariat via e-mail (*GSP-GSOC-Map@fao.org*) for the sample datasets.

5.4 References

Borut Vrščaj, Laura Poggio, Duško Muaketov and Ronald Vargas, Utilizing the Legacy Soil Data of Macedonia: The Creation of the Macedonian Soil Information System and its use for digital soil mapping and assessment applications (Pedometrics 2017).

Burrough, P.A., R.A. McDonnell, 1998. Principles of Geographical Information Systems, 2nd Edition. Oxford University Press.

6. UPSCALING METHODS

6.1. CONVENTIONAL UPSCALING USING SOIL MAPS

6.1.1 OVERVIEW

The two conventional upscaling methods, in the context of SOC mapping, are described by Lettens *et al.* (2004). Details about weighted averaging can be found in Hiederer (2013). Different conventional upscaling approaches were applied in many countries (Baritz *et al.* 1999 (Germany), Cruz-Gaistardo (Mexico), Greve *et al.* 2007 (Denmark), Koelli *et al.* 2009 (Estonia), Arrouay *et al.* 2001 (France), Bhatti *et al.* 2002 (Canada)). Because the structure of soil map databases differs between countries (definition of the soil mapping unit, stratification, soil associations, dominating and co-dominating soils, typical and estimate soil properties for different depths), it is difficult to define a generic methodology for the use of these maps for upscaling soil property information.

However, the essential principle which is commonly used, is to combine soil property data from local observations with soil maps via class- and geomatching.

DEFINITION

DIVERSITY OF NATIONAL SOIL LEGACY DATA SETS

IN ORDER TO DEVELOP A REPRESENTATIVE AND LARGE NATIONAL SOIL DATABASE, VERY OFTEN, DATA FROM DIFFERENT SOURCES (E.G. SOIL SURVEYS OR PROJECTS IN DIFFERENT PARTS OF THE COUNTRY AT DIFFERENT TIMES) ARE COMBINED. THE FOLLOWING CASE OF BELGIUM DEMONSTRATES HOW AVAILABLE LEGACY DATABASES COULD BE COMBINED. THREE DIFFERENT SOURCES ARE USED TO COMPILE AN OVERVIEW OF NATIONAL SOC STOCKS:

DATA SOURCE 1: SOIL PROFILE DATABASE WITH 13,000 POINTS OF GENETIC HORIZONS; FOR EACH SITE, THERE IS INFORMATION ABOUT THE SOIL SERIES, MAP COORDINATES AND LAND USE CLASS; FOR EACH HORIZON, THERE IS INFORMATION ABOUT DEPTH AND THICKNESS, TEXTURAL FRACTIONS AND CLASS, VOLUME PERCENTAGE OF ROCK FRAGMENTS; ANALYTICALLY, THERE IS THE ORGANIC CARBON CONTENT AND INORGANIC CARBON CONTENT.

DATA SOURCE 2: FOREST SOIL DATA BASE WHICH INCLUDES ECTORGANIC HORIZONS. ACCORDING TO THEIR NATIONAL DEFINITION, THE TERM "ECTORGANIC" DESIGNATES THE SURFACE HORIZONS WITH AN ORGANIC MATTER CONTENT OF AT LEAST 30%, THUS, IT INCLUDES BOTH THE LITTER LAYER AND THE ORGANIC SOIL LAYERS. FOR THE CALCULATION OF SOC STOCKS FOR THE ECTORGANIC LAYER, NO FIXED-DEPTH WAS USED, INSTEAD THE MEASURED THICKNESS OF THE ORGANIC LAYERS AND LITTER LAYERS WAS APPLIED.

DATA SOURCE 3: 15,000 SOIL SURFACE SAMPLES WERE USED (UPPER 20 CM OF MINERAL SOIL); CARBON MEASUREMENTS ARE AVAILABLE PER DEPTH CLASS.

FROM ALL DATA SOURCES, SOC STOCKS FOR PEAT SOILS WERE CALCULATED SEPARATELY.

6.1.2 TECHNICAL STEPS

STEP 1: DATA PREPARATION

- Separate the data base for forests, peat and other land uses
- If only horizons are provided: derive or estimate average depth of horizons per soil type; add upper and lower depth.
- Check completeness of parameters per depth using the solum depth to code empty cells
- Correction of organic carbon in case total carbon was determined (total carbon minus inorganic carbon concentration)
- Correction of Walkley and Black method for incomplete oxidation (1.32)
- If BD measured is lacking, select proper pedotransfer functions (PTF) and estimate BD. There are many PTF. At best, publications about the choice of the best suited PTF for specific physio-geographic conditions are available.
- If stone content is missing, investigate using other data sources or literature, to which a correction for stones should be applied
- if possible, derive the standard average stone content for different soils/horizons/depths, or used published soil profiles, as a simple correction factor.
- Calculate SOC stocks for all mineral and peat soils over 0-30 cm, and optionally for forest organic layers and, peat &gt;30 &lt;100 cm.

STEP 2: PREPARATORY GIS OPERATIONS

- Prepare Covariates
- Identify properties of covariates for each point observation using geo-matching
- Upscaling using geo-matching of all points: *Extract the covariate information to all georeferenced sample sites*. The SOC values from all

points within the unit are then averaged. It is assumed that the points represent the real variability of soil types within the units

STEP

STEP 3: UPSCALING

- Upscaling using class-matching of points in agreement with classes

Through *class-matching*, only those points or profiles are attributed to a soil or landscape unit if both the soil and the land use class are the same. Class-matching thus can be performed regardless of the profile location. Before averaging, a weighing factor can be introduced according to the area proportions of dominant, co-dominant and associated soils. Each profile needs to be matched to its soil type/landscape type, and the SOC value averaged.

1. Determine a soil or landscape unit (e.g. national soil legend stratified by climate area and main land cover type (forest, grassland, cropland)
2. Calculate average SOC stocks from from all soils which match the soil/ landscape unit
3. Present the Soil/landscape map with SOC stocks, do not classify SOC stocks into groups (e.g. < 50, 50-100, > 100).

Note: Pre-classified SOC maps cannot be integrated into a global GSOCmap legend.

- Upscaling using geo-matching

Because of its importance, geo-matching is described in more detail (section 6.1.3).

6.1.3 GEO-MATCHING

it is important to first prepare the working environment pre-processed all input data. The following section presents different Geo-matching procedures;

1. Setting up software and working environment
2. Geo-matching SOC with WRB Soil map (step-by-step, using the Soil Map of Macedonia and the demonstration data presented above)

3. Geo-matching SOC with other environmental variables: Land use

4. Finally, the development of Landscape Units (Lettens *et al*. 2004) is outlined.

This example was developed for QGIS and focusses on SOC mapping using vector data. QGIS 2.18 with GRASS 7.05 will be used. For more information, see also:

- *https://gis.stackexchange.com*
- *http://www.qgis.org/*
- *http://www.qgisforum.org/*

STEP 1: SETTING UP A QGIS PROJECT.

1. Install QGIS and supporting software; download the software at *http://www.qgis.org/en/site/forusers/download.html* (select corrent version for Windows, Mac or Linux, 32 or 64 bit).

2. Create a work folder, e.g. D:\GSOC\practical_matching. Copy the folder with the Macedonian demonstration data into this folder.

3. Start 'QGIS desktop with GRASS'

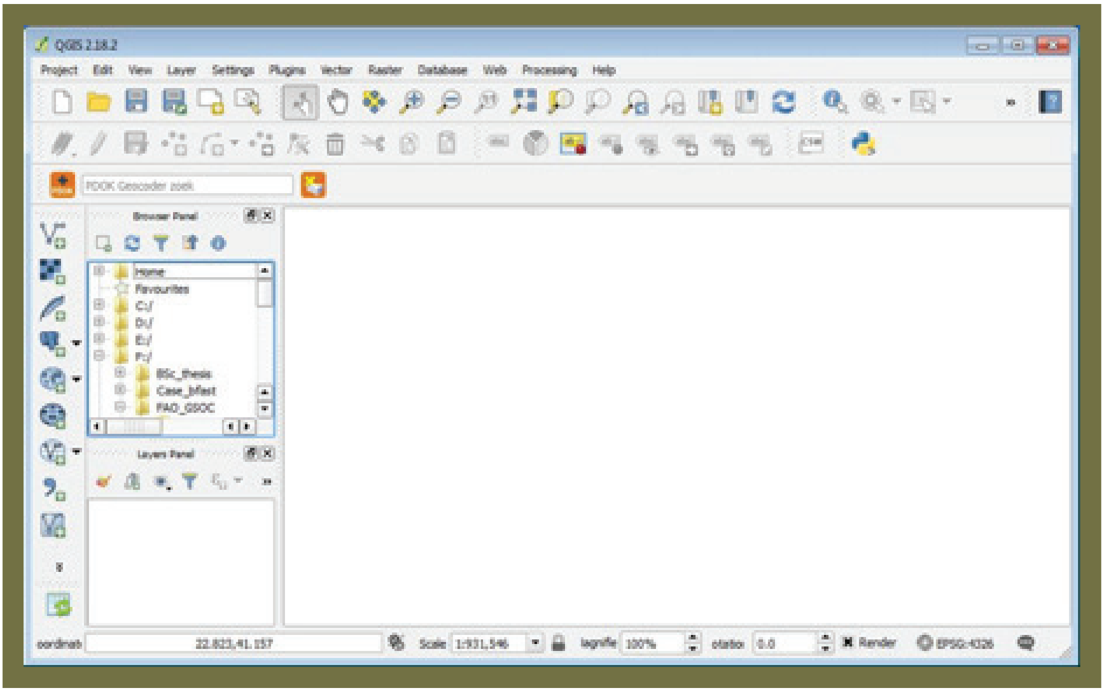
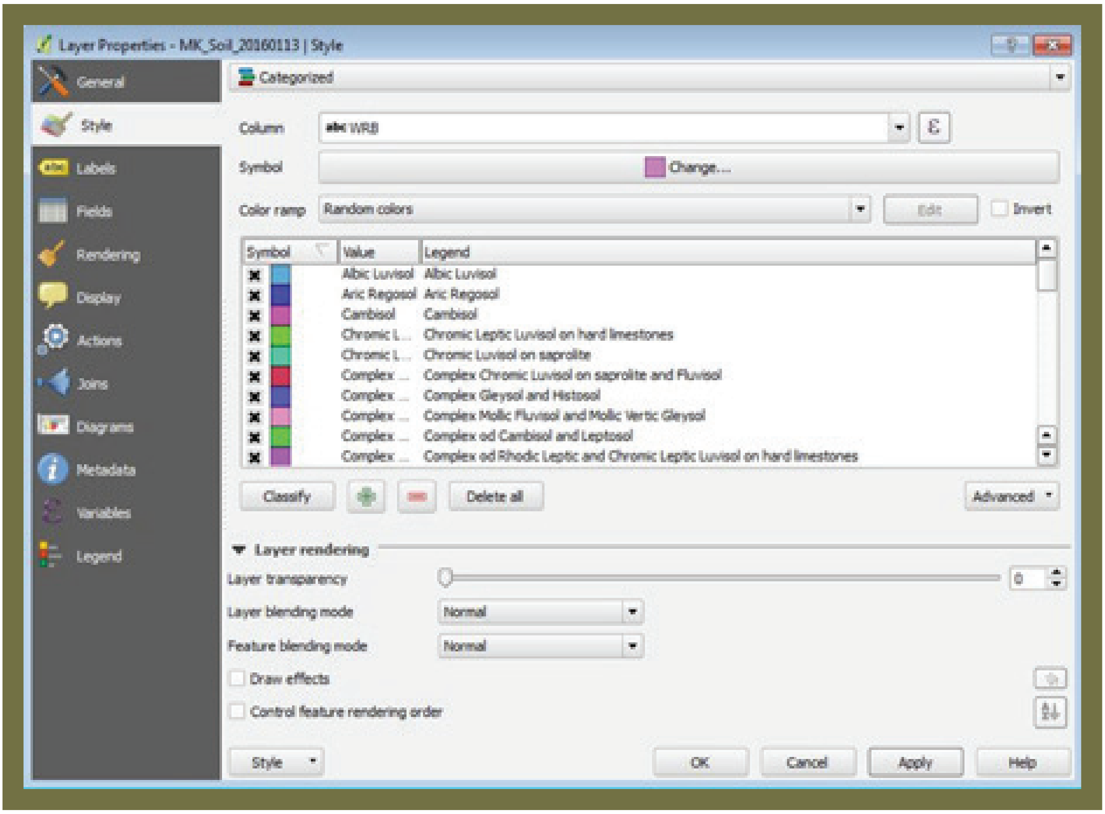


FIGURE 6.1 SHOWS THE START SCREEN OF QGIS DESKTOP. IN THE UPPER LEFT PANEL THERE IS THE BROWSER PANEL, WHICH LISTS THE GEODATA USED FOR THIS EXAMPLE. IN THE BOTTOM LEFT, THE LAYER INFORMATION IS GIVEN FOR THE LAYERS DISPLAYED ON THE RIGHT.

4. Load the Macedonian soil map. Right-click the file in the Browser panel and add the map to your project.

5. Display the soil classes. Right-click on the file in the Layers Panel, properties. Go to Style and change from 'Single symbol' to 'Categorized' (Fig. 6.2). Select the column 'WRB' and press the icon 'Classify' and change the colours if you want. Next, apply the change and finish with clicking the OK-button.


Layer Properties - MK_Soil_20160113 | Style

Categorized

Column: abc WRB

Symbol: Change...

Color ramp: Random colors

☐ Invert

Symbol	Value	Legend
☒ 🟦	Albic Luvisol	Albic Luvisol
☒ 🟦	Aric Regosol	Aric Regosol
☒ 🟪	Cambisol	Cambisol
☒ 🟩	Chromic L...	Chromic Leptic Luvisol on hard limestones
☒ 🟩	Chromic L...	Chromic Luvisol on saprolite
☒ 🟥	Complex ...	Complex Chromic Luvisol on saprolite and Fluvisol
☒ 🟦	Complex ...	Complex Gleysol and Histosol
☒ 🟪	Complex ...	Complex Mollic Fluvisol and Mollic Vertic Gleysol
☒ 🟩	Complex ...	Complex od Cambisol and Leptosol
☒ 🟪	Complex ...	Complex od Rhodic Leptic and Chromic Leptic Luvisol on hard limestones

Layer rendering

Layer transparency: 0

Layer blending mode: Normal

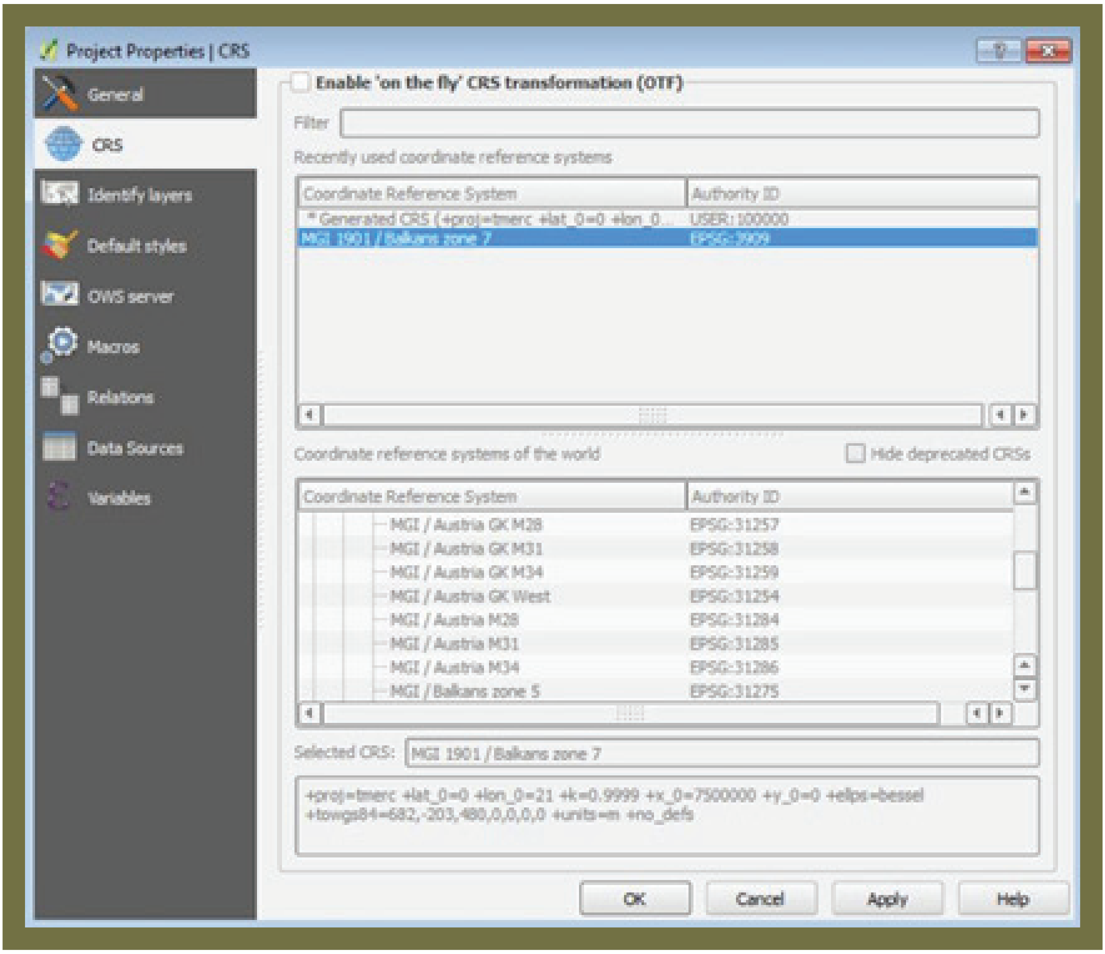
Feature blending mode: Normal

☐ Draw effects

☐ Control feature rendering order

6. Ensure the correct projection for this project. Go to: *Project -> Project properties -> CRS*

In this case, you automatically use the local projection for Macedonia. The EPSG code is 3909 which corresponds to MGI 1901/ Balkans zone *7* (Figure 6.3).



□ Enable 'on the fly' CRS transformation (OTF)

Filter

Recently used coordinate reference systems

Coordinate Reference System	Authority ID
* Generated CRS (+proj=tmerc +lat_0=0 +lon_0...	USER:100000
MGI 1901 / Balkans zone 7	EPSG:3909

Coordinate reference systems of the world

□ Hide deprecated CRSs

Coordinate Reference System	Authority ID
MGI / Austria GK M28	EPSG:31257
MGI / Austria GK M31	EPSG:31258
MGI / Austria GK M34	EPSG:31259
MGI / Austria GK West	EPSG:31254
MGI / Austria M28	EPSG:31284
MGI / Austria M31	EPSG:31285
MGI / Austria M34	EPSG:31286
MGI / Balkans zone 5	EPSG:31275

Selected CRS: MGI 1901 / Balkans zone 7

```
+proj=tmerc +lat_0=0 +lon_0=21 +k=0.9999 +x_0=7500000 +y_0=0 +ellps=bessel
+towgs84=682,-203,480,0,0,0,0 +units=m +no_defs
```

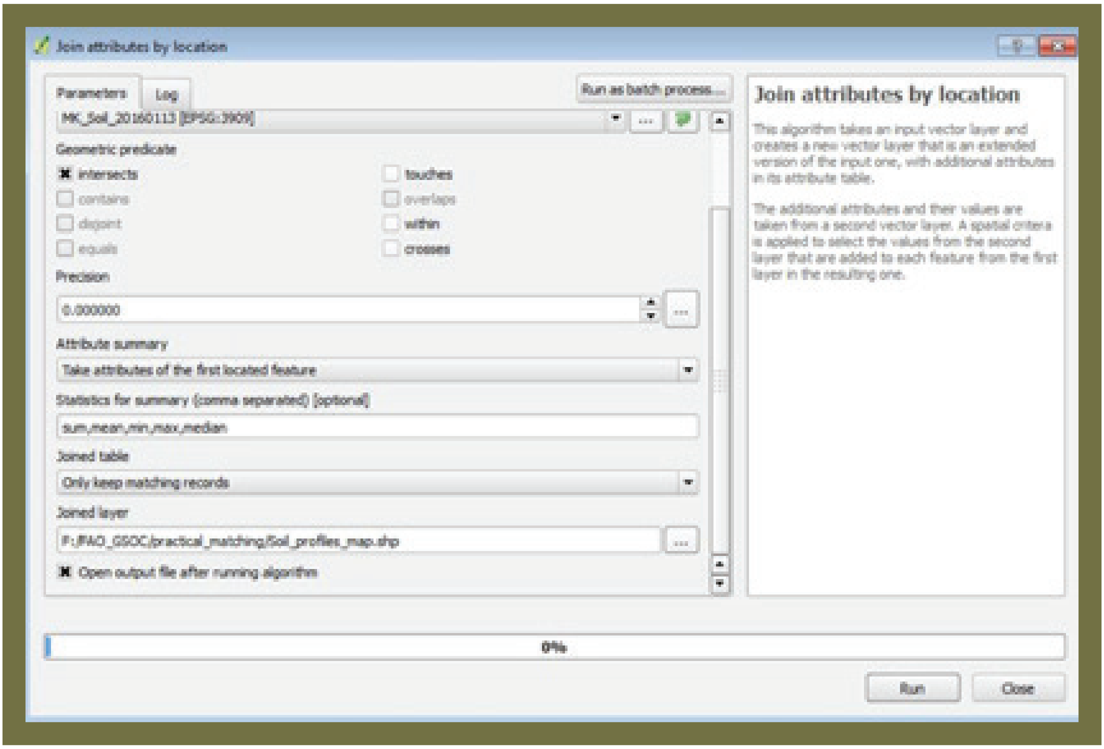
7. Save the project in the created folder

Load and display the pre-processed SOC point data. If a shapefile already exists, this is done the same way as described in Step 4. If you have the data as a text file, you need to create a vector layer out of that file. Go to *Layer -> Add Layer -> Add Delimited Text layer*. Select the correct file and proper CRS projection. The layer should be added to your Layers Panel and displayed on top of the Soil Map.

STEP 2. GEO-MATCHING SOC WITH WRB SOIL MAP.

In this section you will make a SOC map, based on the Macedonian Soil Map and the SOC values at the sampled points, following 3 steps: 1) Extract the soil map information for the point data, 2) obtain the mean and standard deviation of the SOC stocks per soil class, based on the point data and 3) assign these values to the corresponding soil map units. The steps are detailed below:

1. Extract the soil map information to the soil profile data by 'Join Attributes by location'. *Vector -> Data Management Tools -> Join Attributes by location*. Here, the target vector layers are the soil point data, and the join vector layer is the Macedonian Soil Map. The geometric predicate is 'intersects'. Specify at the 'joined table' to keep only matching records and save the 'joined layer' as a new file (Fig. 6.4).



Join attributes by location

MK_Soil_20160113 [EPSG:3909]

Geometric predicate

☒ intersects ☐ touches
☐ contains ☐ overlaps
☐ disjoint ☐ within
☐ equals ☐ crosses

Precision

0.000000

Attribute summary

Take attributes of the first located feature

Statistics for summary (comma separated) [optional]

sum,mean,min,max,median

Joined table

Only keep matching records

Joined layer

F:/FAO_GSOC/practical_matching/Soil_profiles_map.shp

☒ Open output file after running algorithm

Join attributes by location

This algorithm takes an input vector layer and creates a new vector layer that is an extended version of the input one, with additional attributes in its attribute table.

The additional attributes and their values are taken from a second vector layer. A spatial criteria is applied to select the values from the second layer that are added to each feature from the first layer in the resulting one.

0%

FIGURE 6.4 JOIN ATTRIBUTES BY LOCATION

2. Check the newly generated file, open the attribute table. The new file is added to the '*Layers Panel*' . Right-click on the file and open the attribute table. The information from the Macedonian Soil Map is now added to the soil point data.

3. Most likely, the SOC values in the table are not numeric and thus statistics cannot be calculated. Check the data format, right-click on the file in the 'Layers Panel' and check the Type name of the SOC field under the tab '*Fields*'. If they are not integer then change the format.

4. Change of the data format: Open the attribute table and start editing (the pencil symbol in the upper left corner of your table). Open the field calculator and follow these instructions (Fig. 6.5):

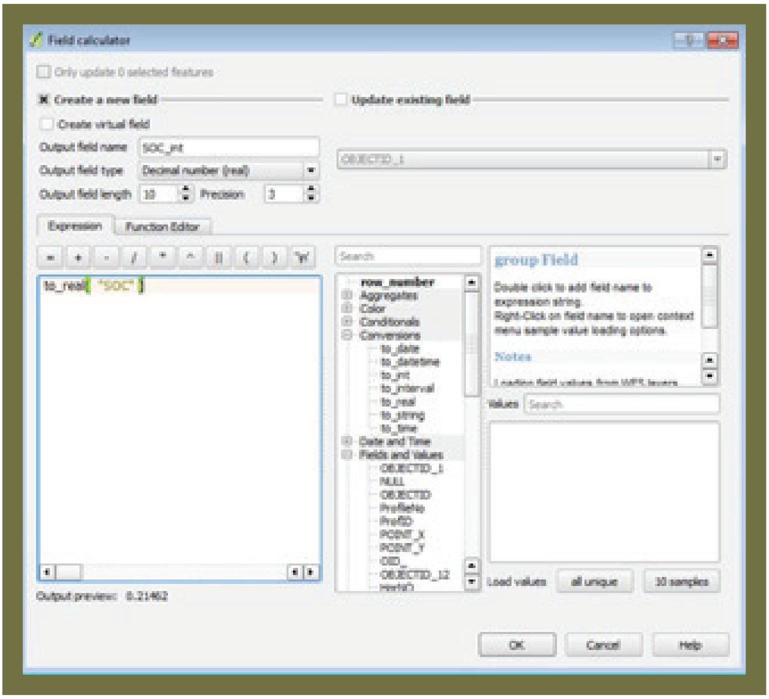
   a. Check box: Create a new field

   b. Output field name: Specify the name of your field

   c. Output field type: Decimal Number (real)

   d. Output field length: 10, precision: 3

   i. Expression: to_real('SOC'), the *to_real* function can be found under '*conversions*' and the 'SOC' field is found under '*Fields and Values*'



Field calculator

□ Only update 0 selected features

☒ Create a new field

□ Update existing field

□ Create virtual field

Output field name: SOC_int

Output field type: Decimal number (real)

Output field length: 10 Precision: 3

OBJECTID_1

```
to_real( "SOC" )
```

Output preview: 0.21462

group Field

Double click to add field name to expression string.
Right-Click on field name to open context menu sample value loading options.

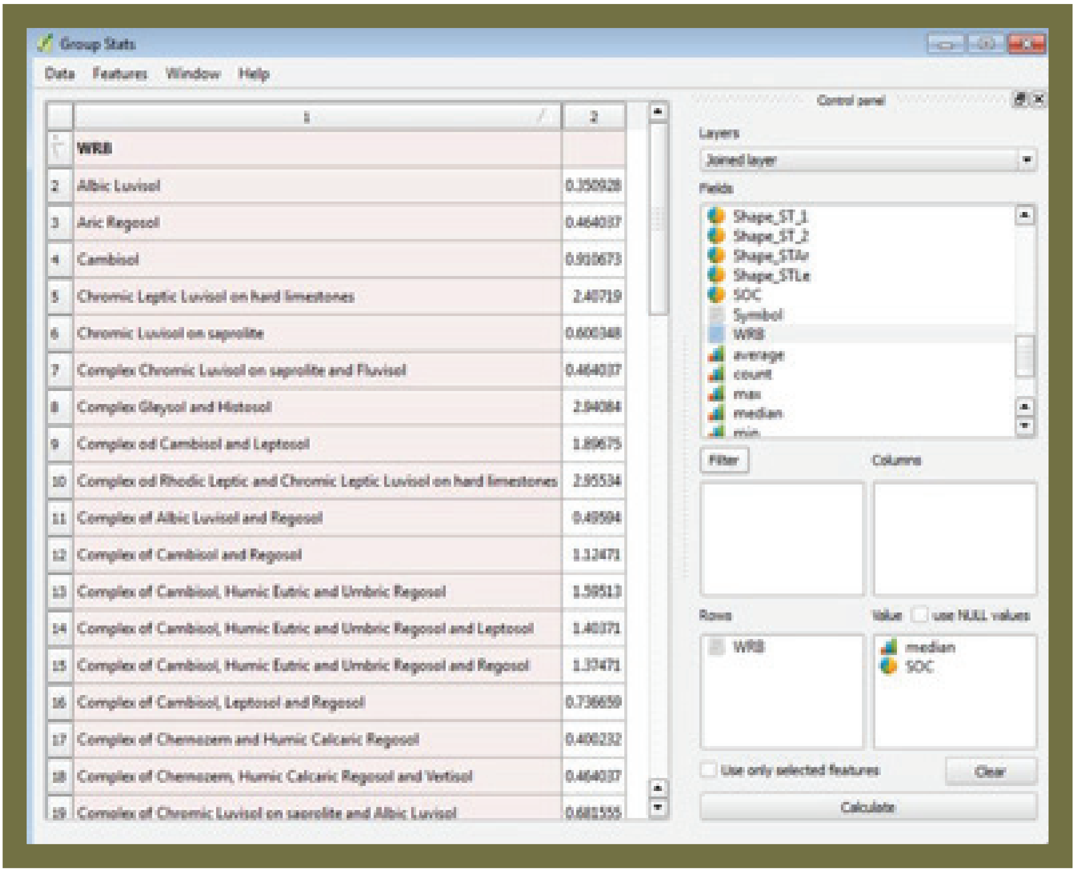
Notes

Loading field values from WFS layers

FIGURE 6.5 EXAMPLE FIELD CALCULATOR

5. after calculating the field, save edits and leave the editing mode prior to closing the table. if changes are not saved, the added field will be lost.

6. calculate the median soc stock per soil type. go to the tab *'Vector'-> group stats*. select the layer from the spatial join you made in *Step 2*. add the field 'soc' and median to the box with 'values' and the field 'wrb' to the 'rows'. make sure the box with 'use only selected features' is not checked. now calculate the statistics. a table will be given in the left pane (figure 6.6). save this file as .csv and repeat the same for the standard deviation.



	1	2
1	WRB	
2	Albic Luvisol	0.350928
3	Aric Regosol	0.464037
4	Cambisol	0.910673
5	Chromic Leptic Luvisol on hard limestones	2.40719
6	Chromic Luvisol on saprolite	0.600348
7	Complex Chromic Luvisol on saprolite and Fluvisol	0.464037
8	Complex Gleysol and Histosol	2.94084
9	Complex od Cambisol and Leptosol	1.89675
10	Complex od Rhodic Leptic and Chromic Leptic Luvisol on hard limestones	2.95534
11	Complex of Albic Luvisol and Regosol	0.49594
12	Complex of Cambisol and Regosol	1.12471
13	Complex of Cambisol, Humic Eutric and Umbric Regosol	1.59513
14	Complex of Cambisol, Humic Eutric and Umbric Regosol and Leptosol	1.40371
15	Complex of Cambisol, Humic Eutric and Umbric Regosol and Regosol	1.37471
16	Complex of Cambisol, Leptosol and Regosol	0.736659
17	Complex of Chernozem and Humic Calcaric Regosol	0.408232
18	Complex of Chernozem, Humic Calcaric Regosol and Vertisol	0.464037
19	Complex of Chromic Luvisol on saprolite and Albic Luvisol	0.681555

Layers: Joined layer

Fields: Shape_ST_1, Shape_ST_2, Shape_STAr, Shape_STLe, SOC, Symbol, WRB, average, count, max, median, min

Rows: WRB

Value: median, SOC

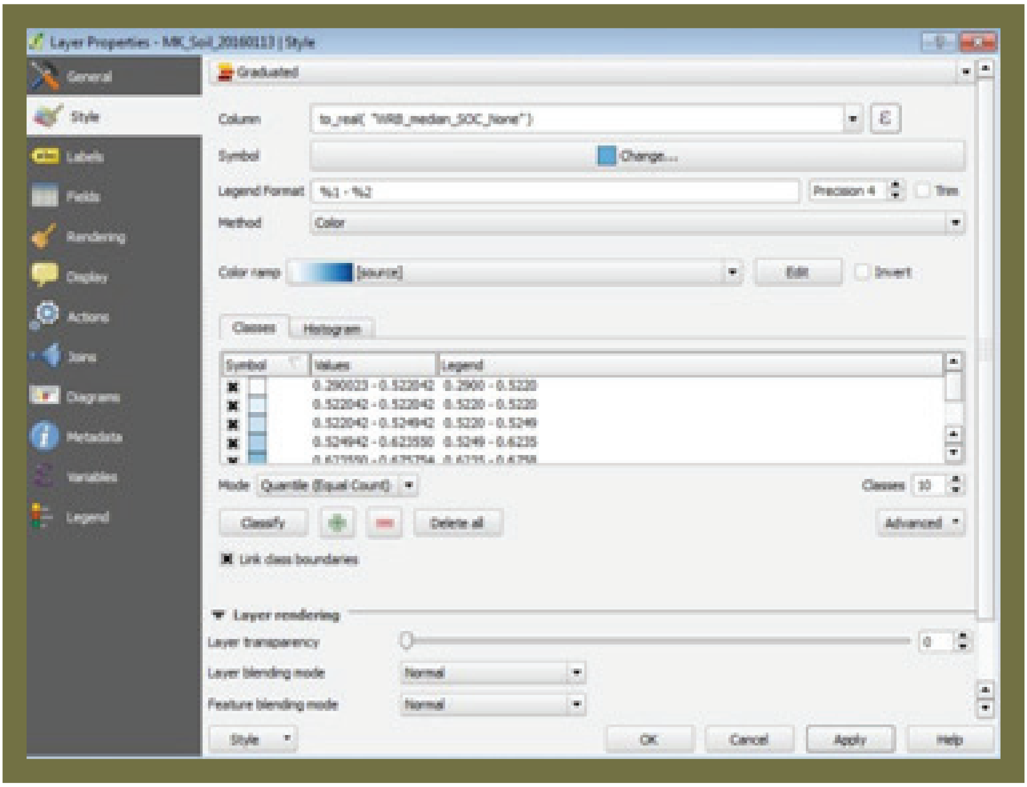
□ use NULL values

□ Use only selected features

FIGURE 6.6 CALCULATE GROUP STATISTICS

7. Join the mean and standard deviation of SOC to the Soil Map. First add the files generated during step 6 to the Layers Panels. In the Layers Panel, right-click on the Macedonian Soil Map. Go to *Properties -> Joins* and add a new join for both the median and standard deviation of SOC. The Join and Target Field are both 'WRB'.

8. Display the SOC maps. Go to the layer properties of the Macedonian Soil Map. Go to Style and change the legend to a graduated legend. In the column you indicate the assigned SOC values. Probably this is not a integer number and so you have to convert this number again to a numeric values. You can do this with the box next to the box (Fig. 6.7). Change the number of classes to e.g. 10 classes, change the mode of the legend and change the color scheme if you want and apply the settings. Now you have a map with the median SOC stocks per WRB soil class.



Layer Properties - MK_Soil_20160113 | Style

Graduated

Column: to_real("WRB_median_SOC_None")

Legend Format: %1 - %2 Precision 4 □ Trim

Method: Color

Color ramp: [source] □ Invert

Symbol	Values	Legend
☒	0.290023 - 0.522042	0.2900 - 0.5220
☒	0.522042 - 0.522042	0.5220 - 0.5220
☒	0.522042 - 0.524942	0.5220 - 0.5249
☒	0.524942 - 0.623550	0.5249 - 0.6235
☒	0.623550 - 0.675754	0.6235 - 0.6758

Mode: Quantile (Equal Count) Classes: 10

☒ Link class boundaries

Layer rendering

Layer transparency: 0

Layer blending mode: Normal

Feature blending mode: Normal

FIGURE 6.7 CHANGE THE LEGEND STYLE TO DISPLAY THE SOC VALUES

9. In order to generate a proper layout, go to *Project -> New Print Composer*

a. Add map using *Layout -> Add Map*. Define a square on the canvas and the selected map will be displayed.

b. Similarly, title, scale bar, legend and a north arrow can be added. Specific properties can be changed in the box ‘Item properties’.

c. When the map is finished, it can be exported as an image or pdf.

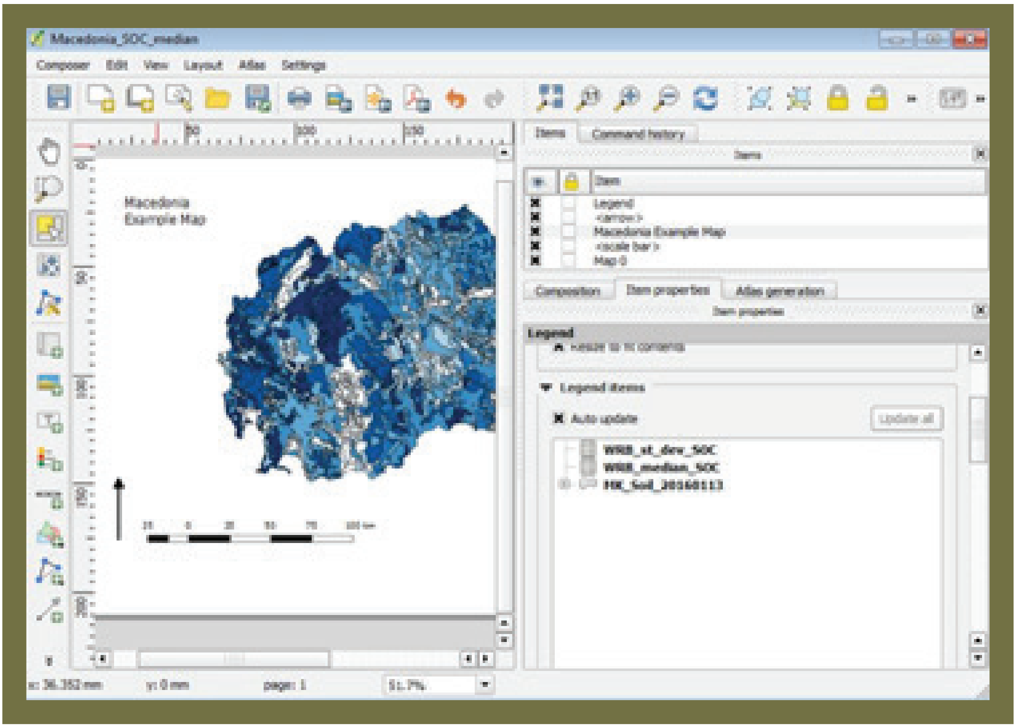


FIGURE 6.8 EXAMPLE OF THE MAP COMPOSER

10. Repeat step 2-8 but now for the standard deviation of the SOC stocks.

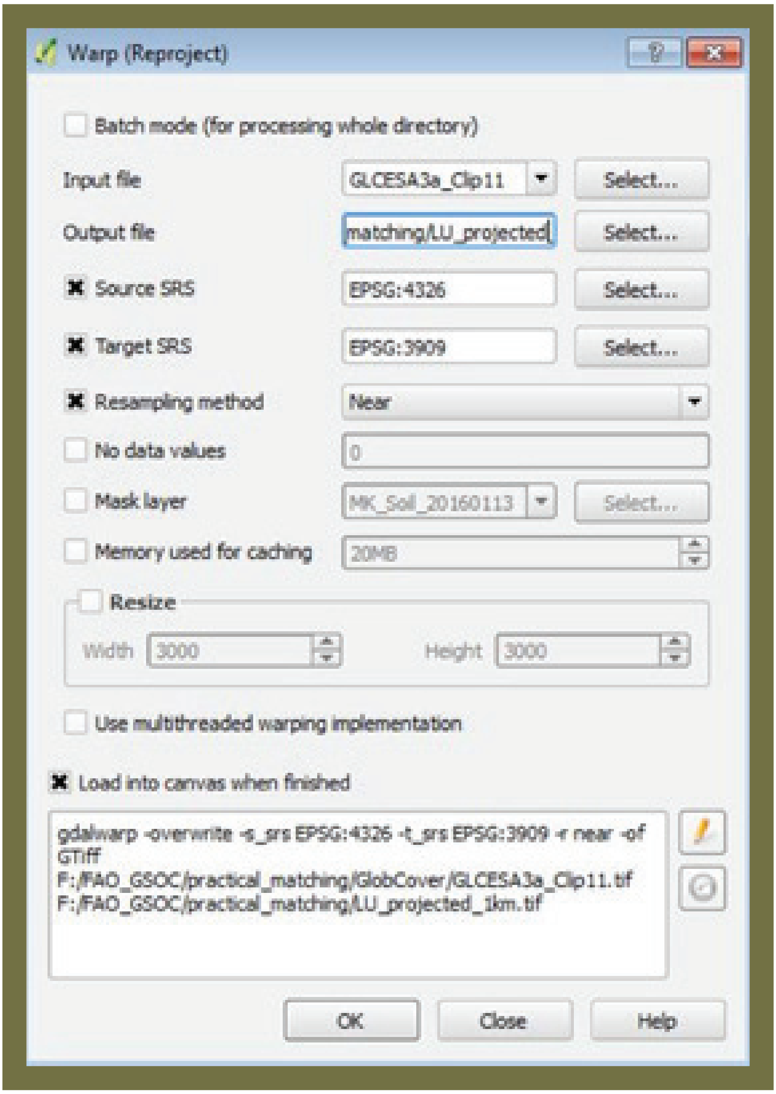
11. Save the file as a new shapefile: Go to '*Layer Panels -> Save as -> ESRI ShapeFile* and make sure that you define the symbology export: Feature Symbology. Now, a shapefile is generated, with both the median and standard deviation SOC stock per soil type. Redundant fields can be removed after the new file is created.

STEP 3: GEO-MATCHING SOC WITH OTHER ENVIRONMENTAL VARIABLES: LAND USE

1. Start a new project and add the soil point data and Macedonia Soil Map layers from the Browser panel

2. Add the Land Use raster file to the Layers Panels. This is a raster file with 1 kilometre resolution and projected in lat long degrees (WGS84). For more information about this product see the online information from worldgrids: *http://worldgrids.org/doku.php/wiki:glcesa3*

3. Change the projection to the MGI 1901/ Balkans region7. Go to *Raster* -> *Projections* -> *Warp* and select the proper projection and a suitable file name, e.g. LU_projected_1km. Tick the checkbox for the resampling method and choose Near. This is nearest neighbour and most suitable for a transformation of categorical data, such as land use (Fig. 6.9).



Warp (Reproject)

□ Batch mode (for processing whole directory)

Input file: GLCESA3a_Clip11

Output file: matching/LU_projected

☒ Source SRS: EPSG:4326

☒ Target SRS: EPSG:3909

☒ Resampling method: Near

□ No data values: 0

□ Mask layer: MK_Soil_20160113

□ Memory used for caching: 20MB

□ Resize — Width: 3000 Height: 3000

□ Use multithreaded warping implementation

☒ Load into canvas when finished

```
gdalwarp -overwrite -s_srs EPSG:4326 -t_srs EPSG:3909 -r near -of GTiff
F:/FAO_GSOC/practical_matching/GlobCover/GLCESA3a_Clip11.tif
F:/FAO_GSOC/practical_matching/LU_projected_1km.tif
```

FIGURE 6.9 CHANGE THE PROJECTION OF A RASTER FILE

4. In order to geomatch the soil point data with Land Use, the raster file needs to be converted into a vector file. Go to Raster -> Conversions -> Polygonize. Set a proper output filename, e.g. LU_polygon_1km, and check the tickbox for Fieldname.

5. Change the legend style into categories (Step 1-5):

Now, the steps from the previous section need to be repeated, using the land use polygon map instead of using the Macedonian Soil Map.

6. Join attributes by location using the soil point data and the polygon land use map.
7. Calculate the median and standard deviation of SOC by using the Group Statistics for SOC and the Land Use classes and save the files as .csv.
8. Add the generated .csv files to the Layers Panel.
9. Join the files with the LU polygon map, generated at step 3-4.
10. Change the classes in the legend and inspect the histogram with the median SOC values. Try to find a proper definition of the class boundaries (Step 2-8).

[EXTRA] STEP 4: JOINING LANDSCAPE UNITS AND SOIL MAPPING UNITS TO SUPPORT CLASS- AND GEO-MATCHING

In this section it is outlined how SOC stocks can be mapped following the method outlined by Lettens *et al.*(2004; DOI:10.1079/SUM2003221). The general idea is that the landscape is stratified into more or less homogenous units and subsequently, the SOC stocks are obtained following the procedure outlined earlier in this practical. Lettens *et al.* (2004) outlines a method to stratify the landscape into homogeneous strata with respect to Land Use and Soil Type, as was explained earlier. In order to obtain such strata, the Soil Map and the Land Use map need to be combined. This can be done using various types of software, e.g. ArcMap, GRASS, QGIS or R.

When using the GIS software, the only thing that needs to be done is intersecting the vector files and dissolving the newly created polygon features. Depending on the software and the quality of your shapefile you may experience problems with the geometry of your shapefile. Generally, ArcMap and GRASS correct the geometry when the shapefile is loaded, while QGIS does not do this automatically. There are various ways to correct the geometry, however, correcting the geometry falls outside the scope of this training. Therefore, we give some hints on how to correct your geometry prior to using the functions 'Intersect' and 'Dissolve'.

1. Change the LU raster map to 5 kilometer resolution: Right-click the Lu_project_1km file and select Save as. Change the resolution to 5000 meters. Scroll down, check the Pyramids box, and change the resampling method to Nearest Neighbour.

2. Convert the raster map to a polygon map and add the file to the Layers Panel

3. Check the validity of the Soil Map and Land Use Map: Vector -> Geometry Tools -> Check Validity

4. Below you find the instructions in case you have no problems with your geometry:

5. Intersect the Soil Map and the Land Use Map. In ArcGIS and QGIS you can use this function. Go to Vector -> Geoprocessing tools -> Intersection. (In GRASS you have to use the function 'Overlay' from the Vector menu)

6. Dissolve the newly generated polygons. Vector -> Geoprocessing tools -> Dissolve

7. Next, this layer can be used to continue with the classmatching or geomatching procedures.

WHEN ENCOUNTERING PROBLEMS WITH THE GEOMETRY THERE ARE AT LEAST THREE WAYS TO CORRECT YOUR GEOMETRY:

- RUN THE V_CLEAN TOOL FROM GRASS WITHIN QGIS. OPEN THE PROCESSING TOOLBOX -> GRASS GIS 5 COMMANDS -> VECTOR -> V.CLEAN

- INSTALL THE PLUGIN 'PROCESSING LWGEOM PROVIDER'. GO TO THE PLUGINS MENU AND SEARCH FOR THE PLUGIN AND INSTALL. YOU CAN FIND THE NEWLY INSTALLED TOOL IN THE PROCESSING TOOLBOX BY TYPING THE NAME IN THE SEARCH FUNCTION

- MANUALLY CORRECT THE ERROR NODES OF THE VECTOR FEATURES

6.1.4 References

Lettens, S., J. Van Orshoven, B. Van Wesemael and B. Muys (2004). Soil organic and inorganic carbon contents of landscape units in Belgium derived using data from 1950 to 1970. Soil Use and Management 20: 40–47.

6.2. REGRESSION-KRIGING

6.2.1 OVERVIEW

Regression-kriging is a spatial interpolation technique that combines a regression of the dependent variable (target variable) on predictors (i.e. the environmental covariates) with kriging of the prediction residuals. In other words, Regression-Kriging is a hybrid method that combines either a simple or a multiple-linear regression model with ordinary kriging of the prediction residuals. The Multiple regression analysis models the relationship of multiple predictor variables and one dependent variable, i.e. it models the deterministic trend between the target variable and environmental covariates. The modelled relationship between predictors and target are summarized in regression equation, which can then be applied to a different data set in which the target values are unknown but the predictor variables are known. The regression equation predicts the value of the dependent variable using a linear function of the independent variables.

In this section, we review the regression kriging method. First, the deterministic part of the trend is modelled using a regression model. Next, the prediction residuals are kriged. In the regression phase of a regression-kriging technique, there is a continuous random variable called the dependent variable (target) Y (in our case SOC) and a number of independent variables which are selected covariates, x1, x2,...,xp. Our purpose is to predict the value of the dependent variable using a linear function of the independent variables. The values of the independent variables (environmental covariates) are known quantities for purposes of prediction, the model is:

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta pxp + \varepsilon$$

Where Y is the response variable, X is a predictor variable and ε is the residual or error.

DEFINITION

IN PARTICULAR CASES, STEPWISE MULTIPLE LINEAR REGRESSION (SMLR) OR PRINCIPAL COMPONENT REGRESSION CAN BE USED. STEPWISE MULTIPLE LINEAR REGRESSION (SMLR) USUALLY SELECTS PREDICTORS THAT HAVE THE STRONGEST LINEAR CORRELATION WITH THE DEPENDENT VARIABLE, AND EXPLAINS THE VARIABILITY IN THE TARGET VARIABLE BEST. IN PCR, INSTEAD OF REGRESSING THE DEPENDENT VARIABLE ON PREDICTORS DIRECTLY, THE ORIGINAL REGRESSOR VARIABLES ARE REPLACED BY THEIR PRINCIPAL COMPONENTS, THUS ORTHOGONALIZING THE REGRESSION PROBLEM AND MAKING COMPUTATIONS EASIER AND MORE STABLE. THE PCR IS STILL BASED ON A STANDARD LINEAR REGRESSION MODEL, BUT USES PCA FOR ESTIMATING THE UNKNOWN REGRESSION COEFFICIENTS IN THE MODEL. UNLIKE MULTIPLE LINEAR REGRESSION (MLR) AND STEPWISE MULTIPLE LINEAR REGRESSION (SMLR), PRINCIPAL COMPONENTS REGRESSION (PCR) CAN COPE WITH DATA CONTAINING LARGE NUMBERS OF COVARIATES THAT ARE HIGHLY COLLINEAR.

DEFINITION

THE LITERATURE AND THE GEOSTATISTICAL TOOLS USE DIFFERENT TERMS FOR REGRESSION-KRIGING. BOTH KRIGING WITH EXTERNAL DRIFT (KDE), AND REGRESSION-KRIGING (RK) ARE BASICALLY THE SAME TECHNIQUES. THESE PROCEDURES ARE WIDELY AVAILABLE IN COMMERCIAL OR OPEN SOURCE STATISTICAL SOFTWARE PACKAGES, SUCH AS SPSS, MINITAB OR R.

6.2.2 ASSUMPTIONS

Standard linear regression models with standard estimation techniques make a number of assumptions about the predictor variables, the response variables and their relationship. One must review the assumptions made when using the model.

Linearity: The mean value of Y for each specific combination of the X's is a linear function of the X's. In practice this assumption can virtually never be confirmed; fortunately, multiple regression procedures are not greatly affected by minor deviations from this assumption. If curvature in the relationships is evident, one may consider either transforming the variables, or explicitly allowing for nonlinear components.

Normality Assumption: It is assumed in multiple regression that the residuals (predicted minus observed values) are distributed normally (i.e., follow the normal distribution). Again, even though most tests (specifically the F-test) are quite robust with regard to violations of this assumption, it is always a good idea, before drawing final conclusions, to review the distributions of the major variables of interest. You can produce histograms for the residuals as well as normal probability plots, in order to inspect the distribution of the residual values.

> NORMALITY OF THE RESIDUALS IS CRUCIAL BECAUSE IT IS AN ESSENTIAL FOR REGRESSION-KRIGING...

Collinearity: There is not perfect collinearity in any combination of the X's. A higher degree of collinearity, or overlap, among independent variables can cause problems in multiple linear regression models. Collinearity (also multicollinearity) is a phenomenon in which two or more predictors in a multiple regression model are highly correlated. Collinearity causes increase in variances and relatedly increases inaccuracy.

Distribution of the Errors: The error term is normally distributed with a mean of zero and constant variance.

Homoscedasticity: The variance of the error term is constant for all combinations of X's. The term homoscedasticity means "same scatter." Its antonym is heteroscedasticity ("different scatter").

6.2.3 PRE-PROCESSING OF COVARIATES

Before using the selected predictors, multicollinearity assumption must be reviewed. As an assumption, there is not perfect collinearity in any combination of the X's. A higher degree of collinearity, or overlap, among independent variables can cause problems in multiple linear regression models. The multicollinearity of number of variables can be assessed using Variance Inflation Factor (VIF). In R, the function vif() from caret package can estimate the VIF. There are several rules of thumb to establish when there is a serious multi-collinearity (e.g. when the VIF square root is over 2). The Principal component analysis can be used to overcome multicollinearity issues.

Principal components analysis can cope with data containing large numbers of covariates that are highly collinear which is the common case in environmental predictors. Often the principal components with higher variances are selected as regressors. However, for the purpose of predicting the outcome, the principal components with low variances may also be important, in some cases even more important.

The PCA + Linear Regression (PCR) method may be coarsely divided into three main steps:

1. Run PCA on the data matrix for the predictors to obtain the principal components, and then select a subset of the principal components for further use.

2. Regress the dependent variable on the selected principal components as covariates, linear regression to get estimated regression coefficients.

3. Transforming the data back to the scale of the actual covariates, using the selected PCA loadings.

6.2.4 THE TERMINOLOGY

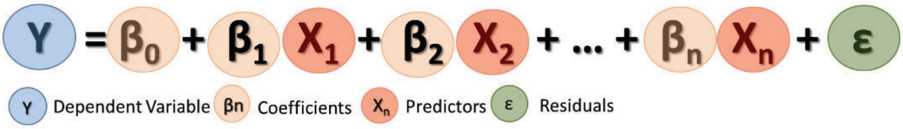

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$
Y Dependent Variable
βn Coefficients
X_n Predictors
ε Residuals

FIGURE 6.1 REGRESSION MODEL

- **Dependent variable (Y):** What we are trying to predict (e.g. soil organic carbon content).

- **Independent variables (Predictors) (X):** Variables that we believe influence or explain the dependent variable (Covariates: environmental covariates - DEM derived covariates, soil maps, land cover maps, climate maps). The data sources for the environmental predictors are provided in Chapter 3.

- **Coefficients (β):** values, computed by the multiple regression tool, reflect the relationship and strength of each independent variable to the dependent variable.

- **Residuals (ε):** the portion of the dependent variable that cannot be explained by the model; the model under/over predictions.

6.2.5 TECHNICAL STEPS

Before we proceed with the regression analysis, it is advisable to inspect the histogram of the dependent/target variable, in order to see if it needs to be transformed before fitting the regression model. The data for the selected soil property is normal when the frequency distribution of the values follow a bell-shaped curve (Gaussian distribution) which is symmetric around its mean. Normality tests may be used to assess normality. If a normality test indicates that data are not normally distributed, it may be necessary to transform the data to meet the normality assumption.

BOTH, THE NORMALITY TESTS AND THE DATA TRANSFORMATION CAN BE EASILY PERFORMED USING ANY COMMERCIAL OR OPEN SOURCE STATISTICAL TOOL (R, SPSS, MINITAB...)

The main steps for the multiple linear regression analysis are shown in the Figure 6.10.

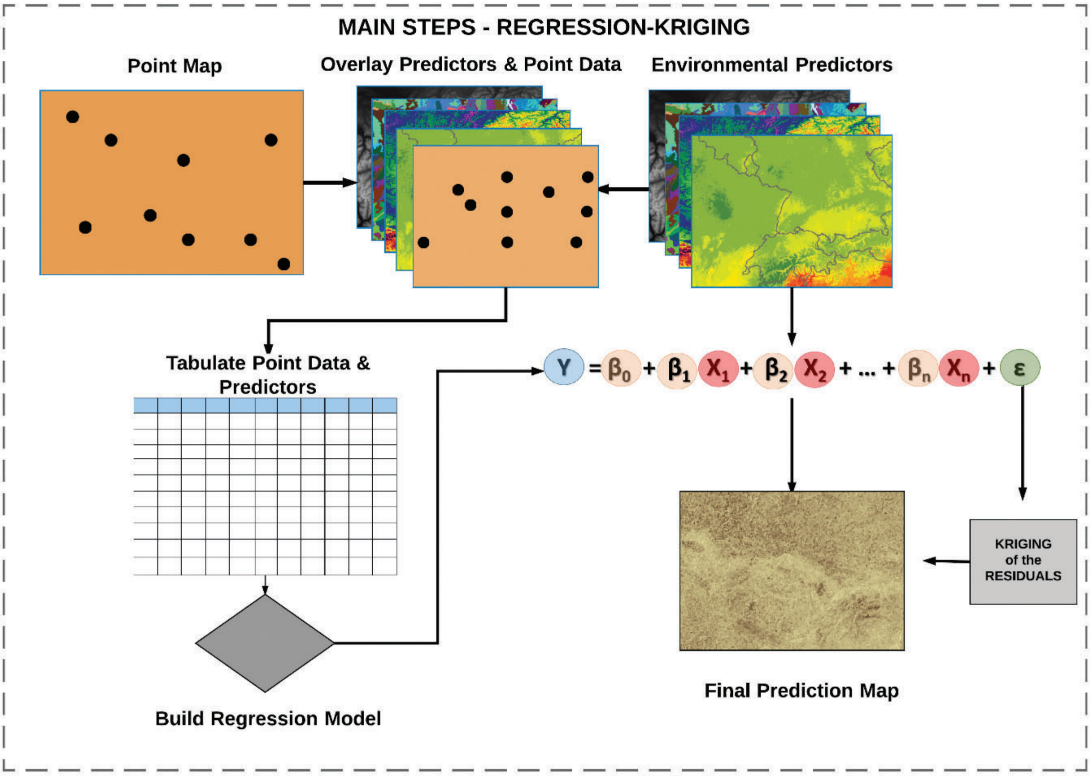

MAIN STEPS - REGRESSION-KRIGING
Point Map
Overlay Predictors & Point Data
Environmental Predictors
Tabulate Point Data & Predictors
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$
KRIGING of the RESIDUALS
Final Prediction Map
Build Regression Model

FIGURE 6.10 WORKFLOW FOR REGRESSION KRIGING

- THE FIRST STEP IS TO PREPARE A MAP SHOWING THE SPATIAL DISTRIBUTION OF THE SAMPLE LOCATIONS AND THE CORRESPONDING SOIL PROPERTY INFORMATION, E.G. SOIL ORGANIC MATTER AND ENVIRONMENTAL PROPERTIES. THE FIRST CAN BE ACHIEVED AS OUTLINED IN SECTION 4.1. THE OVERLAYING OPERATION CAN BE PERFORMED IN R, ARCGIS, SAGA GIS OR QGIS.

- THE ESSENTIAL PART OF MULTIPLE REGRESSION ANALYSIS IS TO BUILD A REGRESSION MODEL BY USING THE ENVIRONMENTAL PREDICTORS. AFTER EXTRACTING THE VALUES OF EXPLANATORY MAPS AND TARGET VARIABLES INTO THE SINGLE TABLE, WE CAN NOW START FITTING MULTIPLE REGRESSION MODEL USING THE TABLE THAT CONTAINS DATA FROM DEPENDENT VARIABLE AND PREDICTORS.

- IN PARTICULAR CASES, STEPWISE MULTIPLE LINEAR REGRESSION (SMLR) CAN BE USED TO ELIMINATE INSIGNIFICANT PREDICTORS. STEPWISE MULTIPLE LINEAR REGRESSION (SMLR) USUALLY SELECTS PREDICTORS THAT HAVE THE STRONGEST LINEAR CORRELATIONS WITH THE TARGET VARIABLE, WHICH REFLECT THE HIGHEST PREDICTIVE CAPACITY.

- KRIGING OF THE RESIDUALS (PREDICTION ERRORS): IN THE REGRESSION-KRIGING, THE REGRESSION MODEL DETRENDS THE DATA, PRODUCES THE RESIDUALS WHICH WE NEED TO KRIGE AND TO BE ADDED TO THE REGRESSION MODEL PREDICTIONS.

6.2.7 INTERPRET THE KEY RESULTS OF MULTIPLE REGRESSION

Regression analysis generates an equation to describe the statistical relationship between one or more predictor variables and the response variable. he r-squared, p-values and coefficients that appear in the output for linear regression analysis must also be reviewed. Before accepting the result of a linear regression it is important to evaluate its suitability at explaining the data. One of the many ways to do this is to visually examine the residuals. If the model is appropriate, then the residual errors should be random and normally distributed.

R-sq

R2 is the percentage of variation in the response that is explained by the model. The higher the R2 Value, the better the model fits your data. R-squared is always between 0% and 100%. R2 usually increases when additional predictors are added in the model.

P Values

To determine whether the association between the dependent and each predictor in the model is statistically significant, compare the p-value for the term to your significance level to assess the null hypothesis. Usually, a significance level of 0.05 works well.

P-value ≤ significance level: The relationship is statistically significant. If the p-value is less than or equal to the significance level, we can conclude that there is a statistically significant relationship between the dependent variable and the predictor.

P-value > significance level: The relationship is ***not*** statistically significant, If the p-value is greater than the significance level, you cannot conclude that there is a statistically significant relationship between the dependent variable and the predictor. You may want to refit the model without the predictor.

Residuals

We can plot the residuals which can help us determine whether the model is adequate and meets the assumptions of the analysis. If the model is appropriate, then the residual errors should be random and normally distributed. We can plot residuals versus fits to verify the assumption that the residuals are randomly distributed and have constant variance. Ideally, the points should fall randomly on both sides of "0", with no recognizable patterns in the points.

The diagnostic plots for the model should be evaluated to confirm if all the assumptions of linear regression are met. After the abovementioned assumptions are validated, we can proceed with making the prediction map using the model with significant predictors.

6.2.8 USING THE RESULTS OF A REGRESSION ANALYSIS TO MAKE PREDICTIONS

The purpose of a regression analysis, of course, is to develop a model that can be used to make the prediction of a dependent variable. The derived regression equation is to be used to create the prediction map for dependent variable.

Tip: Raster calculation can be easily performed using "raster" Package in R or ArcGIS using the "Raster Calculator" tool (It's called Map Algebra in the prior versions).

6.2.9 THE SOFTWARE

R: R and R Packages can be downloaded from *https://cran.r-project.org/*

SAGA GIS

The following modules are available for multiple regression analyses in SAGA GIS environment;

Multiple Linear Regression Analysis, Menu Access: Spatial and Geostatistics|Regression|Table

Multiple Regression Analysis (Grid/Grids), Menu Access: Spatial and Geostatistics|Regression

Multiple Regression Analysis (Points/Grids), Menu Access: Spatial and Geostatistics|Regression

SAGA GIS is available at *https://sourceforge.net/projects/saga-gis/files/*

QGIS

The following analyses are available in QGIS;

Multiple regression analysis (points/grids)

Multiple regression analysis (grid/grids)

QGIS is available at: *http://www.qgis.org/en/site/forusers/download.html*

ArcGIS: 60 day trial can be downloaded at *http://www.esri.com/software/arcgis/free-trial* (needs registration)

Further Reading

A Practical Guide to Geostatistical Mapping of Environmental Variables (Hengl T., 2009)

A Practical Guide to Geostatistical Mapping 2. Edition (Hengl T., 2009)

6.2.10 EXAMPLE: REGRESSION KRIGING

Requirements

The following are required to implement Regression Kriging in R;

1. Latest version of R software, network connection and sufficient RAM, storage capacity (Chapter 4)
2. Latest version of RStudio (Chapter 4)
3. R packages (sp, raster,rgdal, gstat, ithir) (Chapter 4)
4. Point Dataset (.txt or .csv) (Chapter 2)
5. Environmental predictors (covariates) (Chapter 3)

   a. Relief (e.g. DEM, Slope, TWI)

   b. Organism map (e.g. land use, NDVI, land cover)

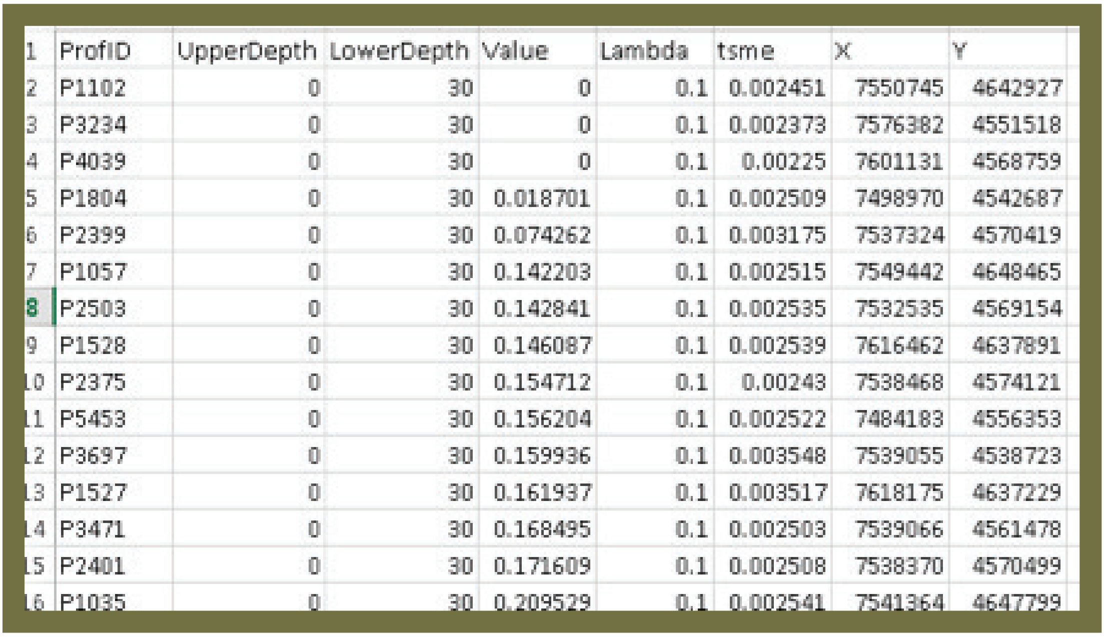
   c. Climate Data (e.g. mean precipitation, mean temperature)

   d. Parent material (parent material, geology)

STEP 1: DATA PREPARATION

Point Dataset

We previously applied spline function to produce continuous soil information to a given soil depth (0-30 cm) in the section 2.4. Spline function basically imports soil profile data (including instances where layers are not contiguous), fits it to a mass-preserving spline and outputs attribute means for a given depth. The output file should contain profile id, upper (surface) and lower depth (30cm), estimated value for the selected soil attribute (Value) and tmse (estimated mean squared error of the spline). If you used the Spline Tool V2, the coordinates were not kept in the output file. The coordinates should be added back in the data table. You can use Profile IDs to add the X, Y columns back. Once your point dataset is ready, copy this table into your working directory as a .csv file.



1	ProfID	UpperDepth	LowerDepth	Value	Lambda	tsme	X	Y
2	P1102	0	30	0	0.1	0.002451	7550745	4642927
3	P3234	0	30	0	0.1	0.002373	7576382	4551518
4	P4039	0	30	0	0.1	0.00225	7601131	4568759
5	P1804	0	30	0.018701	0.1	0.002509	7498970	4542687
6	P2399	0	30	0.074262	0.1	0.003175	7537324	4570419
7	P1057	0	30	0.142203	0.1	0.002515	7549442	4648465
8	P2503	0	30	0.142841	0.1	0.002535	7532535	4569154
9	P1528	0	30	0.146087	0.1	0.002539	7616462	4637891
10	P2375	0	30	0.154712	0.1	0.00243	7538468	4574121
11	P5453	0	30	0.156204	0.1	0.002522	7484183	4556353
12	P3697	0	30	0.159936	0.1	0.003548	7539055	4538723
13	P1527	0	30	0.161937	0.1	0.003517	7618175	4637229
14	P3471	0	30	0.168495	0.1	0.002503	7539066	4561478
15	P2401	0	30	0.171609	0.1	0.002508	7538370	4570499
16	P1035	0	30	0.209529	0.1	0.002541	7541364	4647799

FIGURE 6.11 POINT DATA

Environmental Predictors (Covariates)

In the Chapter 3, several global and continental datasets and access information can be found. In addition to these datasets, numerous covariate layers have been prepared by ISRIC for the GSOC Map project. These are GIS raster layers of various biophysical earth surface properties for each country in the world. Some of these layers will be used as predictors in this section. Please download the covariates for your own study area from GSOCMap Data Repository

STEP 2: SETTING WORKING SPACE AND INITIAL STEPS

One of the first steps should be setting our working directory. If you read/write files from/ to disk, this takes place in the working directory. If we don't set the working directory we could easily write files to an undesirable file location. The following example shows how to set the working directory in R to our folder which contains data for the study area (point data, covariates).

```
# Set the working directory
setwd("C:/masis")
```

Note that we must use the forward slash / or double backslash \\ in R! Single backslash \ will not work. Now we can check if the working directory has been correctly set by using the function:

```
# Check the working directory

> getwd()
[1] "C:/masis"
```

Now load the necessary R packages (you may need to install them onto your computer first):

```
# Install required packages if you have not installed them yet.

install.packages("raster")
install.packages("rgdal")
install.packages("gstat")
install.packages("ithir")

# Load required packages into the current R session.

library(sp)
library(raster)
library(rgdal)
library(gstat)
library(ithir)
```

STEP 3: DATA IMPORT (POINT DATA, COVARIATES)

Now we will import our point dataset using read.csv() function. The easiest way to create a data frame is to read in data from a file—this is done using the function read.csv, which works with comma delimited files. Data can be read in from other file formats as well, using different functions, but read.csv is the most commonly used approach. R is very flexible in how it reads in data from text files (read.table, read.csv, read.csv2, read.delim, read.delim2). Please type ?read.table() for help.

```
# Import pointdata into the session and assign the object as SOC

SOC <- read.csv("pointdata/mac-soc.csv")

# first 6 rows of our SOC object

> head(SOC)
     ProfID      X          Y          SOC
1    P1804       7498970    4542687    0.01870136
2    P2399       7537324    4570419    0.07426247
3    P1057       7549442    4648465    0.14220319
4    P2503       7532535    4569154    0.14284083
5    P1528       7616462    4637891    0.14608661
6    P2375       7538468    4574121    0.15471242
> coordinates(SOC) <- ~X + Y

# We can can use str()  for exploring the format and contents of any object created/
imported.

> str(SOC)
'data.frame': 3298 obs. of 4 variables:
 $ ProfID: Factor w/ 3224 levels "P0004","P0007",..: 771 1254 478 1349 606 1232 2708 1994 605
 1790 ...
 $ X     : int  7498970 7537324 7549442 7532535 7616462 7538468 7484183 7539055 7618175
 7539066 ...
 $ Y     : int  4542687 4570419 4648465 4569154 4637891 4574121 4556353 4538723 4637229
 4561478 ...
 $ SOC   : num  0.0187 0.0743 0.1422 0.1428 0.1461 ...
```

Since we will be working with spatial data we need to define the coordinates for the imported data. Using the coordinates() function from the sp package we can define the columns in the data frame to refer to spatial coordinates—here the coordinates are listed in columns X and Y.

```
> coordinates(SOC) <- ~X + Y
> str(SOC)
Formal class 'SpatialPointsDataFrame' [package "sp"] with 5 slots
 ..@ data     :'data.frame':     2308 obs. of  2 variables:
 .. ..$ ProfID: Factor w/ 3224 levels "P0004","P0007",..: 1037 770 1301 2162 1769 2127 1754
2667 208 1327 ...
 .. ..$ SOC   : num [1:2308] 2.655 1.246 0.713 0.997 0.875 ...
 ..@ coords.nrs : int [1:2] 2 3
 ..@ coords    : num [1:2308, 1:2] 7531078 7498201 7529385 7533904 7538405 ...
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:2308] "2701" "1240" "335" "803" ...
 .. .. ..$ : chr [1:2] "X" "Y"
 ..@ bbox      : num [1:2, 1:2] 7455723 4526565 7665953 4691342
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:2] "X" "Y"
 .. .. ..$ : chr [1:2] "min" "max"
 ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
 .. .. ..@ projargs: chr NA
```

SpatialPointsDataFrame structure is essentially the same data frame, except that additional "spatial" elements have been added or partitioned into slots. Some important ones being the bounding box (sort of like the spatial extent of the data), and the coordinate reference system proj4string(), which we need to define for the sample dataset. To define the CRS, we must know where our data are from, and what was the corresponding CRS used when recording the spatial information in the field. For this data set the CRS used was: Macedonia_State_Coordinate_System_zone_7 (EPSG:6316).

To clearly tell R this information we define the CRS which describes a reference system in a way understood by the PROJ.4 projection library http://trac.osgeo.org/proj/. An interface to the PROJ.4 library is available in the rgdal package. Alternative to using Proj4 character strings, we can use the corresponding yet simpler EPSG code (European Petroleum Survey Group). rgdal also recognizes these codes. If you are unsure of the Proj4 or EPSG code for the spatial data that you have, but know the CRS, you should consult *http://spatialreference.org/ for assistance*.

Please also note that, when working with spatial data, it's very important that the CRS (coordinate reference system) of the point data and covariates are the same.

Now, we will define our CRS;

```
> proj4string(SOC) <- CRS("+init=epsg:6316")
> str(SOC)
Formal class 'SpatialPointsDataFrame' [package "sp"] with 5 slots
  ..@ data      :'data.frame':     2308 obs. of  2 variables:
  .. ..$ ProfID: Factor w/ 3224 levels "P0004","P0007",..: 1037 770 1301 2162 1769 2127 1754
 2667 208 1327 ...
  .. ..$ SOC   : num [1:2308] 2.655 1.246 0.713 0.997 0.875 ...
  ..@ coords.nrs : int [1:2] 2 3
  ..@ coords     : num [1:2308, 1:2] 7531078 7498201 7529385 7533904 7538405 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:2308] "2701" "1240" "335" "803" ...
  .. .. ..$ : chr [1:2] "X" "Y"
  ..@ bbox      : num [1:2, 1:2] 7455723 4526565 7665953 4691342
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:2] "X" "Y"
  .. .. ..$ : chr [1:2] "min" "max"
  ..@ proj4string:Formal class 'CRS' [package "sp"] with 1 slot
  .. .. ..@ projargs: chr "+init=epsg:6316 +proj=tmerc +lat_0=0 +lon_0=21 +k=0.9999
 +x_0=7500000 +y_0=0 +ellps=bessel +towgs84=682,-203,480,0,0,0,0 +units"| __truncated__
```

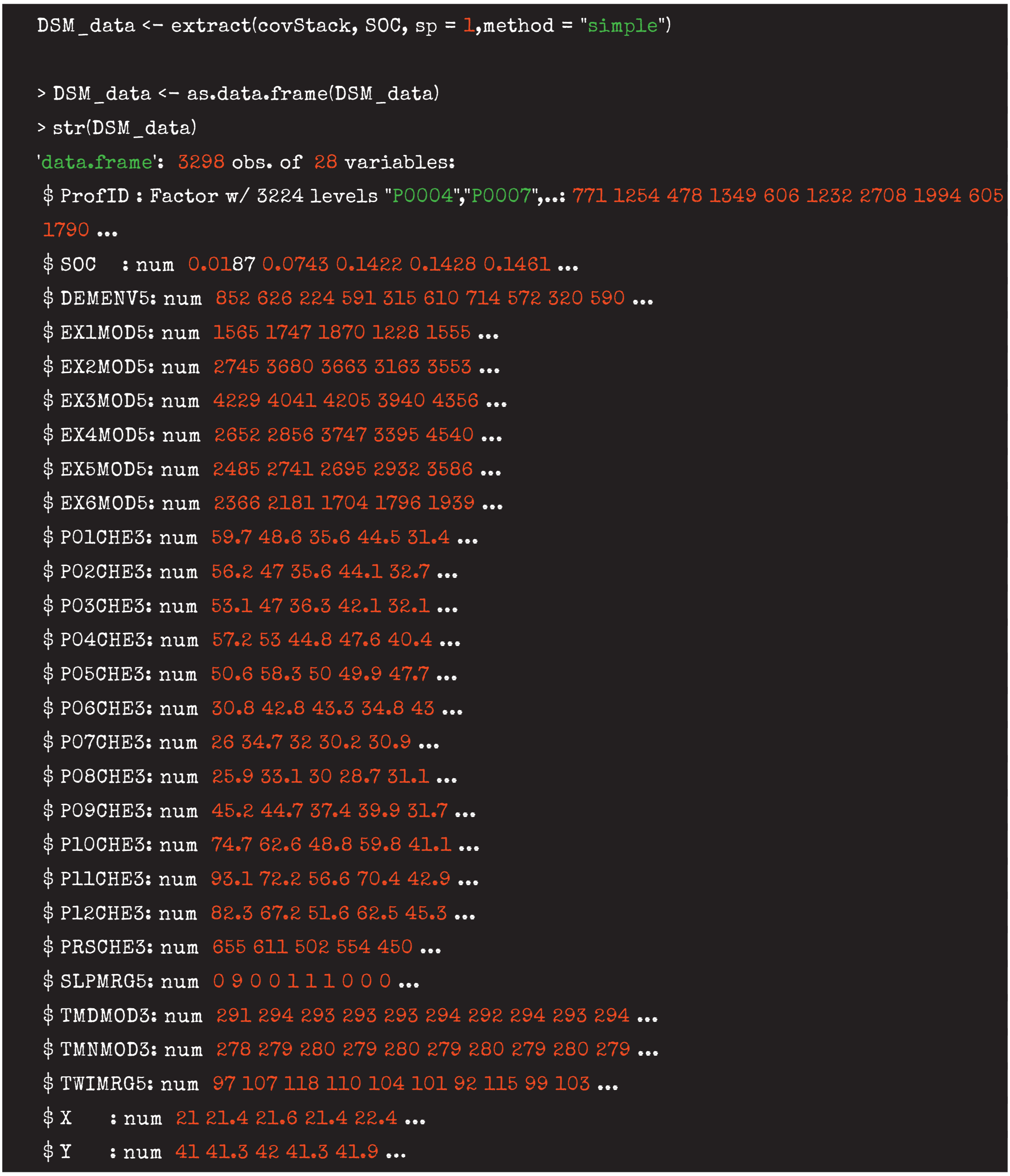
Now we will import the covariates. When the covariate layers are in common resolution and extent, rather than working with individual rasters it is better to stack them all into a single R object. We will use stack() function from raster package. In this example we use 12 covariates from the GSOCMap Data Repository.

```
> Covariates <- list.files(path = "c:/masis/cov/", pattern = "\\.tif$", full.names = TRUE)
> Covariates
 [1] "c:/masis/cov/DEMENV5.tif" "c:/masis/cov/EX1MOD5.tif"
 [3] "c:/masis/cov/EX2MOD5.tif" "c:/masis/cov/EX3MOD5.tif"
 [5] "c:/masis/cov/EX4MOD5.tif" "c:/masis/cov/EX5MOD5.tif"
 [7] "c:/masis/cov/EX6MOD5.tif" "c:/masis/cov/P01CHE3.tif"
 [9] "c:/masis/cov/P02CHE3.tif" "c:/masis/cov/P03CHE3.tif"
[11] "c:/masis/cov/P04CHE3.tif" "c:/masis/cov/P05CHE3.tif"
[13] "c:/masis/cov/P06CHE3.tif" "c:/masis/cov/P07CHE3.tif"
[15] "c:/masis/cov/P08CHE3.tif" "c:/masis/cov/P09CHE3.tif"
[17] "c:/masis/cov/P10CHE3.tif" "c:/masis/cov/P11CHE3.tif"
[19] "c:/masis/cov/P12CHE3.tif" "c:/masis/cov/PRSCHE3.tif"
[21] "c:/masis/cov/SLPMRG5.tif" "c:/masis/cov/TMDMOD3.tif"
[23] "c:/masis/cov/TMNMOD3.tif" "c:/masis/cov/TWIMRG5.tif"
```

Now we can stack covariates into an object;

```
covStack <- stack(Covariates)
```

In order to carry out digital soil mapping in terms of examining the statistical significance of environmental predictors for explaining the spatial variation of soil organic carbon, we should link both sets of data together and extract the values of the covariates at the locations of the soil point data. Note that the stacking of rasters can only be possible if they are in the same resolution and extent. If they are not, raster package resample and projectRaster functions are for harmonising all your different raster layers. With the stacked rasters (Covstack), we can now perform the intersection and extraction.



```
DSM_data <- extract(covStack, SOC, sp = 1,method = "simple")

> DSM_data <- as.data.frame(DSM_data)
> str(DSM_data)
'data.frame':  3298 obs. of  28 variables:
 $ ProfID : Factor w/ 3224 levels "P0004","P0007",..: 771 1254 478 1349 606 1232 2708 1994 605
1790 ...
 $ SOC    : num  0.0187 0.0743 0.1422 0.1428 0.1461 ...
 $ DEMENV5: num  852 626 224 591 315 610 714 572 320 590 ...
 $ EX1MOD5: num  1565 1747 1870 1228 1555 ...
 $ EX2MOD5: num  2745 3680 3663 3163 3553 ...
 $ EX3MOD5: num  4229 4041 4205 3940 4356 ...
 $ EX4MOD5: num  2652 2856 3747 3395 4540 ...
 $ EX5MOD5: num  2485 2741 2695 2932 3586 ...
 $ EX6MOD5: num  2366 2181 1704 1796 1939 ...
 $ PO1CHE3: num  59.7 48.6 35.6 44.5 31.4 ...
 $ PO2CHE3: num  56.2 47 35.6 44.1 32.7 ...
 $ PO3CHE3: num  53.1 47 36.3 42.1 32.1 ...
 $ PO4CHE3: num  57.2 53 44.8 47.6 40.4 ...
 $ PO5CHE3: num  50.6 58.3 50 49.9 47.7 ...
 $ PO6CHE3: num  30.8 42.8 43.3 34.8 43 ...
 $ PO7CHE3: num  26 34.7 32 30.2 30.9 ...
 $ PO8CHE3: num  25.9 33.1 30 28.7 31.1 ...
 $ PO9CHE3: num  45.2 44.7 37.4 39.9 31.7 ...
 $ P10CHE3: num  74.7 62.6 48.8 59.8 41.1 ...
 $ P11CHE3: num  93.1 72.2 56.6 70.4 42.9 ...
 $ P12CHE3: num  82.3 67.2 51.6 62.5 45.3 ...
 $ PRSCHE3: num  655 611 502 554 450 ...
 $ SLPMRG5: num  0 9 0 0 1 1 1 0 0 0 ...
 $ TMDMOD3: num  291 294 293 293 293 294 292 294 293 294 ...
 $ TMNMOD3: num  278 279 280 279 280 279 280 279 280 279 ...
 $ TWIMRG5: num  97 107 118 110 104 101 92 115 99 103 ...
 $ X      : num  21 21.4 21.6 21.4 22.4 ...
 $ Y      : num  41 41.3 42 41.3 41.9 ...
```

It would be better to progress with a data frame of just the data and covariates required for the modelling. In this case, we will subset the columns SOC, the covariates and the the spatial coordinates (X and Y).

```
> DSM_data <- DSM_data[, c(2:28)]
```

After the extraction, It's useful to check if there are missing values (NAs) both in the target variable and covariates. In these cases, these data should be excluded. A quick way to assess if there are missing or NA values in the data is to use the complete.cases() function.

```
> which(!complete.cases(DSM_data))
[1] 1693 2196 2328 2460 2643 2747
> DSM_data <- DSM_data[complete.cases(DSM_data), ]
> which(!complete.cases(DSM_data))
integer(0)
```

After removing NAs now there do not appear to be any missing data as indicated by the integer(0) output above. It means we have zero rows with missing information.

STEP 4: FITTING THE MLR MODEL

Fitting the MLR Model

Let's fit a linear model using with all available covariates.

```
> MLR.Full <- lm(SOC ~ DEMENV5+EX1MOD5+EX2MOD5+EX3MOD5+EX4MOD5+EX5MOD5+EX6MOD5+
P01CHE3+P02CHE3+P03CHE3+P04CHE3+P05CHE3+P06CHE3+P07CHE3+P08CHE3+P09CHE3+P10CHE3+
P11CHE3+P12CHE3+PRSCHE3+SLPMRG5+TMDMOD3+TMNMOD3+TWIMRG5, data = DSM_data)
> summary(MLR.Full)

Call:
lm(formula = SOC ~ DEMENV5 + EX1MOD5 + EX2MOD5 + EX3MOD5 + EX4MOD5 +
    EX5MOD5 + EX6MOD5 + P01CHE3 + P02CHE3 + P03CHE3 + P04CHE3 +
    P05CHE3 + P06CHE3 + P07CHE3 + P08CHE3 + P09CHE3 + P10CHE3 +
    P11CHE3 + P12CHE3 + PRSCHE3 + SLPMRG5 + TMDMOD3 + TMNMOD3 +
    TWIMRG5, data = DSM_data)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-3.820 -0.672 -0.182  0.418 47.116 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.139e+01  1.884e+01   1.135  0.25640    
DEMENV5      2.932e-04  4.417e-04   0.664  0.50692    
EX1MOD5     -1.437e-04  1.298e-04  -1.107  0.26845    
EX2MOD5     -9.832e-05  8.395e-05  -1.171  0.24165    
EX3MOD5      2.160e-04  7.195e-05   3.002  0.00270 ** 
EX4MOD5      5.510e-04  9.026e-05   6.104 1.15e-09 ***
EX5MOD5     -6.156e-04  1.210e-04  -5.086 3.87e-07 ***
EX6MOD5      2.377e-04  1.296e-04   1.834  0.06670 .  
PO1CHE3     -1.243e-01  5.235e-01  -0.237  0.81238    
PO2CHE3      3.139e-02  5.159e-01   0.061  0.95149    
PO3CHE3      3.151e-02  5.139e-01   0.061  0.95110    
PO4CHE3     -2.486e-03  5.195e-01  -0.005  0.99618    
PO5CHE3      4.119e-02  5.191e-01   0.079  0.93675    
PO6CHE3     -8.796e-03  5.156e-01  -0.017  0.98639    
PO7CHE3      1.904e-01  5.119e-01   0.372  0.70989    
PO8CHE3     -2.415e-01  5.275e-01  -0.458  0.64704    
PO9CHE3      1.088e-01  5.196e-01   0.209  0.83418    
P10CHE3     -7.561e-02  5.158e-01  -0.147  0.88347    
P11CHE3      1.015e-02  5.205e-01   0.020  0.98444    
P12CHE3      7.301e-02  5.148e-01   0.142  0.88724    
PRSCHE3     -3.012e-03  5.170e-01  -0.006  0.99535    
SLPMRG5      2.699e-03  5.789e-03   0.466  0.64116    
TMDMOD3     -1.013e-01  3.606e-02  -2.810  0.00498 ** 
TMNMOD3      2.548e-02  5.175e-02   0.492  0.62254    
TWIMRG5      5.749e-03  4.560e-03   1.261  0.20748    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.874 on 3267 degrees of freedom
Multiple R-squared:  0.1845,  Adjusted R-squared:  0.1785 
F-statistic: 30.79 on 24 and 3267 DF,  p-value: < 2.2e-16
```

From the summary of our fill model (MRL.full) above, it seems only a few of the covariates are significant in describing the spatial variation of the target variable. To determine the most predictive model we can run a stepwise regression using the step() function. With this function we can also specify the directions that we want to step.

```
> MLR.Step <- step(MLR.Full, trace = 0, direction="both")
> summary(MLR.Step)

Call:
lm(formula = SOC ~ EX1MOD5 + EX3MOD5 + EX4MOD5 + EX5MOD5 + EX6MOD5 +
   P01CHE3 + P02CHE3 + P05CHE3 + P07CHE3 + P08CHE3 + P09CHE3 +
   P10CHE3 + P12CHE3 + TMDMOD3, data = DSM_data)

Residuals:
  Min     1Q Median     3Q    Max
-3.680 -0.673 -0.184  0.421 47.063

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.394e+01  7.687e+00   4.416 1.04e-05 ***
EX1MOD5     -2.021e-04  1.166e-04  -1.734 0.083041 .
EX3MOD5      1.831e-04  6.171e-05   2.967 0.003033 **
EX4MOD5      5.700e-04  8.709e-05   6.545 6.89e-11 ***
EX5MOD5     -6.191e-04  1.179e-04  -5.253 1.59e-07 ***
EX6MOD5      1.921e-04  1.126e-04   1.706 0.088018 .
P01CHE3     -1.021e-01  3.021e-02  -3.379 0.000735 ***
P02CHE3      4.224e-02  2.173e-02   1.944 0.052039 .
P05CHE3      4.298e-02  1.065e-02   4.034 5.60e-05 ***
P07CHE3      1.774e-01  3.017e-02   5.882 4.47e-09 ***
P08CHE3     -2.564e-01  3.609e-02  -7.104 1.49e-12 ***
P09CHE3      1.051e-01  2.123e-02   4.948 7.86e-07 ***
P10CHE3     -6.325e-02  1.739e-02  -3.638 0.000279 ***
P12CHE3      5.390e-02  2.663e-02   2.024 0.043083 *
TMDMOD3     -1.182e-01  2.511e-02  -4.706 2.64e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.873 on 3277 degrees of freedom
Multiple R-squared:  0.1836,  Adjusted R-squared:  0.1801
F-statistic: 52.63 on 14 and 3277 DF,  p-value: < 2.2e-16
```

Now we can evaluate the test statistics of the calibration model using the goof() function from the "ithir" package.

```
> MLR.pred.rhC <- predict(MLR.rh, DSM_data[training, ])
> MLR.pred.rhV <- predict(MLR.rh, DSM_data[-training, ])
> goof(observed = DSM_data$SOC[training], predicted= MLR.pred.rhC)
       R2 concordance     MSE    RMSE        bias
1 0.1738919   0.2966572 4.029811 2.007439 -1.019185e-13
> goof(observed = DSM_data$SOC[-training], predicted= MLR.pred.rhV)
       R2 concordance     MSE    RMSE     bias
1 0.2205209   0.3955953 2.255178 1.501725 0.0857483
```

STEP 5: PREDICTION AND RESIDUAL KRIGING

Now we can make the predictions and plot the map. We can use either our DSM_data table for covariate values or covStack object for making our prediction. Using stack avoids the step of arranging all covariates into table format. If multiple rasters are being used, it is necessary to have them arranged as a rasterStack object. This is useful as it also ensures all the rasters are of the same extent and resolution. Here we can use the raster predict function such as below using the covStack raster stack as we created in the Step 3.

```
> par(mfrow = c(3, 1))
> map.MLR.r.pred <- predict(covStack, MLR.rh, "SOC_030cm_MLR_pred.tif",
format = "GTiff", datatype = "FLT4S", overwrite = TRUE)
> plot(map.MLR.r.pred, main = "MLR predicted SOC Map of Macedonia") #Figure 6.12
```

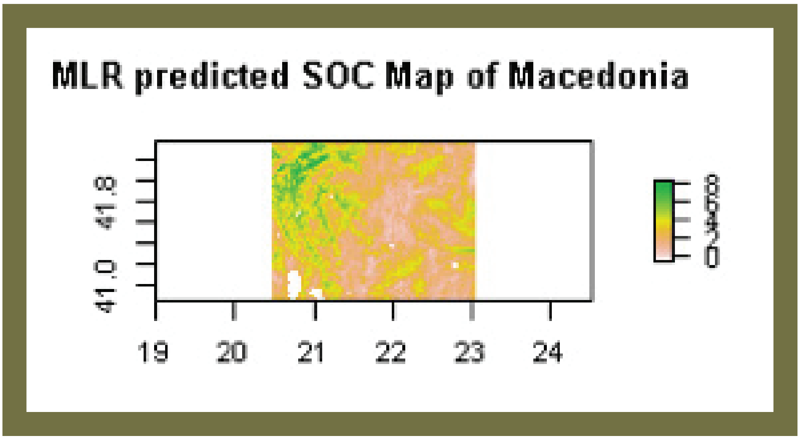


FIGURE 6.12 MULTIPLE LINEAR REGRESSION PREDICTED SOC MAP

Residual Kriging

Now, we can derive the model residual which is the model prediction subtracted from the residual.

```
> Mdata <- DSMTable[training, ]
> Mdata$residual <- Mdata$SOC - predict(Model.1.Stepwise, newdata = Mdata)
> mean(Mdata$residual)
[1] -0.3364321

> map.RK2 <- interpolate(covStack, gRK, xyOnly = TRUE, index = 1,filename
= "SOC0-30cm_residualRK.tif", format = "GTiff",datatype = "FLT4S",
overwrite = TRUE)
[using ordinary kriging]
> map.RK1 <- map.MLR.r.pred
> map.RK2 <- interpolate(covStack, gRK, xyOnly = TRUE, index = 1,filename
= "SOC0-30cm_residualRK.tif", format = "GTiff",datatype = "FLT4S",
overwrite = TRUE)
[using ordinary kriging]
>
> pred.stack <- stack(map.RK1, map.RK2)
Error in compareRaster(x) : different CRS
> proj4string(map.RK1) <- CRS("+init=epsg:6316")
> pred.stack <- stack(map.RK1, map.RK2)
> map.RK3 <- calc(pred.stack, fun = sum,filename = "Macedonia SOC Final
Map", format = "GTiff", progress = "text", overwrite = T)
> par(mfrow = c(3, 1))
> plot(map.RK2, main = "Kriged residual")
> plot(map.RK1, main = "Regression Map")
> plot(map.RK3, main = "Macedonia SOC Map 0-30cm")
#Figure 6.13
```

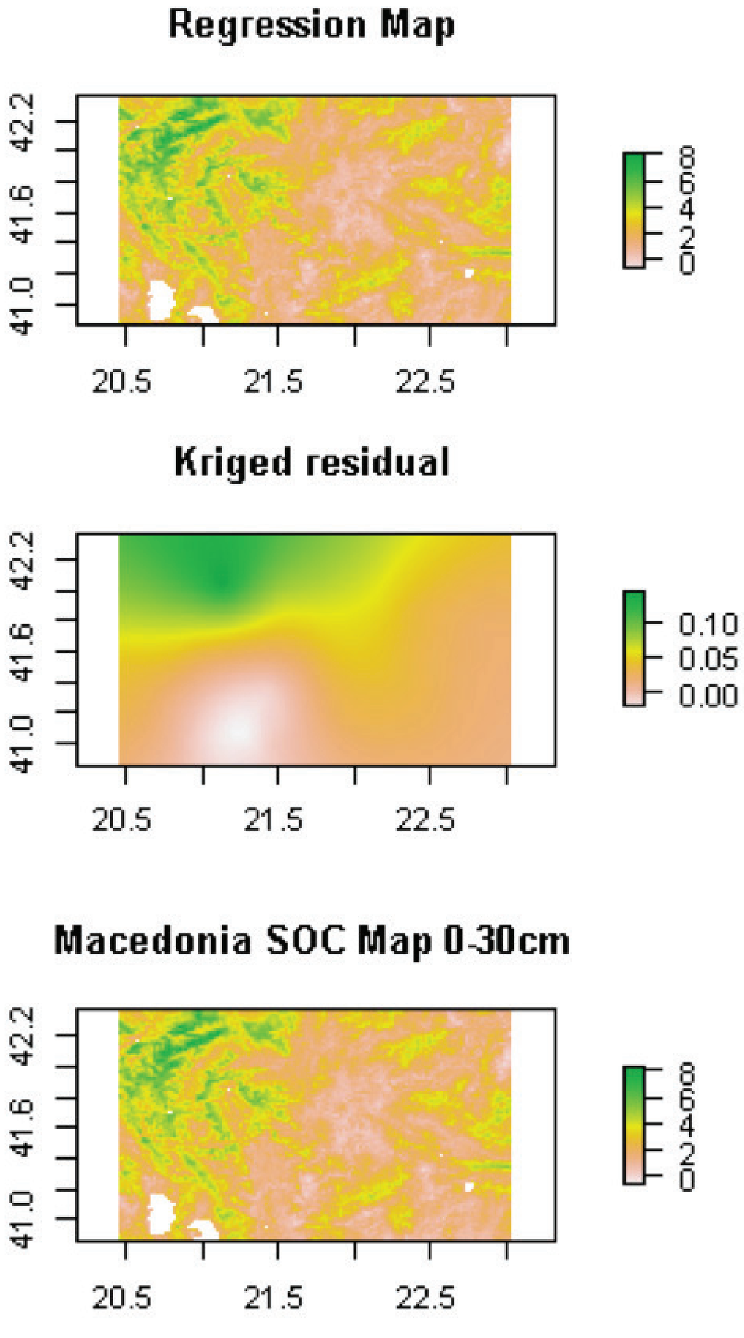


FIGURE 6.13 MLR PREDICTED SOC MAP, RESIDUAL MAP AND FINAL SOC MAP

6.2.12 References

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7618-4

Bontemps, S., Defourny, P., Van Bogaert, E., Arino, O., Kalogirou, V., Perez, J.R., 2011. chapter 8 "Interpolation and Geostatistics" in Bivand, R., Pebesma, E., Rubio, V., (2008). Applied Spatial Data Analysis with R. Use R Series, Springer, Heidelberg, pp. 378.

GLOBCOVER 2009. Product Description and Validation Report.

Gupta, S. (2015). A Multiple Regression Technique in Data Mining. International Journal of Computer Applications, 126(5).

Hengl T., Heuvelink G.B.M., Kempen B., Methods to fit a regression-kriging model, http://gsif.r-forge.r-project.org/fit.gstatModel.html. Accessed, April, 2017.

Hengl T., Heuvelink G.B.M., Rossiter D.G., 2007. About regression-kriging: from equations to case studies. Computers and Geosciences, 33(10): 1301-1315.

Hengl, T. (2009). A Practical Guide to Geostatistical Mapping, 2nd Edt. University of Amsterdam, www.lulu.com, 291 p.

Hoffmann, J. P., & Shafer, K. (2005). Linear regression analysis: Assumptions and applications. Department of Sociology Brigham Young University.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. Applied Statistics, 300-303.

Malone, Brendan P., Budiman Minasny, and Alex B. McBratney. "Using R for Digital Soil Mapping." (2016).

Meinshausen, N. (2006). Quantile regression forests. The Journal of Machine Learning Research, 7, 983-999.

Vasques, G. M., Grunwald, S., & Sickman, J. O. (2009). Modeling of soil organic carbon fractions using visible–near-infrared spectroscopy. Soil Science Society of America Journal, 73(1), 176-184.

6.3. DATA MINING: RANDOM FOREST

6.3.1 OVERVIEW

Random forest is a type of machine learning for uncovering statistical relationship between a dependent variable (e.g. soil property) and its predictors. It belongs to the decision-tree class of models in which the models (also known as classifiers) are like trees with stem, many branches, and leaves. The leaves are the prediction outcomes (final decisions) that flow from the roots through the stem to the branches (Breiman *et al.*, 1984). The decision tree model recursively splits the data into final uniform groups (classes) or unique values based on a set of rules. In random forest, there are many decision trees and each tree recursively splits randomly selected sub-samples from the data (Figure 6.14). The name random forest originates from the fact that the original data is first randomly split into sub-samples, and many decision trees (or forest) are used to model the sub-samples.

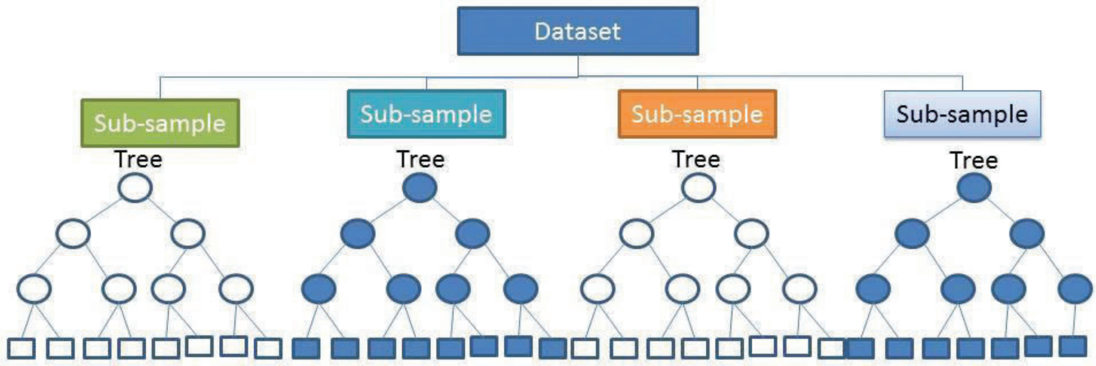


FIGURE 6.14 THE CONCEPT OF RANDOM FOREST AND DECISION TREES

Random forest has been tested by many researchers in digital soil mapping (see for example Poggio *et al.*, 2013; Pahlavan Rad *et al.*, 2014, and references therein). Specifically in soil carbon mapping, there are authors who have shown that it holds a lot of promise when compared to other prediction models. They have demonstrated that it has a relatively improved accurate spatial prediction, is a better approach to dealing with model over-fitting and data noise, and is capable of handling both dimensionally linear and nonlinear relationships (Wiesmeier *et al.*, 2011). Furthermore, with the advent of open-source platforms and freely downloadable ancillary data, the application of random forest and other such models has increasingly become more appealing in digital soil mapping.

The objective of this chapter is to demonstrate how random forest can be implemented in freely downloadable R software for spatial prediction of soil organic carbon. The R package of random forest, known as *randomForest*, was used (Breiman and Cutler, 2017).

6.3.2 REQUIREMENTS

The following are required to implement the Random Forest modelling of SOC in R:

1. R packages (randomForest,ggplot2, fBasics, nortest, car, sp, rgdal, Hmisc)
2. Georeferenced SOC data (in spreadsheet or GIS database)
3. Georeferenced spatial predictors (covariates)
    a. Relief map (e.g. DEM, landform)
    b. Organism map (e.g. land use, NDVI, land cover)
    c. Climate map (e.g. mean precipitation, mean temperature)
    d. Parent material (e.g. geology)
4. Latest version of R software and sufficient RAM and HDD storage capacity
5. Latest version of RStudio (optional but important)

6.3.3 EXAMPLE: RANDOM FOREST

STEP

STEP 1: DATA PREPARATION

The following sample data demonstrate the data requirement characteristics and application of random forest in mapping SOC (Figure 6.15) The soil data was obtained from a study of SOC in north-eastern Kenya. The data was collected using a Y-shape sampling frame (Omuto, 2008) for topsoil (0-30 cm) (Figure 6.16).

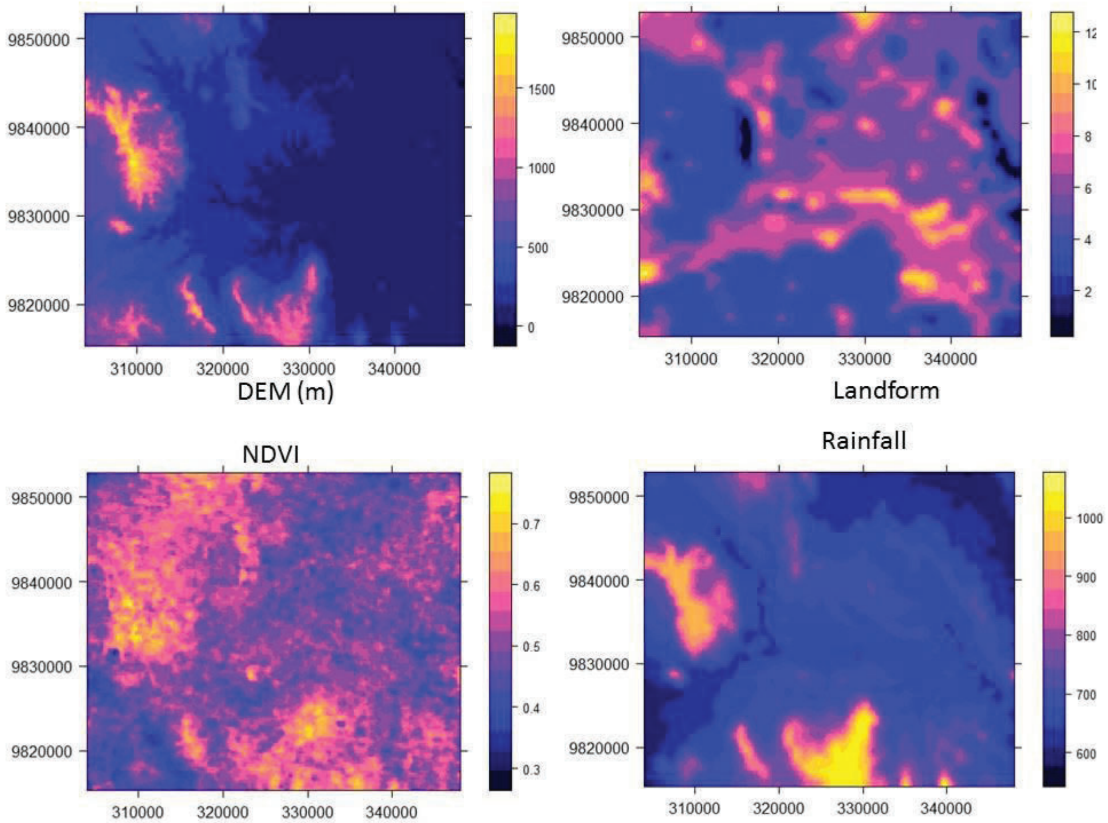


FIGURE 6.15 EXAMPLE COVARIATES FOR MAPPING SOC

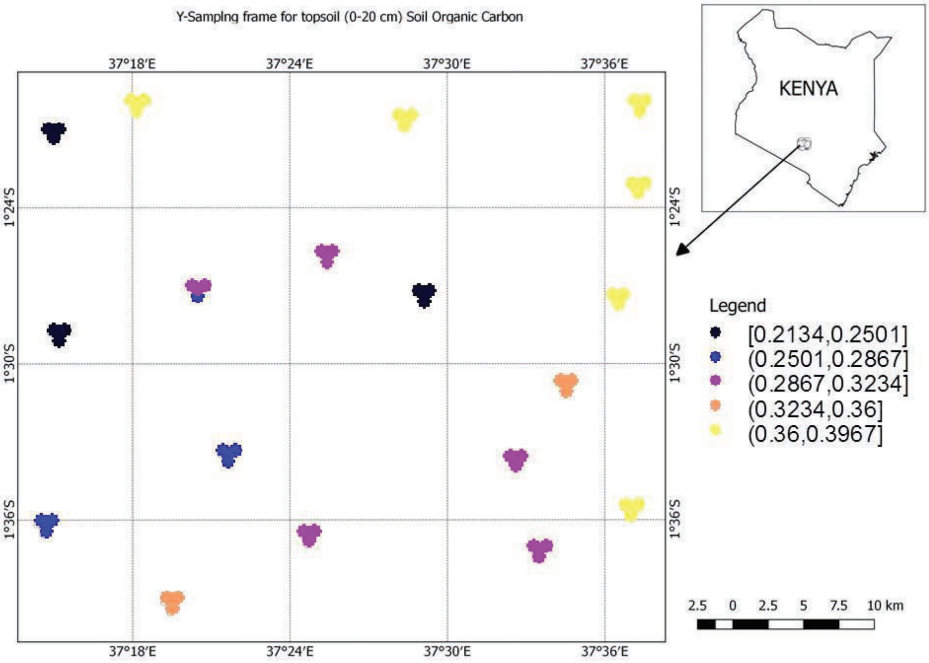
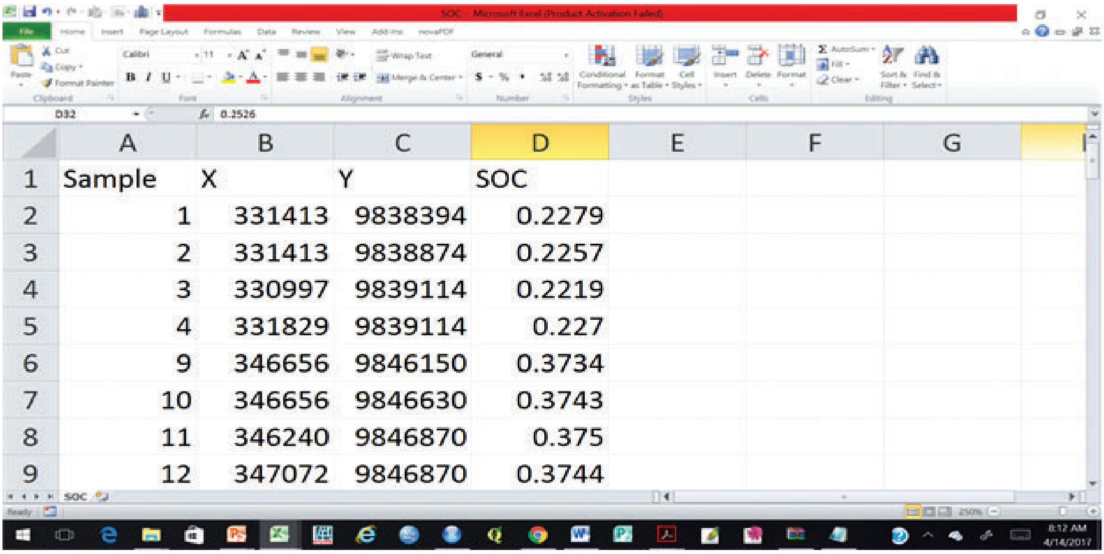


FIGURE 6.16 LOCATION OF SAMPLE DATASET AND SPATIAL DISTRIBUTION OF ITS SOC VALUES

The following table shows how the data should be arranged in the spreadsheet database such as MS Excel or Arc-Shapefile. Note that the first row should contain the header with names for the columns. Although the database can have many columns, the necessary columns are: Sample name, spatial coordinates (latitudes and longitudes), and the SOC values. In the example in the Figure 6.17, the three columns are Sample (for sample name), X (for longitude), Y (for latitude), and SOC (for SOC values in g/kg). This data can be saved as text file (such as Tab delimited or CSV text file) in MS Excel or it can be a GIS vector data (such as shapefile). The illustration given in this chapter uses Tab delimited text-file (in which the saved data is denoted as SOC.txt).



	A	B	C	D
1	Sample	X	Y	SOC
2	1	331413	9838394	0.2279
3	2	331413	9838874	0.2257
4	3	330997	9839114	0.2219
5	4	331829	9839114	0.227
6	9	346656	9846150	0.3734
7	10	346656	9846630	0.3743
8	11	346240	9846870	0.375
9	12	347072	9846870	0.3744

FIGURE 6.17 ARRANGEMENT OF SOC SOIL DATA IN SPREADSHEET DATABASE

6.3.4 TECHNICAL STEPS

The following scripts and steps are used for implementing Random Forest approach in R.

STEP

STEP 2: SET THE WORKING DIRECTORY

```
>setwd("C:/DSM/soildata")
```

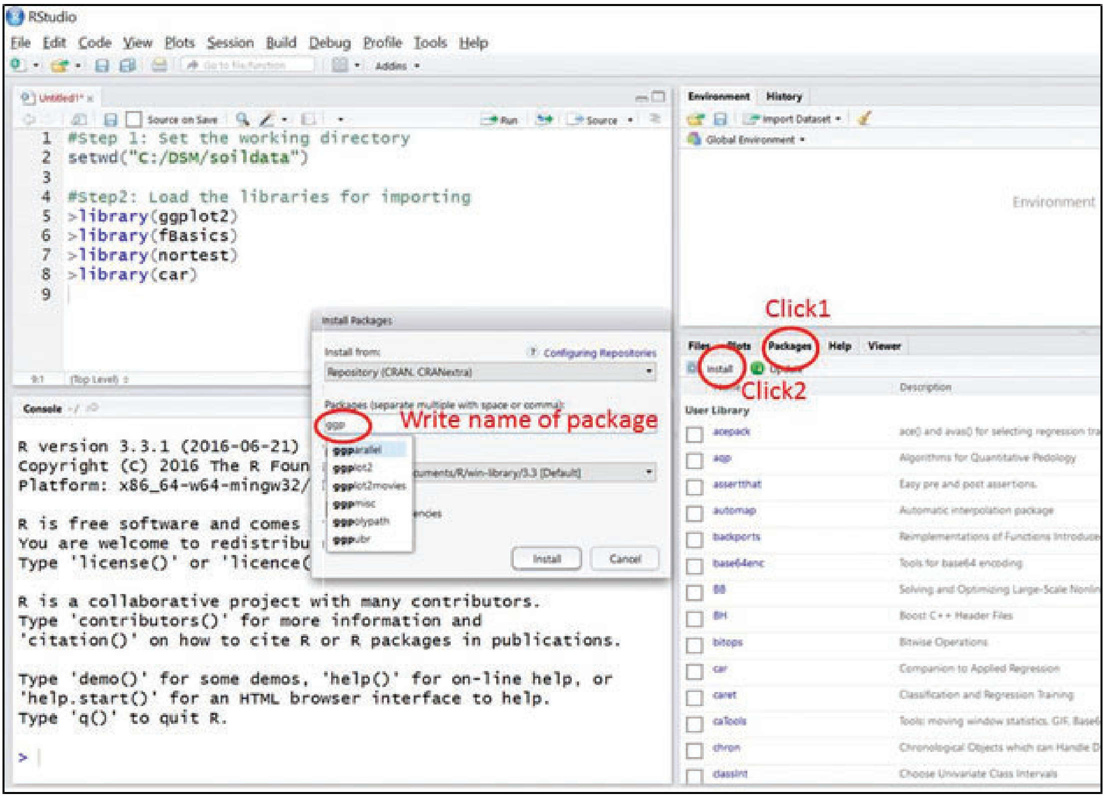
This first step is important for creating the path to the working directory where the data is stored. It's important to note the single-forward-slash between the directory path items. In the next step, the R packages for data exploration are supposed to have been installed in R (from CRAN repository) before loading them.

STEP

STEP 3: LOAD THE LIBRARIES FOR IMPORTING

```
library(ggplot2)
library(fBasics)
library(nortest)
library(car)
```

In case the libraries are not yet installed in the R environment, this can be done from R-Studio as shown in Figure 6.18 Internet connectivity is required to download the packages.



```
#Step 1: Set the working directory
setwd("C:/DSM/soildata")

#Step2: Load the libraries for importing
>library(ggplot2)
>library(fBasics)
>library(nortest)
>library(car)
```

```
R version 3.3.1 (2016-06-21)
Copyright (C) 2016 The R Foun
Platform: x86_64-w64-mingw32/

R is free software and comes
You are welcome to redistribu
Type 'license()' or 'licence(

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Install Packages

Install from: Repository (CRAN, CRANextra)

Packages (separate multiple with space or comma): ggp

ggparallel
ggplot2
ggplot2movies
ggpmisc
ggpolypath
ggpubr

…cuments/R/win-library/3.3 [Default]

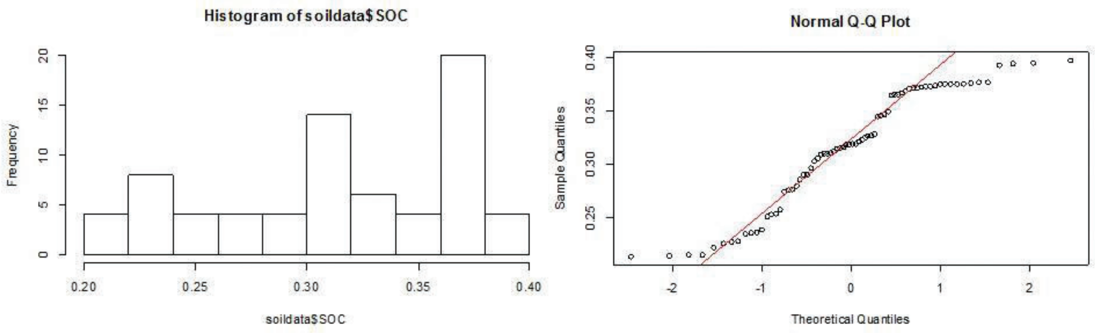
	Description
User Library	
□ acepack	ace() and avas() for selecting regression tra
□ aqp	Algorithms for Quantitative Pedology
□ assertthat	Easy pre and post assertions.
□ automap	Automatic interpolation package
□ backports	Reimplementations of Functions Introduce
□ base64enc	Tools for base64 encoding
□ BB	Solving and Optimizing Large-Scale Nonlin
□ BH	Boost C++ Header Files
□ bitops	Bitwise Operations
□ car	Companion to Applied Regression
□ caret	Classification and Regression Training
□ caTools	Tools: moving window statistics, GIF, Base6
□ chron	Chronological Objects which can Handle D
□ classInt	Choose Univariate Class Intervals

FIGURE 6.18 INSTALLING R PACKAGES IN RSTUDIO ENVIRONMENT

STEP 4: EXPLORE THE DATA

```
>soildata=read.table("soc.txt",header=T)
>summary(soildata)
 Sample              X              Y                   SOC
1st Qu.: 22.75      1st Qu.:313500  1st Qu.  :9823827   1st Qu.  :0.2759
Median: 100.50      Median :327265  Median   :9837261   Median   :0.3181
Mean: 93.39         Mean   :326966  Mean     :9835610   Mean     :0.3152
3rd Qu. : 150.25    3rd Qu.:341500  3rd Qu.  :9846690   3rd Qu.  :0.3703
Max. : 180.00       Max.   :347178  Max.     :9852724   Max.     :0.3967

>hist(soildata$SOC, breaks = 10)#Figure 6.17a
>qqnorm(soildata$SOC, plot.it = T)#Figure 6.17b
>qqline(soildata$SOC,col="red")
>ggplot(soildata, aes(x = X, y = Y)) + geom_point(aes(size = soildata$SOC))
>sampleSKEW(soildata$SOC)#Coefficient of skew
  SKEW
0.1052213
>sampleKURT(soildata$SOC)#Kurtosis
  KURT
1.101113
>ad.test(soildata$SOC)#Anderson-Darling Test
Anderson-Darling normality test
data:  soildata$SOC
A = 1.6854, p-value = 0.0002298
```



(a) Histogram

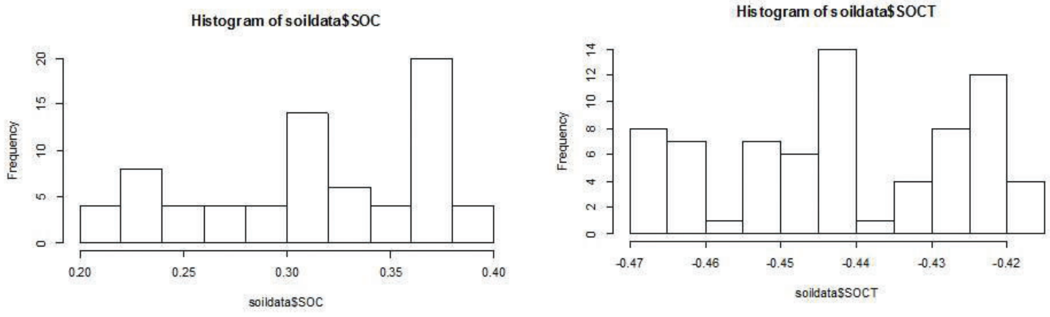
(b) Quantile-quantile plot

FIGURE 6.19 HISTOGRAM AND QQ-PLOT OF SOC DATA

The exploratory analysis of the data showed that Soil Organic Content (g/kg) is not normally distributed (Anderson-Darling test<0.05), positively skewed (Skew>0) and has a high degree of peakedness (Kurtosis > 1). Furthermore, the data has high values in the northeast corner and low values in the western side; giving the impression of west-northeast low-high pattern (Figure 6.5). In general, the exploratory data analysis shows that the data need transformation to normalize it before subjecting it to spatial modelling. The Box-Cox transformation (Box-Cox, 1964), can be used to transform the data in the next step.

STEP 5: TRANSFORM THE DATA USING BOX-COX TRANSFORMATION

```
>soildata$SOCT=(soildata$SOC^(as.numeric(powerTransform(soildata$SOC,
family ="bcPower")["lambda"]))-
1)/(as.numeric(powerTransform(soildata$SOC, family
="bcPower")["lambda"]))
>hist(soildata$SOCT,breaks = 15)# Figure 6.20b
```



(a) Histogram of non-transformed SOC

(b) Histogram of transformed SOC

FIGURE 6.20 HISTOGRAM OF RAW AND TRANSFORMED SOC

The spatial covariates for mapping soil data need also to be loaded into R and aligned with the soil data. According Jenny (1941) and McBratney *et al.* (2003), the covariates for mapping are the following soil forming factors: other available and correlated soil properties, climate data, land use/cover, relief, spatial reference, and geology. Many researchers have used varied forms and combinations of these soil forming factors to predict soil organic carbon. For example, Grimm *et al.* (2008)

used relief attributes (curvature, topographic wetness index, slope, aspect, etc.), soil attributes (colour and texture), forest history, and geology to predict soil carbon concentrations in Barro Colorado Island in Panama. Adhikari *et al.* (2014) used relief attributes (elevation, topographic wetness index, and valley bottom flatness), precipitation, land use, soil type, and wetlands to predict soil organic carbon in Denmark. In the present example, the following covariates were used: landform, rainfall, Normalized Difference Vegetation Index (NDVI), elevation, and spatial coordinates (latitudes and longitudes). These covariates we resampled to 250 m spatial resolution. The following step shows how these covariates are imported into R and aligned with the soil data. The R packages for spatial data have the facility for specifying the projection of the GIS data. This projection is used to align the datasets and it has to be known a priori. QGIS software (http://qgis.org/) can be used to obtain this information in case it is not readily known.

STEP 6. DATA PROCESSING

```
>library(sp)
>library(rgdal)
> library(Hmisc)
>predictors=readGDAL("dem.asc")
>predictors$landform=readGDAL("landform.asc")$band1
>predictors$rain=readGDAL("rain.asc")$band1
>predictors$latitude=readGDAL("latitude.asc")$band1
>predictors$ndvi=readGDAL("ndvi.asc")$band1
>predictors$longitude=readGDAL("longitude.asc")$band1
>predictors$dem=predictors$band1
>predictors$band1=NULL
>proj4string(predictors)=CRS("+proj=utm +zone=37 +south +datum=WGS84+units=m +no_defs")
>coordinates(soildata)=~X+Y
>proj4string(soildata)=CRS("+proj=utm +zone=37 +south +datum=WGS84 +units=m +no_defs")
>predictors.ov=over(soildata, predictors)
>soildata$dem=predictors.ov$dem
>soildata$ndvi=predictors.ov$ndvi
>soildata$rain=predictors.ov$rain
>soildata$longitude=predictors.ov$longitude
>soildata$latitude=predictors.ov$latitude
>soildata$landform=predictors.ov$landform
>SOCT.histbb=histbackback(soildata$landform, predictors$landform, prob=TRUE)#Figure 6.21
```

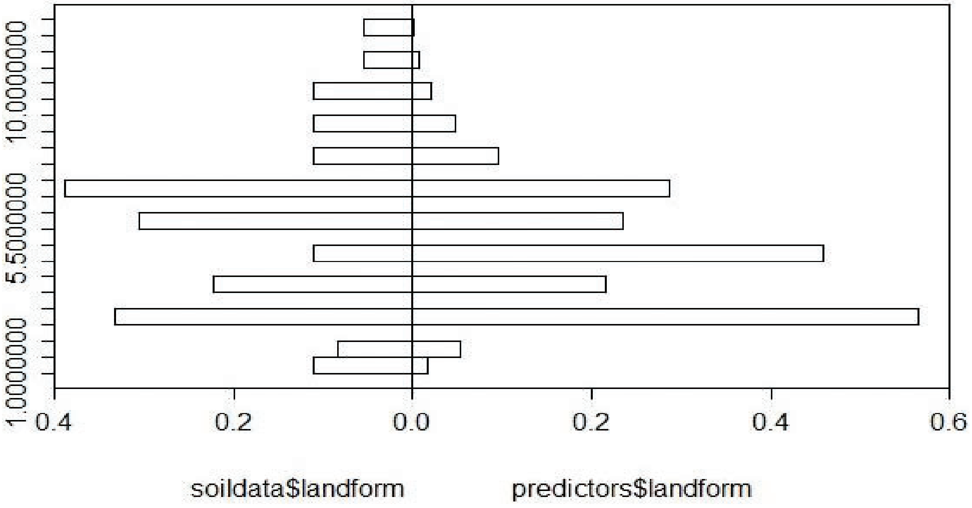


FIGURE 6.21 HISTOGRAM REPRESENTATION OF SAMPLED AND POPULATION CHARACTERISTICS

Apart from seeing that the sample locations are evenly distributed in the study area, it could also be important to assess how the points are distributed in the feature space of each covariate (e.g. landform feature space in Figure 6.10). If the distribution is not even or uniform then potential errors could arise and hamper the model training. Nothing much can be done to increase the number of samples in each feature space if the cost of adding more samples is inconceivable at this stage. However, it's important to note how this facility can be used to plan sampling in DSM.

While building the random forest models, if it's necessary to assess the predictive performance of the model, one do so using by splitting the data into two: a hold-out sample spart on which to build the model and the other part for model testing. After testing the model and accepting the achieved accuracy level, it's important to develop a final model using the whole data (NB: refer to the validation section of this cookbook for more in-depth discussions). In the following scripts, we use the *sample* function to randomly split the data into two parts: training and testing parts.

STEP 7: SPLITTING THE SOIL DATA FOR MODEL TESTING AND TRAINING AND SUBSEQUENT PERFORMANCE EVALUATION.

```
>library(randomForest)
>training <- sample(nrow(soildata), 0.66 * nrow(soildata))#randomly split the data
>modl.rf=randomForest(SOCT~landform+dem+ndvi+rain+longitude+latitude,
data=soildata[training,], importance=TRUE, ntree=500)
>print(modl.rf)
Call:
     randomForest(formula = SOCT ~ landform + dem + ndvi + rain +
              longitude + latitude, data = soildata[training, ], importance =TRUE,
                        ntree = 500)
Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 2
Mean of squared residuals: 3.177888e-05
% Var explained: 88.98

>RF.predSOCT <- predict(modl.rf, newdata = soildata[-training, ])
>RMSE.modl <- sqrt(mean((soildata$SOCT[-training] - RF.predSOCT)^2))
>RMSE.modl
[1] 0.004463855
>R2.modl <- lm(RF.predSOCT ~ soildata$SOCT[-training])
>as.matrix(summary(R2.modl)$adj.r.squared)
     [,1]
[1,] 0.9107788
>bias.modl <- mean(RF.predSOCT) - mean(soildata$SOCT[-training])
>bias.modl
[1] -0.002577775
plot(soildata$SOCT[-training],RF.predSOCT) #Figure 6.22
abline(a=0,b=1,lty=2, col="red")# 1:1 comparison
abline(R2.modl, col="blue")# regression on predicted and measured values
```

The above results appear like the predictive performance of the random forest model was good. However, a closer look at the plot of predicted versus observed values reveal that the model over-predicted low values and under-predicted high values (Figure 6.22). Thus, high values and low values in the resultant map may need to be treated with caution.

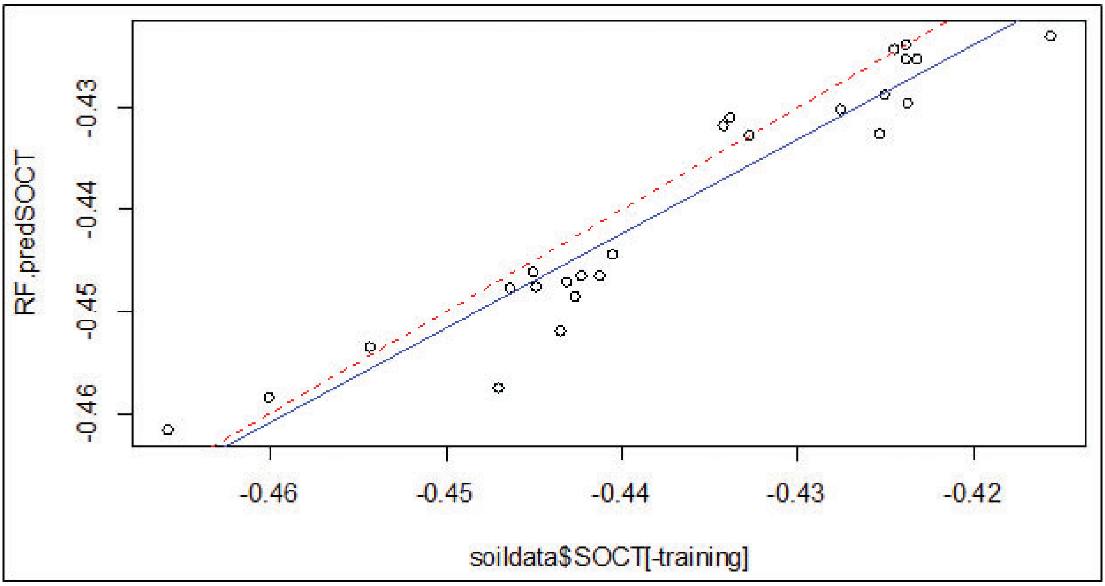


FIGURE 6.22 DIAGNOSTIC COMPARISON OF THE FITTED AND ACTUAL VALUES

STEP 8: SPATIAL PREDICTION OF THE SOIL ORGANIC CARBON.

```
>predictors$SOCT=predict(modl.rf,newdata=predictors)
>lmbda=(as.numeric(powerTransform(soildata$SOC, family
="bcPower")["lambda"]))
>predictors$SOCr=(predictors$SOCT*lmbda+1)^(1/lmbda)
>pred.plt=spplot(predictors["SOCr"], scales=list(draw=TRUE,cex=1))
>print(pred.plt, more=TRUE)# Figure 6.23
```

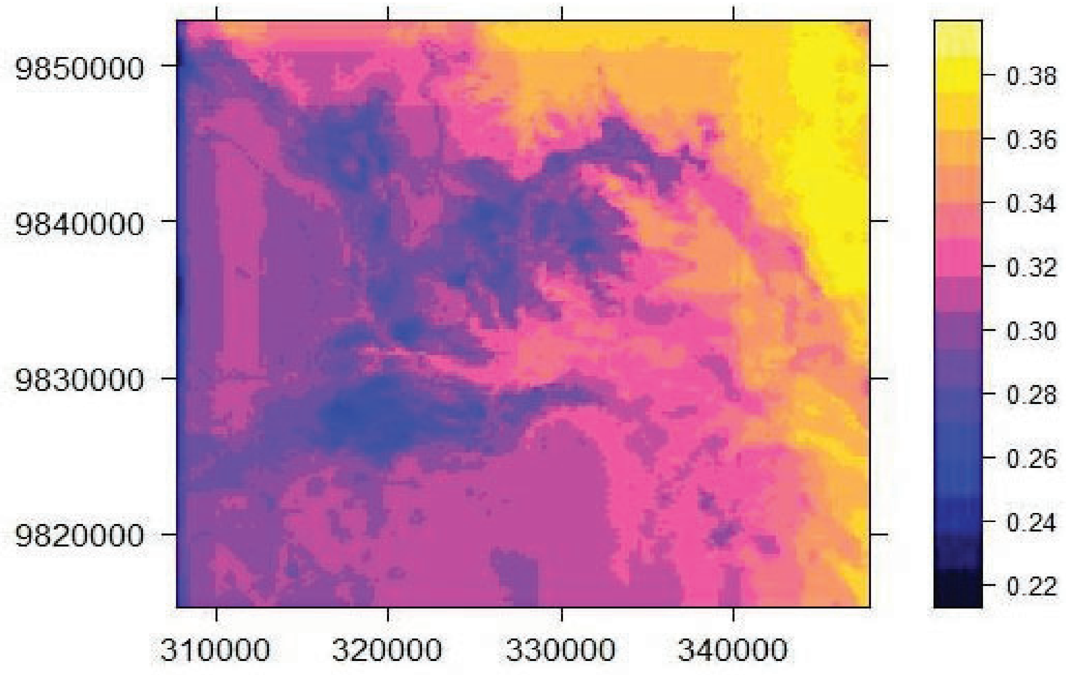


FIGURE 6.23 RANDOM-FOREST PREDICTED SOC MAP

6.3.5 References

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press LLC, Boca Raton, FL.

Wiesmeier, M., Barthold, F., Blank F.B., Kögel-Knabner, I. 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and Soil 340(1):7-24

Box, G. E. P., Cox, D. R. 1964. An analysis of transformations (with discussion). Journal of the Royal Statistical Society B, 26, 211-252

McBratney, A.B., Mendonca-Santos, MX., Minasny, B. 2003. On digital soil mapping. Geoderma 117, 3-52.

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York.

Adhikari K, Hartemink AE, Minasny B, Bou Kheir R, Greve MB, Greve MH. 2014. Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark. PLoSONE 9(8): e105519. doi:10.1371/journal.pone.0105519

Post, TM., Freijer, JI, Ploeger BA, Danhof M. 2008. Extensions to the visual predictive check to facilitate model performance evaluation. Journal of Pharmacokinetics and Pharmacodynamics, 35: 185. doi.10.1007/s10928-007-9081-1

7. UNCERTAINTY

Soil mapping involves making predictions at locations where no soil measurements were taken. This inevitably leads to prediction errors because soil spatial variation is complex and cannot be modelled perfectly. It also implies that we are uncertain about the true soil class or true soil property at prediction locations. We only have the predictions, which differ from the true values in an unpredictable way, and hence we are uncertain about the true value. In fact, we may even be uncertain about the soil at the measurement locations because no measurement method is perfect and uncertainty also arises from measurement errors.

This chapter describes how uncertainty may be characterised by probability distributions. It also explains how the parameters of these distributions may be derived, leading to quantification of uncertainty. We will see that this can become quite complex, because soil properties vary in space and are often cross-correlated, which the uncertainty model must take into account. A further complication is that there are many different sources of uncertainty. In some cases it may be too difficult to arrive at a spatially explicit characterisation of uncertainty, and in such case statistical validation may be used to derive summary measures of the accuracy of soil maps. We begin this chapter with a description of uncertainty sources.

7.1 SOURCES OF UNCERTAINTY

Consider a case in which soil samples were taken from a large number of measurement locations in a study area, taken to the laboratory and analysed for various soil properties. Let us further assume that the measurement locations were indicated on a topographic map and that the soil was also classified at each measurement location. Next, the soil property and soil type observations were used to create maps of soil properties and soil type using digital soil mapping techniques. These techniques not only make use of the soil observations but also benefit from maps of environmental variables that are correlated with the soil, and hence help explain the soil spatial variation. Which sources of uncertainty contribute to uncertainty about the final soil maps? We distinguish four main categories.

7.1.1 ATTRIBUTE UNCERTAINTY OF SOIL MEASUREMENTS

Soil measurements suffer from measurement errors in the field and laboratory. Perhaps the soil was not sampled at the right depth, perhaps the organic layer was not removed completely before collecting soil material, or perhaps by accident bags were interchanged or numbered wrongly. Field estimates of soil type and soil properties are also not error-free, especially when estimation is difficult, such as estimation of organic carbon content or texture. Field estimates may also be subjective, because soil scientists may be trained differently and so there may be systematic differences between their field estimates of soil properties. Similarly, it is also not uncommon for soil scientists to disagree about the soil type when classifying a soil in the field.

Laboratory analysis adds error too. Soil samples may not be perfectly mixed prior to taking a much smaller subsample that is actually measured; instruments have limited precision and may have systematic errors, climate conditions in the lab vary, and there can be differences between procedures used by laboratory personnel. Differences between laboratories are even bigger and may be of the same order of magnitude as the soil variation itself. It is strongly advised to always take sufficient duplicates and randomise the order in which soil samples are analysed in the laboratory. This allows to quantify the combined field and laboratory measurement errors.

7.1.2 POSITIONAL UNCERTAINTY OF SOIL MEASUREMENTS

When collecting soil data in the field we would generally note the geographic coordinates of the measurement locations. Nowadays this is easy with GPS instruments and depending on the device, modest to high positional accuracy can be achieved. But it may still be too large to be negligible. For instance, consider the case where the soil data are used to train a digital soil mapping model that predicts soil properties from covariates. Let these covariates be available at high spatial resolution and have substantial fine-scale spatial variation. Then it is clear that positional uncertainty in the soil measurements may link these measurements to the wrong covariates, which will weaken the strength of relationship between the soil variable and covariates and deteriorate the quality of the final soil map.

Many soil legacy data suffer from large positional uncertainty. Locations may only be traced from vague descriptions such as “near village A” or “east of the road from B to C”. In such case, researchers should consider whether using such data for calibration of a DSM model and for spatial prediction using the calibrated DSM model is wise. It may do more harm than good. This depends on the specific DSM model used and the degree of spatial variation of the covariates. It also depends on the degree of spatial variation of the soil property itself. If it has negligible fine-scale spatial variation and hence has similar values at the registered and actual geographic location, then little harm is done. For instance, in the Sahara desert many soil properties will show little spatial variation over distances of hundreds or perhaps thousands of meters, so in such case poor geographic positional accuracy will not seriously affect DSM predictions.

7.1.3 UNCERTAINTY IN COVARIATES

Maps of covariates that are used in DSM can also suffer from errors and uncertainties. For instance, a Digital Elevation Model (DEM) is a major source of geomorphological covariates but DEMs are only approximations of the real elevation. DEM errors will propagate and cause uncertainty in geomorphological properties such as slope, aspect and topographic wetness index. As a result, the DSM model must be trained on covariate data that are merely approximations of the intended covariates, which will generally lead to weakened relationships and larger DSM prediction errors. Land cover is another example; soil properties may be strongly influenced by land cover, but such relationship may come out quite weak if the DSM model is trained with a land cover map that represents land cover wrongly for a large part of the study area.

Covariates also come in a specific spatial resolution which may be quite coarse in specific cases. In order to use the covariate in a fine-scale DSM model, the coarse-scale grid cell value will be copied to all fine-scale grid cells contained in it, but clearly fine-scale spatial variation implies that uncertainties will be introduced. A possible solution might be to smooth the coarse-scale covariate prior to entering it to DSM calibration but clearly this will not remedy all problems.

Uncertainty in covariates leads to weaker DSM models, but this weakening is not hidden to the developer because the deterioration of predictive power is implicitly included in the DSM model. For instance, the amount of variance explained by a DSM model that uses the true land cover as measured on sampling sites may be much higher than that of a model that uses a land cover map. Users may then be tempted to calibrate the DSM model with the true land cover data, but if they next apply that model using the land cover map to predict the soil at non-measurement locations they would systematically underestimate the uncertainty of the resulting map.

7.1.4 UNCERTAINTY IN MODELS THAT PREDICT SOIL PROPERTIES FROM COVARIATES AND SOIL POINT DATA

Even if the soil point data and covariate data were error-free, the resulting DSM predictions would still deviate from the true soil properties. This is because the DSM model itself also introduces uncertainties. Models are merely simplified representations of the real world. The real world is too complex and approximations are needed. For instance, even though we know that physical, chemical and biological processes determine the soil as given by the state equation of soil formation soil=f(cl,o,r,p,t), the function f is too complex to be fully understood and implemented in a computer model. Instead, we use crude approximations such as multiple linear regression and machine-learning algorithms. These empirical models have the additional burden that extrapolation beyond conditions represented by the calibration data is difficult and risky. For extrapolation purposes it is advised to use DSM models that better represent the mechanisms behind soil formation, but again it is practically impossible to build mechanistic models that represent the real world perfectly. This is not only because we may not understand all processes and their interactions well, but also because dynamic mechanistic models need much information, such as the initial state, boundary conditions and driving forces. Such detailed information is generally lacking.

Model uncertainty is generally subdivided into model parameter uncertainty and model structural uncertainty. The first can be reduced by using models with fewer parameters or by using a larger calibration data set. The latter can be reduced by using a more complex model, but this will only work if there are enough data to calibrate such model. Thus, in general a compromise has to be sought by choosing a level of model complexity that matches the amount of information available.

7.2 UNCERTAINTY AND SPATIAL DATA QUALITY

Research into spatial accuracy in Geographic Information Science has listed five main elements of spatial data quality:

1. lineage
2. positional accuracy
3. attribute accuracy
4. logical consistency
5. completeness

We have already discussed positional and attribute accuracy. Lineage refers to documenting the original sources for the data and the processing steps. This is strongly related to the principle of reproducible research. Logical consistency addresses whether there are any contradictory relationships in the database. For instance, it checks whether all data have the same geographic projection and that measurement units are consistent. Completeness refers to whether there are any missing data. For instance, covariate maps must cover the entire study area if they are to be used as explanatory variables in a DSM model. Soil profile data need not capture all relevant soil properties and tend to have fewer soil measurements at greater depths.

In summary, there are many sources of uncertainty that affect the quality of DSM products. This section has reviewed these sources but was purposely descriptive. The next section selects a few major uncertainty sources and works out quantitatively how these cause uncertainty in the resulting soil map. Perhaps it is useful to mention that focussing attention on errors and uncertainties may give the wrong impression that soil maps are generally inaccurate and of poor quality. This is not the message that we wish to convey here. But producers and users of soil maps should be aware of the sources of uncertainty and should ideally identify how these uncertainties affect the final product. Thus, quantification of the uncertainty in DSM maps, be it through explicit modelling or independent validation is important.

7.3 QUANTIFYING PREDICTION UNCERTAINTY

Uncertainties in soil measurements, covariates and DSM models propagate to resulting soil maps. The uncertainty propagation can fairly easily be traced provided that the uncertainty sources are characterised adequately. The most appropriate way of doing that is by making use of statistics and probability distributions. This section also takes that approach and starts by providing a brief overview of probability distributions and how these may be used to represent uncertainty. Next it analyses how the four sources of uncertainty distinguished in Section 5.1 lead to uncertainty in soil maps produced using DSM.

7.3.1 UNCERTAINTY CHARACTERISED BY PROBABILITY DISTRIBUTIONS

If we are uncertain about the value of a soil property at some location and depth this means that we cannot identify one single, true value for that soil property (Goovaerts 2001, Heuvelink 2014). Instead, we may be able to provide a list of all possible values for it and attach a probability to each. In other words, we represent the true but unknown soil property by a probability distribution. For instance, suppose that we estimate the sand content of a soil sample in the field as 35%, while recognising that a field estimate is quite crude and that the true sand content may very well be less or more than the estimated 35%. We might be confident that the estimation error is unlikely to be greater than 8%, and hence it would be reasonable to represent the sand content by a normal distribution with a mean of 35% and a standard deviation of 4%. For the normal distribution, 95% of the probability mass lies within two standard deviations from the mean, so we would claim that there is a 5% probability that the sand content is smaller than 27% or greater than 43%.

In the example above we had chosen the normal distribution because it is the most common probability distribution but we might as well have used a different distribution, such as the uniform or lognormal distribution. Indeed many soil properties, such as soil nutrient concentrations are better described by lognormal distributions, because values below zero cannot occur and because very high positive values (i.e. outliers) are not unlikely. For instance, we may estimate the organic carbon concentration (OC) of a soil sample as 1.2% and identify with it an asymmetric 95% credibility interval ranging from 0.8% to 2.5%. In general, statistical modelling is easier if the variables under study can be described by normal distributions. This explains why we usually apply a transformation to skewed variables prior to statistical modelling. For instance, when building a DSM model of OC, it may be wise to develop such model for the logarithm of OC and do a back-transform on the DSM predictions.

There are many different soil properties that in addition vary in space and possibly time. Thus, the characterisation of uncertainty about soil properties needs to be extended and include cross- and space-time correlations. It is beyond the scope of this chapter to explain this in detail, for this we refer to standard textbooks such as Goovaerts (1997) and Webster and Oliver (2007). If we assume a joint normal distribution, then a vector of soil properties (be it different soil properties or the same soil property at multiple locations, depths or times) **Z** is fully characterised by the vector of means **m** and variance-covariance matrix **C**. Figure 7.1 shows three examples of 500 paired soil property values that were simulated from different bivariate normal distributions. The left panel shows an uncorrelated case with equal standard deviations for both properties. The centre and right panels show a case where soil property 2 has a greater standard deviation than soil property 1. The difference between these two cases is that the centre panel has a zero correlation between the two soil properties while it is positive in the right panel.

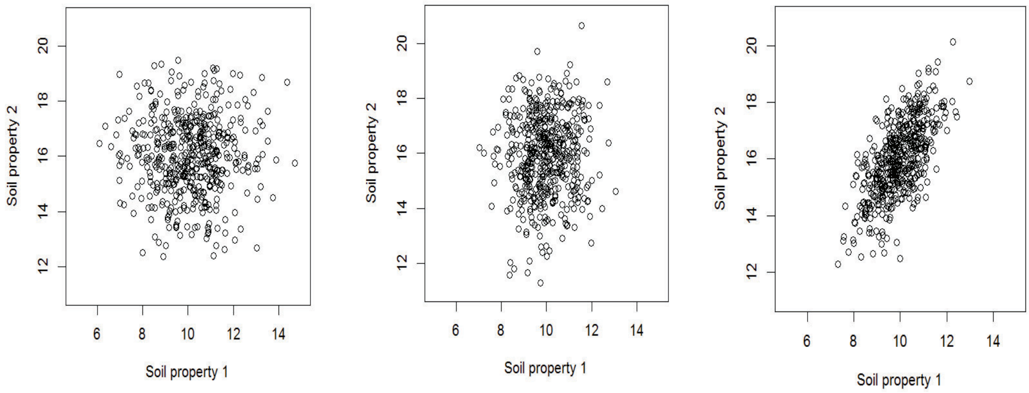


FIGURE 7.1 SCATTER PLOTS OF 500 PAIRED SOIL PROPERTY VALUES DRAWN FROM A TWO-DIMENSIONAL NORMAL DISTRIBUTION. LEFT: M=[10,16], C=[2,0;0,2], CENTRE: M=[10,16], C=[1,0;0,2], RIGHT: M=[10,16], C=[1,1;1,2].

7.3.2 PROPAGATION OF MODEL UNCERTAINTY

Now that we have clarified how uncertainty in soil properties may be characterised by probability distributions, let us consider what these distributions look like in DSM and how these are influenced by the uncertainty sources described in Section 7.1. We begin with uncertainty source 4, uncertainty in DSM models.

We noted before that uncertainty in DSM models may be separated in model parameter and model structural uncertainty. A typical example of this is a multiple linear regression model:

$$Z(s)=\beta_0+\beta_1 \cdot X_1(s)+\beta_2 \cdot X_2(s)+\varepsilon(s) \qquad (7.1)$$

Note that here for simplicity we assumed two environmental covariates X_1 and X_2 while in practice we are likely to use many more. Parameter uncertainty of this model occurs because the parameters β_0 , β_1 and β_2 are merely estimated using calibration data. Under the assumptions made by the linear regression model, these estimation errors are normally distributed and have zero mean, while their standard deviations and cross-correlations can also be computed (e.g. Snedecor and Cochran 1989, Section 17.5). The standard deviations become smaller as the size of the calibration dataset increases. Both the standard deviations and cross-correlations are standard output of statistical software packages. Thus, we could sample from the joint distribution of the parameter estimation errors in a similar way as displayed in Figure 7.1.

The model structural uncertainty associated with the multiple linear regression model Eq. (7.1) is represented by the stochastic residual ε. It too is normally distributed and has zero mean, while its standard deviation depends on the (spatial) variation of the soil property Z and the strength of the relationship between and the covariates X_1 and X_2. If the covariates explain a great deal of the variation of the soil property then the standard deviation of the residual will be much smaller than that of the soil property, as expressed by the goodness-of-fit characteristic R^2, also termed 'amount of variance explained'. It will be close to 1 in case of a strong linear relationship between soil property and covariates. In that case the standard deviation of the stochastic residual will be much smaller than that of the soil property, because a large part of the variation is explained by the model. If the covariates bear no linear relationship with the soil property (i.e., $R^2 = 0$), the stochastic residual will have the same standard deviation as the soil property.

Since the joint probability distributions of the parameter estimation errors and the stochastic residual can analytically be computed and are routinely provided by statistical software, it is not difficult to analyse how these uncertainties propagate through the DSM model Eq. (7.1). This can be done analytically, because Eq. (7.1) is linear in the stochastic arguments (note that the covariates are treated known and deterministic). If we predict the soil property Z at a prediction location s_0 using the calibrated regression model as:

$$\hat{Z}(s_0) = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1\,(s_0) + \hat{\beta}_2 \cdot X_2\,(s_0) \qquad (7.2)$$

then the prediction error will be normally distributed with zero mean and variance (i.e., the square of the standard deviation) given by:

$$\begin{aligned} Var(\hat{Z}(s_0)-Z(s_0)) = {} & Var(\hat{\beta}_0) + Var(\hat{\beta}_1) \cdot X_1(s_0)^2 + Var(\hat{\beta}_2) \cdot X_2(s_0)^2 + \\ & 2\,Cov(\hat{\beta}_0, \hat{\beta}_1) \cdot X_1(s_0) + 2\,Cov(\hat{\beta}_0, \hat{\beta}_2) \cdot X_2(s_0) + \\ & 2\,Cov(\hat{\beta}_1, \hat{\beta}_2) \cdot X_1(s_0) \cdot X_2(s_0) + Var(\varepsilon(s_0)) \end{aligned} \qquad (7.3)$$

This is a complicated expression but all entries are known and hence it can be easily calculated.

In many DSM applications an additional step will be included that makes use of the fact that the stochastic residual in Eq. (7.1) is spatially autocorrelated, as characterised by a semivariogram. If this is the case the residual spatial correlation can be exploited by incorporating a kriging step (Hengl *et al.* 2004). Kriging has been explained in Chapter 6, where it was also explained that the uncertainty in the predictions is quantified by the kriging variance. We will not repeat the theory here, but simply note that the kriging variance computes the prediction error variance just as was done in Eq. (7.3), but that in case of kriging the $Var(\varepsilon(s_0))$ term in Eq. (7.3) is replaced by a smaller term, because kriging benefits from residual spatial correlation. In fact, in case of a pure nugget variogram, the kriging variance would be identical to Eq. (7.3), because in such case there is no spatial autocorrelation that one can benefit from. Note also that here we refer to Kriging with External Drift because we included a non-constant mean (i.e., covariates X_1 and X_2). If no covariates were included Eq. (7.3) would simplify dramatically leaving only uncertainty in the estimated (constant) mean and the stochastic residual. This might then be compared with the ordinary kriging variance.

So far we considered uncertainty in DSM models that are linear in the covariates and that represent the model structural uncertainty by an additive stochastic term. This was relatively easy because tracing how uncertainty in model parameters and model structure propagate to the model output could be done analytically. However, using linear models also poses serious restrictions. The relationship between soil properties and covariates are typically not linear but much more complex.

This has led to the development and use of complex non-linear DSM models, such as regression trees, artificial neural networks, support vector machines and random forests approaches, all summarised under the term ‘machine learning’ (e.g. Hengl *et al.* 2015). These more complex models typically yield more accurate soil predictions but quantification of the associated uncertainty is more difficult. In most cases, one resorts to validation and cross-validation statistics that summarise the prediction accuracy over the entire study area. How this is done will be explained in detail in 7.3. Such summary validation measures are very valuable but are no substitute for spatially explicit uncertainties such as the kriging variance and the prediction error variance presented in Eq. 7.3. Research into quantification of location-specific uncertainties when using machine learning algorithms is therefore important. However, it is beyond the scope of this chapter to review this area of ongoing research. One particular approach makes use of quantile regression forests. We refer to Meinshausen (2006) for a general text and to Vaysse and Lagacherie (2017) for a DSM application of this promising, albeit computationally challenging approach.

7.3.4 PROPAGATION OF ATTRIBUTE, POSITIONAL AND COVARIATE UNCERTAINTY

In Section 7.1 we noted that next to uncertainties in model parameters and model structure there may also be uncertainties in the attribute values and positions of the soil point data, and in the covariates. These sources of uncertainty will also affect the outcome of DSM model predictions.

Uncertainties in soil attribute values effectively mean that the DSM model is calibrated with error-contaminated observations of the dependent variable. Let us consider the multiple linear regression model Eq. (7.1) again. True values of the dependent variable Z (i.e., the target soil property, such as pH, clay content or total nitrogen concentration) are no longer for calibration of this model. Instead, we must make do with measurements Y of Z:

$$Y(s_i) = Z(s_i) + \delta(s_i), \qquad i = 1 \cdots n \qquad (7.4)$$

where n is the number of measurement locations and $\delta(s_i)$ is a random variable representing measurement error. It is custom to assume that all $\delta(s_i)$ are normally distributed, have zero mean and are mutually independent, although these assumptions are not strictly necessary. Their standard deviations may vary between cases and depend on the accuracy and precision of the measurement method. For instance, field estimates tend to be more uncertain than laboratory measurements and so the corresponding measurement errors will have a larger standard deviation. The consequence of the presence of measurement errors is that the estimates of the model parameters will be more uncertain. This is no surprise because the calibration data are of poorer quality. The prediction error variance will be greater too, for the same reason. If spatial correlation of the model residual ε is included and an extension to Kriging with External Drift is made, uncertainty due to measurement errors is further increased because the conditioning of predictions to observations cannot benefit as much as when the observations were error-free. For mathematical details we refer to Cressie (1993). Finally, we should also note that if different observations have different degrees of measurement error, then this will influence the weights that each measurement gets in calibration and prediction. Measurements with larger measurement errors get smaller weights. This is automatically incorporated in multiple linear regression and Kriging with External Drift, but how this can be incorporated in machine-learning approaches is less clear.

Positional uncertainty of soil point observations will also deteriorate the quality of the predictions of calibrated DSM models. However, it is difficult to predict how much the prediction accuracy is affected. It largely depends on the degree of fine-scale spatial variation of the soil property and covariates. For instance, if both the soil property of interest and the covariates are spatially smooth and hardly change over distances within the range of spatial displacement due to positional uncertainty, then little damage is afflicted by positional uncertainty. But otherwise much harm can be done, because the soil observations will be paired with covariate values from displaced locations that can be very different. So far, this interesting and important topic has received only little attention in the DSM literature. Grimm and Behrens (2010) and Nelson *et al.* (2011) are two examples of studies that assessed the effect of positional error on the accuracy of digital soil maps.

Finally, there are also uncertainties in covariates that affect the accuracy of DSM predictions. In fact, these uncertainties are already incorporated in the model structural uncertainty discussed before, because offering covariates that are poor approximations of the true soil forming factors will explain little of the spatial variation and lead to low goodness-of-fit statistics. From a statistical point of view, the covariates used in Eq. (7.1) need not be the 'true' soil forming factors but could as well be proxies of those. This does not harm the theory and quantification of the prediction error variance such as through Eq. (7.3) in the multiple linear regression case or using the kriging variance in a KED approach remain perfectly valid. This does not mean that digital soil mappers should not look for the most accurate and informative covariates, because clearly weak covariates leads to poor predictions of the soil (e.g. Samuel Rosa et al. 2015).

7.4 References

Cressie, N.A.C. (1993), Statistics for Spatial Data, Revised Edition. Wiley.

Goovaerts, P. (1997), Geostatistics for Natural Resources Evaluation. Oxford University Press.

Goovaerts, P. (2001), Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3-26.

Grimm, R., Behrens, T., 2010. Uncertainty analysis of sample locations within digital soil mapping approaches. Geoderma 155(3-4), 154-163.

Hengl, T., G.B.M. Heuvelink and A. Stein (2004), A generic framework for spatial prediction of soil properties based on regression-kriging. Geoderma 120, 75-93.

Hengl, T., J. Mendes de Jesus, G.B.M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M.N. Wright, X. Geng, B. Bauer-Marschallinger, M.A. Guevara, R. Vargas, R.A. MacMillan, N.H. Batjes, J.G.B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel and B. Kempen (2016), SoilGrids250m: Global gridded soil information based on machine Learning. PLoS ONE 12(2): e0169748.

Heuvelink, G.B.M. (1998), Error Propagation in Environmental Modelling with GIS. Taylor & Francis.

Heuvelink, G.B.M. (2014), Uncertainty quantification of GlobalSoilMap products. In: GlobalSoilMap. Basis of the Global Spatial Soil Information System. D. Arrouays, N. McKenzie,

J. Hempel, A. Richer de Forges and A. McBratney (Eds.), pp. 335-340.

Meinshausen, N. (2006), Quantile Regression Forests. Journal of Machine Learning Research 7, 983-999.

Nelson, M.A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011. An error budget for different sources of error in digital soil mapping. European Journal of Soil Science 62(3), 417-430.

Samuel-Rosa, A., G.B.M. Heuvelink, G.M. Vasques and L. Anjos (2015), Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243-244, 214-227.

Snedecor, G.W. and W.G. Cochran (1989), Statistical Methods (Eight Edition). Iowa State University Press.

Vaysse, K. and P. Lagacherie (2017), Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma, http://dx.doi.org/10.1016/j.geoderma.2016.12.017.

Webster, R. and M.A. Oliver (2007), Geostatistics for Environmental Scientists. Wiley.

8. VALIDATION

8.1 WHAT IS VALIDATION?

No map is perfect. All maps, including soil maps, are representations of reality that are often based on an underlying model. This means that there will always be a deviation between the phenomenon depicted on the map and the phenomenon observed in the real world, i.e. each map will contain errors. The magnitude of the errors determine the quality of the map. If a map matches reality well (the error is small), the quality or accuracy of the map is high. On the other hand, if a map does not match reality well, map accuracy is low.

Soil maps are used for many purposes. For example to report on (changes in) soil organic carbon stocks, as input in agro-environmental models, to determine land use suitability or for decision- and policy-making. It is therefore, important that the quality of a map is determined and quantified. This is achieved through (statistical) validation.

Validation is defined here as an activity in which the soil map predictions are compared with observed values. From this comparison, the map quality can be quantified and summarized using map quality measures. These measures indicate how accurate the map is on average for the mapping area, i.e. what is the expected error at a randomly selected location in the mapping area. This means that map quality measures obtained through validation are global measures: each quality measure gives one value for the entire map. Note that this is different from results obtained through uncertainty assessment. Such assessment provides local, location-specific (i.e. for each individual grid cell) estimates of map quality as we saw in the previous sections. Another important difference between validation and uncertainty assessment is that validation can be done using a model-free approach. We saw in section 7.2 that uncertainty assessment takes a model-based approach by defining a geostatistical model of the soil property of interest and deriving an interpolated map and the associated uncertainty from that, or by constructing a geostatistical model of the error in an existing map. The approach yields a complete probabilistic characterisation of the map uncertainty, but such characterisation is only valid under the assumptions made;for instance, the stationarity assumptions required for kriging. Validation, when done properly as explained hereafter, does not assume a geostatistical model of the error, and hence is model- or assumption-free. This is an important property of validation since we do not want to question the objectivity and validity of the validation results.

We distinguish internal and external map accuracy. Statistical methods typically produce direct estimates of map quality, for instance the kriging variance or the coefficient of determination (R^2) of a linear regression model. These we refer to as internal accuracy measures since these rely on model assumptions and are computed from data that are used for model calibration. Preferably, validation is done with an independent dataset not used in map making. Using such dataset gives the external map accuracy. One will often see that the external accuracy is poorer than the internal accuracy.

In section 8.3.2 we will present the most common accuracy measures used to quantify map quality of quantitative (continuous) soil maps and qualitative (categorical) soil maps. In section 8.3.3 we will introduce three commonly used validation methods and show how to estimate the map quality measures from a sample. This chapter is largely based on Brus *et al.* (2011). For details, please refer to this paper.

8.2. MAP QUALITY MEASURES

8.2.1 QUALITY MEASURES FOR QUANTITATIVE SOIL MAPS

All map quality measures considered here are computed from the *prediction error*. For quantitative soil maps of continuous soil properties (e.g. organic carbon content, pH, clay content) the prediction error is defined as the difference between the predicted value at a location and the true value at that location (which is the value that would be observed or measured by a preferably errorless measurement instrument) (Brus *et al.*, 2011):

$e(s) = \hat{z}(s) - z(s)$

where $\hat{z}(s)$ is the predicted soil property at validation location s, and $z(s)$ is the true value of the soil property at that location. We consider six map quality measures that are computed from the prediction error here: the mean error, the mean absolute error, the mean squared error and root mean squared error, the model efficiency and the mean squared deviation ratio.

Before we introduce the map quality measures and show how to estimate these, it is important to understand the difference between the population and a sample taken from the population. The population is the set of all locations in a mapping area. For digital soil maps, this is the set of all pixels or grid cells of a map. A sample is a subset of locations, selected in some way from the set of all locations in the mapping area. With validation we want to assess the map accuracy for the entire population, i.e. for the map as a whole; we are not interested in the accuracy at the sample of locations only. For instance, we would like to know the prediction error averaged over all locations of a map and not merely the average prediction error at a sample of locations. Map quality measures are therefore, defined as population means. Because we cannot afford to determine the prediction error at each location (grid cell) of the mapping area to calculate the population means, we have to take a sample of a limited number of locations in the mapping area. This sample is then used to estimate the population means. It is important to realize that we are uncertain about the population means, because we estimate it from a sample. Ideally this uncertainty is quantified and reported together with the estimated map quality measures. In this section we will introduce the definitions of the map quality measures.

In the next section, we show how we can estimate these measures from a sample.

Mean error

The mean error (ME) measures bias in the predictions. The ME is defined as the population mean (spatial mean) of the prediction errors:

$$ME = e = \frac{1}{N}\sum_{i=1}^{N} e(s_i)$$

where i indicates the location, i = 1,2, … , N, and N is the total number of locations or grid cells/pixels in the mapping area. The mean error should be (close to) zero, which means that predictions are unbiased meaning that there is no systematic over- or under-prediction of the soil property of interest.

Mean absolute error and (root) mean squared error

The mean absolute error (MAE) and mean squared error (MSE) are measures of map accuracy and indicate the magnitude of error we make on average. The MAE is defined by the population mean of the absolute errors:

$$MAE = |\underline{e}| = \frac{1}{N}\sum_{i=1}^{N} \underline{e}(s_i)$$

and the MSE by the population mean of the squared errors:

$$MSE = \underline{e}^2 = \frac{1}{N}\sum_{i=1}^{N} \underline{e}^2(s_i)$$

Many authors report the root mean squared error (RMSE) instead of the MSE, which is computed by taking the square root of the MSE. The RMSE can be a more appealing quality measure since it has the same unit of measurement as the mapped property and can therefore more easily be compared to it. If the squared error distribution is strongly skewed, for instance when several very large errors are present, then this can severely inflate the (R)MSE. In such case, the (root) median squared error is a more robust statistic for the 'average' error (Kempen *et al.*, 2012).

Brus *et al.* (2011) argue that instead of using a single summary statistic (the mean) to quantify map quality measures, one should preferably express quality measures for quantitative soil maps through cumulative distribution functions (CDFs). Such functions provide a full descriptions of the quality measures from which various parameters can be reported, such as the mean, median or percentiles. Furthermore, they argue that it can be of interest to define CDFs or its parameters for sub-areas, for instance geomorphic units, soil or land cover classes. Brus *et al.* (2011) give examples of estimating CDFs for validation of digital soil maps.

Amount of variance explained

The model efficiency, or Amount of Variance Explained (AVE) (Angelini *et al.*, 2016; Samuel-Rosa *et al.*, 2015), quantifies the fraction of the variation in the data that is explained by the prediction model. It measures the improvement of the model prediction over using the mean of the data set as predictor and is defined as follows (Krause *et al.*, 2005):

$$AVE = 1 - \frac{\sum_{i=1}^{N} (\hat{Z}(s_i) - Z(s_i))^2}{\sum_{i=1}^{N} (Z(S_i) - \underline{Z})^2}$$

where $\underline{Z}$ is the population mean of soil property Z. The quantity in the numerator is the sum of the squared prediction errors (for each location the prediction error is computed and squared; the squared prediction errors are summed over all locations in the area). In linear regression this quantity is known as the *residual sum of squares* (RSS). The quantity in the denominator is also a sum of squared prediction errors, but here the mean of the area is used as predictor. In linear regression this quantity is known as the *total sum of squares* (TSS). Note that if we would divide the quantity in the denominator by the number of locations in the mapping area N we would obtain the population variance (spatial variance) of the soil property Z.

If the numerator and denominator are equal, meaning the AVE is zero, then the model predictions are no improvement over using the mean of the data set as predictor for any location in the mapping area. An AVE value larger than zero (RSS smaller than TSS) means that the model predictions are an improvement over using the mean as predictor (this is what we hope for). In case the AVE is negative, then the mean of the data set is a better predictor than the prediction model.

Mean squared deviation ratio

Finally, we introduce the mean squared deviation ratio (MSDR) as a map quality measure (Kempen *et al.*, 2010; Lark, 2000; Voltz and Webster, 1990; Webster and Oliver, 2007). Contrary to the quality measures discussed so far, the MSDR assesses how well the prediction model estimates the prediction uncertainty (expressed as the prediction error variance). The MSDR is defined as:

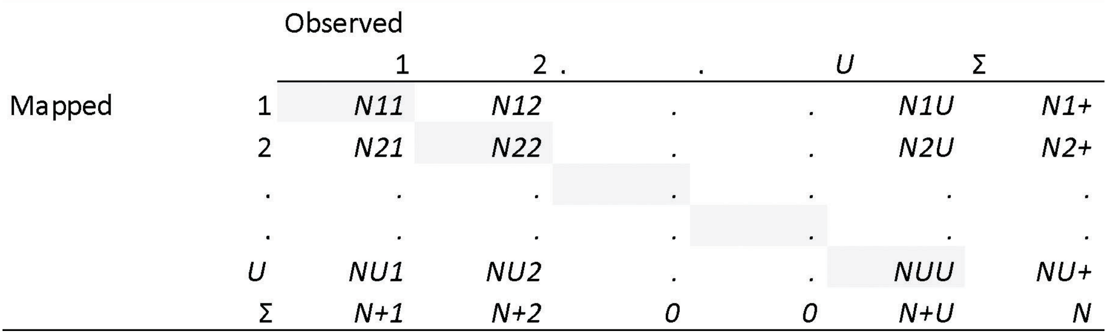
$$MSDR = \frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{Z}(s_i)-Z(s_i))^2}{\sigma^2(s_i)}$$

where $\sigma^2(s_i)$ is the prediction error variance at location s_i, $i = 1,2, \ldots , N$. The numerator is the squared error at location . The fraction represents the squared Z_{score}. In case of kriging, the prediction error variance is the kriging variance. In case of linear regression, the prediction error variance is the prediction variance of the linear regression predictions that can be obtained by the statistical software R by running the predict function with argument se.fit=TRUE. This function returns for each prediction location the standard error of the predicted value as well as the residual standard deviation (the residual.scale value). By squaring both values and then summing these, the prediction error variance is obtained. If the prediction model estimates the error variance well, then the MSDR should be close to one. A value smaller than one suggests that the prediction error variance overestimates the variance; a value larger than one suggests that the prediction error variance underestimates the variance.

Lark (2000) notes that outliers in the prediction data will influence the squared Z_{score} and suggests to use the median squared Z_{score} instead of the mean since it is a more robust estimator. A median squared Z_{score} equal to 0.455 suggests that the prediction model estimates the prediction uncertainty well.

8.2.2 QUALITY MEASURES FOR QUALITATIVE SOIL MAPS

Like the quality measures for quantitative soil maps, the quality measures for qualitative or categorical soil maps (e.g. soil classes) are defined for the population, i.e. all locations in the mapping area. The basis for map quality assessment of qualitative maps is the error matrix (Brus *et al.*, 2011; Lark, 1995). This matrix is constructed by tabulating the observed and predicted class for all locations in the mapping area in a two-way contingency table (Figure 8.1). The population error matrix is a square matrix of order *U*, with *U* being the number of soil classes observed and mapped. The columns of the matrix correspond to observed soil classes and the rows to predicted soil classes (the map units). *N* is the total number of locations of the mapping area. Elements N_{ij} are the number of locations mapped as class *i* with observed class *j*. The row margins N_{i+} are the locations mapped as class *i*, and column margins N_{+j} the locations for which the observed soil class is *j*. Note that the elements of the population error matrix can also be interpreted as surface areas. In that case element N_{ij} is the surface area mapped as class *i* with observed class *j*.



		Observed					
		1	2 .	.		U	Σ
Mapped	1	*N11*	*N12*	.	.	*N1U*	*N1+*
	2	*N21*	*N22*	.	.	*N2U*	*N2+*
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	U	*NU1*	*NU2*	.	.	*NUU*	*NU+*
	Σ	*N+1*	*N+2*	*0*	*0*	*N+U*	*N*

FIGURE 8.1 POPULATION ERROR MATRIX.

From the population error matrix several quality measures can be summarized, though it is strongly recommended that the error matrix is included in a validation assessment. Brus *et al.* (2011) follow the suggestion by Stehman (1997) that quality measures for categorical maps should be directly interpretable in terms of the probability of a misclassification and therefore recommend the use of three map quality measures: the overall purity, the map unit purity and class representation. We follow this recommendation here. Note that the map unit purity often is referred to as *user's accuracy*, and class representation as *producer's accuracy* (Stehman, 1997; Adhikari *et al.*, 2014). Lark (1995) however, questions the appropriateness of these terms since both quality measures can be important for users as well as producers. He proposes to use map unit purity and class representation instead, which is adopted by Brus *et al.* (2011) and followed here.

A fourth frequently used group of quality measures are Kappa indices, which adjust the overall purity measure for hypothetical chance agreement (Stehman, 1997). How this chance agreement is defined differs between the various indices. Some authors however, conclude that Kappa indices are difficult to interpret, not informative, misleading and/or flawed and suggest to abandon their use (Pontius and Millones, 2011). These authors argue that Kappa indices attempt to compare accuracy to a baseline of randomness, but randomness is not a reasonable alternative for map construction. We therefore do not consider kappa here.

The overall purity is the fraction of locations for which the mapped soil class equals the observed soil class and is defined as (Brus *et al.*, 2011):

$$p = \sum_{i=1}^{U} N_{uu} / N$$

which is the sum of the principal diagonal of the error matrix divided by the total number of locations in the mapping area. The overall purity can be interpreted as the areal proportion of the mapping area that is correctly classified.

Alternatively, an indicator approach can be used to compute the overall purity. A validation site gets a '1' if the observed soil class is correctly predicted and a '0' otherwise. The overall purity is then computed by taking the average of the indicators.

Map unit purity

The map unit purity is calculated from the row marginals of the error matrix. It is the fraction of validation locations with mapped class u for which the observed class is also u. The map unit purity for class u is defined as (Brus *et al.*, 2011):

$$p_u = \frac{N_{uu}}{N_{u+}}$$

The map unit purity can be interpreted as the proportion of the area of the map unit that is correctly classified. The complement of p_u, $1 - p_u$, is referred to as the error of commission for mapped class u.

Class representation

The class representation is calculated from the column marginals of the error matrix. It is the fraction of validation locations with observed class u for which the mapped class is u. The class representation for class u is defined as (Brus *et al.*, 2011):

$$r_u = \frac{N_{uu}}{N_{+u}}$$

The class representation can be interpreted as the proportion of the area where in reality class u occurs that is also mapped as class u. The complement of r_u, $1 - r_u$, is referred to as the error of omission for mapped class u.

8.2.3 ESTIMATING THE MAP QUALITY MEASURES AND ASSOCIATED UNCERTAINTY

In validation, we estimate the population means of the map quality measures from a sample taken from a limited number of locations in the mapping area. After all, we cannot afford to sample all locations, i.e. each grid cell of our soil map. Because the map quality measures are estimates, we are uncertain about these: we infer the quality measures from only a limited number of observations taken from the population. We do not know the true population means. The estimation uncertainty can be quantified with the sampling variance.

From the variance, the lower and upper boundary of a confidence interval, typically the 95%, can be computed using basic statistical theory:

$$CI = (\underline{\hat{x}} - 1{,}96x\,\frac{\sigma}{\sqrt{n}}\;;\;\underline{\hat{x}} + 1{,}96\,x\,\frac{\sigma}{\sqrt{n}})$$

where $\underline{\hat{x}}$ is the estimated map quality measure, for instance the ME, MSE or overall purity, σ is the estimated standard deviation of the map quality measure and σ is the validation sample size.

Quantified information about the uncertainty associated to map quality measures is useful and required for statistical testing. For instance, if one wants to test if one mapping method performs better than the other method one needs quantified information about uncertainty. Because we are uncertain about the estimated quality measures, an observed difference in map quality between two methods does not necessarily mean that one method is better than the others, even when there is a substantial difference. The difference might be attributed to chance because we infer the quality measures from a limited sample from the population. With statistical hypothesis testing we can calculate how large the probability is that observed difference is caused by chance. Based on the outcome we can accept or reject the hypothesis that there is no difference between the performance of two mapping methods (this would be the null hypothesis for statistical testing) for a given significance level, usually 0.05.

8.3. GRAPHICAL MAP QUALITY MEASURES

In addition to quantifying map accuracy statistically, one can also present validation results obtained from a sample graphically. This can be done by creating scatter plots of predicted against observed values and spatial bubble plots of validation errors. Figure 8.2 shows an example of a scatterplot and bubble plot. Both plots can be easily made with R (R Development Core Team, 2016). Use the function plot(x,y) to generate a scatter plot. The 1:1 line (black line in Figure 8.2) can be added to the plot with the command abline(0,1). The spatial bubble plot can be generated with the bubble function of the sp package (Pebesma and Bivand, 2005).

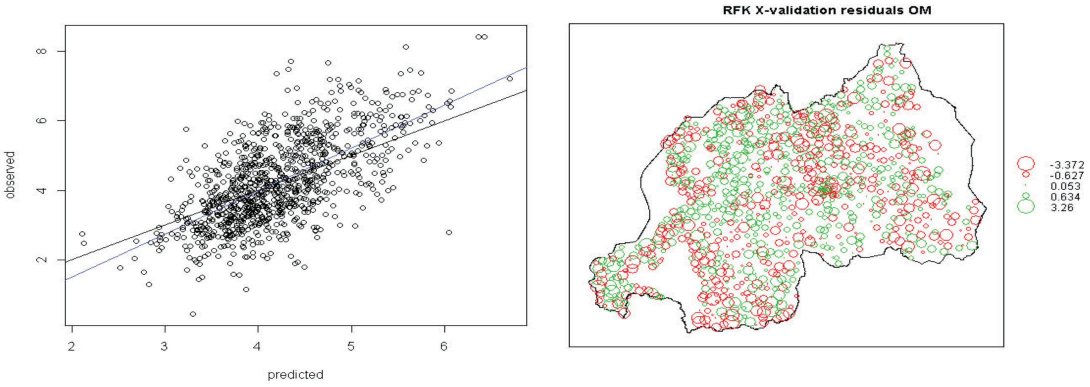


FIGURE 8.2 SCATTERPLOT OF PREDICTED VERSUS OBSERVED SOIL ORGANIC MATTER CONTENT FOR RWANDA (LEFT) AND SPATIAL BUBBLE PLOT OF CROSS-VALIDATION ERROR FOR SOIL ORGANIC MATTER (RIGHT) (KEMPEN *ET AL.*, 2015). THE BLACK LINE IN THE SCATTER PLOT REPRESENTS THE 1:1 LINE OF PREDICTION VERSUS OBSERVED, THE BLUE LINE REPRESENTS THE REGRESSION BETWEEN OBSERVED AND PREDICTED VALUES.

8.4. VALIDATION METHODS AND STATISTICAL INFERENCE

Following Brus *et al.* (2011), we introduce and discuss three common validation methods: *additional probability sampling*, *data-splitting* and *cross-validation*, and show how to estimate the map quality measures introduced in previous section from a sample.

With additional probability sampling an independent dataset is collected from the sampling population (all grid cells of a digital soil map) for the purpose of validation. This dataset is used in addition to a dataset that is used to calibrate a prediction model. Such dataset is often a legacy dataset collected with a purposive sampling design.

Data-splitting and cross-validation are applied in situations where one has only one data set available for prediction model calibration and validation. This can be a dataset collected with probability sampling, but in practice this typically is a legacy dataset collected with some purposive sampling design.

We warn here that if one uses data-splitting or cross-validation with a dataset collected with purposive sampling, then this has severe implications on the validity and interpretation of the estimated map quality measures as we will explain below.

8.4.1 ADDITIONAL PROBABILITY SAMPLING

The most appropriate approach for validation is by additional probability sampling. This means that an independent validation dataset is collected in the field on basis of a probability sampling design. Validation based on probability sampling ensures one obtains *unbiased* and *valid* estimates of the map quality measures (Brus *et al.*, 2011; Stehman, 1999). Additional probability sampling has several advantages compared to data-splitting and cross-validation using non-probability sample data. These are:

- no model is needed for estimating map quality estimates. We can apply *design-based estimation*, meaning that model-free unbiased and valid estimates of the map quality measures can be obtained;
- discussions on the validity of the estimated map quality are avoided;
- model-free, valid estimates of the variance of the map quality measures can be obtained that allow for hypothesis testing, e.g. for comparison of model performance.

Disadvantages can be extra costs involved in collecting an additional sample or terrain conditions that make it difficult to access all locations in the mapping area.

Probability sampling is random sampling such that:

- all locations in the mapping area have a probability larger than 0 of being selected
- the inclusion probabilities are known but need not be equal.

It should be noted that random sampling is often used for arbitrary or haphazard sampling. Such sampling is not probability sampling because the inclusion probabilities are not known. Design-based, model-free estimation of map quality measures is not possible in this case. All probability samples are random samples but not all random samples are probability samples. The term *probability sampling* should therefore only be used for random sampling with known inclusion probabilities.

There are many different probability sampling designs: simple, stratified, systematic, two-stage, clustered random sampling. We will not give an exhaustive overview here of all these designs. A good resource is de Gruijter *et al.* (2006). For reasons of simplicity we focus here on *simple random sampling*.

In simple random sampling, no restrictions are imposed on random selection of sampling sites except that the sample size is fixed and chosen prior to sampling (de Gruijter *et al.*, 2006). All sampling locations are selected with equal probability and independently from each other. This can for instance be done as follows (de Gruijter *et al.*, 2006):

1. Determine the minimum and maximum X and Y coordinates of the mapping area (the *bounding box*).
2. Generate two independent random coordinates X and Y from a uniform probability distribution on the interval (x_{min}, x_{max}) and (y_{min}, y_{max})
3. Check if the selected sampling site falls within the mapping area. Accept the sampling site if it does; discard the sampling site if it does not.
4. Repeat steps 2 and 3 until the locations have been selected.

If a sampling location cannot be visited because of inaccessibility for instance, then this location should be discarded and be replaced by a location chosen from a reserve list. Always the location at the top of the list should be selected for this purpose; not an arbitrarily chosen location from the list such as the closest one. It is not allowed to shift an inaccessible sampling location to a location nearby that can be accessed. Irregularity, clustering and open spaces characterise the simple random sampling design (de Gruijter *et al.*, 2006).

Estimation of quantitative map quality measures: For each validation location we compute the error, $e(s_i) = \hat{z}(s_i)-z(s_i)$, the absolute error, $|e|(s_i) = |\hat{z}(s_i)-z(s_i)|$, or squared error, $e^2(s_i) = (\hat{z}(s_i)-z(s_i))^2$. The spatial mean of the mapping area for map quality measure x is then estimated by:

$$\underline{\hat{x}} = \frac{1}{N}\sum_{i=1}^{N} x(s_i)$$

where i indicates the validation location, i = 1,2, ... ,n, n the validation sample size, and $x(s_i)$ the estimated population mean of map quality measure x at location $s_i \cdot x$ is the prediction error in case of the ME, absolute error in case of the MAE, squared prediction error in case of the MSE. Note that the estimator is the unweighted sample mean. This unweighted mean is an unbiased estimator because all sampling locations were selected with equal probability.

The MSDR is estimated by:

$$\widehat{MSDR} = \frac{1}{N}\sum_{i=1}^{n} \frac{(\hat{z}(s_i)-z(s_i))^2}{\sigma^2(s_i)}$$

and the AVE by:

$$\widehat{AVE} = 1 - \frac{\sum_{i=1}^{n} (\hat{z}(s_i)-z(s_i))^2}{\sum_{i=1}^{n} (z(s_i)-\underline{\hat{z}})^2}$$

where $\hat{z}$ is the mean of the target soil property estimated from the validation sample.

One should be careful when assessing the proportion of variance explained by computing the R^2 from a linear regression of the predicted value on the observed value (Krause *et al.*, 2005), as is often done in practice. The R^2 quantifies the dispersion around the regression line; not around the 1:1 line in which we are interested in validation. So it does not directly compare the predicted with observed value as does the AVE; i.e. it is not based on the prediction error. A high R^2-value therefore, does not automatically mean a high AVE. For instance, in case of strongly biased predictions the R^2 can be high but the AVE will be low. The blue line in Figure 8.2 is the regression line that one obtains when regression the observed value on the predicted value. This line slightly differs from the 1:1 line. In this example the R^2 of the regression is 0.42 while the AVE is 0.40.

The uncertainty associated to the estimated map quality measures is quantified with the sampling variance, which for the ME, MAE and MSE is estimated by:

$$Var(\underline{\hat{x}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n} (x(s_i)-\underline{\hat{x}})$$

and the 95% confidence interval (CI) of is given by:

$$CI_{95} = \underline{\hat{x}} \pm 1{,}96\, x \sqrt{Var(\underline{\hat{x}})}$$

We should warn here that the calculation of the CI is based on the assumption that the estimated map quality measure means have a normal distribution (the central limit theorem). For the squared errors this assumption can be unrealistic, especially for small sample sizes.

Estimation of qualitative map quality measures: For validation of qualitative soil maps, a sample error matrix is constructed from the validation data (Figure 8.3). n is the total number of validation locations in the sample. Element nij of the matrix corresponds to the number of validation locations that have been predicted as class *i*, *i* = 1,2, … , *U* and belong to class *j*, *j* = 1,2, … ,*U* (Lark, 1995). The matrix summarizes correct predictions and incorrect predictions within the validation data.

		Observed					
		1	2	.	.	*U*	Σ
Mapped	1	*n11*	*n12*	.	.	*n1U*	*n1+*
	2	*n21*	*n22*	.	.	*n2U*	*n2+*
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	U	*nU1*	*nU2*	.	.	*nUU*	*nU+*
	Σ	*n+1*	*n+2*	*0*	*0*	*n+U*	*n*

FIGURE 8.3 SAMPLE ERROR MATRIX.

From the sample error matrix the overall purity, map unit purity and class representation are estimated by:

$$\hat{p} = \sum_{i=1}^{U} n_{uu}/n$$

$$\hat{p}_u = \frac{n_{uu}}{n_{u+}}$$

$$\hat{r}_u = \frac{n_{uu}}{n_{+u}}$$

Alternatively, the overall purity can be estimated by defining a purity indicator variable for each validation location that takes value 1 if the mapped soil class equals the observed soil class at that location, and 0 else. The overall purity is then estimated by:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \partial(s_i)$$

where is the indicator variable at validation location s_i. The variance of the estimated overall purity is estimated by:

$$Var(\hat{p}) = \frac{1}{n(n-1)}\sum_{i=1}^{n} (\partial(s_i) - \hat{p})^2$$

Alternatively, the variance is estimated by:

$$Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$$

which is the variance of a binomial probability distribution. The 95% confidence interval of $\hat{p}$ is given by:

$$CI_{95} = \hat{p} \pm 1{,}96\, x \sqrt{Var(\hat{p})}$$

We warn that the CI as calculated here is a rough approximation which only holds when n x $\hat{p}$ and n x (1 - $\hat{p}$) are large (5 as a rule of thumb). Otherwise the binomial distribution should be used to compute the CI.

Figure 8.4 shows a hypothetical example of a sample error matrix for soil class map. For this example, the overall purity is estimated by: (19 + 33 + 25 + 42 + 19)/240 = 0.575, meaning that for an estimated 57.5% of the mapping area the mapped soil class is equal to the true soil class.

		Observed					
		Anthrosol	Cambisol	Gleysol	Luvisol	Podzol	Σ
Mapped	Anthrosol	19	5	3	0	1	28
	Cambisol	5	33	9	13	5	65
	Gleysol	2	8	25	3	5	43
	Luvisol	3	15	9	42	2	71
	Podzol	1	3	8	2	19	33
	Σ	30	64	54	60	32	240

FIGURE 8.4 SAMPLE ERROR MATRIX FOR A HYPOTHETICAL SOIL CLASS MAP.

Table 8.1 gives the map unit purities and class representations for this example. The map unit purity of the Gleysol is 0.581, meaning that at 58.1% of the validation locations for which a Gleysol is predicted, a Gleysol is observed. Assuming the validation data were collected by simple random sampling, we could conclude that for 58.1% of the area mapped as Gleysol we would find a Gleysol in the field. The class representation of the Gleysol is 0.463, meaning that for 46.3% of the validation locations classified as Gleysol, we map a Gleysol. The majority of the Gleysol locations is thus mapped as a different soil class. Again, assuming the validation data were collected by probability sampling, we would estimate that 22.5% ($\frac{54}{240} x\ 100\ \%$) of our mapping area is covered by Gleysols. We map Gleysols for 17.9% of the area ($\frac{43}{240} x\ 100\ \%$). It can happen that a soil class has a high map unit purity and a low class representation. This means that if we map a Gleysol we will likely find a Gleysol there, but that a large extent of the true Gleysol area is not mapped as such.

Table 8.1. Map unit purity and class representation statistics for the hypotheticalexample given in Figure 8.4.

	map unit purity	class representation
Anthrosol	0.679	0.633
Cambisol	0.508	0.516
Gleysol	0.581	0.463
Luvisol	0.592	0.700
Podzol	0.576	0.594

8.4.2 DATA-SPLITTING

In data-splitting the sample data set is split in two subsets. One subset is used to calibrate the prediction model. The other subset is used for validation. A frequently used splitting criterion is 70-30, where 70% of the sample data are used for calibration and 30% for validation. The choice of a splitting criterion however, is arbitrary and it is not evident how to split a data set in such a way that unbiased and valid estimates of the map accuracy can be obtained. For sparse data sets, data-splitting can be inefficient since the information in the data set is not fully exploited for both calibration and validation.

It is important to note here that a random subsample of (legacy) data that are collected with a purposive (non-probability) design, is *not* a probability sample of the study area. This means that design-based estimation of map quality measures is not possible.

If a validation (sub)sample is a non-probability sample of the mapping area, then we must account for possible spatial autocorrelation of the prediction errors when estimating the map quality measures. One can imagine that when two validation locations are close together and the prediction errors are correlated that there is less information in these two locations (there is information redundancy because of autocorrelation) than in two isolated locations. This information redundancy has to be accounted for when estimating map quality measures and implies that we have to rely on model-based estimation: a model for the spatially autocorrelated prediction error has to be assumed. Thus, we will not obtain model-free, unbiased and valid estimates of the quality measures from non-probability sample validation data. In a case study, Knotters and Brus (2013) showed that model-based predictions of producer's accuracies from two models differed strongly, indicating that with the model-based approach the validation results strongly depend on model assumptions.

In most studies however, spatial correlation is not accounted for when estimating map quality measures using the estimators presented above under 'Simple random sampling' from non-probability sample data. In such case, the quality measures cannot be considered unbiased and valid estimates of the population means of the map quality measures. In addition, the estimated variance of the map quality measures is not valid and statistical testing of mapping methods to assess which method gives the most accurate predictions cannot be done.

In other words, if the simple random sampling estimators are used to estimate map quality measures then these are only valid for the validation data points. The map quality measures do not give a valid estimate of the quality of the map as a whole (the population). For instance, the overall purity cannot be interpreted as an areal proportion of correctly mapped soil classes, only as the proportion of the validation data points for which the soil class is correctly predicted.

8.4.3 CROSS-VALIDATION

In *K*-fold cross-validation (CV), the dataset is split into *K* roughly equal sets. One of these sets is set aside for validation. The model is then calibrated using the data from the *K*-1 sets and used to predict the target variable for the data points set aside. From this prediction the prediction error is calculated. This procedure is repeated *K* times, each time setting a different set aside for validation. In this way we obtain *K* estimates of the prediction error: one for each validation sample site. In this way, all data are used for validation and model calibration. It is thus much more efficient than data-splitting.

K is typically chosen as 5 or 10, or as *N* the number of data points. The latter is referred to as leave-one-out cross-validation (LOOCV) in which only one validation site is set aside in each iteration. The model is then calibrated with *N*-1 observations. Some repeat *K*-fold cross-validation a number of times and average the results to obtain a more robust estimate of the map quality measures.

Note that the problem of spatially correlated errors remains when data are non-probability sample data. Cross-validation using a non-probability sampling dataset suffers from the same drawbacks with respect to unbiasedness and validity of the estimates of the map quality measures as data-splitting. The estimates cannot be interpreted as being valid for the mapping area, but only for the validation locations.

In R, the caret package (Kuhn, 2015) offers functionality for data-splitting and cross-validation.

8.5 References

Adhikari, K., Minasny, B., Greve, M.B., Greve, M.H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214–215: 101-113.

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., Morrás, H.J.M., 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. Geoderma 281, 102-118.

Brus, D.J., Spätjens, L.E.E.M., de Gruijter, J.J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89(1–2), 129-148.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62(3), 394-407.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for natural resource monitoring. Springer.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., De Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci Soc Am J 76(6), 2097-2115.

Kempen, B., Heuvelink, G.B.M., Brus, D.J., Stoorvogel, J.J., 2010. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. European Journal of Soil Science 61, 333-347.

Kempen, B., Vereijken, P., Keizer, P., Ruiperez Gonzalez, M., Bindraban, P., Wendt, J., 2015. Preliminary evaluation of the feasibility of using geospatial information to refine soil fertility recommendations., VFRC Report 2015/6. Virtual Fertilizer Research Centre, Washington D.C.

Knotters, M., Brus, D.J., 2013. Purposive versus random sampling for map validation: a case study on ecotope maps of floodplains in the Netherlands. Ecohydrol. 6(3), 425-434.

Krause, P., Boyle, B.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5, 89-97.

Lark, R.M., 1995. Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. International Journal of Remote Sensing 16(8), 1461-1480.

Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science 51, 137-157.

Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R News 5(2).

Pontius, R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32(15), 4407-4429.

Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244, 214-227.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment 62(1), 77-89.

Stehman, S.V., 1999. Basic probability sampling designs for thematic map accuracy assessment. International Journal of Remote Sensing 20(12), 2423-2441.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. Journal of Soil Science 41(3), 473-490.

Webster, R., Oliver, M.A., 2007. Geostatistics for environmental scientists. Statistics in practice. Second edition ed. John Wiley & Sons, Chichester

9. DATA SHARING

This chapter reviews possibilities and "good practices" of exchanging produced soil data. Once the analysis, spatial prediction and quality control have been all completed, it is useful to follow some minimum steps and export and prepare the data for distribution so that its potential users can easily access it, use it, and make correct interpretation of data. We consider geo-publishing options for soil data either based on using third-party web services or by using one's own installation of the software. We put a clear focus on using the Open Source software solutions: GDAL[1], R[2], GeoServer[3], OpenLayers[4] and Leaflet[5], and public domain data and metadata standards.

The authors have 15+ years of experience with producing, publishing and sharing soil maps and have been involved in large soil mapping projects where data volumes often exceed standard desktop GIS capacities. For information on specific software please refer to the provided links. Even more information on using GDAL and similar GIS tools through a command line can be found via the Global Soil Information Facilities tutorials of ISRIC at *http://gsif.isric.org*. The text is illustrated with example scripts of the statistical software R in combination with GDAL.

9.1 EXPORT FORMATS

9.1.1 TYPE OF SOIL DATA AND THEIR FORMATTING

Before we start reviewing soil data formats, it is useful to understand which types of soil variables, soil maps and soil DBs are most commonly generated and used, and what are their specific advantages and limitations. Soil science works with many variables common to ecology and/or physical geography (e.g. soil temperature), but it also works with several variables specific to soil science only. Some soil factor-type variables specific to soil science only are for example:

[1] HTTP://WWW.GDAL.ORG / HTTPS://EN.WIKIPEDIA.ORG/WIKI/GDAL
[2] HTTPS://CRAN.R-PROJECT.ORG/WEB/VIEWS/SPATIAL.HTML
[3] HTTP://GEOSERVER.ORG/
[4] HTTPS://OPENLAYERS.ORG/
[5] HTTP://LEAFLETJS.COM/

Soil taxa or soil classes (this includes taxonomic systems and connected diagnostic soil properties and horizons).

- Soil texture-class systems.
- Soil color classification systems e.g. Munsell color codes.
- Soil drainage classes (hydrological classifications).
- Soil diagnostic horizons.

Consider for example the following soil texture data:

```
> library(soiltexture)
> tex <- data.frame(
+  CLAY = c(05,60,15,05,25,05,25,45,65,75,13,47),
+  SILT = c(05,08,15,25,55,85,65,45,15,15,17,43),
+  SAND = c(90,32,70,70,20,10,10,10,20,10,70,10)
+ )
>
> TT.plot(class.sys = "USDA.TT", tri.data = tex, main = "", cex.axis=.7, cex.lab=.7)
```

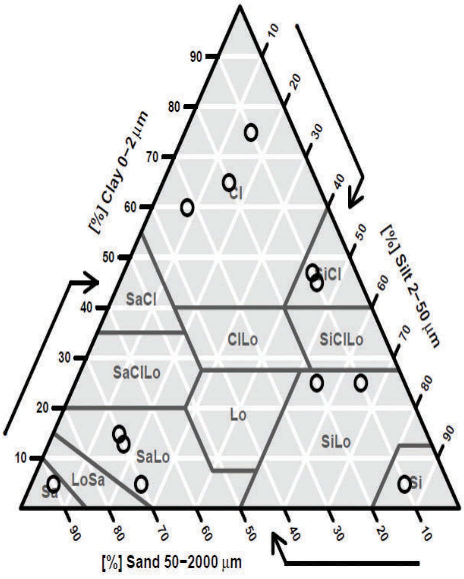


Figure 1: Soil texture triangle plot with 12 measurements of fine earth soil texture fractions (sand, silt and clay) (**soiltexture** package).

FIGURE 9.1 SOIL TEXTURE TRIANGLE PLOT. AN EXAMPLE OF SOIL SCIENCE SPECIFIC DATA.

The way soil texture data is displayed and texture classes (SaLo, Lo, Sa etc.) used in a texture triangle is specific to soil science. The way this data is formatted and presented can be, likewise, specific to soil science only.

Most of soil data is in fact spatial. **"Spatial"** implies that spatial (and temporal) reference is attached to each measured / estimated value, i.e. it is location specific. Spatio-temporal references typically includes for example:

- Geographic location in local or geographic coordinates (ideally longitude and latitude in the WGS84 coordinate system);
- Depth interval expressed in cm from land surface (upper and lower depth);
- Support size or referent soil volume (or voxel) i.e. the horizontal sampling area multiplied by the thickness of the sampling block;
- Temporal reference i.e. begin and end date/time of the period of measurements/ estimations.

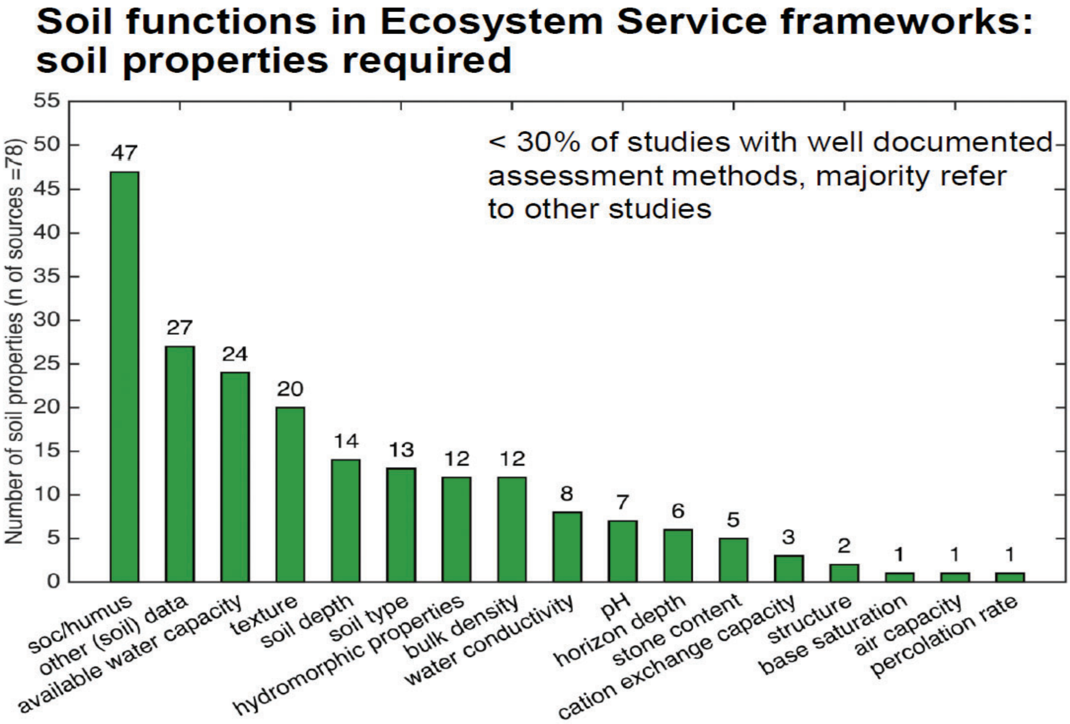
Spatial data formats are used to represent spatial objects. This can be (Bivand *et al.* 2013; Neteler and Mitasova, 2013):

- Points (2D or 3D): used to represent sampling locations, soil horizons, soil profiles etc.
- Lines (2D): used to represent soil transects, streams, administrative boundaries etc.
- Polygons (2D): used to represent soil mapping units and/or geomorphological units, landforms, administrative areas, farms, plot trials etc.
- Grids or rasters (2D or 2.5D): used to represent soil spatial predictions (spatially complete) of soil properties and classes etc.
- 3D grids or Voxels: used to represent soil spatial predictions (spatially complete) of soil properties in 3D.

It is also important to be able to distinguish between sampled or predicted soil data:

1. Soil samples (usually points or transects) are spatially incomplete. They are used to generate spatial predictions.

2. Spatial predictions of soil variables (soil maps) are spatially complete. They are used for decision making and further modeling i.e. they are used to construct a Soil Information System.



[6]

FIGURE 9.2 SOME FREQUENTLY REQUIRED SOIL VARIABLES (SORTED BY NUMBER OF STUDIES) BASED ON THE STUDY BY KELLER *ET AL.* (2014). THIS LIST IS PROBABLY COUNTRY/PROJECT SPECIFIC BUT ILLUSTRATES THE DIFFERENCES CONSIDERING THE INTEREST IN SOIL DATA.

A collection of spatially exhaustive soil grids of various soil properties (physical and chemical soil properties, soil water, soil classification etc) make a **Soil Information System** (SIS). SIS are often complemented with soil sample data and serve both data formats[6]. A Soil Information System should preferably be a **Database** (DB), so that users can access and query data using some standard DB languages (for example SQL). Steps to export soil data into a DB format are explained in later sections.

[6] E.G.:HTTP://WWW.LANDIS.ORG.UK/, HTTP://WWW.ASRIS.CSIRO.AU

9.1.2 GENERAL GIS DATA FORMATS: VECTOR, RASTER, TABLE

All soil data we produce through soil mapping can be in principle distributed using one of the two basic GIS formats of data:

- *Vector format*: this format is often more suitable for exporting point, line and polygon (areal) data,
- *Raster or gridded format*: this format is often more suitable for exporting spatial predictions of soil variables,

Data in vector format can be converted to raster (see e.g. rasterize function in the raster R package[7]) and vice versa — raster data can be converted to vector formats. For example, rasters can be converted to polygons (see e.g. rast2vect function in the plotKML package[8]). If the conversion is done carefully and if all the relations between scale and pixel size have been considered (see Hengl, 2006 for more details), then information loss due to conversion from raster to vector and vice versa should be minimal.

Both vector and raster GIS data can also be converted to tabular data format. By converting a GIS layer to a table, spatial geometry and spatial relations will be 'stripped off', so that only limited spatial analysis operations can be applied. To convert raster layers to tabular data, consider using the **SpatialPixelsDataFrame-class** in the sp package[9] and/or the **RasterLayer-class** from the raster package[10] in combination with the rgdal package (Bivand *et al.* 2013):

```
> library(rgdal)
> library(plotKML)
> ?readGDAL
> spnad83 <- readGDAL(system.file("pictures/erdas_spnad83.tif", package = "rgdal")[1])
... erdas_spnad83.tif has GDAL driver GTiff
and has 658 rows and 571 columns
> spnad83.tbl <- as.data.frame(spnad83)
> str(spnad83.tbl)
'data.frame': 375718 obs. of 3 variables:
 $ band1: int 0 0 0 0 0 0 0 0 0 0 ...
 $ x    : num 79019 79059 79099 79139 79179 ...
 $ y    : num 1439248 1439248 1439248 1439248 1439248 ...
```

[7] HTTPS://WWW.RDOCUMENTATION.ORG/PACKAGES/RASTER/VERSIONS/2.5-8/TOPICS/RASTERIZE
[8] HTTPS://WWW.RDOCUMENTATION.ORG/PACKAGES/PLOTKML/VERSIONS/0.5-6/TOPICS/VECT2RAST
[9] HTTPS://WWW.RDOCUMENTATION.ORG/PACKAGES/SP/VERSIONS/1.2-4/TOPICS/SPATIALPIXELSDATAFRAME
[10] HTTPS://WWW.RDOCUMENTATION.ORG/PACKAGES/RASTER/VERSIONS/2.5-8/TOPICS/RASTER

where as.data.frame is a function converting a raster object to a table. Note that the output table now contains coordinates for each cell (center of the grid node), which is in fact memory inefficient as coordinates are provided for each row in the table.

Likewise, to convert a vector layer to tabular formats one can use the **Simple Features** functionality of the sf package[11]. The SF standard is widely implemented in spatial databases (PostGIS, ESRI ArcGIS) and forms the vector data basis for libraries such as GDAL and web standards such as GeoJSON (*http://geojson.org/*). To convert for example spatial polygons layer to a tabular format we would use:

```
> library(sf); library(plotKML)
> data(eberg_zones)
> class(eberg_zones)
[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"
> eberg_zones.tbl <- as(eberg_zones, "sf")
> str(eberg_zones.tbl)
Classes ‘sf’ and 'data.frame':  11 obs. of  2 variables:
 $ ZONES   : Factor w/ 4 levels "Clay_and_loess",..: 2 1 2 3 4 2 2 2 3 3 ...
 $ geometry: List of  11 , printing List of 1
  ..$ : num [1:313, 1:2] 3570250 3570250 3570262 3570275 3570288 ...
  ..- attr(*, "class")= chr  "XY" "POLYGON" "sfg"
 - attr(*, "sf_column")= chr "geometry"
 - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",..: NA
  ..- attr(*, "names")= chr "ZONES"
```

Note that using spatial layers in simple tabular formats can be cumbersome because many spatial relationships and properties are likely lost (although these can be assigned reversibly). In addition, the size of tabular objects is much bigger than if we use data in the original GIS data formats, especially if those formats support compression. On the other hand, having data in tabular format can be often the only way to exchange the data from spatial to non-spatial databases or from software without any data communication bridge. Also, tabular data is human-readable which means that it can be opened in text editors, spreadsheet programs or similar.

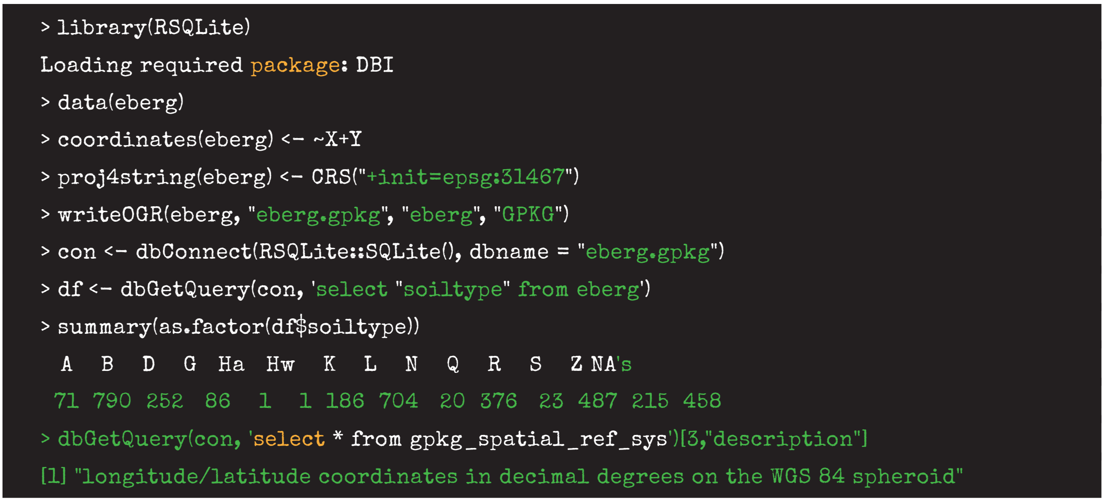
[11] HTTPS://CRAN.R-PROJECT.ORG/WEB/PACKAGES/SF/VIGNETTES/SF1.HTML

9.1.3 RECOMMENDED GIS DATA EXCHANGE FORMATS

As a general recommendation producers of soil data should primarily look at using the following data formats for exchanging soil data (points, polygons and rasters):

- **GPKG**[12] (an Open Format for Geospatial Information): platform-independent, portable, self-describing, compact format for transferring geospatial information.
- **GeoTIFF** (for rasters): a TIFF (image) file that allows embedding spatial reference information, metadata and color legends. It also supports internal compression algorithms and hierarchical indexing.

Both formats can be read easily in R or similar data processing software. Vectors are also commonly exported and shared in ESRI Shapefile (SHP) format. The advantage of GPKG format versus somewhat more common ESRI SHP format is that **GPKG files are basically a portable database** (SQLite container) so that the user does not have to import the whole data into a program but also fetch parts of data by using SQL queries and it can handle vector and raster data in it. The following example demonstrates how to create a GPKG file and how to query it:

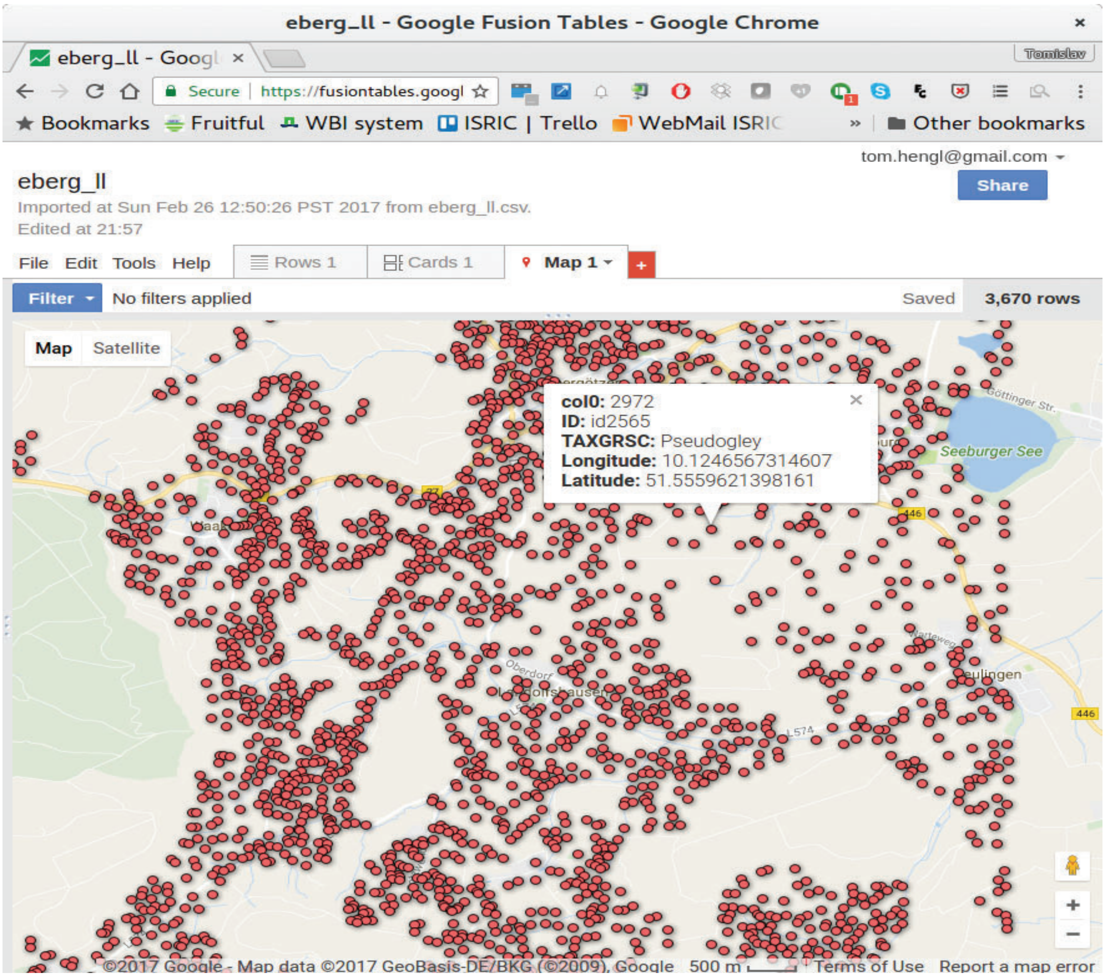


```
> library(RSQLite)
Loading required package: DBI
> data(eberg)
> coordinates(eberg) <- ~X+Y
> proj4string(eberg) <- CRS("+init=epsg:31467")
> writeOGR(eberg, "eberg.gpkg", "eberg", "GPKG")
> con <- dbConnect(RSQLite::SQLite(), dbname = "eberg.gpkg")
> df <- dbGetQuery(con, 'select "soiltype" from eberg')
> summary(as.factor(df$soiltype))
  A   B   D   G  Ha  Hw   K   L   N   Q   R   S   Z NA's
 71 790 252  86   1   1 186 704  20 376  23 487 215  458
> dbGetQuery(con, 'select * from gpkg_spatial_ref_sys')[3,"description"]
[1] "longitude/latitude coordinates in decimal degrees on the WGS 84 spheroid"
```

Note that the RSQLite package is a generic package for connecting to SQLite DBs. This means that GPKG files can be accessed and updated in its native storage format without intermediate format translations. Just putting a GPKG file on server with read and execute access allows users to connect and fetch data.

[12] HTTP://WWW.GEOPACKAGE.ORG/

Alternatively, it is also good idea to store point data in non-spatial format such as simple tables. For example, in comma-separated file format (.csv). A fast way to publish and share tabular data is to use Google Fusion Tables™. Google Fusion Tables have an API[13] that allows accessing and using tabular data through various programming platforms. The limitation of using Google Fusion tables is however, data size (currently about 1GB per user) and similar data volume limits, so this platform should be only used as intermediate solution for smaller data sets.



eberg_ll

Imported at Sun Feb 26 12:50:26 PST 2017 from eberg_ll.csv.
Edited at 21:57

No filters applied

3,670 rows

col0: 2972
ID: id2565
TAXGRSC: Pseudogley
Longitude: 10.1246567314607
Latitude: 51.5559621398161

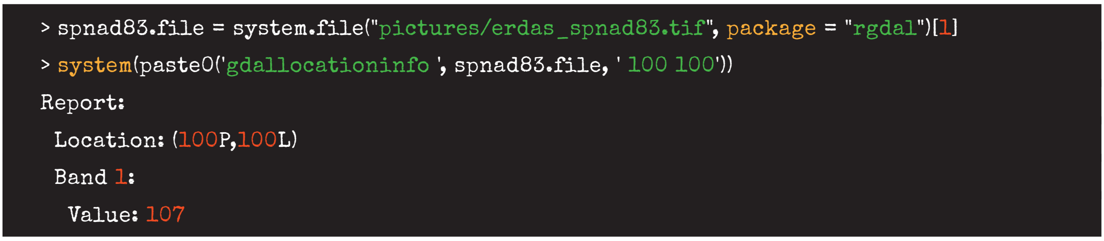
FIGURE 9.3 DISPLAYING POINT DATA SET EBERG (USED IN THE PREVIOUS EXAMPLE) IN GOOGLE FUSION TABLES.

[13] HTTPS://DEVELOPERS.GOOGLE.COM/FUSIONTABLES/

GeoTIFF format is highly recommended for sharing raster data for the following reasons:

1. It is GDAL's default data format and much functionality for subsetting, reprojecting, reading and writing GeoTIFFs already exists (see GDAL utils).
2. It supports internal compression via creation options (e.g. "COMPRESS=DEFLATE").
3. Extensive overlay, subset, index, translate functionality is available via GDAL and other Open Source software. Basically GeoTiff functions as a raster DB.

Consider for example the **gdallocationinfo**[14] function which allows spatial queries following some indexing system such as row and column number:



```
> spnad83.file = system.file("pictures/erdas_spnad83.tif", package = "rgdal")[1]
> system(paste0('gdallocationinfo ', spnad83.file, ' 100 100'))
Report:
 Location: (100P,100L)
 Band 1:
  Value: 107
```

Such type of overlay operations, thanks to GDAL (Warmerdam, 2008), are extremely fast and efficient. Likewise, **gdalwarp** function can be used subset rasters based on spatial extent or grid index. Rasters imported to GeoServer and shared through Web Coverage Service (see next section) or similar likewise function as a spatial raster DB.

As a general recommendation, and to avoid large file sizes, we recommend, however, that you always use integers inside GeoTiffs because floating point formats can lead to up to 4+ times larger sizes (without any gains in accuracy). This might mean you have to multiply the values of the soil property of interest by 10 or 100, in order not to lose accuracy (e.g. multiply pH values by 10 before exporting your raster as a GeoTiff).

[14] HTTP://WWW.GDAL.ORG/GDALLOCATIONINFO.HTML

9.2 WEB SERVICES - SERVING SOIL DATA USING WEB TECHNOLOGY

9.2.1 THIRD-PARTY SERVICES

If you are a data producer but with limited technical capacity and/or financial resources, then publishing geo-data through a third-party service could be very well that the easiest and most professional solution for you. Some commonly used commercial web-services to share geo-data are:

- Google MyMaps (https://www.google.com/mymaps)
- ArcGIS Online (https://www.arcgis.com/home)
- MapBox (https://www.mapbox.com)
- CARTO (https://carto.com)

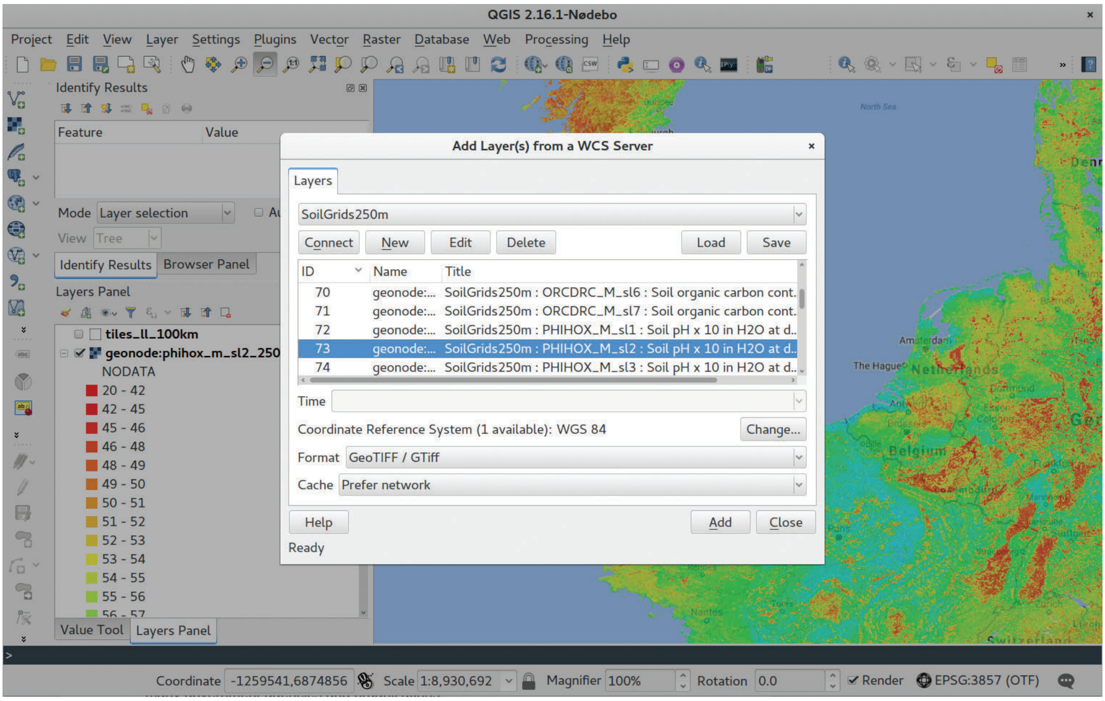
All these have limitations and primarily suitable for sharing vector type data only. Their free functionality is very limited so before you start uploading any larger data sets, please check the size limits based on your account. Upgrading your license will allow you to increase storage and functionality so that even with few hundred dollars per year you could have a robust solution for sharing your data to thousands of users.

Soil data producers can also contact ISRIC, as World Data Centre for Soils, to request support for hosting and/or distributing their soil data in case they lack the technical capacity to do so themselves, while adhering to the data sharing agreement and licence set by the data producer.

9.2.2 GEOSERVER (WEB SERVING + WEB PROCESSING)

GeoServer (http://geoserver.org) is Open Source software solution for serving raster or vector data. It includes majority of the Open Geospatial Consortium Service standards: the Web Map Service, Web Coverage Service and Web Processing Service (Youngblood, 2013). Installation and maintenance of GeoServer is however not trivial and requires specialized technical staff. Web services can also entail significant costs depending on the amount of web-processing and web-traffic. For every medium to large size organization, it is probably a better idea to use the out-of-box solution for GeoServer which is the GeoNode (*http://geonode.org*). GeoNode also includes a web-interface and user-management system so that new layers can be uploaded to GeoServer through a web-form type interface.

Very important functionality of GeoServer are the OGC standard services such as the **Web Coverage Service** (WCS) and the **Web Feature Service** (WFS). WCS means that not only data views are available to users, but WCS can also do data translation, aggregation or resampling, overlay etc. Consider for example the SoilGrids WCS (which can also be opened in QGIS or similar software supporting WCS). Users can direct this WCS and request only a subset of data for some area, aggregated to some preferred resolution / pixel size by using gdal_translate or similar. This means that, in few steps, users can download a subset of data on GeoServer in a preferred format without a need to download the whole data set.



SoilGrids250m

ID	Name	Title
70	geonode:...	SoilGrids250m : ORCDRC_M_sl6 : Soil organic carbon cont.
71	geonode:...	SoilGrids250m : ORCDRC_M_sl7 : Soil organic carbon cont.
72	geonode:...	SoilGrids250m : PHIHOX_M_sl1 : Soil pH x 10 in H2O at d..
73	geonode:...	SoilGrids250m : PHIHOX_M_sl2 : Soil pH x 10 in H2O at d..
74	geonode:...	SoilGrids250m : PHIHOX_M_sl3 : Soil pH x 10 in H2O at d..

Time

Coordinate Reference System (1 available): WGS 84 Change...

Format GeoTIFF / GTiff

Cache Prefer network

FIGURE 9.4 SOILGRIDS (HENGL *ET AL.* 2017) WCS OPENED IN QGIS.

9.2.3 VISUALIZING DATA USING LEAFLET AND/OR GOOGLE EARTH

A quick way to visualize produced soil maps and then share them to users without GIS capacities is to use Leaflet package[15]. Leaflet is basically a stand-alone web-page that contains all information (including some popular Web Mapping Services) so that users can visually explore patterns without having to install and use any desktop GIS. Consider the following example:

```
> library(leaflet)
> library(htmlwidgets)
> library(GSIF)
> library(raster)
> demo(meuse, echo=FALSE)
> omm <- autopredict(meuse["om"], meuse.grid[c("dist","soil","ffreq")],
method="ranger", auto.plot=FALSE, rvgm=NULL)
> meuse.ll <- reproject(meuse["om"])
Reprojecting to +proj=longlat +datum=WGS84 ...
> m = leaflet() %>% addTiles() %>%
addRasterImage(raster(omm$predicted["om"]), colors = SAGA_pal[[1]][4:20])
%>% addCircles(lng = meuse.ll@coords[,1], lat = meuse.ll@coords[,2],
color = c('black'), radius=meuse.ll$om)
> saveWidget(m, file="organicmater_predicted.html")
```

[15] HTTPS://RSTUDIO.GITHUB.IO/LEAFLET/

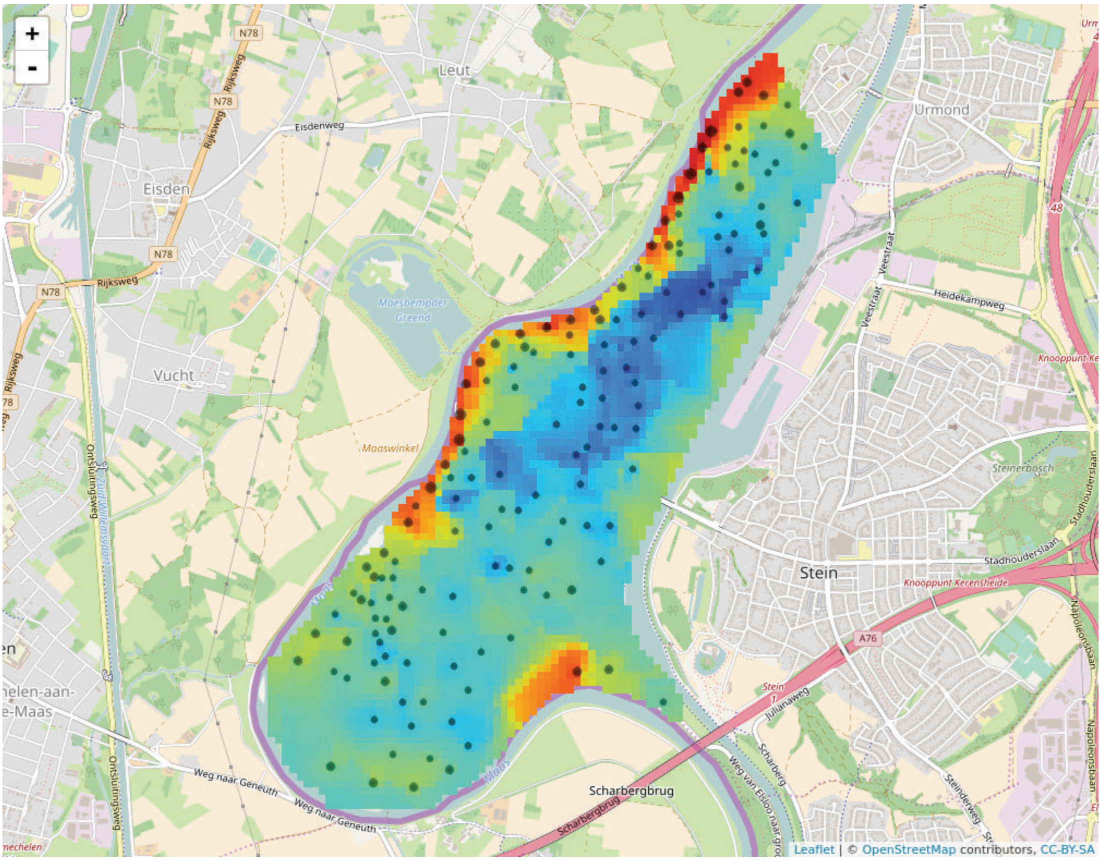


FIGURE 9.5 SAMPLED LOCATIONS AND PRODUCED PREDICTIONS VISUALIZED USING LEAFLET PACKAGE.

Note that the whole data set including styling and legends is basically available through a single html file (organicmater_predicted.html). Anyone opening that html in their browsers will get an interactive web-map that contains both samples and spatial predictions.

An alternative to using Leaflet is to put all data, including documents and multimedia, about your project in a KML (Keyhole Markup Language) file, so the data is available for viewing in Google Earth (Hengl *et al.* 2015). KML is very rich in what it can incorporate: textual data, photographs, documents, animations, videos etc. In fact, probably whole projects can be put into a single KML files so that the users only need to open it in Google Earth and then explore interactively. Note that KML files with ground overlays will be generated by GeoServer by default, although further customization is up to the data producer.

9.3. PREPARING SOIL DATA FOR DISTRIBUTION

9.3.1 METADATA

One important thing to consider prior to data distribution is construction of metadata (explanation of data, how was it produced and what are the exact technical specifications). There are several metadata standards that can be used to prepare metadata. More recently, complete and consistent metadata is a requirement by many government agencies and organizations. There are now several public metadata validators[16] that run all possible consistency and completeness checks before the metadata (and data) can be accepted.

Typical metadata should (at least) contain:

DEFINITION

- DETAILED DESCRIPTION OF THE VARIABLES AVAILABLE IN THE DATA.
- DATA LICENSE AND TERMS OF USE (URL).
- EXPLANATION OF MEASUREMENT METHODS AND UNITS USED.
- MENTION OF THE REFERENCE SUPPORT SIZE INCLUDING REFERENT DEPTH INTERVALS TO WHICH THE SOIL DATA REFERS TO (E.G. 0–30 CM DEPTH INTERVAL).
- MENTION OF THE REFERENT TIME PERIOD IN WHICH THE CALIBRATION DATA WAS COLLECTED.
- LINK TO LITERATURE (REPORT, BOOK OR SCIENTIFIC ARTICLE) WHERE THE DATA PRODUCTION IS EXPLAINED IN DETAIL. USING A PUBLISHED AND PEER-REVIEWED SCIENTIFIC ARTICLE AS THE MAIN REFERENCE FOR DATA IS A GOOD PRACTICE SINCE IT ALSO SHOWS THAT THE DATA PRODUCTION PROCESS HAS BEEN EVALUATED BY INDEPENDENT RESEARCHERS.
- PROJECT HOMEPAGE I.E. URL CONTAINING MORE INFORMATION AND ESPECIALLY UP-TO-DATE CONTACTS WHERE USERS CAN FIND ORIGINAL DATA PRODUCERS AND REQUEST SUPPORT.

[16] E.G. HTTPS://MRDATA.USGS.GOV/VALIDATION/

Metadata (including color legends) can be also directly embedded into the GeoTiff file by using the gdal_edit command[17] available in GDAL. The following example shows how to add a simple explanation of the data and a URL to find more info about the GeoTiff:

```
> data("eberg_grid")
> gridded(eberg_grid) = ~ x+y
> proj4string(eberg_grid) <- CRS("+init=epsg:31467")
> writeGDAL(eberg_grid["DEMSRT6"], "eberg_DEM.tif", options="COMPRESS=DEFLATE")
> ?eberg
> system(paste0('gdal_edit.py -mo \"DESCRIPTION=elevation values from the
SRTM DEM\" -mo \"DOWNLOAD_URL=http://geomorphometry.org/content/ebergotzen\"
eberg_DEM.tif'))
> system('gdalinfo eberg_DEM.tif')
Driver: GTiff/GeoTIFF
Files: eberg_DEM.tif
Size is 100, 100
Coordinate System is:
PROJCS["DHDN / 3-degree Gauss-Kruger zone 3",
...
Origin = (3570000.000000000000000,5718000.000000000000000)
Pixel Size = (100.000000000000000,-100.000000000000000)
Metadata:
  AREA_OR_POINT=Area
  DESCRIPTION=elevation values from the SRTM DEM
  DOWNLOAD_URL=http://geomorphometry.org/content/ebergotzen
Image Structure Metadata:
  COMPRESSION=DEFLATE
  INTERLEAVE=BAND
Corner Coordinates:
Upper Left  ( 3570000.000, 5718000.000) ( 10d  0'36.93"E, 51d35'36.25"N)
Lower Left  ( 3570000.000, 5708000.000) ( 10d  0'29.77"E, 51d30'12.69"N)
Upper Right ( 3580000.000, 5718000.000) ( 10d  9'16.39"E, 51d35'31.46"N)
Lower Right ( 3580000.000, 5708000.000) ( 10d  9' 8.20"E, 51d30' 7.92"N)
Center      ( 3575000.000, 5713000.000) ( 10d  4'52.82"E, 51d32'52.16"N)
Band 1 Block=100x20 Type=Float32, ColorInterp=Gray
```

Similarly, all necessary metadata can be added into GeoTiff so that future users have all information at one place i.e. inside the data file.

[17] HTTP://WWW.GDAL.ORG/GDAL_EDIT.HTML

9.3.2 EXPORTING DATA – FINAL TIPS

As we have shown previously, if you export soil data into either GPKG and/or GeoTiff, these data can be accessed using DB operations. In fact, by exporting the data to GPKG and GeoTiffs, you have created a soil spatial DB or a soil information system. This does not necessarily mean that its targeted users will be able to find all information without problems and/or questions.

How usable and how popular a data set is, is a function of many aspects, not only data quality. You could create maps of perfect quality, but have no users at all. Some things you should definitively consider, as a way to boost usability of your data are:

DEFINITION

- MAKE A LANDING PAGE FOR YOUR DATA THAT INCLUDES: (1) SIMPLE ACCESS/ DOWNLOAD INSTRUCTIONS, (2) SCREENSHOTS OF YOUR DATA IN ACTION (PEOPLE PREFER VISUAL EXPLANATIONS WITH EXAMPLES), (3) LINKS TO KEY DOCUMENTS EXPLAINING HOW THE DATA WAS PRODUCED, AND (4) WORKFLOWS EXPLAINING HOW TO REQUEST SUPPORT (WHO TO CONTACT AND HOW).

- MAKE DATA ACCESSIBLE FROM MULTIPLE SYSTEMS E.G. BOTH VIA WCS, FTP AND THROUGH A MIRROR SITE. THIS MIGHT BE INEFFICIENT CONSIDERING THERE WILL BE MULTIPLE COPIES OF THE SAME DATA, BUT SOMETIMES IT QUADRUPLES DATA USAGE.

- EXPLAIN THE DATA FORMATS USED TO SHARE DATA, AND POINT TO TUTORIALS THAT EXPLAIN HOW TO ACCESS AND USE DATA TO BOTH BEGINNERS AND ADVANCED USERS.

- CONSIDER INSTALLING AND USING A VERSION CONTROL SYSTEM (OR SIMPLY USE GITHUB OR SIMILAR REPOSITORY) SO THAT THE USERS CAN TRACK BACK VERSIONS OF DATA.

- CONSIDER CLOSELY FOLLOWING PRINCIPLES OF REPRODUCIBLE RESEARCH[18] (ALL PROCESSING STEPS, INPUTS AND OUTPUTS ACCESSIBLE). THIS TUTORIAL COMES WITH R CODE[19] THAT IS AVAILABLE VIA GITHUB SO THAT EVERYONE SHOULD BE ABLE TO REPRODUCE THE EXAMPLES SHOWN IN THE TEXT.

[18] HTTPS://ROPENSCI.ORG/BLOG/2014/06/09/REPRODUCIBILITY/

[19] HTTPS://GITHUB.COM/ISRICWORLDSOIL/GSIF_TUTORIALS/BLOB/MASTER/SOILDATA/DATA_FORMATS_SOILDATA.R

9.4 References

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). Classes for Spatial Data in R. In Applied Spatial Data Analysis with R (pp. 21-57). Springer New York.

Hengl, T., Roudier, P., Beaudette, D., & Pebesma, E. (2015). plotKML: scientific visualization of spatio-temporal data. Journal of Statistical Software, 63(5), 1-25.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ... & Guevara, M. A. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PloS one, 12(2), e0169748.

Keller, A., Della Peruta, R., Schaepman, M., Gomez, M., Mann, S., & Schulin, R. (2014, May). An integrated Modelling framework to monitor and predict trends of agricultural management (iMSoil). In EGU General Assembly Conference Abstracts (Vol. 16, p. 6854).

Mitchell, T. and GDAL Developers (2014). Geospatial Power Tools: GDAL Raster & Vector Commands. Locate Press LLC, 346 p.

Neteler, M., & Mitasova, H. (2013). Open source GIS: a GRASS GIS approach (Vol. 689). Springer Science & Business Media.

Warmerdam, F. (2008). The geospatial data abstraction library. In Open source approaches in spatial data handling (pp. 87-104). Springer Berlin Heidelberg.

Youngblood, B. (2013). GeoServer Beginner's Guide. Packt Publishing Ltd.

9.5 EXPORT FORMATS

The produced results need to be exported in formats that can be easily read by a variety of GIS software. Two widely used formats for raster data are GeoTIFF and KML.

GeoTIFF is a public domain metadata standard which allows georeferencing information to be embedded within a TIFF file. The potential additional information includes map projection, coordinate systems, ellipsoids, datums, and everything else necessary to establish the exact spatial reference for the file. Keyhole Markup Language (KML) is an XML notation for expressing geographic annotation and visualization within Internet-based, two-dimensional maps and three-dimensional Earth browsers. KML became an international standard of the Open Geospatial Consortium in 2008.

Raster data in GeoTIFF format need a defined geographic projection. Each country has its own national system (or systems). In order to construct a mosaic of national datasets a common projection has to be defined and national data need to be re-projected in the common projection. Data projections can be managed using open source tools such as GIS softwares, GDAL tools suite (*http://www.gdal.org*) and various packages of the R software (*https://www.r-project.org*). Projections can be defined according to different standards. A common way is the EPSG database. EPSG Geodetic Parameter Dataset is a collection of definitions of coordinate reference systems and coordinate transformations which may be global, regional, national or local in application. A numeric code is assigned to each of the most common projections, making easier to refer to them and to switch between them. One of the most common global projections is WGS84 (EPSG 4336), used for maps and by the GPS satellite navigation system.

Each file should be accompanied by a set list of metadata. Geospatial metadata is a type of metadata that is applicable to objects that have an explicit or implicit geographic extent. While using GeoTIFF files it is possible to define a metadata field in which information can be recorded. Metadata can be edited with most GIS software and directly with GDAL tools (*http://www.gdal.org/gdal_edit.html*).

GDAL: gdalwarp -t_srs 'xxx' input.tif output.tif where t_srs is the target spatial reference set, i.e. the coordinate systems that can be passed are anything supported by the GRSpatialReference.SetFromUserInput() call, which includes EPSG PCS and GCSes (i.e. EPSG:4326), PROJ.4 declarations or the name of a .prj file containing well known text. For further information see http://www.gdal.org/gdalwarp.html

9.6 References

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), http://cran.r-project.org/doc/Rnews

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30: 683-691.

Robert J. Hijmans (2016). raster: Geographic Data Analysis and Modelling. R package version 2.5-8.

Julien Moeys (2016). soiltexture: Functions for Soil Texture Plot, Classification and Transformation. R package version 1.4.1. https://CRAN.R-project.org/package=soiltexture

Beaudette, D.E., Roudier P., and A.T. O'Geen. Algorithms for Quantitative Pedology: A Toolkit for Soil Scientists. 2012

Tomislav Hengl (2016). GSIF: Global Soil Information Facilities. R package version 0.5-3. https://CRAN.R-project.org/package=GSIF

Tomislav Hengl, Pierre Roudier, Dylan Beaudette, Edzer Pebesma (2015). plotKML: Scientific Visualization of Spatio-Temporal Data. Journal of Statistical Software, 63(5), 1-25. URL http://www.jstatsoft.org/v63/i05

10. TECHNICAL OVERVIEW AND THE CHECKLIST

STEP

10.1 POINT DATASET

DID YOU REMOVE NON-GEOREFERENCED OBSERVATIONS FROM YOUR DATASET?

DID YOU CHECK FOR OUTLIERS OR ANY UNUSUAL VALUES FOR THE MEASURED SOC, PH, BD, STONINESS/GRAVEL CONTENT, AND MIN/MAX DEFINITIONS OF YOUR SOIL HORIZONS?

IS THERE SPATIAL CORRELATION IN YOUR SOC VALUES, AS OBSERVED FROM THE VARIOGRAM?

DID YOU CHECK THE PROBABILITY DISTRIBUTION AND APPLIED A TRANSFORMATION IN CASE THE SAMPLES WERE NOT NORMALLY DISTRIBUTED?

BE AWARE WHETHER YOU ARE GOING TO PREDICT SOC OR SOM VALUES!

STEP

10.2 COVARIATES

DID YOU CHOOSE AND APPLY THE PROPER PROJECTION, ONE THAT IS SUITABLE FOR YOUR COUNTRY AND IS SUITABLE FOR SPATIAL STATISTICS?

DO ALL THE COVARIATES HAVE A RESOLUTION OF 1 KM, AND DID YOU USE EITHER NEAREST NEIGHBOUR RESAMPLING FOR THE CATEGORICAL VARIABLES AND IDW/CUBIC SPLINE FOR CONTINUOUS DATA?

DID YOU CORRECTLY SET THE NODATA VALUES AS NODATA, I.E. NOT A STANDARD ASSIGNED VALUES SUCH AS -9999, 256 ETC.

DID YOU CHECK FOR OUTLIERS OR ANY UNUSUAL VALUES, ESPECIALLY IN YOUR DEM LAYER(S)?

DID YOU SET ANY CATEGORICAL DATASET AS 'FACTOR' INSTEAD OF BEING 'NUMERIC' OR 'INTEGER'?

10.3 STATISTICAL INFERENCE

DID YOU CHOOSE A PROPER MODEL WHICH IS CAPABLE TO MODEL THE VARIABILITY IN YOUR SOC POINT DATA BEST? (MULTIPLE REGRESSION OR DATA MINING WITH OR WITHOUT INTERPOLATION OF THE RESIDUALS USING KRIGING)?

DID YOU MAKE SURE THAT THE RANDOM FOREST DID NOT OVER FIT YOUR DATA?

DID YOU APPLY A VALIDATION SCHEME, E.G. K-FOLD CROSS-VALIDATION OR AN INDEPENDENT VALIDATION, IF SO REPORT THE R2 AND RMSE AS ACCURACY MEASURES

DO THE MODEL SUMMARIES MAKE SENSE? I.E. MOST IMPORTANT PREDICTOR VARIABLES AND MODEL FIT?

IS THERE SPATIAL STRUCTURE LEFT IN YOUR RESIDUALS, IF SO MAKE SURE YOU INTERPOLATE THE MODEL RESIDUALS USING KRIGING?

10.4 SPATIAL INTERPOLATION

DID YOU OBTAIN AN EXHAUSTIVE MAP OR ARE THERE STILL GAPS?
IF SO, CHECK IF YOUR RASTER HAS THE CORRECT 'FACTOR' VALUES.

DO THE PATTERNS MAKE SENSE OR IS THERE A COVARIATE THAT CAUSES AN UNREALISTIC PATTERN, BASED ON EXPERT JUDGEMENT. IF SO, CONSIDER REMOVING THIS COVARIATE?

IN CASE YOU DID KRIGING, DON'T FORGET TO LOOK AT THE KRIGING VARIANCE. THIS IS A VERY IMPORTANT INDICATOR OF THE ACCURACY. OTHERWISE, CONSIDER MODELLING THE 90% CONFIDENCE INTERVALS OF THE PREDICTIONS!

DON'T FORGET TO BACK-TRANSFORM YOUR PREDICTED VALUES!

10.6 CALCULATION OF STOCKS

YOU MIGHT WANT TO CALCULATE STOCKS PER LU TYPE, MANAGEMENT TYPE OR BY MUNICIPALITY. IF YOU DO THAT MAKE SURE YOU GIVE AN INFORMED NUMBER, I.E. AN ESTIMATE PLUS AN ESTIMATE OF THE UNCERTAINTY

10.7 EVALUATION OF OUTPUT/ QUALITY ASSESSMENT

REPORT THE MODEL CALIBRATION AND VALIDATION STATISTICS!

REPORT SOME MAP QUALITY MEASURES!

EVALUATE TO WHICH EXTENT THE MODEL AND MAP IS EITHER UNDERESTIMATING OR OVERESTIMATING SOC/SOM!

DESCRIBE THE SPATIAL PATTERNS AND RELATE THEM TO THE LANDSCAPE CHARACTERISTICS!

11. DELIVERABLES

Data shared by countries will be collected by the GSP Secretariat. The GSP data policy (see GSP GSOC Guidelines, Chapter 8.5) will ensure that the national terms of condition are fully respected. Data will be shared using common GIS formats, and metadata should be compiled in an excel file. The Soil Organic Carbon Map will be delivered as grid using the 30 arc-seconds grid.

The following data will be delivered;

- **The Soil Organic Carbon Map:** will be digitally delivered in grid with 30 arc-seconds resolution. An empty grid is provided by the GSP Secretariat (*ftp://gsp.isric2.org/&lt;countryname&gt;/mask/mask.tif*). The SOC values should be transferred to this empty grid from the produced SOC data. To transfer values to the empty grid, QGis (Raster Calculator), ArcGIS (Mosaicing tool, Raster Calculator) can be used.

- **Uncertainty Layers and prediction quality figures:** a) Qualitative assessment (Conventional Upscaling) b) Quantitative assessment for DSM Methods. The uncertainty associated to the estimated map quality measures will be provided (Mean Error (ME), Mean Absolute Error (MAE), and MSE (Mean Squared Error).

- **Metadata** : Metadata is data describing data sets. It provides standardized information about a data set, for example, the maintaining institution through which the data can be accessed. It thus helps a user to find spatial data sets and services and to indicate for which purpose it can be used.

In the case of the GSOC maps, a project-specific metadata template has been prepared. It deviates from the metadata elements listed in common standards such as ISO 19115. Following the guideline for SOC mapping[18], the importance of soil-specific methodical elements is given. This includes metadata describing the data sources used for SOC mapping (Annex I -Table A) and the upscaling method (Annex I - Table B).

[18] PILLAR 4 WORKING GROUP (2017). GSP GUIDELINES FOR SHARING NATIONAL DA-TA/INFORMATION TO COMPILE A GLOBAL SOIL ORGANIC CARBON (GSOC) MAP. VERSION 1. FAO, ROME, 2017.

ANNEX I

Table A Documentation of soil carbon measurements (soil profile/auger data) and pre-processing (SOC stock assessment)

Metadata elements soil carbon		Description/examples
Soil sampling	Type of sampling	soil profile or auger
	Programme for data collection	soil mapping (provide scale if scale-specific), soil monitoring (any repetitions), other (<specify>)
	Sampling period	e.g. 1960-1975
	Total number of soil profiles (auger locations)	*<number>* Provide map of sample locations (e.g. soil profiles)
	Georeferencing	e.g. coordinates in ETR89
	Depth	Depth classes, e.g. 0-5 cm, 5-10 cm, etc. Soil horizons, cite national soil mapping/sampling guide
	Distribution of locations	random, systematic (e.g. transect, catena, toposequence), land use
Soil (organic) carbon	Measuring unit	e.g. g/kg
if estimated:	estimation method (e.g. estimated using a soil color chart, calibrated for SOC classes)	provide method, tables (codes and classes), example
if analyzed	SOC analysis method: Loss ignition (LoI), Wet Oxidation (wet ox), Dry combustion (DC)	- provide name of the method - sample preparation (air dried, oven dried, grounded, sieved), - for LoI, DC: temperature - for wet ex: agents, concentrations, recovery factor

Soil inorganic carbon	analysis method	
	measuring unit	
Bulk density	Measuring/estimation unit	e.g. g/cm^3
if estimated:	Pedotransfer function, literature default values	provide method details
if analysed		provide details about the sampling: size, number, location of cylinders, stones in the cylinders were accounted for yes/ no
Coarse fragments	Measuring/estimation unit	e.g. % volume
Organic layers	Sampling and description method	provide details
Peat	Sampling and description method	provide details

Table B Documentation of the national soil carbon stock map (upscaling, spatial input layers)

Metadata elements upscaling		Description/examples

Contact information for this resource	Name(s)
	Institute(s)
	Phone(s)
	E-mail(s)	

Keywords		
Upscaling method	conventional upscaling, regression kriging, random forest, other	*<specify>*
Representation type	Vector/Raster	

Update frequency	In Years/Months etc or Not Planned	
Scale/Resolution	Metre, Kilometre, Arc-Seconds	- -
Coordinate reference system		
Citation		

This cookbook was prepared thanks to the financial support of
The Kingdom of the Netherlands and the Swiss Confederation

Ministry of Foreign Affairs of the Netherlands

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation