

Progress in Soil Science

Alexandre M.J.-C. Wadoux  
Brendan Malone · Budiman Minasny  
Mario Fajardo · Alex B. McBratney

# Soil Spectral Inference with R

Analysing Digital Soil Spectra using  
the R Programming Environment



Springer

# **Progress in Soil Science**

*Series Editors:*

Alfred E. Hartemink, *Soil Science, University of Wisconsin, Madison, WI, USA*

Alex. B. McBratney, *Sydney Institute of Agriculture School of Life and  
Environmental Sciences, The University of Sydney, Sydney, NSW, Australia*

## **Aims and Scope**

Progress in Soil Science series aims to publish books that contain novel approaches in soil science in its broadest sense – books should focus on true progress in a particular area of the soil science discipline. The scope of the series is to publish books that enhance the understanding of the functioning and diversity of soils in all parts of the globe. The series includes multidisciplinary approaches to soil studies and welcomes contributions of all soil science subdisciplines such as: soil genesis, geography and classification, soil chemistry, soil physics, soil biology, soil mineralogy, soil fertility and plant nutrition, soil and water conservation, pedometrics, digital soil mapping, proximal soil sensing, soils and land use change, global soil change, natural resources and the environment.

More information about this series at <http://www.springer.com/series/8746>

Alexandre M.J.-C. Wadoux  
Brendan Malone • Budiman Minasny  
Mario Fajardo • Alex B. McBratney

# Soil Spectral Inference with R

Analysing Digital Soil Spectra  
using the R Programming Environment

Alexandre M.J.-C. Wadoux  
The University of Sydney  
Sydney, NSW, Australia

Brendan Malone  
CSIRO  
Canberra, ACT, Australia

Budiman Minasny  
The University of Sydney  
Sydney, NSW, Australia

Mario Fajardo  
The University of Sydney  
Sydney, NSW, Australia

Alex B. McBratney  
The University of Sydney  
Sydney, NSW, Australia

ISSN 2352-4774

ISSN 2352-4782 (electronic)

Progress in Soil Science

ISBN 978-3-030-64895-4

ISBN 978-3-030-64896-1 (eBook)

<https://doi.org/10.1007/978-3-030-64896-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Foreword

The soil, a key factor in agricultural production and in particular in agro-ecology, is a phenomenon that is still largely under-measured. It is essential to develop methods to know the soil better, understand how it works, assess its potential, and estimate its state in order to guide corrective actions or estimate the ecosystem services it provides (carbon storage, geochemical cycle closing, etc.). These data will be indispensable in the development of a 'digital agro-ecology', which is largely based on understanding the interactions between all factors of production – soil, climate, species, varieties and their interactions (the basis of agro-ecology!), and other inputs – in order to optimize them. These soil analysis methods must be quantitative, fast, easy to implement, precise, and, if possible, realized with devices that can be easily transported by operators, on-ground machines, or drones for in-field measurements with minimum disturbance and handling of soil samples. Spectral methods meet several of these criteria and can therefore revolutionize the quantitative characterization of soils, for the development and realization of 'digital agro-ecology'. Their main advantage is their ease of implementation with, for some of them, minimal sample preparation. Moreover, while the basic principle is the same for all of these methods (i.e. based on the interaction of electromagnetic radiation with the object of analysis – here the soil), the wide electromagnetic range (UV, visible, NIR, MIR, LIBS, Raman, XRF, etc.) produces various types of digital spectra, which give access to a range of soil properties. Their main drawback is that they are indirect methods which need a calibration – that is, a model – to be constructed between the spectrum and the soil property of interest. Conventionally, these calibrations are built by 'black-box' software, based on algorithms which are generally not known or understood and which are limited in terms of adjustment capacities. Initiatives are emerging to help researchers to better master chemometric techniques in order to build processing pipelines in which they control everything: for example, here in Montpellier, INRAE has launched such an initiative, Chemhouse, led by Jean-Michel Roger. No doubt, there are similar initiatives around the world. Indeed, this book, which focuses on quantitative chemometric methods, is another eagerly awaited response to this

need for knowledge-guided spectral analyses to make the most of digital spectra and improve the quality and usefulness of information from soil spectroscopy.

**Véronique Bellon-Maurel**

**Director of #DigitAg, the Digital Agriculture Convergence Laboratory, Montpellier, France**

**Deputy head of the Mathnum Department, INRAE (mathematics, computer and data sciences, digital technologies)**

# Preface

Digital spectroscopy is one of the new tools of the state-of-the-art soil scientist. Properly processed spectral data satiate the demand for cheap and accurate soil information required for precision agriculture and food production, earth system modelling, climate change mitigation, and general soil process parametrization. Our understanding of soil spectroscopy has advanced rapidly in the last two decades. The technological developments of cheaper and more accurate sensors coupled with the advent of new numerical tools have contributed to this significant improvement.

The focus of the book is on the techniques of using spectral data for characterizing soil. Spectral data may come from different sensors and wavelengths,  $\gamma$  rays, X rays, and infrared, among others, and from scans made either close to the soil material – in the field or laboratory, or remotely, for example, when the sensor is mounted on a plane or satellite. Most of the examples here are based on the infrared part of the spectrum, largely because of its demonstrated utility for soil science. We present explanation and code in a didactic way that can handle all kinds of spectral data however, in the hope that this book will contribute to the development of common procedures for soil spectral analysis and data sharing whatever the wavelength range. We also hope that this book can be used for developing training courses and capacity building.

Sydney, NSW, Australia  
July 2020

Alexandre M.J-C. Wadoux  
Brendan Malone  
Budiman Minasny  
Mario Fajardo  
Alex B. McBratney



# Acknowledgements

We thank those who have contributed, directly or indirectly, to the development of the materials presented in this book. Leonardo Ramirez-Lopez (BUCHI Labortechnik, Zurich) is pretty much solely responsible for building and maintaining packages for spectral similarity and modelling of complex spectral datasets using local calibration algorithms. Colleagues at the University of Sydney, especially Edward Jones, have given feedback and helped in the development of materials over the years. Finally, we thank also the students and the workshop participants for continuous feedbacks and questions.

# Endorsements

*Soil Spectral Inference with R* offers an introduction and hands-on practical approach on soil spectroscopy for anyone who wants to extend their understanding and capabilities in soil spectroscopy. Since the basics of soil spectral analysis but also new developments are addressed, the book is suitable for beginners and more experienced scientists, either as study material or as a basis for capacity building.

As such, it addresses an important step in the entire soil spectroscopy workflow that stretches from sampling, wet and dry chemistry measurements, quality assessment and control, spectral library development, calibration transfer, spectral data analysis, and use of the resulting soil data for monitoring or mapping of soil properties and functions to data serving. This entire workflow (and best practice guidelines for its parts) is currently addressed in two international initiatives. The initiative to build a Global Soil Spectral Calibration Library and Estimation Service by the Global Soil Laboratory Network ([GLOSOLAN](#)), for now mainly focused on mid-infrared lab analysis and prediction including capacity building for both, and the IEEE initiative [P4005](#) Standard Protocol and Scheme for Measuring Soil Spectroscopy, focusing on near-infrared lab analysis. The GLOSOLAN initiative aims to provide a standard global dataset and easy-to-use tools for labs and scientists for spectral data processing. This book can foster understanding of the spectral analysis used and as such can be instrumental to a proper use of the tool or service and help in correct interpretation of the results.

**Fenny van Egmond, ISRIC – World Soil Information, Wageningen – the Netherlands, co-lead of the GLOSOLAN initiative for a Global Soil Spectral Calibration Library and Estimation Service**

*Soil Spectral Inference with R* provides a step-by-step description on how soil spectroscopic data can be modelled in the R modelling environment. Starting with a detailed description on soil spectroscopy, this book provides a comprehensive set of tools and techniques used in diffuse reflectance spectroscopy approach for assessing soils. Theoretical concepts and equations are presented in each chapter along with relevant R codes and sample outputs both in the form of data and figures. The references used in this book are up to date and the parts on (a) noise

removal, (b) different similarity measures, (c) subsampling approaches, and (d) spectral transformations are going to help students and soil professionals explore new ways and means of analyzing spectral data. With the graphical illustrations of results from almost every segment of example R codes, this book retains its visual presentation style and is expected to serve as a perfect sequel to the earlier book *Using R for Digital Soil Mapping* published from the same group. I highly recommend this book to my students and researchers who are exploring the use of diffuse reflectance spectroscopy for soil analysis.

**Bhabani S. Das, Agricultural and Food Engineering Department, Indian Institute of Technology Kharagpur-India**

This timely and opportune book will help readers to deal with the increasingly complex kaleidoscope of tools and lines of codes applied to soil spectroscopy. It touches aspects of conversion from wavelength to wavenumber up to the sharing of soil spectral libraries and the need for spectral standardization between laboratories. In this way, the text comes in handy for neophytes and professionals alike. Within its concise contents, the book covers a getting started with R, a list of useful spectroscopy packages, and data handling parts. Yet, it also covers topics from pre-processing to exploratory soil spectral analysis, which will be of interest to competent users, leading towards more proficient levels. Model calibration and estimating soil properties are also implemented and will allow readers to develop their skills up to the expert level with full application of soil spectroscopy. Besides ready-to-use lines of code, datasets with soil spectra and laboratory analytical data were also made available through computer coding. And authors have also compiled all data and functions, used in the book, in a single R package called *soilspec*, available in the open source software development environment and social network GitHub. Therefore, it's my pleasure to endorse the book *Soil Spectral Inference with R*, which embodies a complete guidance for lecturing and learning soil spectral inference using the statistical computing environment R.

**Alexandre ten Caten, Department of Agriculture, Biodiversity and Forests, Federal University of Santa Catarina-Brazil**

*Soil Spectral Inference with R* can help us to build an integrated application of a soil spectral inference system from scratch using the R platform. This book elaborates on the whole process of soil spectral inference with detailed and practical R codes, from importing and pre-processing of spectra to model calibration and validation. There are many valuable routines in this book, especially for the vis-NIR spectra, such as bagging PLSR for calibration and EPO for the removal of moisture effect. These methods can be examined via the use of example datasets and readily transferred to real-life applications. I, thus, highly recommend this book to anyone who is engaged in the exploration and application of soil spectra.

**Changkun Wang, Institute of Soil Science, Chinese Academy of Sciences, Nanjing-China**

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	Spectroscopy in Soil Science	2
1.2	Populating a Soil Database	5
1.3	Objectives of This Book	7
	References	7
<b>2</b>	<b>Getting Started with R</b>	11
2.1	Use of R and RStudio	11
2.2	Simple Manipulations	14
2.3	Data Structure	16
2.4	Programming Tools	19
2.5	Plotting	21
2.6	Documentation and Help	24
	References	25
<b>3</b>	<b>Materials</b>	27
3.1	Datasets	27
3.2	R Packages	29
3.3	The soilspec Book Package	34
	References	36
<b>4</b>	<b>Data Handling of Spectra</b>	37
4.1	Importing Data	37
4.2	Loading ASD Data	39
4.3	Plotting the Spectra	41
4.4	Averaging the Replicates	43
4.5	Converting Units of Measurement	45
4.6	Exporting the Spectra	47
	References	48
<b>5</b>	<b>Pre-processing of Spectra</b>	49
5.1	Noise Removal	52
5.2	Scatter Correction	58

5.3	Derivatives .....	64
5.4	Centring and Standardizing .....	66
5.5	Spectral or Dimension Reduction .....	67
5.6	Other Specific Transformations.....	73
	References.....	78
<b>6</b>	<b>Exploratory Soil Spectral Analysis .....</b>	<b>81</b>
6.1	Feature Selection .....	82
6.2	Principal Component Analysis .....	97
6.3	Spectral Prediction Domain.....	104
6.4	Soil Colour .....	108
	References.....	113
<b>7</b>	<b>Similarity Between Spectra and the Detection of Outliers .....</b>	<b>115</b>
7.1	Similarity/Dissimilarity Measures .....	117
7.2	Detecting Outlier Spectra .....	128
	References.....	140
<b>8</b>	<b>Selection of the Samples for Laboratory Analysis .....</b>	<b>143</b>
8.1	Sampling Design .....	145
8.2	Sample Size.....	155
	References.....	164
<b>9</b>	<b>Estimating Soil Properties and Classes from Spectra .....</b>	<b>165</b>
9.1	Goodness of Fit Measures .....	166
9.2	Models for Quantitative Variables .....	174
9.3	Models for Categorical Variables.....	204
9.4	Soil Spectral Inference Systems .....	211
	References.....	213
<b>10</b>	<b>Spectral Transfer and Transformation.....</b>	<b>215</b>
10.1	Spectral Transfer Between Instruments Using a Standard Sample .....	216
10.2	Direct Standardization .....	225
10.3	Piecewise Direct Standardization.....	229
10.4	Removing External Effects, such as Soil Moisture (EPO) .....	232
	References.....	246

## About the Authors



**Alexandre Wadoux** is research associate in soil science at the University of Sydney and member of the Sydney Institute of Agriculture, Australia.

**Brendan Malone** is a senior research scientist in soil science at CSIRO Canberra, Australia.

**Budiman Minasny** is professor of soil-landscape modelling at the University of Sydney, Australia.

**Mario Fajardo** is a postdoctoral research fellow at the Precision Agriculture Laboratory, University of Sydney, Australia.

**Alex McBratney** is professor of digital agriculture and soil science at the University of Sydney and director of the Sydney Institute of Agriculture, Australia.

# Chapter 1

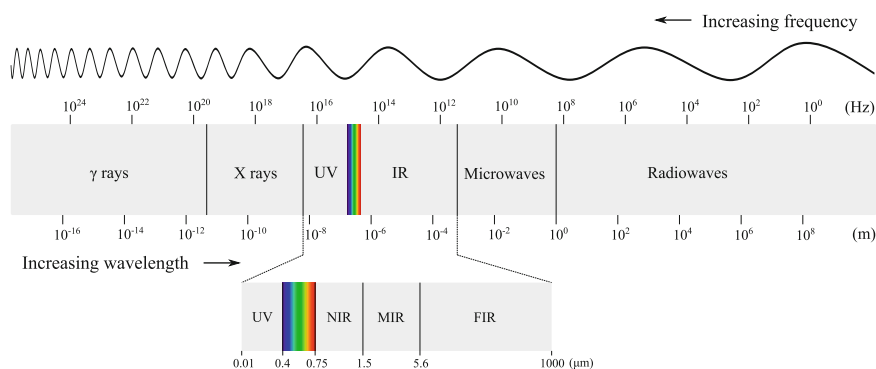
## Introduction



Soil provides a multitude of key ecosystem services such as food production, climate change adaptation, nutrient and water cycling and carbon sequestration (Dominati et al. 2010). Ongoing global environmental change has put unprecedented pressure on soil, resulting in significant and widespread degradation and erosion. The soil science community is tasked to deliver timely, nuanced and high-quality thematic soil data and knowledge to assess and monitor soil change (Sanchez et al. 2009). This is reflected by recent initiatives to provide soil information to populate regional, national and worldwide soil databases (Grunwald et al. 2011). Soil data are conventionally acquired through soil surveys coupled with laboratory analyses. The methods to obtain soil information are often impractical because they are expensive, require time-consuming field campaigns and use chemical reagents for soil analysis (McCarty et al. 2002; Brown et al. 2006; Ben-Dor et al. 2009; Stenberg et al. 2010). The use of sensors for characterizing chemical, mineralogical, biological and physical properties of the soil has thus gained lots of traction in soil research. Advances in sensors and software are occurring at a rapid pace. Soil sensing, in particular the use of soil spectroscopy, is now widely available using a range of modalities and wavelengths across the electromagnetic spectrum.

Soil spectroscopy can characterize soil properties efficiently. Soil spectroscopy can be simply defined as the study of the spectral signature of a soil material (Nocita et al. 2015). The spectral signature relates to soil characteristics such as organic and mineral components. Spectroscopic measurements are fast, cost-effective and non-destructive and can be made both in the laboratory and *in situ* in the field. Soil composition and characteristics are encoded in the spectrum at specific wavelengths of the electromagnetic spectrum. For example, mid-infrared spectra have encoded information on soil mineralogy or soil organic matter composition, which can be assessed quantitatively or qualitatively using the absorption or reflectance at specific wavelengths (Viscarra-Rossel et al. 2016).

Figure 1.1 summarizes the electromagnetic spectrum over an extensive range of wavelengths and frequencies. The electromagnetic spectrum ranges from the  $\gamma$  rays to radiowaves. The visible (to the human) portion of the electromagnetic spectrum is between 0.4 and 0.75  $\mu\text{m}$ . Radiowaves have long wavelengths that can reach several hundreds of metres, while high-energy  $\gamma$  rays have wavelengths shorter than  $10^{-13}$  m. Each portion of the electromagnetic spectrum relates to specific soil properties or characteristics. For example, the visible part contains information on soil colour, while the  $\gamma$  rays, X rays and infrared spectra are useful to estimate soil properties, especially elemental composition and soil mineralogy.



**Fig. 1.1** Components of the electromagnetic spectrum. (After Lillesand et al. 2015)

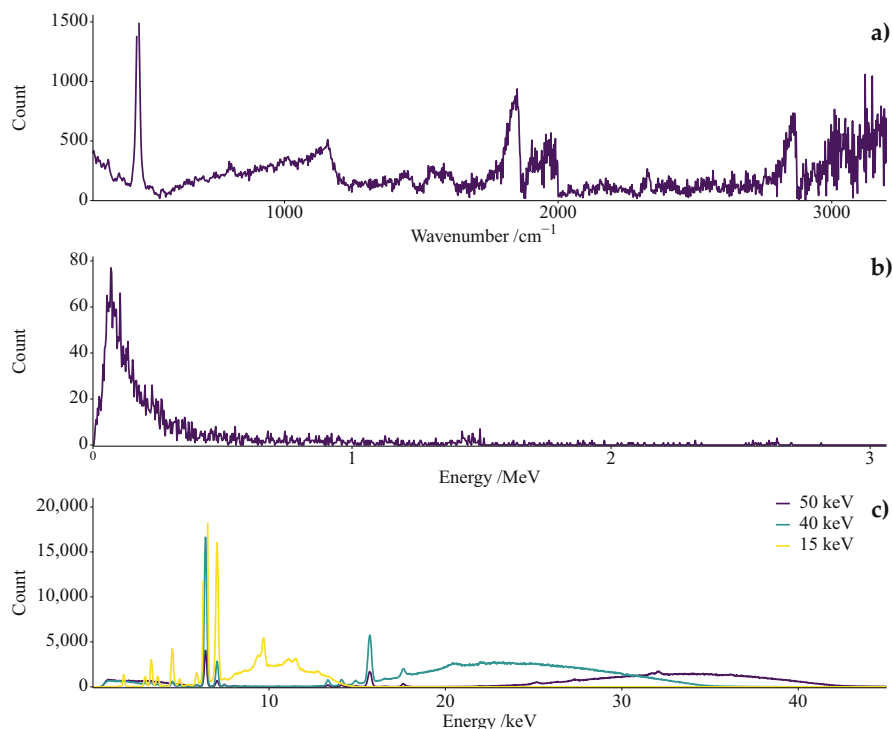
## 1.1 Spectroscopy in Soil Science

Interest in spectroscopy for soil started as early as the 1920s with studies analysing the mineral (Hendricks and Fry 1930) and later organic composition (e.g. Hunt et al. 1950 or Holmes and Toth 1957) of soils, but it was in the 1970s that scientists started to investigate the direct relationships between soil spectral information and soil properties. Condit (1970), for instance, developed a soil spectral library which quickly became a classical tool for soil scientists (Ben-Dor et al. 2009). The large use of spectral information in soil science was made possible with the advancement of computer and information technology. A big change came in the 1980s and 1990s when spectral instruments were transformed from analogue (chart-recording) to digital devices producing long bivariate data streams of wavelength (or its homologue) and intensity representing the analogue spectra. These data streams are digital spectra usually of length  $2^9$  to  $2^{12}$ .

Spectroscopy is currently used in a large number of applications: to characterize soil minerals (Viscarra-Rossel and Webster 2012), organic matter (Gerzabek et al. 2006; Ertlen et al. 2010), colour (Viscarra-Rossel and Webster 2011), but also texture (Minasny et al. 2008), iron oxides (Malengreau et al. 1996), carbonates (Grinand et al. 2012), salinity (Nawar et al. 2014) and soil quality indicators



(Cécillon et al. 2009). Soil colloids such as clay minerals are also detected by X ray ( $\sim 10\text{ nm}$ – $10\text{ pm}$ ) diffraction (Wilson and Cradwick 1972), by detection of peaks in the pure minerals and comparing them to those recorded on the soil sample. Soil aggregates can be characterized by visible, near and mid-infrared spectroscopy (Cañasveras et al. 2010; Askari et al. 2015). At a larger spatial scale, soil variation and diversity can be characterized by  $\gamma$  rays or microwaves such as radar remote sensing (Cook et al. 1996; Weihermüller et al. 2007).



**Fig. 1.2** Examples of Raman (a),  $\gamma$  rays (b) and X ray fluorescence (c) spectra of a soil sample from the Geeves (Geeves et al. 1994) dataset. The x-axis unit is in either wavenumber ( $\text{cm}^{-1}$ ), megaelectron volts (MeV) or kiloelectron volts (keV)

Figure 1.2 is an example of three spectra from Raman,  $\gamma$  ray and X ray fluorescence (XRF) spectroscopies. Raman vibrational spectroscopy is useful to characterize soil substances and requires minimal soil sample preparation.  $\gamma$  ray spectroscopy relates to soil mineralogical properties and geochemistry of the soil sample by measuring the natural emission of  $\gamma$  rays and anthropogenic radionuclides, e.g. Caesium-137. An XRF spectrum relates to soil elemental composition.

The visible and infrared range of the electromagnetic spectrum has garnered much interest in soil science. The measurement of the infrared spectrum of soil samples enables the quantification of several soil properties from their spectral response in a faster and cheaper way than by conventional methods of soil analyses

(Stenberg et al. 2010; Bellon-Maurel and McBratney 2011). In addition, recording an infrared spectrum does not make use of any chemical reagents and can be done both in the laboratory or for in-field soil analysis (Ramirez-Lopez et al. 2019). Infrared spectra are sensitive to both organic and inorganic soil materials, making them an excellent tool for quantitative soil assessment. The mid-infrared (MIR) range of the spectrum, in particular, contains more information and direct information on soil organic and mineral components of the soil than the visible and near-infrared (vis-NIR) range. For example, various components of the soil organic matter have very distinct spectral signature in the mid-infrared range. The reason is that the fundamental molecular vibrations occur in the mid-infrared range, while the overtones and combinations occur in the vis-NIR (McCarty et al. 2002). In practice, this means that the absorption features detected in the vis-NIR are fewer, broader and more complex than those recorded in the mid-infrared (Islam et al. 2003).

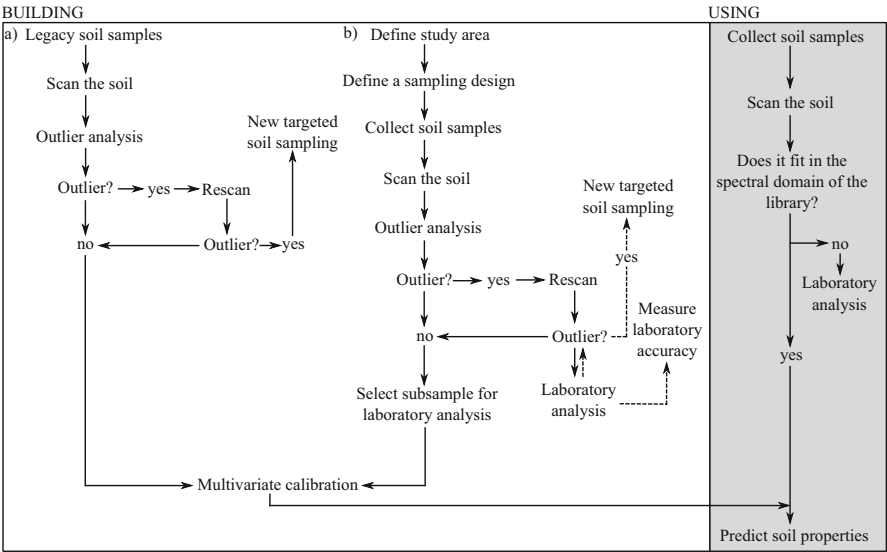
While spectroscopy has been used in soil science since the 1950s, the last two decades have seen an increase in its use, in particular vis-NIR and MIR spectroscopy, to replace and complement soil analyses. This increase was supported by the development of chemometrics (the application of mathematical and statistical methods to the analysis of chemical data (Varmuza and Filzmoser 2016)), multivariate statistical analysis and the increase in computer resources. Soil properties have complex absorption patterns. Infrared spectral bands are largely non-specific (i.e. they are not linearly related to a single soil property) and overlap between properties (Ben-Dor and Banin 1995). This is particularly significant in the vis-NIR range of the spectra (Soriano-Disla et al. 2014). To extract these complex patterns and obtain quantitative estimates of a soil property, soil scientists have used mathematical transfer functions to correlate spectral wavelengths to soil properties (Viscarra-Rossel et al. 2008). The transfer function is calibrated using the spectral wavelengths as independent variables and the laboratory measured values of the soil properties as the dependent variable. Once calibrated on the spectra, the soil property can be predicted using the spectral information only.

Relatively simple statistical models can be built to transfer the spectra to soil information. Early studies on soil spectroscopy used linear regression models on specific wavelengths. For example, Dalal and Henry (1986) fitted a linear model on three user-defined wavelengths to predict soil moisture, organic carbon and total nitrogen. The large number of wavelengths to consider and the correlation between them made the use of linear models complicated. Techniques for variable selection, such as stepwise variable selection or dimension reduction such as principal component regression, quickly emerged as valuable to handle the multivariate spectral data. A variant of principal component regression called partial least squares regression (PLSR, Abdi 2003) is now routinely used. PLSR relates the soil property values and the principal component scores (a dimension reduction analysis) of the spectra. The PLSR models can handle the full spectra as predictors (not only a few wavelengths) and are not sensitive to the correlation between wavelengths (Janik et al. 2007). In addition, they are substantially faster to calibrate than stepwise linear regression models. In the last two decades, other multivariate analysis techniques have been used, in particular machine learning algorithms. For example, Nawar and Mouazen (2019) used random forest to estimate soil organic carbon on soil samples collected

in six fields in the United Kingdom. Viscarra-Rossel and Behrens (2010) compared several linear and non-linear (machine learning) models to calibrate soil spectra on soil properties. Other methods to derive soil information from a spectrum are based on the discrimination on the soil spectral signature such as in absorption feature analysis (Clark and Roush 1984). This book provides implementation to derive soil information from using both multivariate statistical models and absorption feature analysis.

1.2 Populating a Soil Database

The opportunity to retrieve cheap soil information from a spectrum has resulted in the development of soil spectral libraries for the quantification of soil properties at local (Guerrero et al. 2016), regional (Gogé et al. 2012) or global (Viscarra-Rossel et al. 2016) scales. Nowadays, several institutions provide spectral libraries with spectra scanned on pure materials, for example, minerals, vegetation or rocks. The United States Geological Survey (USGS) spectral library version 7 (Kokaly et al. 2017) contains several thousands of spectra of different materials for the ultraviolet to the far infrared (0.2 to 200 microns [μm]). Other libraries are exclusive to soil samples, like the Land Use/Cover Area statistical Survey (LUCAS) compiled in Europe (Orgiazzi et al. 2018). By 2018, this soil spectral library had approximately 45,000 soil samples with spectra in the vis-NIR regions and soil attributes such as pH, organic carbon and cation exchange capacity, among others.



**Fig. 1.3** Simplified scheme for building a soil spectral library. (Adapted from Viscarra-Rossel and McBratney 2008). The steps describe how to build a spectral library using (a) legacy soil samples or (b) a new soil sampling. The scheme explains both the development and the use of the soil spectral library

To build a conventional soil spectral library, Viscarra-Rossel et al. (2008) defined three key requirements: 1. it should contain as many and as representative as possible soil samples to represent the soil spatial variability in the study area, 2. the soil sampling and scanning should be made with caution as all change in the soil sample and scanning procedure is embodied in the spectrum and 3. the laboratory measurement of the soil properties should be accurate. Figure 1.3 illustrates the steps to build a soil spectral library using either legacy soil samples (a) or a new soil sampling (b) and to both build and use the spectral library.

Using legacy soil samples, Fig. 1.3a shows that all soil samples are scanned and the spectra analysed for outliers. If a spectrum is considered as an outlier, the soil is re-scanned. When the re-scanned soil sample is still considered as an outlier, one should consider a new targeted soil sampling for this specific outlier soil sample. After the outlier detection, the spectra are correlated to the values of laboratory-analysed soil properties (e.g. soil organic carbon) using a multivariate statistical model. When, conversely, a spectral library is built using new soil sampling, the sampling design plays a key role. The soil samples are collected in the study area of interest, using a sampling design and for a given sample size. The sample size is decided with budget constraints. The soil samples are then scanned and analysed for outliers. When the spectra contain outliers, they can be re-scanned or analysed in the laboratory. The spectral dataset is subsampled to determine which soil samples are sent to the laboratory for conventional analyses. After this step, the spectra and values of the soil properties are correlated using a multivariate statistical model.

In both cases, when the multivariate statistical model is validated and has sufficiently high accuracy, it can be used for prediction on new soil sample spectra. When new soil samples are scanned and added to the library, they should belong to the same population as the soils in the library. If otherwise, it is likely that the calibrated multivariate models will be inefficient at predicting the soil properties of interest. This book provides implementation for all these steps.

To date, most soil spectral libraries have been built for conventional soil properties. They have been shown a useful means of organizing spectra for small farms or individual fields but also larger, regional areas or continents. Under the scheme shown in Fig. 1.3, it is in principle also possible to build a soil spectral library for properties previously unknown at the time of the soil sampling, provided that the properties are identifiable by spectroscopic techniques. For example, this is the case for microbial biomass carbon (i.e. the carbon contained within the living component of soil organic matter) (Mirzaeitalarposht and Kambouzia 2020), polluting chemicals (Paradelo et al. 2016) or microplastic in soil (Corradini et al. 2019) which are now dynamic research areas but were not considered until recently. This means that spectral libraries and the collection of soil spectra might have use in the future for purposes which we currently disregard.

When soil scientists build a model to link soil properties to the spectra, it is done by using software for statistical analysis. One of the claims made with the availability of spectroscopic measurement devices is the provision of easy-to-follow commercialized software and numerical implementation, which make complicated statistical treatment practicable. While these implementations have provided the

majority with tools to produce soil information, they are often used at the expense of a deeper understanding of the techniques required to treat a specific soil spectral library. Useful research has been made in developing functions and code to perform these tasks in open-source software, such as in R (R Core Team 2018). This book will provide a step towards the implementation of spectroscopic analysis techniques and their use in an open, accessible and comprehensible manner and with a view to improving these methods.

### 1.3 Objectives of This Book

This book is a step-by-step guide to processing soil spectra, particularly from the visible and infrared range of the electromagnetic spectrum. This book is fully reproducible and can serve as a basis for teaching soil spectroscopy to undergraduate students. The examples are implemented in the R programming language, for which the reader is expected to have some basic knowledge. All the data used in the examples, together with the R functions, are provided in a book-associated R package freely accessible online. Instructions to obtain and install the package are provided further on.

Specific topics covered in this book are:

- Importing and plotting spectra in R.
- Pre-processing the spectra.
- Using dimension reduction techniques to visualize the spectra.
- Obtaining soil information (mineralogy, colour) directly from the spectra.
- Outlier detection in the spectral space.
- Similarity measures between spectra.
- Sampling designs and determining the optimal number of soil samples for laboratory analysis.
- Multivariate calibration.
- Soil spectral inference systems.

We also included examples and code for additional (more specific) spectral treatments such as:

- Transferring spectra between instruments.
- Removing the effect of external factors affecting the spectrum, such as soil moisture.

## References

- Abdi H (2003) Partial least square regression (PLS regression). *Encycl Res Methods Soc Sci* 6:792–795
- Askari MS, Cui J, O'Rourke SM, Holden NM (2015) Evaluation of soil structural quality using VIS–NIR spectra. *Soil Tillage Res* 146:108–117

- Bellon-Maurel V, McBratney AB (2011) Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils—Critical review and research perspectives. *Soil Biol Biochem* 43:1398–1410
- Ben-Dor E, Banin A (1995) Near infrared analysis (NIRA) as a method to simultaneously evaluate spectral featureless constituents in soils. *Soil Sci* 159:259–270
- Ben-Dor E, Chabrilat S, Demattê JAM, Taylor GR, Hill J, Whiting ML, Sommer S (2009) Using imaging spectroscopy to study soil properties. *Remote Sens Environ* 113:S38–S55
- Brown DJ, Shepherd KD, Walsh MG, Mays MD, Reinsch TG (2006) Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* 132:273–290
- Cañasveras JC, Barrón V, Del Campillo MC, Torrent J, Gómez JA (2010) Estimation of aggregate stability indices in Mediterranean soils by diffuse reflectance spectroscopy. *Geoderma* 158:78–84
- Cécillon L, Barthès BG, Gomez C, Ertlen D, Génot V, Hedde M, Stevens A, Brun J-J (2009) Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). *Eur J Soil Sci* 60:770–784
- Clark RN, Roush TL (1984) Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *J Geophys Res Solid Earth* 89:6329–6340
- Condit HR (1970) The spectral reflectance of American soils. *Photogramm Eng* 36:955–966
- Cook SE, Comer RJ, Groves PR, Grealish GJ (1996) Use of airborne gamma radiometric data for soil mapping. *Soil Res* 34:183–194
- Corradini F, Bartholomeus H, Lwanga EH, Gertsen H, Geissen V (2019) Predicting soil microplastic concentration using vis-NIR spectroscopy. *Sci Total Environ* 650:922–932
- Dalal RC, Henry RJ (1986) Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. *Soil Sci Soc Am J* 50:120–123
- Dominati E, Patterson M, Mackay A (2010) A framework for classifying and quantifying the natural capital and ecosystem services of soils. *Ecol Econ* 69:1858–1868
- Ertlen D, Schwartz D, Trautmann M, Webster R, Brunet D (2010) Discriminating between organic matter in soil from grass and forest by near-infrared spectroscopy. *Eur J Soil Sci* 61:207–216
- Geeves GW, Cresswell HP, Murphy BW, Gessler PI, Chartres CJ, Little IP, Bowman GM (1994) Physical, chemical and morphological properties of soils in the wheat-belt of southern NSW and northern Victoria. NSW Department of Conservation; Land Management/CSIRO Division of Soils Occasional Report, CSIRO
- Gerzabek MH, Antil RS, Kögel-Knabner I, Knicker H, Kirchmann H, Haberhauer G (2006) How are soil use and management reflected by soil organic matter characteristics: a spectroscopic approach. *Eur J Soil Sci* 57:485–494
- Gogé F, Joffre R, Jolivet C, Ross I, Ranjard L (2012) Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemom Intell Lab Syst* 110:168–176
- Grinand C, Barthes BG, Brunet D, Kouakoua E, Arrouays D, Jolivet C, Caria G, Bernoux M (2012) Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). *Eur J Soil Sci* 63:141–151
- Grunwald S, Thompson JA, Boettinger JL (2011) Digital soil mapping and modeling at continental scales: finding solutions for global issues. *Soil Sci Soc Am J* 75:1201–1213
- Guerrero C, Wetterlind J, Stenberg B, Mouazen AM, Gabarrón-Galeote MA, Ruiz-Sinoga JD, Zornoza R, Viscarra-Rossel RA (2016) Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil Tillage Res* 155:501–509
- Hendricks SB, Fry WH (1930) The results of X-ray and microscopical examinations of soil colloids. *Soil Sci Soc Am J* 11:194–195
- Holmes RM, Toth SJ (1957) Physico-chemical behavior of clay-conditioner complexes. *Soil Sci* 84:479–488
- Hunt JM, Wisherd MP, Bonham LC (1950) Infrared absorption spectra of minerals and other inorganic compounds. *Anal Chem* 22:1478–1497
- Islam K, Singh B, McBratney AB (2003) Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Soil Res* 41:1101–1114

- Janik LJ, Skjemstad J, Shepherd K, Spouncer L (2007) The prediction of soil carbon fractions using mid-infrared-partial least square analysis. *Soil Res* 45:73–81
- Kokaly RF, Clark RN, Swayze GA, Livo KE, Hoefen TM, Pearson NC, Wise RA, Benzel WM, Lowers HA, Driscoll RL, others (2017) USGS spectral library version 7. US Geological Survey
- Lillesand T, Kiefer RW, Chipman J (2015) Remote sensing and image interpretation. Wiley, New York
- Malengreau N, Bedidi A, Muller J-P, Herbillon AJ (1996) Spectroscopic control of iron oxide dissolution in two ferrallitic soils. *Eur J Soil Sci* 47:13–20
- McCarty GW, Reeves JB, Reeves VB, Follett RF, Kimble JM (2002) Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Sci Soc Am J* 66:640–646
- Minasny B, McBratney AB, Tranter G, Murphy BW (2008) Using soil knowledge for the evaluation of mid-infrared diffuse reflectance spectroscopy for predicting soil physical and mechanical properties. *Eur J Soil Sci* 59:960–971
- Mirzaeitalarposht R, Kambouzia J (2020) Development of mid-infrared spectroscopic feature-based indices to quantify soil carbon fractions. *Eurasian Soil Sci* 53:73–81
- Nawar S, Buddenbaum H, Hill J, Kozak J (2014) Modeling and mapping of soil salinity with reflectance spectroscopy and landsat data using two quantitative methods (PLSR and MARS). *Remote Sens* 6:10813–10834
- Nawar S, Mouazen AM (2019) On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. *Soil Tillage Res* 190:120–127
- Nocita M, Stevens A, van Wesemael B, Aitkenhead M, Bachmann M, Barthès B, Dor EB, Brown DJ, Clairrotte M, Csorba A, others (2015) Soil spectroscopy: an alternative to wet chemistry for soil monitoring. In: *Advances in agronomy*. Elsevier, Burlington, pp 139–159
- Orgiazzi A, Ballabio C, Panagos P, Jones A, Fernandez-Ugalde O (2018) LUCAS soil, the largest expandable soil dataset for Europe: a review. *Eur J Soil Sci* 69:140–153
- Paradelo M, Hermansen C, Knadel M, Moldrup P, Greve MH, Jonge LW de (2016) Field-scale predictions of soil contaminant sorption using visible–near infrared spectroscopy. *J Near Infrared Spectrosc* 24:281–291
- Ramirez-Lopez L, Wadoux AMJ-C, Franceschini MHD, Terra FS, Marques KPP, Sayão VM, Demattê JAM (2019) Robust soil mapping at the farm scale with vis-NIR spectroscopy. *Eur J Soil Sci* 70:378–393
- R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Sanchez PA, Ahamed S, Carré F, Hartemink AE, Hempel J, Huising J, Lagacherie P, McBratney AB, McKenzie NJ, Lourdes Mendonça-Santos M de, others (2009) Digital soil map of the world. *Science* 325:680–681
- Soriano-Disla JM, Janik LJ, Viscarra-Rossel RA, Macdonald LM, McLaughlin MJ (2014) The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl Spectrosc Rev* 49:139–186
- Stenberg B, Viscarra-Rossel RA, Mouazen AM, Wetterlind J (2010) Visible and near infrared spectroscopy in soil science. In: *Advances in agronomy*. Elsevier, Burlington, pp 163–215
- Varmuza K, Filzmoser P (2016) Introduction to multivariate statistical analysis in chemometrics. CRC press, Boca Raton
- Viscarra-Rossel RA, Behrens T (2010) Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158:46–54
- Viscarra-Rossel RA, Behrens T, Ben-Dor E, Brown D, Demattê J, Shepherd KD, Shi Z, Stenberg B, Stevens A, Adamchuk V, others (2016) A global spectral library to characterize the world's soil. *Earth-Sci Res* 155:198–230
- Viscarra-Rossel RA, Jeon YS, Odeh IOA, McBratney AB (2008) Using a legacy soil sample to develop a mid-IR spectral library. *Soil Res* 46:1–16
- Viscarra-Rossel RA, McBratney AB (2008) Diffuse reflectance spectroscopy as a tool for digital soil mapping. In: *Digital soil mapping with limited data*. Springer, Berlin, pp 165–172

- Viscarra-Rossel RA, Webster R (2012) Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *Eur J Soil Sci* 63:848–860
- Viscarra-Rossel RA, Webster R (2011) Discrimination of Australian soil horizons and classes from their visible–near infrared spectra. *Eur J Soil Sci* 62:637–647
- Weihermüller L, Huisman JA, Lambot S, Herbst M, Vereecken H (2007) Mapping the spatial variation of soil water content at the field scale with different ground penetrating radar techniques. *J Hydrol* 340:205–216
- Wilson MJ, Cradwick PD (1972) Occurrence of interstratified kaolinite-montmorillonite in some Scottish soils. *Clay Miner* 9:435–437



# Chapter 2

## Getting Started with R



R provides a convenient and flexible data-analytic environment for soil spectral data. R is a programming language and a software facility for data manipulation, statistical analysis and graphics. R is an implementation of the S language developed at Bell Laboratories (Venables et al. 2009) in the 1980s. While R is an integrated environment for data manipulation, it is mostly used for statistical analyses. R builds is a so-called ‘GNU’ project, i.e. it is public domain and all resources are freely accessible, unlike some other programming languages such as Matlab.

The development of R for statistical analyses relies on the users who develop and maintain a large variety of packages. A few of them are built into the R-base system, but all existing packages are accessible online (see Sect. 2.6). One of the main advantages of R is the amount of information and resources that any user can find on the Internet and the constant evolution of resources.

This chapter is a very short introduction to the use of R and to one of the user-friendly graphical user interfaces (GUI) called RStudio. This chapter provides explanations of the main commands, programming tools and graphical functions needed to understand and follow the content of the book. This chapter is written for users with little or without any previous programming experience but is not enough by itself to be proficient in programming using R. In the latter case, the reader is redirected to Sect. 2.6 or to Venables et al. (2009) for further references.

### 2.1 Use of R and RStudio

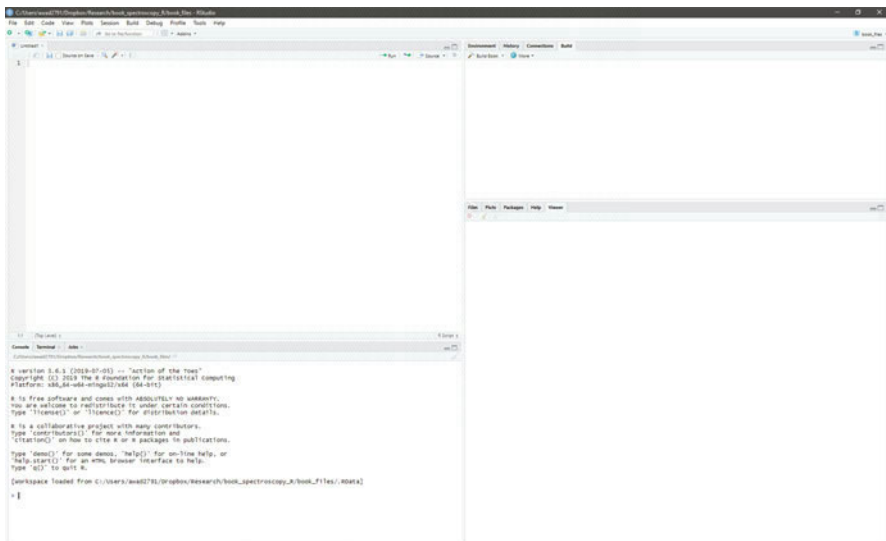
**Installing R** Installing the latest version of R is freely and legally accessible from the Comprehensive R Archive Network (CRAN) website <https://cloud.r-project.org/> with the following steps.

1. Click on *Download R for Windows* assuming you work on Windows; otherwise, select the platform Linux or (Mac) OS X.

2. Click on *base* or *install R for the first time* which redirects to the [base package](#) page for the latest version of R available.
3. Click on *Download R 3.6.2 for Windows* (Note that at the moment of writing, version 3.6.2 is the latest) and save the executable file.
4. Locate and click on the executable file. Select the default answers for all questions. R installs itself automatically. You should see an R icon on your desktop. If you click on this icon, R will open as a command windows, and you can start using it. Most users, however, will find it hard to use R in this way as it does not have a GUI. Many freeware GUI are available for R. In this book, we recommend to use one of the most common, called RStudio.

**Installing RStudio** RStudio is an interface developed to improve the R user experience. RStudio builds the interface on the background R installation and includes some core functionalities such as visualization and code editor panels and of course the R console. Installing the latest version of RStudio is freely accessible on the [RStudio website](#) with the following steps:

1. Go to <https://rstudio.com/products/rstudio/> and click on *RStudio Desktop*.
2. Choose the Open Source Edition (free) of RStudio and click on *Download RStudio Desktop*.
3. Click on *Download* under the RStudio Desktop icon.
4. Choose the correct platform. Assuming you work on Windows, click on *RStudio-1.2.5033.exe* (Note that at the moment of writing, the version 1.2 is the latest), and save the executable file on your computer.
5. Locate and click on the executable file. Select the default answers for all questions. RStudio installs itself automatically. RStudio should now be installed in your computer. Click on the RStudio icon to open the interface presented in Fig. 2.1.



**Fig. 2.1** The RStudio interface with the four windows. The upper left window is hidden by default but can be opened by clicking the file menu, then New File and then R script