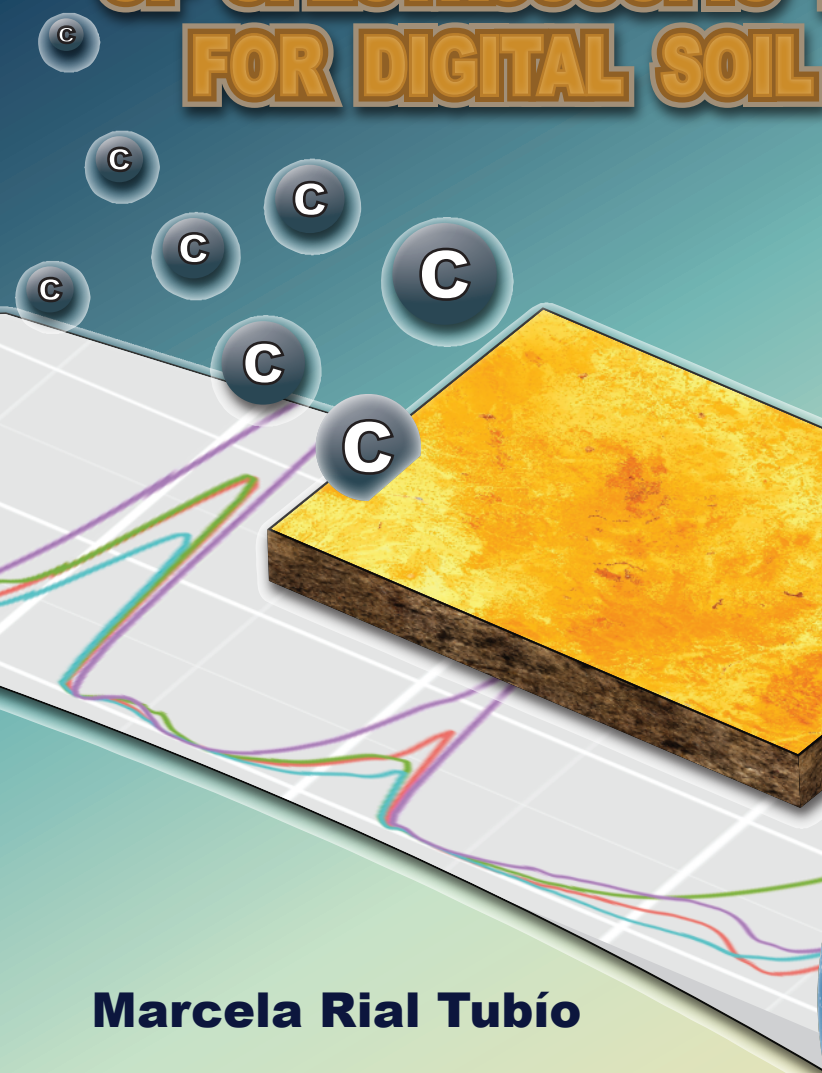
USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Ph.D. THESIS

MODELLING OF THE TOPSOIL ORGANIC CARBON CONTENT BY ANALYSING THE POTENTIAL OF SPECTROSCOPIC TECHNIQUES FOR DIGITAL SOIL MAPPING



Marcela Rial Tubío

Santiago de Compostela
2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Ph.D. THESIS

MODELLING OF THE TOPSOIL ORGANIC CARBON CONTENT BY ANALYSING THE POTENTIAL OF SPECTROSCOPIC TECHNIQUES FOR DIGITAL SOIL MAPPING

Marcela Rial Tubío

MEDIOAMBIENTE E RECURSOS NATURAIS
FACULTADE DE BIOLOXIA

SANTIAGO DE COMPOSTELA
2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Ph.D. THESIS

MODELLING OF THE TOPSOIL ORGANIC CARBON CONTENT BY ANALYSING THE POTENTIAL OF SPECTROSCOPIC TECHNIQUES FOR DIGITAL SOIL MAPPING

Marcela Rial Tubío

Signed: .....

MEDIOAMBIENTE E RECURSOS NATURAIS
FACULTADE DE BIOLOXIA

SANTIAGO DE COMPOSTELA
2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

D. Antonio Martínez Cortizas and D. Luis Rodríguez Lado,

As supervisors of the PhD Thesis titled:
"Modelling of the topsoil organic carbon content by analysing the potential of spectroscopic techniques for digital soil mapping"
Presented by Marcela Rial Tubío, student of the Doctoral Program on Environment and Natural Resources.

Authorize the presentation of the indicated thesis, considering that it meets the requirements demanded in the article 34 of the regulation of doctoral studies, and that as supervisors of the same does not incur the abstention causes established in the law 30/1992.

Signed: .....
D. Antonio Martínez Cortizas

Signed: .....
D. Luis Rodríguez Lado

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Modelling of the topsoil organic carbon content by analysing the potential of spectroscopic techniques for digital soil mapping

Marcela Rial Tubío

Abstract

Soil research is being driven by a need to understand the role of soil in the global climate change. The scientific community and policymakers have been expressed the need for spatially referenced information about soil organic carbon distribution. It represents the largest terrestrial carbon pool, being one of the most relevant components in the carbon cycle budget and climate change feedbacks. The advances in computer science have brought enormous potential for improve the manner in that soil maps were produced. The use of new statistical approaches showed the potential of digital soil mapping to inference the spatial distribution of organic carbon content with a limited sampling point scheme. Recently, digital soil mapping has been complemented with visible and infrared spectroscopy, which provides an effective tool to obtain soil organic carbon concentration of soil samples and overcomes the high cost and time-consuming typical of traditional chemical methods.

The work here realized aims to develop statistical methods to quantify the topsoil organic carbon content and stocks in three study cases by using spectroscopic data as a tool for digital soil mapping frameworks. In the first study case we developed a statistical approach to map organic carbon content at regional scale in topsoils from Galicia (NW Spain) and we explored the capacity of spectroscopy for predict soil organic carbon content. In the second study case we pass to continental scale aimed to improve the accuracy of the current approaches. In such manner, we developed a spatially non-stationary approach that allows to map soil organic carbon content at European scale and also identify the factors more relevant for soil organic accumulation/degradation across Europe. Finally, we evaluated the capacity of digital soil mapping methods for monitoring and for the quantification of the soil organic carbon stocks expected for future times under different climate change scenarios. We used for such purpose legacy data from Santa Cruz Island (Galapagos), a place under a special protection status where the effects of climate change were pinpointed as one of the major threats for the ecosystems of this area.

Acknowledgements

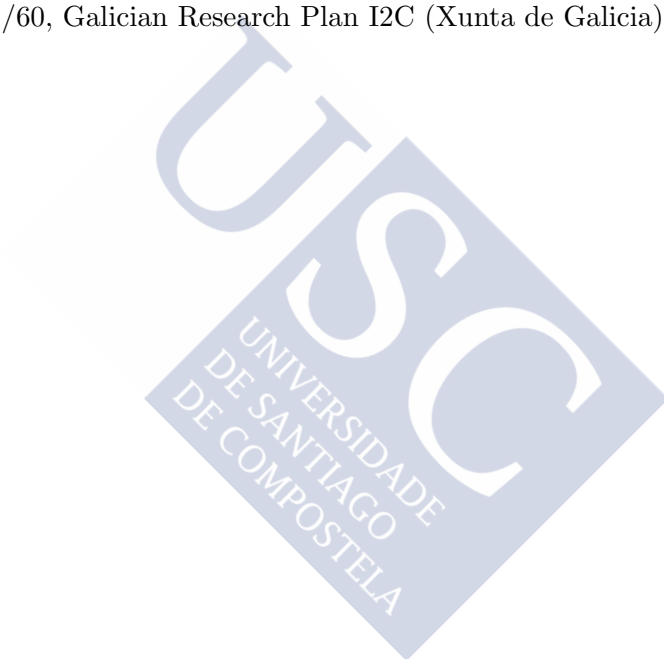
First of all, I just want to thank my supervisors for their support and competent guidance. The grateful I feel is overwhelming. Thanks to Prof. Antonio Martínez Cortizas for give me the opportunity to undertake my thesis in his group making possible to start this great job. I also gotta thank to Dr. Luis Rodríguez Lado for having a major contribution in the development of this research, giving me the necessary tools for carry it out.

I would also like to send a heartfelt acknowledgement to the members of Earth System Science research group who made the working place a fascinating area. Thanks also to Prof. George Stoops for providing the original soil samples from the Galapagos Islands used in this work.

Thanks also to the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP5, and I thank the climate modelling groups (Table 5.2) for producing and making available their model output. For CMIP5 the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portal. I also want

to thank to the Land Resource Management Unit (Institute for Environmental & Sustainability) of the Joint Research Centre (JRC) of the European Commission, the Consortium for Spatial Information (CGIAR-CSI) GeoPortal, the European Environment Agency and the developers of the WorldClim database and SoilGrids products for making available their model outputs enabling the development of this thesis.

This research was supported in part through the research grant EM 2012/60, Galician Research Plan I2C (Xunta de Galicia).



To my family

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Figures

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Tables

Abbreviations

ANN	**A**rticial **N**eural **N**etwork
AR5	Fifth **A**ssesment **R**eport
ATR	**A**ttenuated **T**otal **R**eflectance
AVHRR	**A**dvanced **V**ery **H**igh **R**esolution **R**adiometer
BN	**B**ayesian **N**etwork
BRT	**B**oosted **R**egression **T**ree
CART	**C**lassification **A**nd **R**egression **T**ree
CCE	**C**hicago **C**limate **E**xchange
CDM	**C**lean **D**evelopment **M**echanism
CEC	**C**ation **E**xchange **C**apacity
CLC	**C**orine **L**and **C**over
CMIP5	**C**oupled **M**odel **I**ntercomparison **P**roject
COP	**C**onference **O**f the **P**arties
CR	**C**ontinuum **R**emoval
CSM	**C**onventional **S**oil **M**apping
DB	**D**ouble **B**ond

DEM	**D**igital **E**levation **M**odel
DLaTGS	**D**euterated **L-a**lanine **T**trig**l**ycine **S**ulphate
DTGS	**D**euterated **T**trig**l**ycine **S**ulphate
DRIFT	**D**iffuse Reflectance
DSM	**D**igital **S**oil **M**apping
ECRE	**E**lectrical **C**onductivity/resitivity based on **R**olling **E**lectrodes
EMI	**E**lectro**m**agnetic **I**nduction
ENSO	**E**l **N**ino **S**outhern **O**scillation
ESDB	**E**uropean **S**oil **D**ata**b**ase
FTIR	**F**ourier **T**ransforms **I**nfrared
GA	**G**enetic **A**lgorithm
GAM	**G**eneralised **A**daptive **M**odel
GHG	**G**reen**h**ouse **G**ases
GIS	**G**eographic **I**nformation **S**ystem
GLM	**G**eneralised **L**inear **M**odel
GOES	**G**eostationary **O**rbiting **E**arth **S**atellite
GOF	**G**oodness **O**f **F**it
GPS	**G**lobal **P**ositioning **S**ystem
GRS	**G**amma-**R**ay **S**pectrometry
GWR	**G**eographically **W**eighted **R**egression
HWSD	**H**armonized **W**orld **S**oil **D**atabase
IPCC	**International P**anel on **C**limate **C**hange
ITCZ	**Inter-T**ropical **C**onvergence **Z**one
JRC	**J**oint **R**esearch **C**entre

LAI	**L**eaf **A**rea **I**ndex
LOO	**L**eave-**O**ne-**O**ut
LCC	**L**and **C**over **C**hange
LUCAS	**L**and **U**se/Land **C**over Statistical **A**rea Frame **S**urvey
LV	**L**atent **V**ariable
MAE	**M**ean **A**bsolute **E**rror
MARS	**M**ultivariate **A**daptive **R**egression **S**plines
MIR	**M**edium **I**nfra**r**ed
MLR	**M**ultiple **L**inear **R**egression
MSC	**M**ultiplicative **S**catter **C**orrection
MSE	**M**ean **S**quared **E**rror
NDVI	**N**ormalized **D**ifference **V**egetation **I**ndex
NIR	**N**ear **I**nfra**r**ed
OOB	**O**ut **O**f **B**ag
OK	**O**rdinary **K**riging
OLS	**O**rdinary **L**east **S**quares
OSC	**O**rthogonal **S**ignal **C**orrection
PAS	**P**hoto**a**coustic **S**pectroscopy
PCA	**P**rincipal **C**omponent **A**nalysis
PCR	**P**rincipal **C**omponent **R**egression
PLS	**P**artial **L**east **S**quares
RBFN	**R**adial **B**asis **F**unction **N**etworks
RCP	**R**epresentative **C**oncentration **P**athways
RF	**R**andom **F**orest

RK	**R**egression **K**riging
RMSE	**R**oot **M**ean **S**quared **E**rror
SAR	**S**ynthetic **A**perture **R**adar
SE	**S**tandard **E**rror
SG	**S**avitzky-**G**olay
SNV	**S**tandard **N**ormal **V**ariate
SOC	**S**oil **O**rganic **C**arbon
SSE	**S**um of **S**quared **E**rror
SSPF	**S**patial **S**oil **P**rediction **F**unction
SST	**S**ea **S**urface **T**emperature
SVM	**S**upport **V**ector **M**achines
TB	**T**riple **B**onds
TM	Landsat **T**hematic **M**apper
TOVS	**T**IROS **O**perational **V**ertical **S**ounder
UNEP	**U**nited **N**ations **E**nvironment **P**rogram
UNFCCC	**U**nited **N**ations **F**ramework **C**onvention on **C**limate **C**hange
VNIR	**V**isible and **N**ear **I**nfrared
WMO	**W**orld **M**eteorological **O**rganization

Chapter 1

INTRODUCTION

1.1 Significance of the study

Since the beginning of the Industrial Revolution, human activities exerted an intense pressure on natural systems. The growing demands for non-renewable energy resources related with the fossil fuel industry, altered the natural carbon cycle (Figure 1.1) in a dramatically way [1–5]. It was estimated that human activities involve an atmospheric carbon net annual increase of 4.5 ± 0.1 GtC/year. In fact, the increase in concentration of atmospheric gases, such as carbon dioxide (CO_2) and methane (CH_4), produces the intensification of the greenhouse effect over the Earth [6–8]. Concentrations of these Greenhouse Gases (GHG) reached values that are

unprecedented in at least the last 800000 years. The progressive increase of GHG led to planet to an imminent climate change that will carry effects like global warming, the increase of sea and earth surface temperatures, more droughts and heat waves, more floodings, melting of icecaps, rise of sea levels, the intensification of extreme events like El Niño Southern Oscillation (ENSO), changes in oceanic and atmospheric currents, and the change of actual climate patterns [4]. The threats posed by these effects to human life on Earth prompted the development of international policy agreements that aim to control the advance of this catastrophic situation [9, 10].

The political decisions adopted are based on scientific evidence of the effects of climate change. In 1988, the United Nations Environment Program (UNEP) and the World Meteorological Organization (WMO) created the International Panel on Climate Change (IPCC), which is the leading international body for the assessment of climate change. Its objective is providing the world with a clear scientific view on the current state of knowledge in climate change and its potential environmental and socio-economic impacts. In view of IPCC reports, after the Rio Convention (Earth Summit - 1992) and a series of meetings, the United Nations Framework Convention on Climate Change (UNFCCC) signed the Kyoto Protocol in 1997, the most important political agreement to date about climate change.

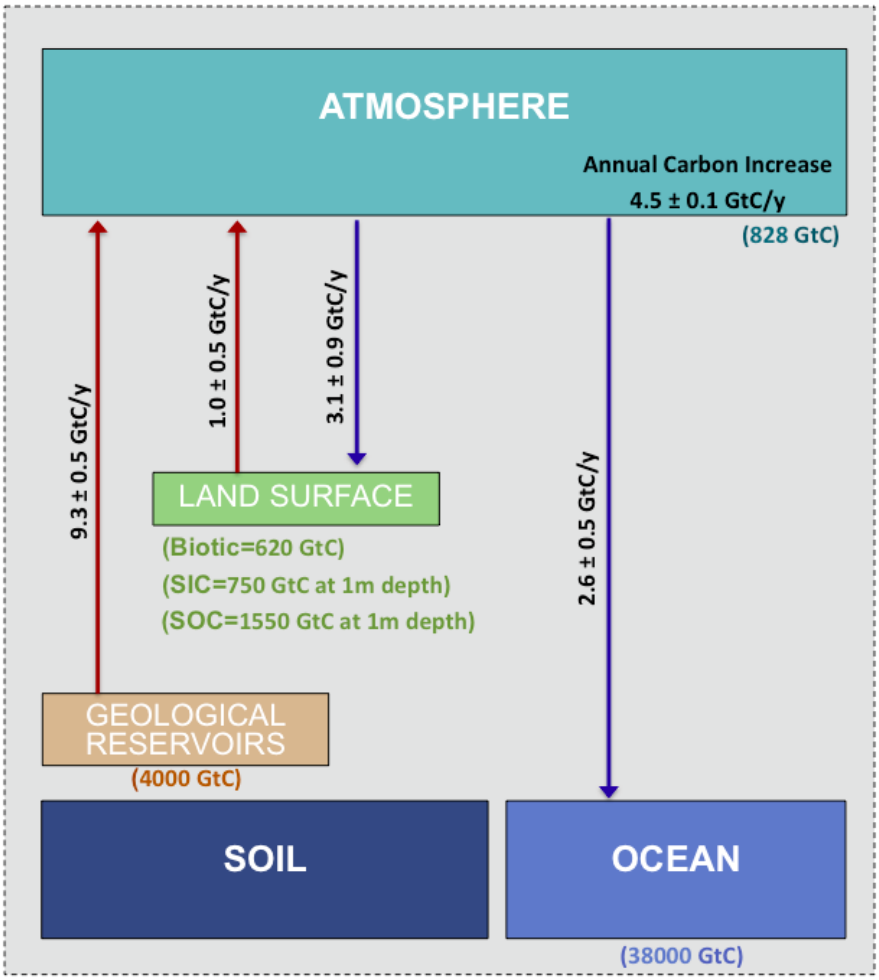


FIGURE 1.1: Global carbon cycle and fluxes. It includes Soil Inorganic Carbon (SIC).
(adapted from Le Quéré et al. (2016) [11] and Lal (2016) [12])

It is tracked and reviewed periodically in the Conference of the Parties (COP). The main goal of the Kyoto Protocol is to reduce the anthropogenic GHG emissions posing specific targets to the industrialized countries [13, 14]. In order to reach this ambitious objective, carbon emissions began to be regulated through a system of carbon credits that can be traded by the World Bank, the Chicago Climate Exchange (CCE), the European Climate Exchange, the national and local industry and the Clean Development Mechanism (CDM) of the own Kyoto Protocol [15–18].

The last release of the IPCC report (AR5) was launched in 2014 [4]. It reflects that GHG emissions suffered a dramatically increase since 1970, when emissions were 27 Gt. In 2010 GHG emissions reached values of 49 Gt, 92% corresponding to carbon gases. In this report the new Representative Concentration Pathways (RCP) climate change scenarios for future times were fixed, which are defined as function of radiative forcing levels in W m^2. Figure 1.2 shows the expected trajectories of GHG levels under each scenario. The report indicates that even if anthropogenic emissions were stopped, many aspects of the climate change will continue for centuries. So there is an imperative need to reduce GHG emissions and to establish adaptation measures that could be able to limit the risks associated to climate change effects.

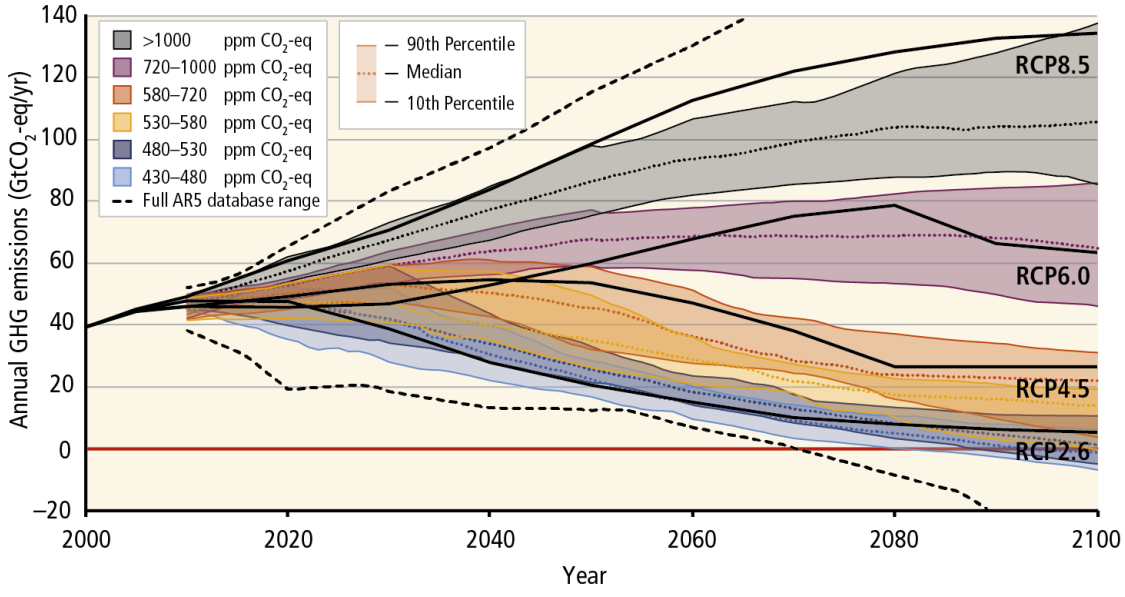


FIGURE 1.2: GHG emission pathways 2000-2100 for all RCP scenarios.
(extracted from the Climate change 2014: Synthesis Report [4] Figure SPM.11 rev1-01)

Soil plays a decisive role into the carbon cycle [19]. Soil Organic Carbon (SOC) represents one of the largest terrestrial carbon pools (Figure 1.1) [7]. Many researches estimated the SOC stored at global scale at 1 m depth by using statistical methods. Scharlemann *et al.* [20] compiled SOC estimates of historic data from 27 studies published between 1951-2011. The SOC stock reported by these studies ranges between 504-3000 GtC, with an average value of 1460.5 GtC. According to the current version of the Harmonized World Soil Database (HWSD), that is the most complete, recent and coherent global dataset of SOC content, the SOC stock at 1 m depth would be 2476 GtC [21] using the original values for bulk density. Adjusting the HWSD bulk density of soils with high organic carbon results in a mass of 1230 GtC and, additionally, setting the bulk density of Histosols to 0.1 g cm^3, results in a mass of 1062 GtC [22]. The largest SOC stocks are located in wetlands and peatlands, most of them located in tropic and permafrost regions [21, 23]. These areas are identified as the major source of systematic error in model estimations and further research is still necessary to adjust the estimates of SOC stocks to more accurate values [24].

The role of SOC in climate change mitigation was further investigated and incorporated to the international treaties. Tian *et al.* [24] analyzed the global patterns of SOC dynamics over the past century showing that SOC mean residence time suffered a reduction of 3.4 years. Climate and land use

changes promoted a decline in SOC stocks at global scale, while CO_2 and nitrogen deposition over intact ecosystems increased it. The modification of soil activities, such as land use or land cover changes, can produce gains of SOC stocks enhancing GHG sequestration [5, 25]. The potential of land management activities to sequester atmospheric carbon in the soil is reflected in articles 3.3 and 3.4 of the Kyoto Protocol (http://unfccc.int/kyoto_protocol/items/2830.php) [26]. In the last Conference of the Parties (COP21) the *"4 per 1000"* initiative was presented. It was estimated that the annual GHG emission is about 8.9 GtC per year, meanwhile the SOC stock estimated at 2 m depth is 2400 GtC. The ratio of anthropogenic emissions and the SOC stock result in a value of 4 per 1000. It means that an increase of global soil organic matter stocks by 0.4 percent on an annual basis is necessary to offset atmospheric GHG emissions (http://newsroom.unfccc.int/lpaa/agriculture/join-the-41000-initiative-soils-for-food-security-and-climate/) [12, 27]. This makes sense in the carbon credits industry. For example, countries that earn carbon credits from the adoption of suitable land management practices can sell them to those that want to offset their emissions. In view of the scientific, economical and political implications of the climate change stage, urges the need to quantify and monitor the SOC content and stocks at different spatial scales [28]. The knowledge of SOC dynamics also has several implications, like

the use of SOC content as input in biogeochemical models [29, 30] or the study of SOC losses associated to soil erosion and soil quality decline [31–34].

1.2 Objectives

The main objective of this thesis is **developing geostatistical approaches aimed to ascertain the spatial distribution of SOC contents and stocks by using infrared data**. The specific objectives are the following:

- To demonstrate the potential of spectroscopic techniques in the Visible-Near Infrared (VNIR) and Medium Infrared (MIR) ranges for SOC quantification at topsoil level.
- To relate spectroscopic data with data from traditional chemical analysis by using statistical models.
- To identify the factors that promote SOC accumulation in three study cases.
- To obtain the spatial distribution of SOC content at regional and continental scales.
- To evaluate the predictive capacity of the developed statistical models.
- To identify the relevant spectroscopic bands providing information on the SOC contents in Galicia, Spain.
- To obtain the spatial distribution of the factors that control SOC accumulation in Europe, from VNIR data.

- To analyze the potential effects of climate change on SOC stocks in Santa Cruz Island, Galapagos, for two future time series.

1.3 Outline

This thesis is based on the following structure. It is divided in six chapters:

Chapter 1 introduces the main topic, the role of SOC in climate change, the state of the regulation of SOC stocks and the political decisions worldwide stablished for enhancing GHG sequestration. The objectives that we pretend to reach with this study and an outline of the thesis structure are also presented. **Chapter 2** provides a literature review of the use of Digital Soil Mapping (DSM) methods to show the spatial distribution SOC content and stocks, the factors that promote SOC accumulation and a review of the DSM algorithms used for mapping purposes. This chapter also describes the role of spectroscopic techniques in DSM frameworks, its theoretical fundaments, its applications in soil science and the statistical methods commonly used to relate SOC content and spectroscopic data.

Chapter 3, 4 and 5 present three particular cases studies based on published or submitted papers to international journals (see Appendices). **Chapter 3** presents a statistical methodology for mapping SOC content in Galicia (NW Spain) and shows the potential of infrared spectroscopy for quantify such contents and the factors that control its accumulation. **Chapter 4** represents a change to continental

scale. In view of the high accuracy obtained at local scale, we developed a spatially non-stationary approach aimed to improve the low accuracy typical of continental scale approaches. This statistical methodology allows mapping SOC content and identifying the factors involved in its accumulation on each area across 23 states of the European Union, based on spectroscopic features. Finally, **Chapter 5** explores the capacity of DSM methods to quantify present SOC stocks and those expected for future times. For that purpose, we use legacy data from soil samples collected in Santa Cruz Island (Galapagos), a place under a special protection status where the human influence is minimal. This work presents a statistical methodological approach for estimating SOC stocks in Santa Cruz in the period 1950-2000 and the stocks expected for two future time series under different climate change scenarios.

Chapter 6 compiles the main findings and conclusions of the thesis.

Chapter 2

LITERATURE REVIEW

2.1 Digital Soil Mapping

DSM can be defined as *"the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge and from related environmental variables"* [35].

Figure 2.1 summarizes the principles of DSM. Geo-referenced soil samples are collected and analyzed to obtain the value of a specific soil property (S_p). This property can be then predicted and mapped over a geographic area by using a function (f) called Spatial Soil Prediction Function. These functions are based on correlations between

environmental variables, in the form of geo-refenced data layers or spatial position, and the soil attributes measured at sampling points [36]. Many soil properties were estimated by DSM, as for example SOC contents and stocks, Cation Exchange Capacity (CEC), pH, sand, clay and silt contents or soil classes. The spatially inferred property can be used to predict other functional soil properties more difficult to measure, such as field or available water capacity, by means of pedotransfer functions [36]. The creation of spatial soil information systems composed of a set of soil property maps, is the appropriate response to the huge demand of quantitative spatial soil information [35], enabling the evaluation of soil functions like biomass production or buffering capacity [36].

Numerous research papers and books described the basis and further reviewed the state of art of DSM [36–41]. It is complicated to date the origin of DSM. Soils maps were made even since 1600, but it is not until the beginning of 20th century when DSM originate. DSM roots derive from Conventional Soil Mapping (CSM), which appeared in USA led by the expert on pedology Eugene Hilgard, who was considered the father of modern soil science and pioneer in soil survey [42]. The advances achieved over the 20th century for making and using soil surveys were described in the standard from the Soil Survey Division Staff [43, 44]. In CSM, the maps were made by experienced soil surveyors that

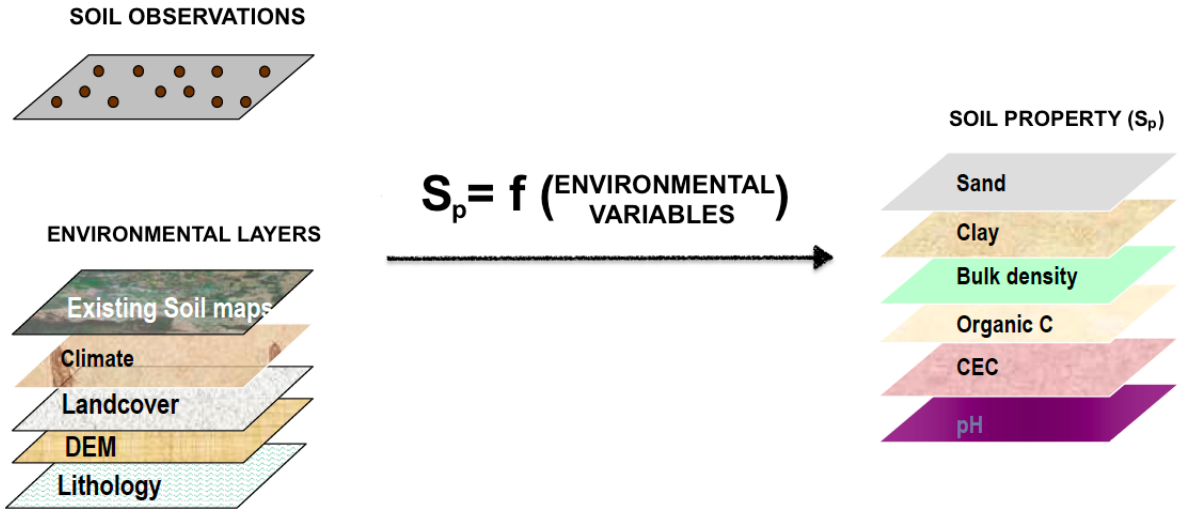


FIGURE 2.1: Principles of DSM.
(adapted from Minasny and McBratney (2008) [36])

delineated soils mapping units with the only help of soil drillings and aerial photographs. DSM makes use of computer science to reduce the need of human resources and field work typical of CSM. First studies mentioning DSM back to the 70s, when Roger Tomlinson, the father of Geographic Information Systems (GIS), labelled digitalized polygon maps as digital soil maps [41, 45]. With the boom of computer era, the research studies on this field boosted gradually since 1990 (Figure 2.2). The development of new technologies such as Global Positioning Systems (GPS), GIS, remote sensing, on-site geophysical instrumentation and associated data loggers, and the development of statistical and geostatistical techniques greatly increased our ability to collect, analyze, and predict spatial information related to soils [46].

In 2004, the interests and skills of DSM research were put in common in the first workshop about DSM that took place in Montpellier. This meeting of international experts led to the creation of the IUSS working group on Digital Soil Mapping and the celebration of more conferences that culminated with the GlobalSoilMap project. The main aim of the project is to predict and up-date soil properties at fine spatial resolution at global scale [47, 48]. To date, the most important goal reached at global scale is the creation of the SoilGrids products, a series of 3D spatial maps at 250 m resolution for various soil properties that include: SOC content (g kg^{-1}), pH, sand, silt and clay fractions (%), bulk

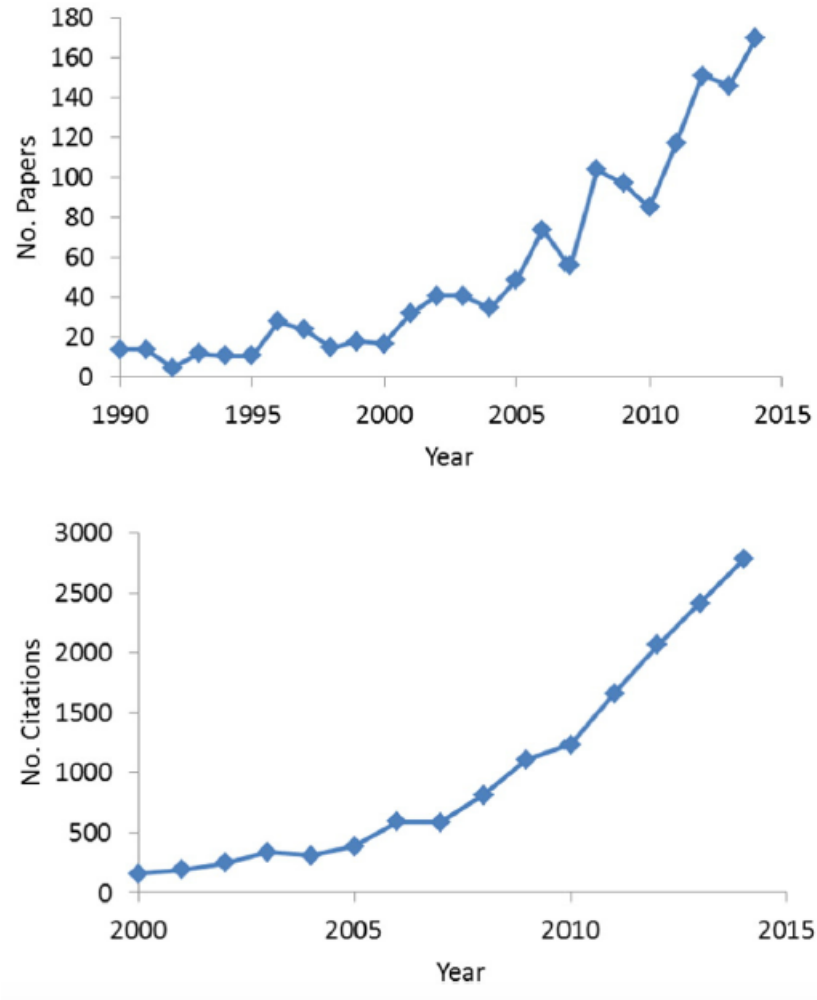


FIGURE 2.2: Evolution of DSM research.
(extracted from Minasny and McBratney (2016) [41])

density (kg m^3), CEC (cmol+ kg^{-1}), coarse fragments (%), SOC stock (t ha^{-1}), depth to bedrock (cm), World Reference Base soil groups, and USDA Soil Taxonomy suborders (https://soilgrids.org/) [49, 50].

2.2 Environmental variables in DSM

The aim of DSM is to develop statistical models that relate soil observations with environmental variables. Hence, the first step is the identification of the most influential environmental variables for spatial prediction. These variables are specific of the study area and the soil property that it tries to predict. The Jenny's equation [51] describes a series of factors that are relevant in soil formation. McBratney *et al.* [40] adapted Jenny's equation to an spatially explicit approach, adding soil and space to the Jenny's factors. According to McBratney's approach (Equation 2.1), the soil classes or attributes to be modeled (S_p) are function of other measured properties of the soil at the sampling points (S), the climatic properties (C), the organisms (O), the topographic and landscape attributes (R), the parent material (P), the age or time (A) and the geographic location (N). The error associated to the prediction of the soil property (ε) is measured as the spatial dependency of the model residuals.

$$S_p = f(S, C, O, R, P, A, N) + \varepsilon \tag{2.1}$$

2.2.1 Soil (S)

Soil maps previously created are often used to built a prediction model of a new soil property. Nowadays, there are specific instruments for recording soil properties such as hyperspectral sensors for mapping mineralogical features (iron oxides, carbonates and sulfates), Synthetic Aperture Radar (SAR) to evaluate the volumetric moisture content, Electromagnetic Induction (EMI) and Electrical Conductivity/resistivity based on Rolling Electrodes (ECRE) for mapping soil bulk electrical conductivity or Gamma-Ray Spectrometry (GRS) to detect radioactive elements.

2.2.2 Climate (C)

Climatic conditions determine the degree of weathering of the parent material, vegetative production and the activity of soil organisms. Climate is mainly represented by temperature and rainfall parameters. It can also be used to derive factors like evapotranspiration or soil moisture and water balance components. The maps of such parameters can be obtained by extrapolation of meteorological stations data or from

remote sensing satellite data such as Advanced Very High Resolution Radiometer (AVHRR), Geostationary Orbiting Earth Satellite (GOES), TIROS Operational Vertical Sounder (TOVS), Landsat Thematic Mapper (TM). SOC content tends to show a strong correlation with climate, especially at global scale. The mineralization of organic matter is carried out by microorganisms, under the influence of climatic and ambient soil conditions [52]. Temperature is a key factor controlling the rate of SOC decomposition that occurs more rapidly in the tropics than in temperate areas and, hence, results in a lager accumulation of organic matter in soils located in areas with cooler climates because of slower mineralization rates. Elevated levels of rainfall and soil moisture also result in higher SOC contents because of the increase in biomass production, which provides more organic residues, and thus more potential food for soil biota [53]. Decomposition rates in areas with the highest SOC stocks, wetlands and permafrost, are low due to low availability of oxygen and low temperatures, respectively. Thus, SOC is vulnerable to changes in the hydrological cycle as well as to changes in permafrost dynamics [54].

Regarding recent climate change, the higher temperatures in the humid tropics are expected to result in an increase of decomposition rates of organic matter. This climate effect will be less pronounced in colder regions. The decrease of rainfall rates expected for the near future in arid subtropical regions,

will result in large differences in decomposition rates between contemporary and future conditions, being tropical mountain forest hotspots of future carbon losses. In boreal and arctic regions, the increase of primary production net balances outweighing the potential carbon losses due the thawing of permafrost [21, 55]. Carvahllais *et al.* [56] highlight the significance of the hydrological cycle in carbon dynamics at global scale, suggesting that biological activity is limited by water availability specially in arid, semi-arid and dry sub-humid ecosystems. The progression of aridity conditions in many sites on the globe due to climate change will produce a reduction of vegetation cover, stimulating processes that imply SOC losses, such as rock weathering, over the dominance of biological processes like litter decomposition that will promote SOC contents [57].

In a recent study, Doettert *et al.* [7] also found that climate factors are relevant for SOC prediction, but they suggest that the interactions between geochemical factors and climate must be considered for getting robust predictions.

2.2.3 Organisms (O)

Organisms play an important role in the development and composition of soil, adding organic matter, aiding to decomposition, weathering and nutrient cycling. Vegetation is

the main component of the biota affecting soil formation and its modification due to human activities such as land cover or land use modification. Vegetation maps are obtained through remote sensing methods as AVHRR or vegetation indices like the Normalized Difference Vegetation Index (NDVI). Salinity, toxicity and extremes in soil pH result in poor biomass production [53]. Harauk *et al.* [58] studied the role of biological processes in the carbon cycle concluding that the addition of microbial biomass dynamics into carbon models improves the predictive capacity of the models. Zhu *et al.* [59] analyzed the greening of the Earth and its drivers using long-term satellite Leaf Area Index (LAI) records and global ecosystem models. Factorial simulations suggested that CO_2 fertilization explain 70% of the observed greening trend, followed by nitrogen deposition (9%), climate change (8%) and Land Cover Change (LCC) (4%). The effect of CO_2 fertilization is more pronounced in the tropics, whereas climate change is more relevant in high latitudes and the Tibetan Plateau, and LCC contributed the most to the regional greening observed in southeast China and the eastern United States. Although net primary production is expected to increase in the future due to this CO_2 fertilization, productivity may still be limited by the availability of nitrogen or other resources [21, 55].

Due to the implications of climate change, there is a need for the adoption of land management practices that enhance

the amount of SOC accumulated in soils [25, 60]. The meta-analysis done by Guo and Gifford [61], from data from 74 publications, summarizes the effects of land use changes on SOC stocks. They found SOC declines after land use changes from pasture to plantation (-10%), native forest to plantation (-13%), native forest to crop (-42%), and pasture to crop (-59%). SOC stocks increase after land use changes from native forest to pasture (+8%), crop to pasture (+19%), crop to plantation (+18%), and crop to secondary forest (+53%). Lugato *et al.* [26] also explored the potential changes induced by land management practices at European scale under different climate change scenarios. Although their results indicated that sequestration potential was strongly dependent on the spatial and temporal scale considered, cropland to grassland conversion was the best effective management to enhance SOC stocks followed by reduced tillage, ley crops presence within the cash rotation and cereal straw incorporation. The influence of management practices on crop residues, SOC decomposition and distribution with depth is still poorly understood and under discussion. For example, meta-data analysis showed that SOC content in the surface layers increased under no-tillage when compared to full-inversion tillage and that this tendency increased with time. Limited effects of no-tillage on SOC stocks were found when the whole soil depth was considered. Cases have also been published where no-tillage did not increase SOC contents

or resulted in SOC increases at great depth when no-tillage was used with legume cover crops [60].

2.2.4 Relief (R)

Topography also plays a significant role on soil formation as it determines runoff of water, and its orientation affects microclimate, which in turn affects vegetation. Maps related with topographic parameters, which include slope, aspect, curvature or the multi-resolution index of valley bottom flatness, among others, are derived from Digital Elevation Models (DEM). Taking into account only the topographic parameters, SOC accumulation is favored at the bottom of hills. Conditions here are wetter than at upper slopes and SOC is transported to the lowest point in landscape through runoff and erosion. However, a high erosion rate could cause stripping of the soil thus preventing parent material to stay in place producing weakly developed soil at the mid- and near the bottom of the slope. At higher altitude it is usual to find higher SOC contents due to the effect of the lower temperatures. Furthermore, in the Northern Hemisphere, SOC levels are higher on north-facing slopes compared with south-facing slopes, because temperatures are lower in the former. This occurs the other way round in the Southern Hemisphere [53].

2.2.5 Parental Material (P)

Soil parent material may be in situ weathered rock or sediments deposited by wind, water or ice. The character and chemical composition of the parent material plays an important role in determining soil properties, especially during the early stages of development. Its influence on soil properties tends to decrease with time, as it is altered, and climate becomes more important. This kind of information is obtained from digitized geological maps used in a DSM framework. Jobaggy *et al.* [23] suggested that climatic conditions dominate SOC accumulation in shallow layers, while clay content do it in deeper layers due to increasing percentages of slow cycle SOC fractions. However, a series of findings also indicates that the decomposition of the organic matter is mainly controlled by biological and environmental conditions [62]. Anyways, SOC tends to increase as clay content increases. The organic matter content in fine textured (clayey) soils is two to four times greater than in coarse texture (sandy) soils under similar climate conditions [63]. In the soil matrix, SOC is stabilized through physico-chemical processes and through interactions with mineral surfaces as well as metal ions, forming aggregates, which translates into its spatial inaccessibility for the soil microbes and limited oxygen availability. This stabilization implies that soil type can limit the capacity to store SOC [60]. But parent material

influences SOC accumulation not only through the effect of soil texture. Soils developed from inherently nutrient-rich materials, such as basalt, are more fertile than soils formed from granitic materials, which contain less mineral nutrients. Moreover, the former experience more organic matter accumulation because of abundant vegetative growth [53].

2.2.6 Age (A)

As time passes, the weathering processes continue to act on soil parent material breaking it down and decomposing it. Horizon development processes continue diferenting layers in the soil mass by their physical and chemical properties. Climate interacts with time during the soil development process. Soil development proceeds much more rapidly in warm and wet climates thus reaching a mature status sooner. In cold climates, weathering is impeded and soil development takes much longer.

2.2.7 Location (N)

Some of the spatial soil prediction functions necessary for mapping use the spatial dependence of the samples as covariate, as it will be explained in the next subsections. The

location of soil samples is recorder with a GPS instrument and latitude and longitude would be then used as covariates.

2.3 Spatial Soil Prediction Functions

A Spatial Soil Prediction Function (SSPF) can be defined as a statistical algorithm that aims to predict and map a soil property or class by using environmental variables. The statistical model created upon the sample dataset is applied to a geographic database in order to create a predictive map [64, 65]. The majority of SSPF use regression algorithms, however a classification model could also be necessary as required, for example, for spatial prediction of soil classes. Much research was published in the last years in the DSM field and the SSPF used for making predictions varied widely from simple geostatistical approaches to more complex methods, such as machine learning techniques and expert systems [40, 64].

A compilation of studies found in the literature indicates that linear models are the most extensively used for SOC predictions due to their simplicity and easy interpretation, followed by the purely spatial approaches and machine learning techniques. The spatial distribution of SOC contents and stocks were mainly determined by using simple purely spatial and linear algorithms, such as Multiple Linear

Regression (MLR), Ordinary Kriging (OK), co-Kriging, Regression-Kriging (RK) and Geographically Weighted Regression (GWR) [66–70]. Despite their complexity, there is an increasing interest in using Machine Learning techniques to generate soil properties maps. Various machine learning approaches were used for SOC prediction, such as tree based models, Generalized Linear Models (GLM) or Artificial Neural Networks (ANN) [49, 71–73].

Although is not a matter of this thesis, during last years more complexed dynamic simulation models were developed, as CANDY [74], C-TOOL [75], CENTURY [76], DAYSY [77], DNDC [78], EPIC [79], ICBN [80], ROMUL [81] or RothC [82], that can overcome the static of SSPF, enabling to simulate the fluxes and turnover of SOC content.

2.3.1 Purely spatial approaches

Spatial approaches predict soil properties from spatial position largely by interpolating between soil observation locations [40]. The most popular spatial approaches are OK and its precedent methods. Kriging is a process of interpolation designed to predict attribute values between locations of measured samples. It uses a measure of spatial dependence, the variogram, to determine the weights applied to the data when computing the averages [64]. It is possible

considering the correlation between the variable of interest and another easily measured ones or even integrate multivariate data into OK, by using its respectively extensions called Co-Kriging and Factorial Kriging [83].

RK is an hybrid method that interpolates by kriging the residuals from a non-spatial model for quantify the predictive errors in a regression model. It is a popular spatial method used among soil researches because it is easy to use and produces good predictions in many cases [84]. Lark *et al.* [85] demonstrated that when the variogram of the residuals was estimated using RK a bias occurs at long lags. Furthermore, purely spatial approaches have a series of limitations like the assumption of spatial correlation, the need of a large number and closely spaced data points and the failure when the knowledge of soil materials and processes are taken into account [86, 87]. These facts advocated researches to find and explore other statistical methods to estimate soil properties. Statistical methods like linear models or machine learning techniques explore the relationship between quantifiable environmental variables and soil properties in order to create predictive soil maps. In comparison with the spatial approaches, these techniques are mostly used at regional scale due the lack of soil data and information about environmental variables at higher scales.

2.3.2 Linear models

Linear models attempt to describe the response variable, which is the property to be predicted, through a linear combination of one or more environmental variables. The linear algorithms differ mainly in the number of variables considered.

Ordinary Least Squares (OLS) was widely used in the prediction of soil attributes because of the easiness and wide availability. OLS is used to model a single response variable with a single predictor variable [40]. OLS regression model can be extended to include multiple explanatory variables by simply adding more variables to the equation. In this case OLS is normally referred as MLR. For large numbers of correlated predictor variables the most suitable regression methods are Principal Component Regression (PCR) and Partial Least Squares (PLS). These methods enable to reduce data dimensionality making linear combinations of the original variables. PCR produces linear combinations of the original variables, so-called principal components, and then use them in the regression model. PLS works much the same way, but its useful when dealing with several response variables. PLS calculates the components for the response and the predictor variables [40, 88]. PCR and PLS were used quite extensively in predicting soil attributes from the electromagnetic spectrum (see Section 2.4). GWR can also be considered a linear method but with traces of the purely

spatial approaches. It is an exploratory technique mainly intended to indicate where non-stationarity is taking place on the map, which is where locally weighted regression coefficients move away from their global values. It is based on the fact that the fitted coefficient values of a global model, fitted to all the data, may not represent adequately detailed local variations in the data; in doing so it follows local regression implementations [89, 90].

To overcome a non-normal response distribution and non-linearities in the linear models, simple methods like GLM and Generalized Adaptive Models (GAM) can be applied [40]. GLM and GAM were not so much used in soil mapping due to soil researchers tend to use machine learning techniques to cope with complex data that does not follow a linear pattern.

2.3.3 Machine learning techniques

Machine learning techniques are suitable for the analysis of complex data, which require to deal with non-linear relationships, high order interactions and missing values [91]. Machine learning techniques like tree based models, which includes Random Forest (RF) algorithm, are the most commonly learners used [92, 93]. RF has several advantages, as it can handle categorical and continuous predictors, it avoids overfitting and implements unbiased measures of error

rate and variable importance [72]. RF algorithm can cope with regression or classification problems. It divides the initial database into multiple decision trees, each one is trained from a randomized bootstrap sample of the entire training set and a subset of predictors randomly selected [94]. ANN is also a machine learning technique widely used as regressor or classifier. ANN works in an analogue way to the human brain: a neural network is formed by a set of interconnected units (neurons) linked to hidden layers, which estimate the non-linear correlations among input variables and predict the outputs values [95]. Genetic Algorithms (GA) are also well suited methods for soil mapping. GA is a randomized search and optimization technique used in classification problems based on the concept of evolution by natural selection as solutions are evolved in a stochastic, iterative manner. In GA only the stronger predictors survive to the training [96]. A few examples of Multivariate Adaptive Regression Splines (MARS) for soil mapping can also be found. MARS is an adaptive procedure aimed to built a function formed by a collection of piecewise linear basic functions. These functions are linear splines with a knot that reflects symmetric pairs [97, 98]. Although it is not yet stablished as a mainstream tool, bayesian methods, as Bayesian Networks (BN), can also be used for classification and regression in soil mapping approaches. BN are graphical probabilistic models in which predictions are obtained using prior probabilities derived from

either measured data or expert opinion. The main advantage of BN is that it avoids the "black-box" design of some previous methods like ANN, RF or GA. BN structure and the internal interactions between nodes are defined by the user, based on prevailing process understanding [99].

Other examples of techniques borrowed by machine learning and used for classification purposes in soil mapping are fuzzy logic for clustering, Support Vector Machines (SVM) or logistic regression [100]. The fuzzy algorithm assigns to each data point a probability of membership in each cluster center resulting in a continuous soil surface map where individual locations can belong to more than one class and no rigid boundaries are delineated to separate the soil [64]. SVM aims to construct a hyperplane to separate various groups by detecting the closest points between classes in space [100–102]. Finally, logistic regression is one of the most frequently used learners for classification. It is well suited for datasets where the dependent variable is categorical. The algorithm describes the relationships between a set of predictor variables and a dichotomous dependent variable expressing the results in probabilistic terms that indicates the probability of occurrence [100].

2.3.4 Expert systems

Expert systems are composed of data, a knowledge base and an inference engine that combines both (i.e. data and the knowledge) to infer logically valid conclusions [64]. For that, the computer-based knowledge can use the human expert or numerical methods. Expert systems are able to exploit soil surveyor knowledge by developing rule-based systems that imitate the surveyor conceptual model of soil variability [64]. Bui [103] and Wielemaker *et al.* [104] proposed methodological frameworks to formalize the landscape knowledge of the soil surveyor by structuring terrain objects in a nested hierarchy followed by inference and formalisation of knowledge rules. The main attempts to formally make such knowledge rules are Prospector [105], Expector [106] and Netica (http://www.norsys.com/netica.html).

2.4 Integration of spectroscopy in DSM

Taking political decisions to achieve a sustainable land use management demands accurate and up-to-date quantitative information on the spatial distribution of essential soil physical and chemical properties at large scales [25, 107]. In this sense, the main drawback of DSM frameworks for

appropriate soil monitoring is the time-consuming effort associated to soil surveys and soil data acquisition [48]. Spectroscopy in the visible and infrared ranges appears as a powerful alternative to support DSM studies due to its advantages compared to conventional soil analysis techniques, as it is a non-destructive technique, cheap, rapid, minimal sample preparation required, highly reproducible, easy to use and the possibility of making in situ measurements using on-the-go devices [108, 109]. Although spectroscopic techniques were widely used in the last 20 years to accurately predict different soil properties, the potential use of this technique for DSM has not yet been extensively explored [110].

The first examples of studies reporting the influence of moisture, particle size and chemical composition on soil spectral reflectance measured as albedo and soil color were done in 20s and 30s [108, 111–114]. Spectroscopic information can be retrieve in three ways: remote sensing platforms, such as satellites or airborne devices, laboratory spectroscopic equipments that allow the direct measurement of the soil sample, and proximal sensing portable devices that enable to get in-situ and on-the-go measurements on the soil surface [41, 60]. Spectroscopy works using remote sensing spectrometers are among the first studies applying soil spectroscopy, back in the 70s [41, 60, 115–118]. Plenty of research studies in the field soil spectroscopy were developed

since 90s. Due to the large amount of data generated by spectrometers and the complexity of the spectra, it is imperative to use computational procedures to analyze them [119]. The large amount of information held by the spectra, as well as recent advances in computation, instrument manufacturing, developments in multivariate statistics and the great number of potential applications in agriculture and soil science, produced an exponential increase in investigations that use spectroscopic methods [108, 120–123]. Spectroscopy was used to study physical, chemical and biological soil properties, such as aggregation, water content, particle size, carbon and organic matter content, CEC, macronutrients, micronutrients, mineralogy, electrical conductivity, pH and lime requirement, contaminants, microbial biomass, microbial respiration, microbial groups and enzymatic activities. Soriano-Disla *et al.* [124], Viscarra *et al.* [122], Stenberg *et al.* [125], Linker [119], Shi *et al.* [126] and Horta *et al.* [127] reviewed the applications of soil spectroscopy with numerous examples from the literature.

The most ambitious project in the soil spectroscopy field appeared in 2008, when a group of international researches began to developed a global spectral soil library in response to the growing interest on this technique [123, 128].

2.4.1 Theoretical basis of spectroscopic methods

Spectroscopy is defined as the set of techniques based on the interaction between matter and electromagnetic radiation. When a soil sample is irradiated with electromagnetic energy, the molecules that form it absorb this energy generating spectral energetic peaks that can be used to identify this molecules and functional groups.

Soil spectroscopy focuses in the visible, Near Infrared (NIR) and MIR ranges (Figure 2.3). From a quantic point of view, when energy from the visible region is imposed on a molecule it induces an electronic transition due to the raise of electrons to an excited electronic state. Infrared energy is no longer sufficient to produce electronic transitions, causing the transition between vibrational states. This transition originates a bending (changes into bond angle) or stretching (changes in the interatomic distance in the bond axis) vibration. Both vibration types can present symmetric, asymmetric, inside or out the plane modes as shows Figure 2.4. In the MIR range appears fundamental molecular vibrations, meanwhile the NIR reflects the overtones and combinations bands of the fundamental vibrations of the MIR spectrum. Overtones are due to transitions forbidden by the selection rules obtained after solving the Schrödinger equation for an harmonic oscillator. According to such selection rules

only $\Delta v = \pm 1$ is permitted. Combination bands are observed when two or more fundamental vibrations are excited simultaneously. Due to the broad and overlapping bands displayed in the NIR spectrum, it is more difficult to interpret than the corresponding MIR spectrum [125].

2.4.2 Spectral data acquirement

The acquirement of spectroscopic data is done with the aid of a spectrophotometer. Nowadays, especially in the MIR region, those that operate using the Fourier Transform are quite popular. This kind of spectroscopy is called FTIR (Fourier Transforms Infrared). A FTIR spectrophotometer includes an interferometer device composed by a system of mirrors (Figure 2.5). The most common one used is the Michelson interferometer, which consists in a beam splitter located between two perpendicular mirrors, one of which can move along an axis perpendicular to its plane [119]. In that way, the radiation beam from source splits into two beams and then these two radiation beams reach the detector with a delay or path-lenght difference due to the movement of the moving mirror. The two beams can undergo constructive interference, destructive interference or a combination of both, depending on the path-lenght difference [119]. After the process, the obtained interferogram is converted into a typical spectra applying the well-known Fourier Transform, a

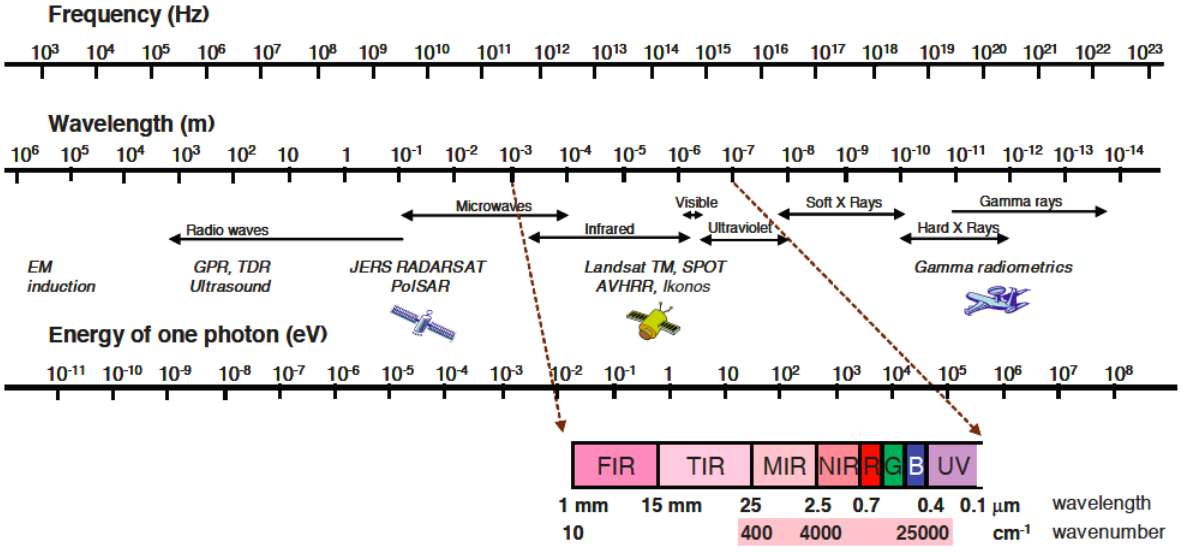


FIGURE 2.3: Electromagnetic spectrum.
(adapted from Viscarra Rossel et al. (2006) [122] and McBratney et al. (2003) [40])

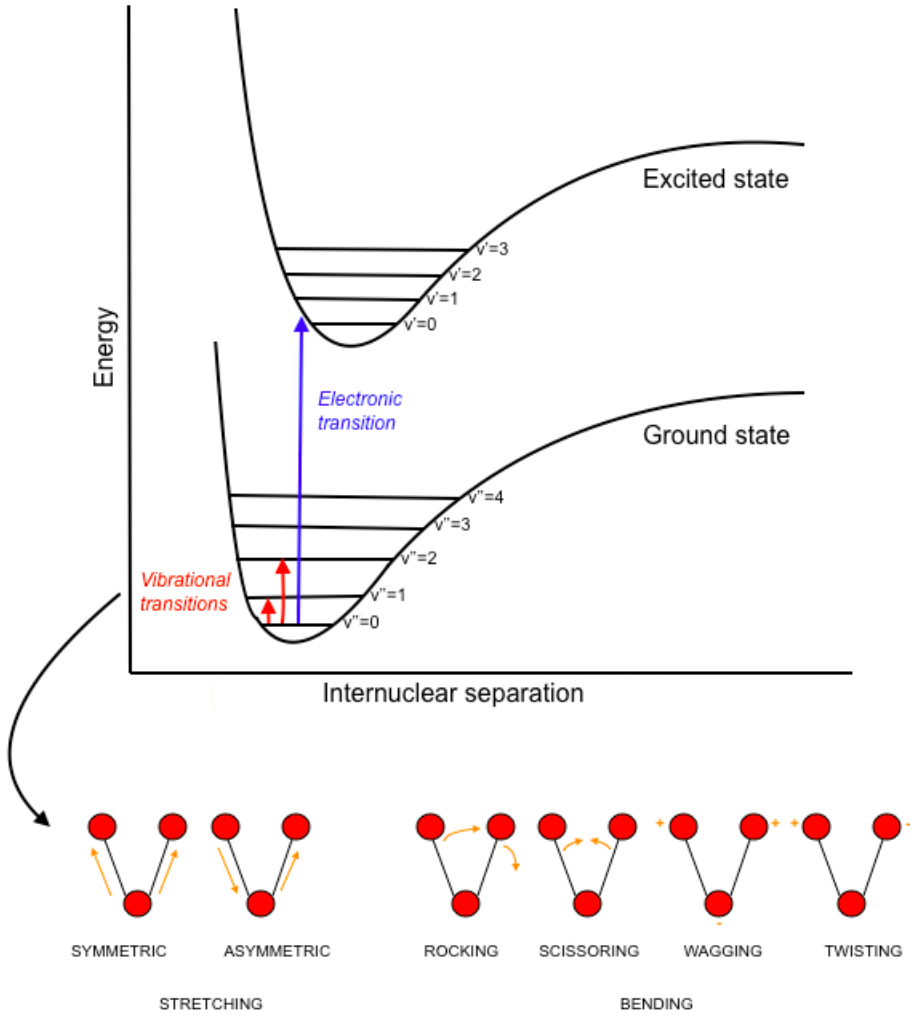


FIGURE 2.4: Electronic transitions in the visible and infrared ranges.

mathematical procedure that transforms the signal in the time domain into the frequency domain.

Spectrophotometers can include three different devices depending on the mode in which they operate: Diffuse Reflectance (DRIFT), Attenuated Total Reflectance (ATR) and Photoacoustic Spectroscopy (PAS) [119]. In DRIFT the radiation beam collides with a powdered sample mixed with KBr powder and the scattered light is reflected to the detector [130]. In ATR spectroscopy the radiation beam propagates through a crystal with a high refractive index that is in contact with the sample and the radiation not absorbed by the sample reaches the detector [119]. In PAS mode, the sample is placed in a sealed enclosure, that is purged with He to avoid atmospheric interferences, connected to a highly sensitive microphone which records the pressure waves that result from the local heating induced by the absorbed radiation.

2.4.3 Spectral pre-processing methods

Once the spectrum of a soil sample is recorded, it is common to apply a pre-processing method to correct for non-linearities, variations between samples, noise in the spectroscopic measurements or simply to highlight the more relevant spectral peaks [125, 131]. Pre-treatment methods can

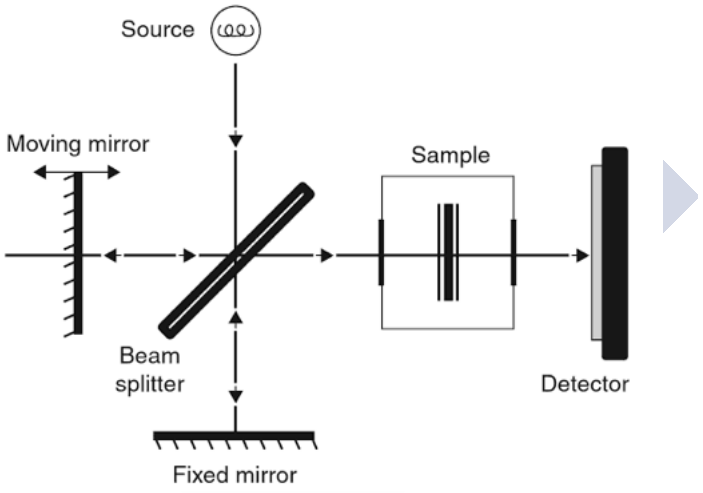


FIGURE 2.5: Schematic representation of FTIR spectrophotometer.
(extracted from Sun (2009) [129])

be classified in two groups: the algorithms that correct baseline effects and those that focus on scattering effects [132].

One on the most popular methods for removing baseline effects, resolve peak overlaping and enhance spectral features is derivative transformation. The first derivative removes only the baseline, while the second derivative removes both baseline and linear trend [132]. Derivatives tend to amplify noise and thus a smoothing method if often required to reduce it. Some commonly used smoothing methods to reduce noise in spectral signals include averaging spectra, moving average and median filters and the Savitzky-Golay (SG) transform [133]. Spectral derivatives can be calculated by obtaining the differences between two consecutive points, by smoothing/differentiating specified gap distance, or SG polynomial fitting [134, 135]. The last method is one of the most popular choices. SG is an averaging algorithm that fits a least squares polynomial to the data points, and then the value to be averaged is predicted from the polynomial [136]. Another well-known pre-processing method to correct baseline effects is Continuum Removal (CR). CR is a normalization method, which generates new spectral data by dividing the envelope curve of a continuum on raw reflectance spectra [137]. It is effective at isolating specific absorption features, and removing the effects of changing slopes and overall reflectance levels [138].

The success of a pre-treatment algorithm that pretends to remove scattering relies in how it can separate light scattering from light absorbed [134]. In the case of soil samples, the particle size distribution affects to the degree of scattering. A coarser structure increases the scatter and consequently the apparent absorbance increases. The simplest way to attempt linearization between absorbance and concentration, occurs when the measured reflectance (R) spectra is transformed to log 1/R units. Other similar transformations include the Kubelka-Munk and the Dahm equation [125]. However, there are more sophisticated methods that have proven to be effective. Among them, Multiplicative Scatter Correction (MSC) [139], Standard Normal Variate (SNV) [140], or Orthogonal Signal Correction (OSC) [141]. MSC corrects spectra according to a simple linear univariate fit to a standard spectrum. Alpha and beta coefficients are estimated by least squares regression using a standard spectrum. An average spectrum is required to correct the offset and multiplicative scatter in further individual spectra [142]. The same basic idea is used in SNV. This means that MSC and SNV are equivalent for most practical applications and the resultant spectra after the application of these algorithms are similar [132, 143]. The advantage of SNV is that, meanwhile MSC is used to normalize the mean of a group of samples, SNV uses only the spectrum of one sample. The concept behind OSC is similar too. In OSC pre-processing, the largest

variation in matrix X having zero correlation with the target matrix Y can be selectively removed from X [132].

Algorithms used to correct the scatter effect may be used alone or in combination with those that correct the baseline [144].

2.4.4 Influential spectroscopic bands for soil analysis

Spectra contain information on the organic and mineral composition of soil, the amount of water present, its particle size and its color [108]. Figure 2.6 shows typical spectra of soil samples in MIR, NIR and visible ranges. MIR spectroscopy is particularly well suited for the analysis of soil organic matter and mineral composition because absorption bands associated with both organic functional groups and soil minerals can be readily identified in the spectra. At higher wavenumbers (4000-2300 cm^{-1}) appear the stretching vibrations of heteroatoms bonding to hydrogen, at medium wavenumbers (2300-1200 cm^{-1}) the vibrations associated to Triple and Double Bonds (TB, DB) and at lower wavenumbers (1200-400 cm^{-1}) it is observed the fingerprint region. Although bands in NIR are broad and difficult to interpret than those of MIR, this region contains useful information on organic and inorganic materials in the soil. Spectroscopic devices often

combine visible and NIR ranges (VNIR). Absorptions in the visible region are mostly due to electronic excitations and are primarily associated with the darkness of soil organic matter and to chromophores of iron containing minerals [145]. Many authors described the bands that appear in the spectra of a soil sample and relate them with compounds and functional groups. Table 2.1 summarizes the main peaks described in bibliography, compiled from Viscarra *et al.* [122], Stenberg *et al.* [125], Viscarra and Behrens [146], Madari *et al.* [147] and Yang *et al.* [148].

Variations into the relative position of the complete spectra or into the angle and convexity of some spectral peaks indicates the prevalence of certain soil components. A shift towards a high energy indicates a high content in iron oxides, organic matter and a more weathered soil. The analysis of a MIR spectrum requires an exhaustive work consisting in the identification of all the peaks listed in the previous table for soil compound identification. However, the VNIR spectrum of a soil sample is more easy to interpret because of the limited number of bands. Bands in the visible range indicate the presence of iron oxides. Soils richer in goethite show narrower concavity shape and less intense bands in this region in comparison with soils with more hematite. The convexity observed between 800-1200 nm indicates the presence of iron oxides. This band often does not appear in samples with high organic matter content due to the masking effect that it

produces. In fact, samples with a high organic matter content even could display a broad band that extends from the visible range to 1800 nm. The intensity of the peaks at 1420, 1920 and 2220 nm, is related with the presence of clay minerals. More intensity is associated to a 2:1 mineralogy, while lower intensity is due to 1:1 mineralogies [125, 149].

TABLE 2.1: Spectroscopic bands for fundamental MIR absorptions and their overtones and combinations in VNIR.

MIR band (cm^{-1})	**VNIR band (nm)**	**Assignment**
	434,480,650, 920	Electronic transitions of goethite
	404,444,529, 650,884	Electronic transitions of haematite
3695,3620	1395,1415	Asymmetic-symmetric O-H from kaolin doublet
3620	2206,2340, 2450	O-H stretching of illite
3620	2206	O-H stretching of smectite
3380		O-H stretching of phenolic compounds
3400-3300		O-H stretching (H bonded OH groups)
3484	1380,1135, 940	O-H symmetric stretching from water
3278	1915	O-H asymmetric stretching from water
3300	1500,1000, 751	N-H stretching of amines
3030	1650,1100, 825	Aromatic C-H stretching
2940-2900		Aliphatic C-H stretching
2930,2850	1706,1754, 1138,1170, 853,877	Alkyl asymmetic-symmetric C-H stretching
2600		O-H stretching of H-bonded -COOH
1725-1720	1930,1449	C=O stretching of carboxylic acids and ketones
1660-1630	2033,1524	C=O stretching of amide groups (amide I band), quinine C=O / C=O of H-bonded conjugated ketones

Continued on next page

Table 2.1 – *Continued from previous page*

MIR band (cm^{-1})	**VNIR band (nm)**	**Assignment**
1645	1455	H-O-H bending from water
1620-1600		Aromatic C=C stretching / asymmetric -COO stretching
1610	2060	N-H stretching of amine
1590-1517		COO- symmetric stretching, N-H deformation+C=N stretching (amide II band)
1525		Aromatic C=C stretching
1460-1450	2275,1706	Aliphatic C-H
1415	2500,2336	CO_3^{2-} asymmetric stretching of carbonates
1400-1390		OH deformation and C-O stretching of phenolic OH, C-H deformation of CH_2 and CH_3 groups, COO- asymmetric stretching
1350		Symmetric COO- stretching / -CH bending of aliphatics
1270	1971	C-OH stretching of phenolics compounds
1280-1200		C-O stretching and OH deformation of COOH, C-O stretching of aryl ethers
1225		C-O stretching and OH deformation of COOH
1170-950	2137	C-O stretching of polysaccharides
1170		C-OH stretching of aliphatic OH, C-C stretching of aliphatic groups
1100-1000		Si-O of quartz
1063		CO_3^{2-} symmetric stretching of carbonates
1050	2381	C-O stretching of carbohydrates
915	2230	Al-OH bending of smectite
915	2160,2208	Al-OH bending of kaolin
885		AlFe-OH stretching of smectite
879		CO_3^{2-} out of plane bending of carbonates
830		Aromatic CH out of plane bending
775		Aromatic CH out of plane bending
680		CO_3^{2-} in plane bending of carbonates

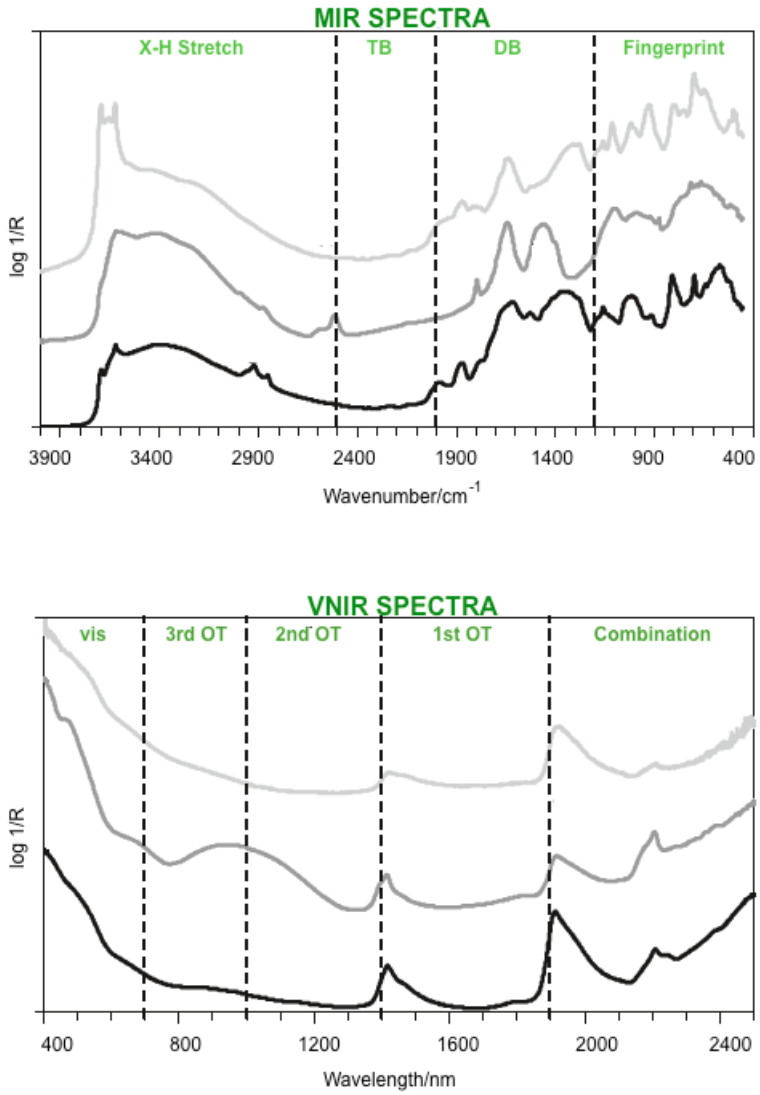


FIGURE 2.6: Main spectroscopic bands located in the visible, near and mid-infrared ranges.
(adapted from Viscarra Rossel and McBratney (2008) [145])

2.4.5 Statistical procedures

Some of the mathematical algorithms used for the spatial prediction of soil properties also showed predictive potential for the quantification of soil properties from spectroscopic data. Early studies of soil spectroscopy used models based only on specific wavelengths that were selected using variable selection techniques such as Step-wise Multiple Linear Regression (SMLR) [120, 150]. With time, some regression techniques such as PCR and PLS, became more popular because of they allow the compression of full spectrum data. Although PCR and PLS are in most cases similar, PLS is often preferred because it relates the response and predictor variables. In such manner, the PLS models explain more of the variance in the response with fewer components, being easily interpretable [60]. Recently, machine learning techniques were also applied to predict soil properties from spectroscopic data. However, linear models remain the most used by far. Some examples of machine learning applications include MARS, ANN and tree based models like RF, Boosted Regression Trees (BRT) or Radial Basis Function Networks (RBFN) [145, 151–154]. Bellon-Maurel and McBratney [155], Linker [119], Reeves [156], Stenberg *et al.* [125] and Viscarra *et al.* [122] reviewed the algorithms commonly used for soil properties prediction, including several examples of SOC quantification.

2.5 Performance of the statistical models

All the statistical models included in this thesis were developed within the statistical environment R 3.1.1 [177], available for download at http://www.R-project.org/. The Goodness Of Fit (GOF) of the models was evaluated by means of the R-squared, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) parameters, which were calculated using the following equations:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2.3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.4}$$

where n represents the number of observations, $\bar{y}$ is the mean of observed values, y_i is the observed value and $\hat{y}_i$ is the predicted value. R-squared values near to 1 and low values of RMSE and MAE are indicators of a good predictive model.

The maps here presented were produced using the Geographic Information System QGIS v.2.12 software (http://www.qgis.org/es/site/).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

STUDY CASE I: Galicia, NW Spain

3.1 Background and study area

In a previous study conducted in our research group, the distribution of SOC content in topsoils from the autonomous region of Galicia (NW Spain) was modeled for first time [157]. The present study is a continuation of this approach and it aims to further explore the potential of infrared spectroscopy for predict and map such SOC contents.

A total of 221 samples from upper soil horizons (0-30 cm depth) were collected with a soil auger along the study area. Each soil sample is a composite of 8 subsamples taken within

a radius of 20 m from a central point, which represents the sampling location, within the same land use. Each sampling location was recorded using a GPS receiver with an accuracy of $\pm$ 5% m. The sampling scheme follows mainly a regular grid of 16x16 km covering 3/4 of the area (n = 124). In addition, 97 samples available from previous studies [158] were also included in the study. This sampling scheme covers most of the climatic, land use and geological diversity of the region (Figure 3.1).

The study region is a transitional climatic area from oceanic hyper-humid to sub-humid conditions with a climate described as temperate subtropic with wet winters [159]. The dominant soil temperature regime is mesic [160]. There is a WE gradient of temperature, with the coldest areas corresponding to those locations in eastern longitudes. The dominant soil moisture regime is udic, with a transition towards xeric in the SE of the region and in eroded areas where shallow soils predominate. The spatial range of the climatic variables presented in this area is shown in Table 3.1. It includes values of mean annual temperature (T), mean annual accumulated precipitation (P), mean annual potential evapotranspiration (ET_0), water balance (P-ET_0) and ombrothermic indices (Ios, Ios_2, Ios_3, Ios_4) [161]. There is a high geological diversity, with granites, shales and slates being the dominant rock types. Soil texture varies from sandy in soils developed over granites to loamy in soils developed from shales and slates. Soils show moderate depth, have low

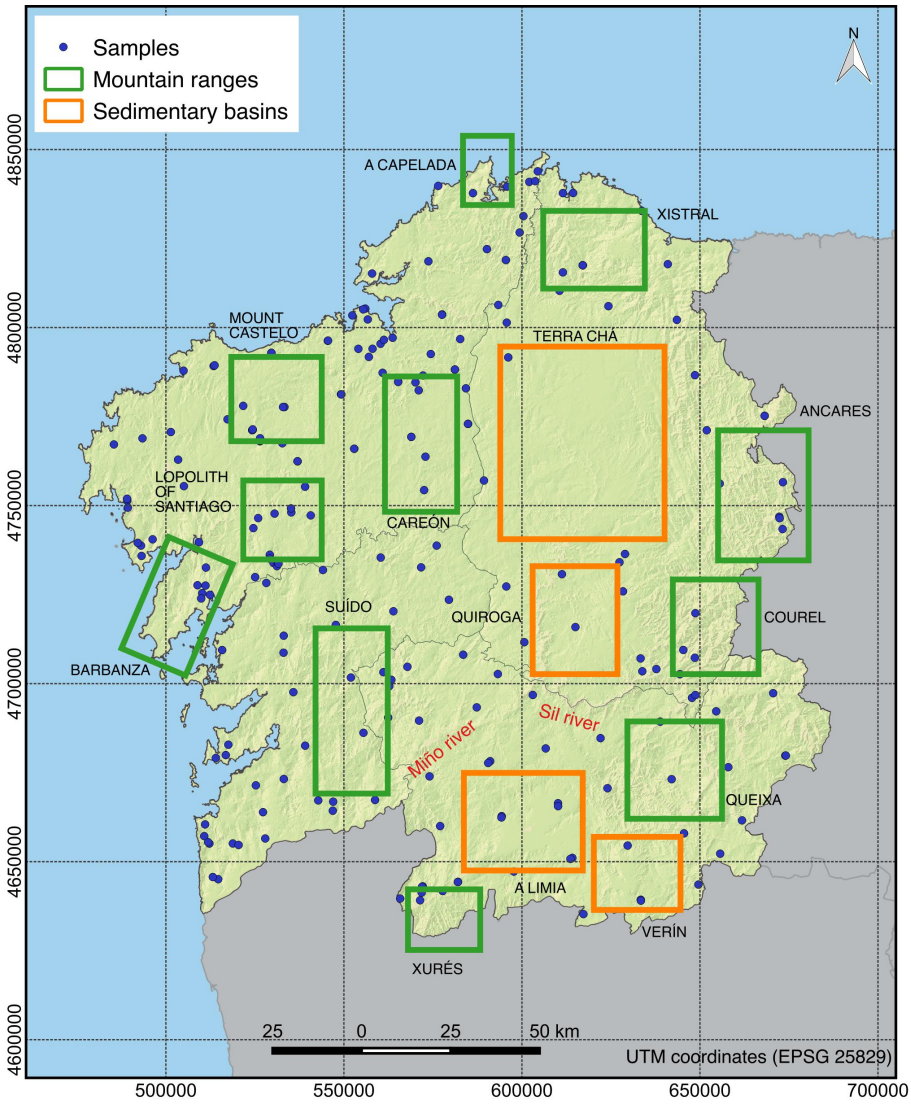


FIGURE 3.1: Geographic location of samples collected in Galicia and main mountain ranges and sedimentary basins in the study area.

fertility, pH values of 4.5-5.5, low effective CEC, low alkaline cations contents and high amounts of exchangeable aluminium [162–165]. Scrubland is the dominant land use type, covering about 30.06% of the territory, followed by cultures, deciduous and coniferous forests (Table 3.2, www.ign.es).

TABLE 3.1: Summary of the climatic variables in the study area.

Variables	**Min-Max**	**Mean**
Temperature - T (°C)	6.6-14.7	12.2
Accumulated precipitation - P (mm)	589-1809	1245
Potential evapotranspiration - ET_0 (mm)	485-814	688
Water balance - P-ET_0 (mm)	-6.1-1094	557
Annual ombrothermic Index - Ios	4.3-20.4	9.0
Ombrothermic index June-July - Ios_2	0.7-3.6	1.6
Ombrothermic index June-August - Ios_3	1.0-4.5	2.2
Ombrothermic index May-August - Ios_4	1.6-6.4	3.1
Index of continentality - Ic	9.8-14.8	12.0
Thermicity index - It	52-342	255

TABLE 3.2: Main land use occupation in the study area.

Land use type	**Total area (%)**
Coniferous	11.82
Cultures	21.98
Eucaliptus	7.63
Deciduous	15.44
Shrubs	30.06
Prairies	5.25
Vineyards	0.68

3.2 Geochemical data

Samples were air-dried and sieved through a 2 mm mesh before analysis. The Walkley-Black method was used to determine the SOC content. This method consists in an oxidative digestion of 0.2 g of soil sample using a solution of $K_2Cr_2O_7$ in H_2SO_4 (98%). $Cr_2O_7^{2-}$ excess, in 5 mL of the digested sample, is then evaluated by titration with $Fe(NH_4)_2(SO_4)_2$ 0.5 N and the SOC content (%) is calculated by determining the amount of $Cr_2O_7^{2-}$ reduced during the reaction with the soil sample [166].

Total C content was also measured by combustion using a LECO carbon analyzer (Model CHNS-932, LECO Corp., St Joseph, MI). This measurement was used to evaluate the fraction of organic carbon not captured by the Walkley-Black method that could lead to an underestimation of the SOC stored in this area.

3.3 Spectroscopic data

Soil samples were finely ground using a Retsch MM 301 Mixer Mill (model 01-462-0201). FTIR-ATR spectra were obtained using a Varian 670-IR spectrometer (Varian Inc., Santa Clara, CA), attached to a thermal detector - deuterated L-alanine doped triglycine sulphate (DLaTGS) and an ATR device with a single-reflection diamond crystal (Pike, Madison, WI). Each

soil sample was placed in contact with the ATR crystal and an infrared radiation, with an angle of 45 degrees, was applied through it. The FTIR-ATR spectrum was recorded with a resolution of 4 cm^{-1}. All FTIR-ATR spectra were baseline corrected to avoid bias in the spectroscopic signal due to scattering, reflection, temperature, concentration or instrument anomalies [167].

3.4 Environmental variables

A total of 31 GIS-based raster maps, at 25 m resolution, were used as environmental covariates to model the spatial distribution of the FTIR-ATR data over the study area. These maps include information on climatic conditions, parent material and land use classes (Figure 3.2).

Climate: We used raster maps of mean annual and monthly temperature (T), accumulated precipitation (P), potential evapotranspiration (ET_0), water balance (P-ET_0), annual ombrothermic index (Ios), ombrothermic indices for June-July (Ios_2), June-August (Ios_3), May-August (Ios_4), index of continentality (Ic) and thermicity index (It) to describe the spatial variability of climate in the region.

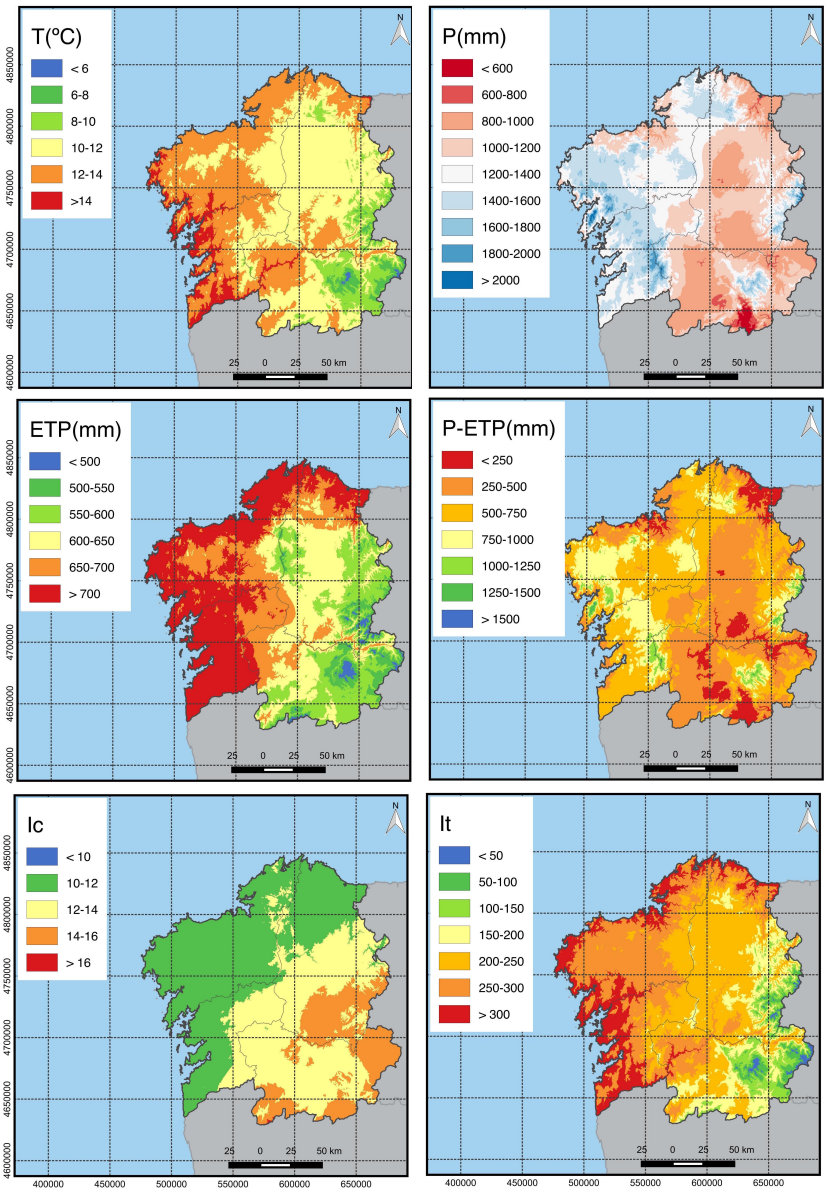


FIGURE 3.2: Environmental variables used for mapping purposes in Galicia.

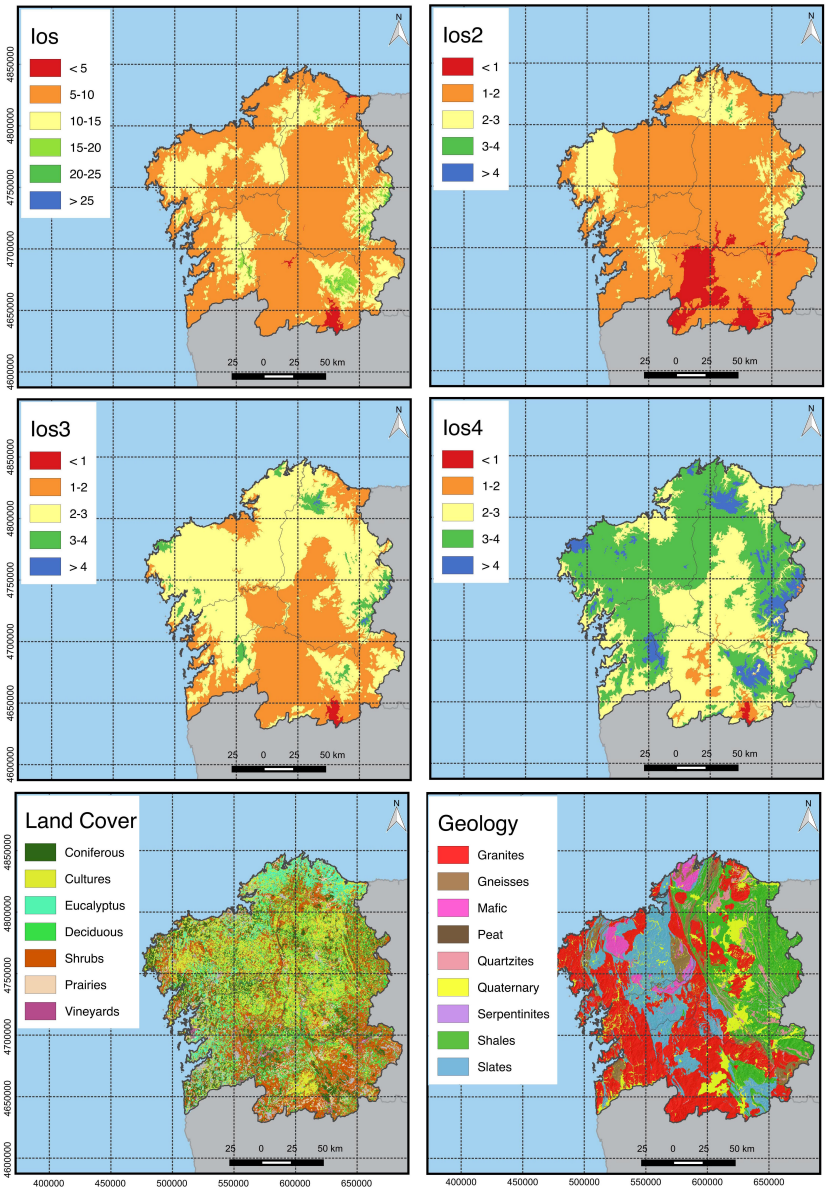


FIGURE 2 (CONT.): Environmental variables used for mapping purposes in Galicia.

Geology: The geologic information within 97 vector maps at 1:50000 scale from the Spanish Institute of Geology (IGME) were synthesized into 9 geologic categories which include granitic rocks, gneisses, mafic rocks, peat soils, quartzites, quaternary deposits, serpentinites, shales and slates. These features were transformed to 9 binary rasters, at the same resolution as that of the climatic rasters, representing the presence/absence of each geologic class in the study area.

Land Cover: Seven binary maps, showing the presence/absence of the land use types within our field samples, were created from the land use vector maps (1:25000 scale) available at the National Geographic Institute of Spain (http://www.ign.es). These land uses correspond to coniferous, deciduous and eucalyptus forest, cultures, shrubs, prairies and vineyards.

3.5 Modelling procedures

Figure 3.3 summarizes the statistical steps used in this study. Walkley-Black SOC content (%) and FTIR-ATR spectra were determined for each soil sample. RF was used to determine the hierarchical relationships between SOC and FTIR-ATR data in order to identify the spectral signatures (spectral bands) that mostly explain the variability in the SOC. Then, a PLS model

was used to determine the statistical relationships between the selected bands and a number of raster maps of climatic, geologic and land use variables in order to create spatial models showing the distribution of the influential absorbance bands along the study area. Finally, MLR was used to model the relationships between the SOC and the influential bands detected and derive a predictive map of the SOC content in the region.

Random Forest (RF): Ensemble learning algorithms like boosting [168] and bagging [169] are widely used regression tree methods [170, 171]. RF is a machine learning method generating multiple randomized Classification And Regression Trees (CART) which are then averaged to create an unique model with higher predictive performance than each single CART alone [92, 172]. Each CART is calibrated on a random set of bootstrap samples from the original data and provides a robust estimation for the remaining data, the so-called Out-Of-Bag (OOB) samples. The best split at each node of the tree is created using a random subset of predictors. During the process, an estimation of the OOB samples is produced and the Mean Square Error (MSE_{OOB}) is calculated by aggregating the predictions from all trees according to the following expression:

$$MSE_{OOB} = n^{(-1)} \sum_{i=1}^{n} (z_i - \hat{z}_i^{OOB})^2 \tag{3.1}$$

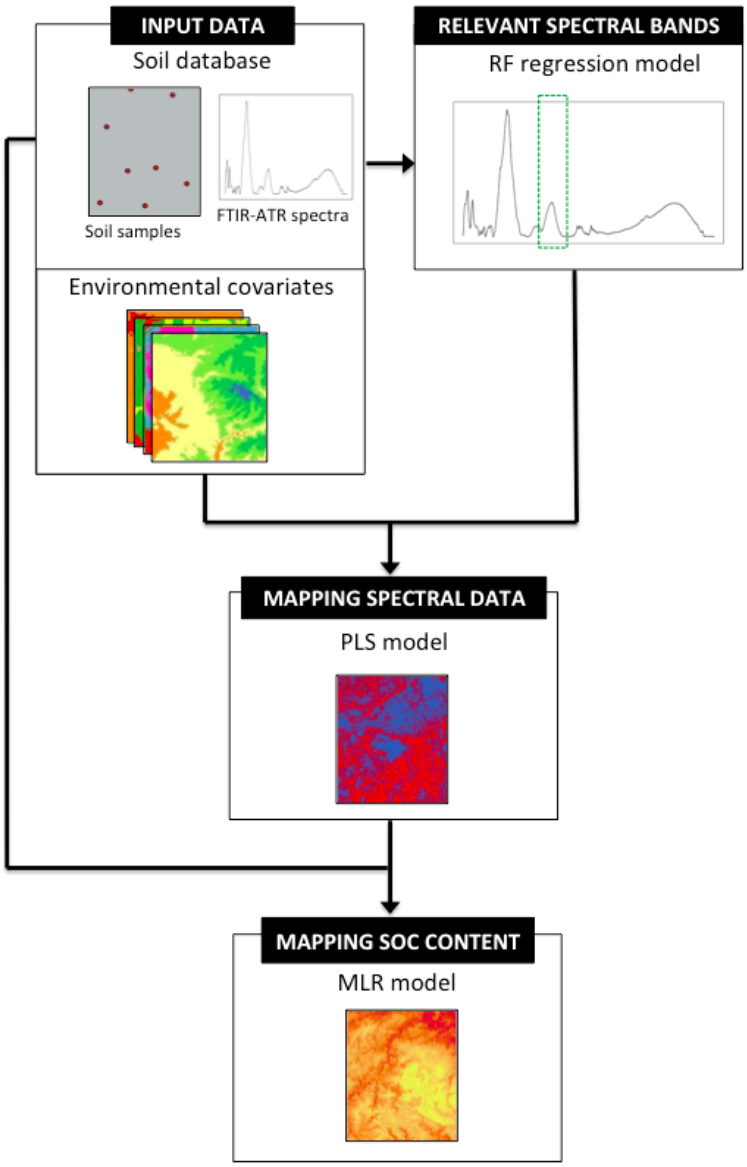


FIGURE 3.3: Statistical framework used for mapping SOC content in Galician topsoils.

where n is number of trees and $\hat{z}_i^{OOB}$ is the average of all OOB predictions for the z_i observations [92, 172, 173]. The GOF of the model can be directly obtained from the MSE_{OOB} and the total variance of the response variable (var_z) as the percentage of explained variance (var_{ex}) [92]:

$$var_{ex} = 1 - (MSE_{OOB}/var_z) \tag{3.2}$$

RF presents a series of advantages over other predictive methodologies, such as no overfitting, higher prediction performance, low correlation of the individual trees, low bias, low variance due to averaging over a large number of trees and robust error estimates by using the OOB dataset. In addition, the RF algorithm can handle either continuous or categorical auxiliary variables [72, 92]. Further details of the RF procedure are available in [94, 174].

Multiple Linear Regression (MLR): MLR is one of the most commonly used methods for modeling the relationship between a single response variable ($\hat{Y}$) and a set of predictor variables (X_1,..., X_n) according to the equation:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{3.3}$$

where β_0 represents the intercept and the β_1,...,β_n terms are coefficients of the regression curve.

Partial Least Squares Regression (PLS): PLS is a robust multivariate linear regression technique for the analysis and modeling of noisy and highly correlated data [175] which combines features of Principal Component Analysis (PCA) and MLR. PLS has a number of advantages in comparison to other statistical methods used for similar purposes such as PCR: i) it reduces the dimensionality of the independent variables by creating a new set of orthogonal parameters, the so-called latent variables, which are linear combinations of the original proxies; ii) these latent variables maximize the amount of variance accounted for the dependent variable; and iii) since the new latent variables are orthogonal, the model classification and predictive ability of PLS are not affected by multicollinearity [122, 176]. PLS summarizes the information of the data matrices of covariates (X) and dependent variables (Y), and relate them through a linear model. The covariate matrix X is decomposed into a product of a matrix of scores T and a matrix of loadings P as follows:

$$X = TP^T \tag{3.4}$$

Then, the scores values are used to predict Y values as:

$$Y = TBC^T \tag{3.5}$$

where B represents a matrix with the regression weights as

diagonal elements and C is the weight matrix of the covariates. In this study, the performance of the PLS models was estimated by cross-validation. More details about the PLS technique can be found in [88].

3.6 Use of infrared data for mapping

Random Forest analysis was used to ascertain the FTIR-ATR bands with higher influence in the prediction of SOC. The RF model was created upon 1000 regression trees, using at each time a random selection of bands as independent variables. During the process, the Mean Squared Error (MSE) and the bands selected in each regression tree were recorded. The increase in MSE due to the exclusion of each band within the regression trees was used to identify the relevant ones to predict SOC. The increase in MSE is calculated as the average increase in squared residuals of the test set when the explanatory variable is randomly permuted. When a given variable has little predictive power, its permutation will not cause substantial difference in model residuals, therefore, for each variable, a high increase in MSE when the variable is excluded indicates that the variable of interest is a good predictive parameter [178].

The RF model here created presents a high predictive performance ($R^2 = 0.95$, RMSE = 1.39, MAE = 1.01) and the

increase of MSE shows that the model is highly dependent on the information contained in bands at 2127, 1697, 1695, 1693, 1691, 1689 1302, 1298, 1296 and 1294 cm^{-1} (Figure 3.4). These bands provide the most relevant information to predict SOC from FTIR-ATR data. The band placed at 1697 cm^{-1} showed the highest increase of MSE when excluded within the Random Forest model. We calculated the coefficients of correlation between the above selected bands to check their statistical independency (Table 3.3).

TABLE 3.3: Correlation coefficients (r) between the band at 1697 cm^{-1} and the remaining influential bands identified after RF.

$\tilde{\nu}$ **(cm^{-1})**	2127	1695	1693	1691	1689	1302	1298	1296	1294
r	0.791	0.999	0.999	0.998	0.996	0.849	0.839	0.833	0.828

The analysis shows that the band at 1697 cm^{-1} was highly correlated to the remaining bands. The higher correlation coefficients correspond to proximal bands placed around 1697 cm^{-1}, indicating that absorbances in these bands probably correspond to vibrations of the same functional groups in the sample. Thus, we retained the band at 1697 cm^{-1} as the best-uncorrelated spectroscopic band able to predict SOC by linear regression (Figure 3.4). According to bibliographic sources this peak corresponds to the stretching vibration of C=O bond from aldehydes, ketones and carboxylic acids in hydrophobic and hydrophilic compounds of soil organic matter [146, 179–182].

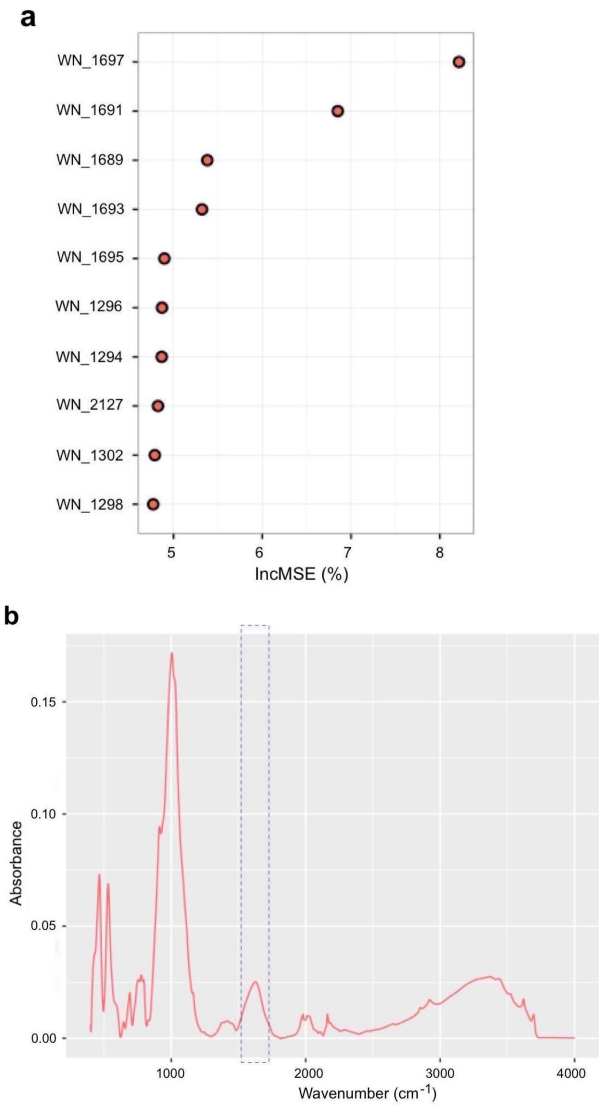


FIGURE 3.4: **RF analysis for identify the relevant spectral bands. a,** Variable importance graph from RF. **b,** Mean FTIR-ATR spectra. The dotted region represents the relevant spectral band identified from RF.

3.7 Factors influencing the amount of SOC

We used a PLS regression model to relate the values of the selected absorbance band to the environmental information stored as raster maps in order to generalize the spectral information obtained at sample scale to the whole study area. The PLS analysis showed that most of the variability in the selected band can be explained by a linear combination of three Latent Variables (LV), obtaining a good predictive performance (R^2 = 0.70, RSME = 0.57, MAE = 0.32). Table 3.4 shows that most of the variability is explained by LV_1, accounting for 55% of variance in the spectroscopic signal at 1697 cm^{-1}. LV_1 is highly correlated to precipitation (P) and water balance variables (P-ET_0, Ios, Ios_2, Ios_3, Ios_4). This LV was interpreted as the accumulation of SOC due to water availability. The correlation coefficient of LV_1 is moderately high and positive for peat soils, confirming that this latent variable is linked to the accumulation of SOC due to climatic factors. The correlation coefficients for the remaining LV are low, only explaining 14 and 2% of the variability in SOC respectively. LV_2 is negatively correlated to the presence of slates while LV_3 is negatively correlated to temperature related variables, which indicates a negative effect of high values of T, ET_0 and It on the accumulation of SOC. The positive coefficient for Ic indicates that

accumulation of SOC is positively affected by continental climates, as may occur in the SE of Galicia. Recent studies relating processes of accumulation/mineralization of SOC in Galicia [157] demonstrated that high SOC is related to high precipitation rates and water balance variables (expressed as Io indices). This spectroscopic signal is also positively correlated to both precipitation and water balance (Table 3.4), indicating that the accumulation of organic matter in soils in Galicia is mainly driven by the availability of water within the soil.

The map relative to the distribution of the spectroscopic signal at 1697 cm^{-1} (Figure 3.5) shows that high values of absorbance are mainly expected at locations with high values of precipitation rates and ombrothermic balances.

3.8 Spatial distribution of SOC content

The relationship between the SOC measurements obtained by the Walkley-Black method (SOC_{WB}) and the total carbon measured by combustion (C_{LECO}) was modeled using a linear regression model (Equation 3.6). The r-squared value (r^2 = 0.88) showed that in our samples almost all the carbon is

TABLE 3.4: Correlation coefficients between spectroscopic data ($\tilde{\nu}_1 = 1697$ cm^{-1}) and the environmental covariates.

Parameters	**LV$_1$**	**LV$_2$**	**LV$_3$**
T	-0.46	0.30	**-0.76**
P	**0.80**	-0.17	-0.47
ET_0	-0.09	0.42	**-0.83**
P-ET_0	**0.86**	-0.32	-0.22
Ic	-0.45	-0.38	**0.63**
It	-0.25	0.38	**-0.84**
Ios	**0.93**	-0.21	0.09
Ios_2	**0.93**	0.03	-0.16
Ios_3	**0.96**	-0.03	-0.07
Ios_4	**0.98**	-0.13	0.01
Coniferous	-0.09	0.18	-0.14
Cultures	-0.06	-0.23	-0.07
Eucaliptus	0.00	0.16	-0.28
Deciduous	-0.16	-0.17	0.08
Shrubs	0.36	0.07	0.29
Prairies	-0.10	-0.01	0.11
Vineyards	-0.11	0.04	-0.03
Gneiss	0.01	-0.04	-0.20
Granites	-0.12	0.06	-0.26
Mafic	0.22	-0.09	-0.09
Quartzites	0.02	-0.12	0.19
Quaternary	-0.14	-0.01	0.27
Serpentinites	0.03	-0.08	-0.05
Shales	-0.19	0.29	0.01
Slates	0.01	**-0.43**	0.16
Peat	**0.62**	0.51	0.42
1697 cm^{-1}	0.74	0.37	0.13

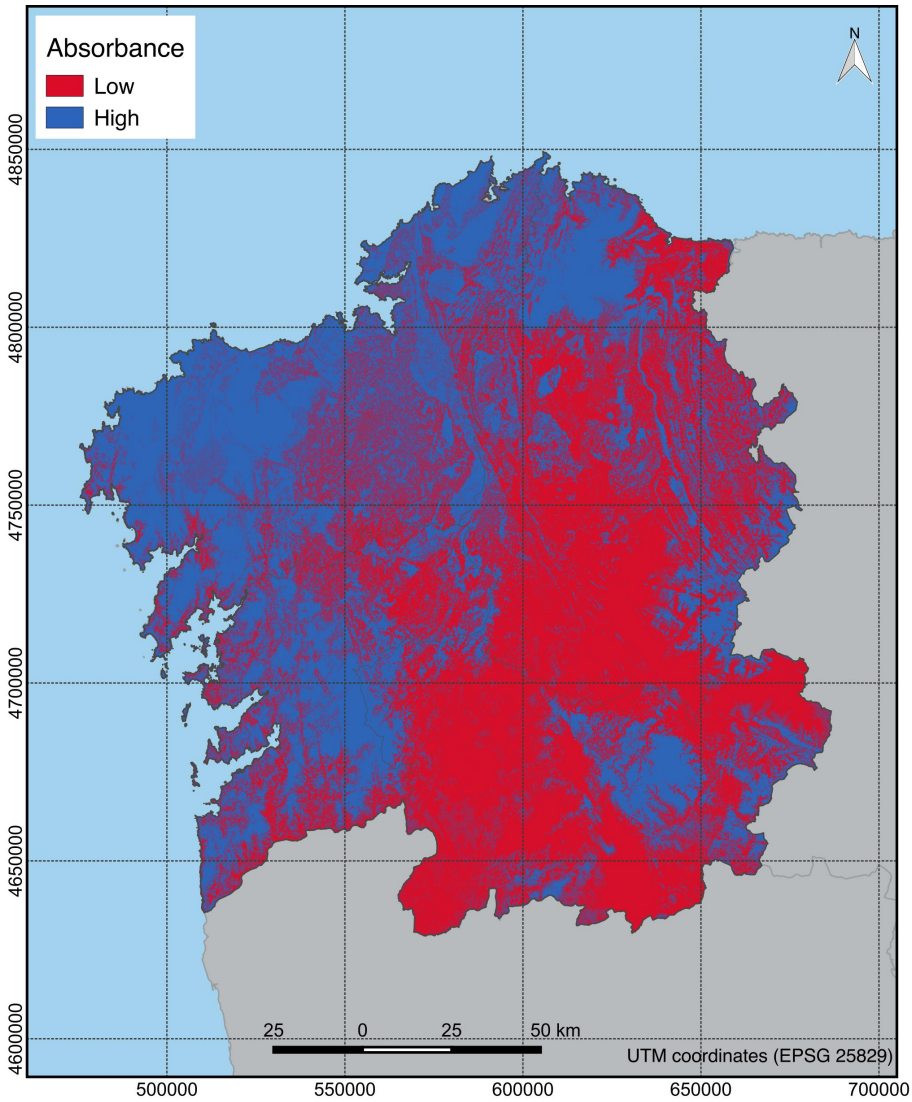


FIGURE 3.5: Map of the spatial distribution of the estimated FTIR-ATR data at 1697 cm^{-1}.

present as oxidizable carbon, thus the SOC can be properly evaluated by the Walkley-Black method.

$$SOC_{WB} = 0.8336 C_{LECO} + 0.279 \tag{3.6}$$

The mean SOC content in our dataset of samples is 7.4%, with a range of variation from 46.2% in peat to 1.7% in cultivated soils developed on shales. Most of the SOC content in the dataset is lower than 10%. Figure 3.6 shows the distribution of SOC by types of geology and land use. Regarding geology, the highest SOC contents were found in peat. Soil developed on mafic rocks, quartzites and serpentinites show SOC contents generally higher than the mean SOC measured in the region. This may indicate that these parent materials positively influence the accumulation of SOC. Lower SOC contents were found in soils developed over shales and quaternary materials. Regarding land use, the higher SOC contents were found in areas with shrub vegetation while minimum SOC contents were measured in vineyard soils. ANOVA tests were performed to check statistical differences in SOC among land uses and geologic classes. These tests indicate that there are significant differences in the SOC content between the land use types ($P < 0.05$), while, for geology types, statistical differences in SOC was only found for peat ($P = 1.855\ 10^{-14}$).

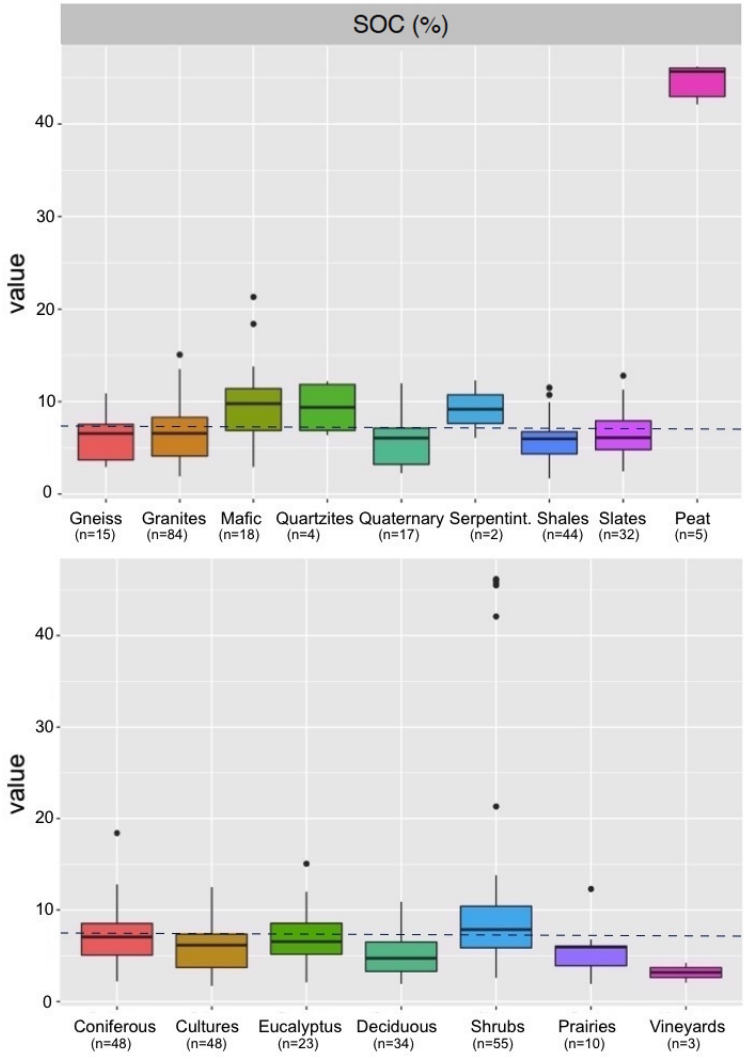


FIGURE 3.6: Boxplots relating SOC content to geology and land use types in samples from Galicia. The dotted line represents the mean SOC value.

The original dataset of samples was randomly split into a calibration dataset, containing 70% of the original samples and used to create the model, and a validation dataset (30%) used to determine its performance. The linear regression model here created follows the relationship:

$$SOC = -4.23 + 1115\tilde{\nu}_1 \tag{3.7}$$

where $\tilde{\nu}_1 = 1697$ cm^{-1}.

The validation of the regression model, upon the validation dataset, shows a high predictive performance (r^2 = 0.88, RMSE = 2.14, MAE = 1.64) indicating that SOC of the analyzed samples can be easily estimated by using the information contained at 1697 cm^{-1} of the FTIR-ATR spectra. The final map of SOC content derived from FTIR-ATR data (Figure 3.7) was created using this relationship and the map of absorbance at 1697 cm^{-1} as the independent spatial proxy. The map shows that the higher SOC contents are located in areas such as the mountain ranges in the North (Xistral and A Capelada), East (Courel and Ancares), West (Barbanza and Suído) and Southeast (Queixa) of Galicia. Slightly high SOC contents were also estimated for areas with moderately high precipitation rates, such as Mount Castelo, Careón, or the lopolith of Santiago, but with presence of mafic rocks.

The weathering products from these materials contain amorphous phases of iron and aluminium oxy-hydroxides that can form stable organo-mineral compounds that lead to an accumulation of SOC [183–186]. Nevertheless, the weights of both parent material and land use in the SOC content are relatively low, in relation to that of precipitation. The lower SOC values occur in areas associated to low precipitation rates and higher temperatures such as those located along the sea coast and regions in the SE of the study area (Quiroga, A Limia, Verín and the valleys of Sil and Miño rivers.

3.9 Uncertainty of SOC estimations

Since these results are similar to those obtained in the previous study conducted in our research group [157], we tested the performance of the methodology here proposed in relation to the results obtained from the methodology used in the previous study. In order to be comparable, we used the same set of samples and covariates to estimate SOC from measurements obtained using the standard Walkley-Black method within a PLS approach with 4 LVs. The spatial differences in the prediction from both methods ($SOC_{(FTIR\text{-}ATR)}$ - $SOC_{(WB)}$) (Figure 3.8) range between -8.95 and +3.84% SOC, with a mean value of -1.71% (SD = 1.13). The differences do not exceed ±2% SOC in 65% of the total

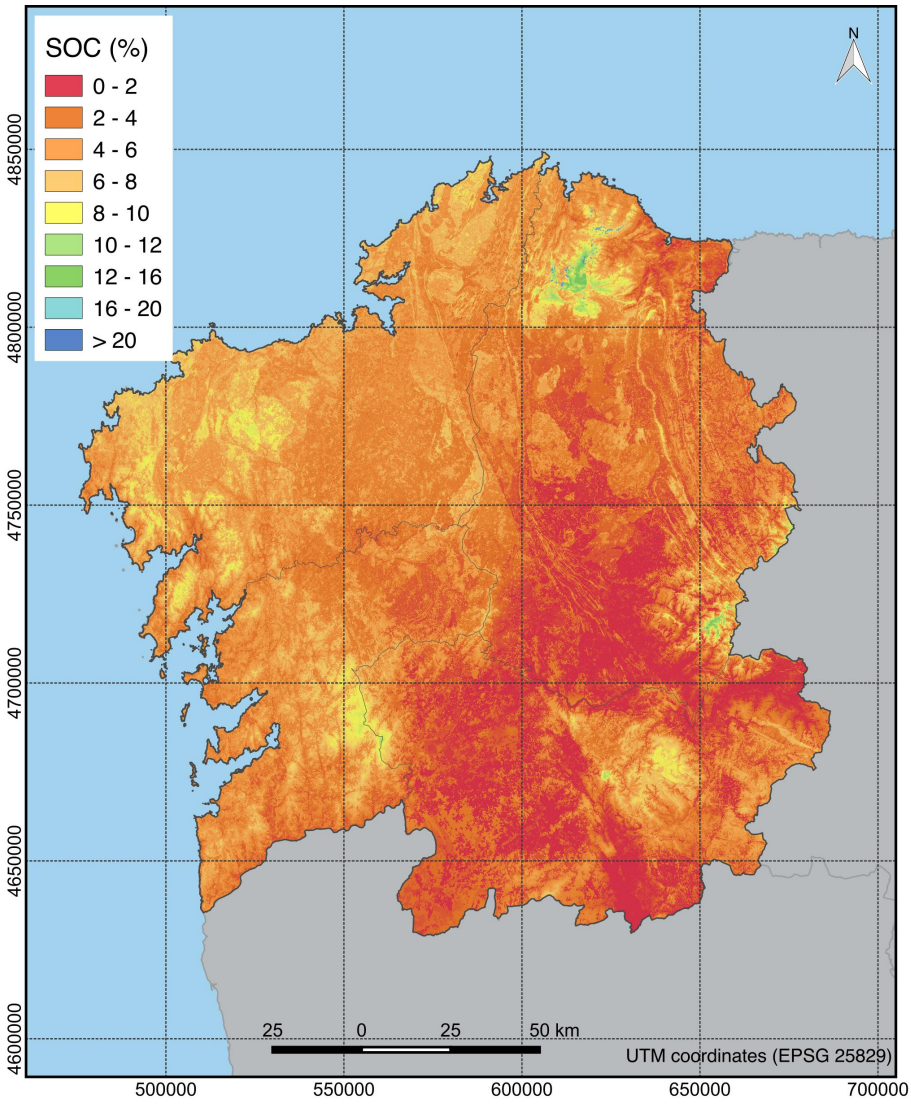


FIGURE 3.7: Map of predicted SOC content in Galician topsoils.

surface. For mountain areas with high precipitation rates and water balance variables (Xistral, Ancares, Courel and Queixa), the estimations obtained from the FTIR-ATR data are lower than those obtained using Walkley-Black data, as observed by negative values in the map of the differences. For areas with high precipitation rates and high temperature variables such as the coastal region and the valleys of the rivers Miño and Sil, and in peat areas in the northern sierras, the model of SOC using FTIR-ATR data tends to produce higher estimates of SOC.

The validation results show a slightly higher predictive performance of the model using Walkley-Black SOC measurements (R^2 = 0.81, RMSE = 2.41) than the model here developed (R^2 = 0.74, RMSE = 2.81). However, the differences in performance are low and, thus, it is feasible to map SOC using FTIR-ATR data obtaining similar results to those obtained by using conventional SOC measurements.

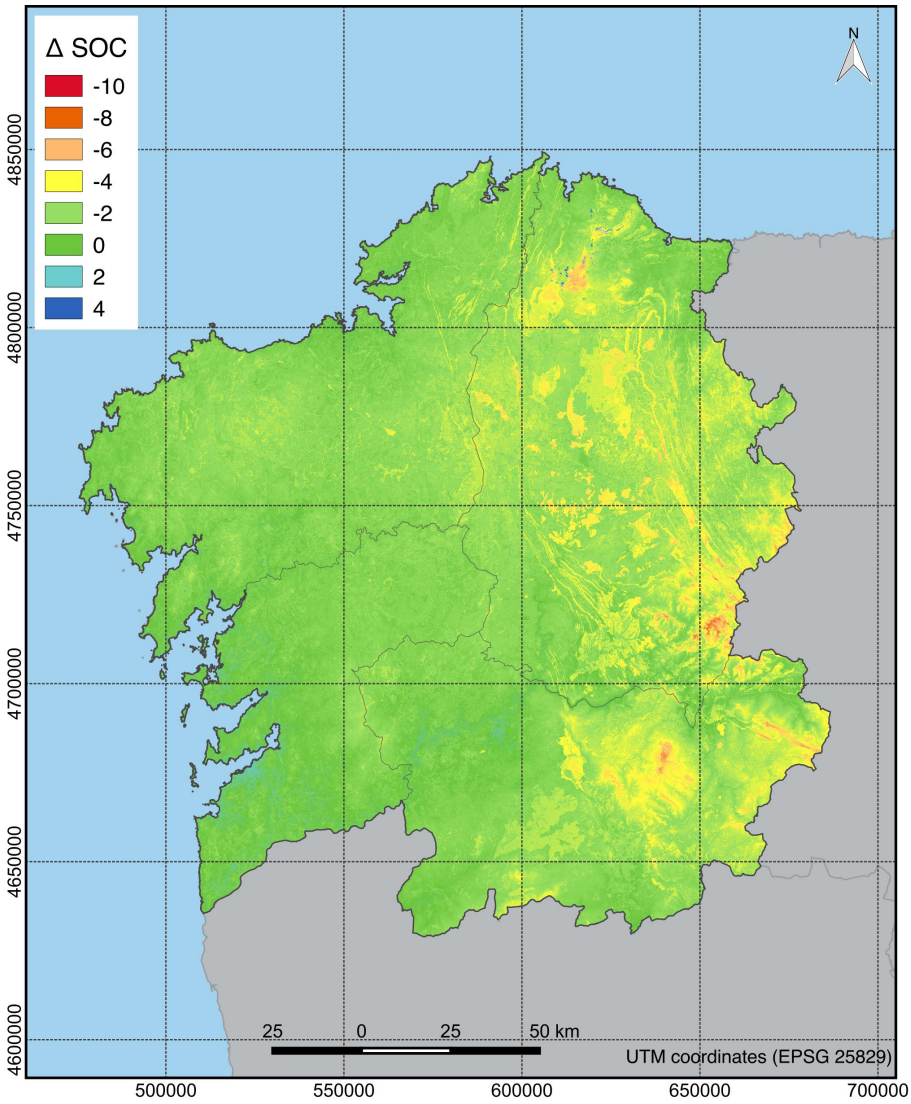


FIGURE 3.8: Differences in the spatial predictions of SOC, expressed as (FTIR-ATR predicted - Walkley-Black predicted) SOC in %.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 4

STUDY CASE II: Europe

4.1 Background and study area

The implementation of efficient policies for CO_2 sequestration requires knowledge on which environmental changes at regional or local scale can lead to the highest sequestering potential of soils [187]. It has been estimated that the biosphere in Europe absorbs the equivalent of about 10% of its total carbon emissions [188, 189]. Reforestation and land use management have been recognized as positive practices to enhance the sequestration of atmospheric carbon [25, 26, 190–192]. Recent studies suggest that European forests are reaching saturation in their carbon sink capacity [193], while other studies demonstrate that reforestation

practices in Europe during the last 250 years did not contribute to mitigate climate change in Europe [194], evidencing that land use management strategies to fight against climate change need to be further revised. Regarding soils, the implementation of measures to control atmospheric CO_2 emissions requires a better understanding of the dynamics and factors that control SOC accumulation. At European scale, the distribution and magnitude of such factors are still uncertainty [195]. Mechanistic understanding of the environmental processes that control the carbon cycle at local scales is essential for interpreting the processes that occur at higher scales [28].

The first estimation of SOC distribution at European scale is that of the European Soil Database (ESDB) [196], that combined pedotransfer rules with information of land cover and air temperature to derive a vector map of SOC content classes. To date, the land use/land cover statistical area frame survey database (LUCAS) constitutes the largest compilation of harmonised georeferenced topsoil data - including geochemical information, land use, and geology - across Europe [197, 198]. The LUCAS database was used to map SOC at European scale by means of different statistical approaches. de Brogniez *et al.* [199] applied a GAM model together with topographic data, land cover, accumulated annual temperature, and net primary productivity as environmental covariates, to create a map of SOC in topsoils

from Europe. This model explained only 28% of the total variability in SOC, thus presenting a poor predictive performance. Yigini and Panagos [200] used Regression Kriging with climatic, topographic, land cover, soil structure, available water capacity, soil classification, cation exchange capacity and parent material layers. Their model explained 40% of SOC content, improving the previous approach. Aksoy *et al.* [201] also used Regression Kriging on data from soil samples of the BIOSOIL, SoilTrEC and LUCAS databases to create a series of maps of SOC content in Europe. In this case, topography, soil type, geology, land cover, texture, parent material and climatic data were used as environmental auxiliary proxies. The predictive ability was similar to that obtained by Yigini and Panagos [200], explaining up to 41% of the SOC variability with LUCAS+SoilTrEC databases and 40% with the LUCAS database only. Yigini *et al.* [22] also employed the LUCAS dataset to predict SOC contents in European agricultural soils by means of Principal Component Analysis followed by Multiple Linear Regression and climate data, land cover, terrain and NDVI as environmental covariates. Their model explained 35% of SOC variability and was used to predict SOC stocks in agricultural areas at global scale.

One important limitation of the regression approaches followed up to now is that they consider the statistical relationships between dependent and independent variables as

linear, thus spatially stationary. Since statistical relationships may vary spatially, the models fail to capture the variability due to the changing relationships between soil properties and environmental parameters. Spatial changes in sign and intensity of the effect of the environmental factors may result in similar SOC content in soils under different environmental conditions. The identification of areas in which SOC accumulation is driven by a common set of environmental factors is crucial to improve the predictive power of the statistical models created. This can partially explain why studies of SOC content at regional and national scales, where the environmental factors governing SOC accumulation are similar over the studied territory, show higher predictive performance than those performed at continental scales [123, 125, 157, 202–207].

The LUCAS topsoil database is available at the European Commission through the European Soil Data Centre at http://esdac.jrc.ec.europa.eu/content/lucas-2009-topsoil-data [197, 198]. The dataset includes about 20000 topsoil samples (0-20 cm) collected within 23 states of the European Union with the aim of providing a harmonized soil dataset on different land use/cover over Europe (Figure 4.1). Coarse fragments ($>$ 2 mm), particle size distribution, pH, cation exchange capacity, extractable phosphorus and potassium contents, total nitrogen, organic carbon and VNIR diffuse reflectance were measured using standard procedures. The

dataset also includes environmental information such as the nature of the dominant parent material from the ESDB, as well as the main land cover from the Corine Land Cover datasets (CLC). We removed from the analyses those samples classified as infrastructures, water and wetlands by their LC1 attribute since they are either misclassified soil samples or samples in which the local conditions could not be attributed to a more general environmental pattern in the area.

4.2 Geochemical data

Total carbon content was measured by dry combustion (ISO 10694:1995) using a VarioMax CN Analyzer (Elementar Analysis, Germany). The carbonate content, measured according to the standard ISO 10693:1995, was subtracted from total carbon values to obtain the SOC content of the samples.

4.3 Spectroscopic data

Soil samples were air-dried and sieved (0.2 mm) before spectral analyses. Infrared spectra (400-2500 nm) were recorded with a spectral resolution of 0.5 nm using a FOSS XDS Rapid Content Analyzer (FOSS NIRSystems Inc., Denmark) equipped with Si

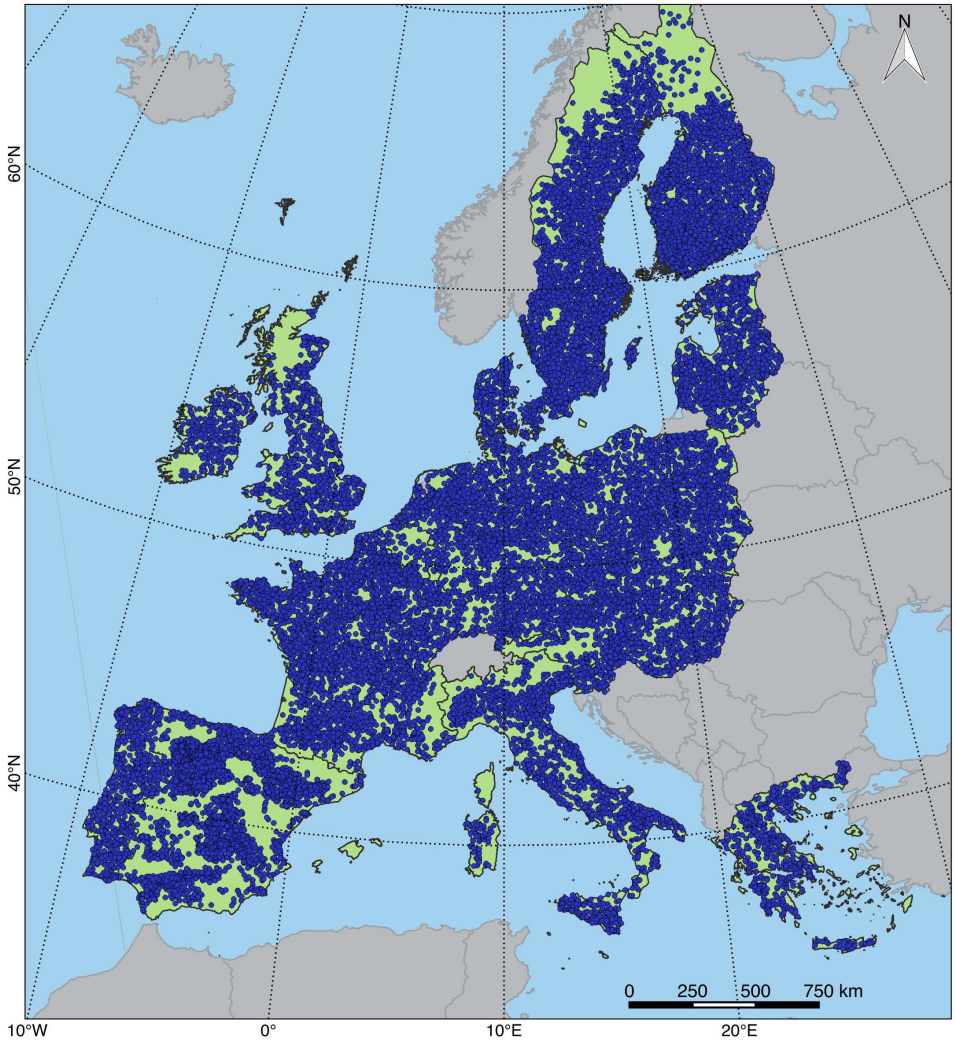


FIGURE 4.1: Geographic location of samples collected in Europe.

and PbS detectors. We have resampled each spectrum at 10 nm in order to reduce computational cost. It is generally admitted that this changes in resolution do not compromise the quality of the information that can be extracted from the spectra. All spectra were corrected by continuum removal method. This is a pre-processing method commonly used in soil spectroscopy that highlights the spectral features of minerals, organic compounds and water, helping in their identification [123].

4.4 Environmental variables

A total of 20 GIS-based raster maps, at 1 km resolution, were used as environmental covariates to model the spatial distribution of groups and SOC content over the study area. These maps include information on climatic conditions, parent material and land cover classes (Figure 4.2).

Temperature: A raster map of mean annual temperature at 1 km resolution was downloaded by tiles from the Global Climate data repository for ecological modeling and GIS (http://www.worldclim.org/current, [208]). These tiles were merged and reprojected to the Lambert Azimuthal Equal Area coordinate system centered at 10°E of longitude and 52°N latitude (EPSG:3035) to comply with the INSPIRE European Directive.

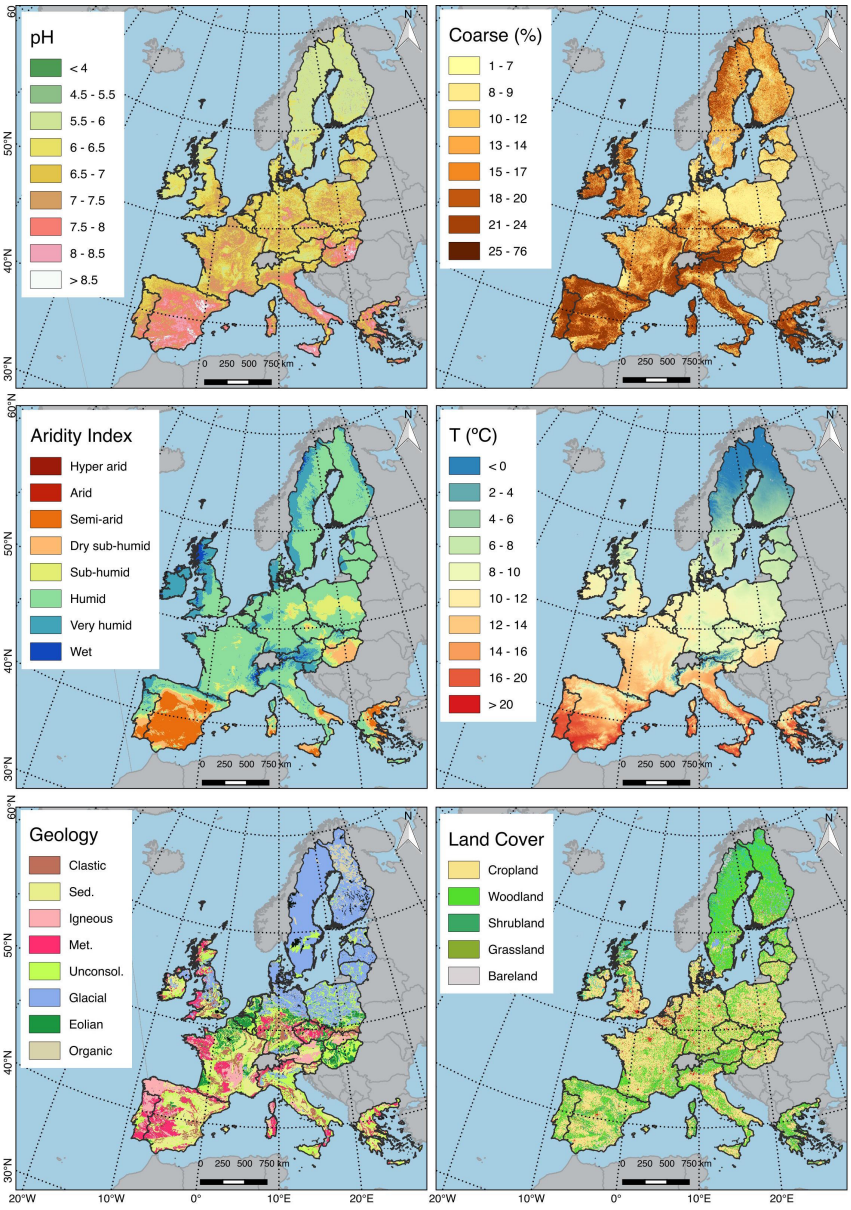


FIGURE 4.2: Environmental variables used for mapping purposes in Europe.

Aridity index: A raster map of the aridity index at 1 km resolution was obtained from the Consortium for Spatial Information (CGIAR-CSI) GeoPortal at *http://csi.cgiar.org/aridityGlobal_AridityPET_Methodolgy.asp* [209] and reprojected to the EPSG: 3035 coordinate system. The values of aridity index indicate the presence of the different climatic zones: hyper-arid (< 0.03), arid (0.03-0.2), semi-arid (0.2-0.5), dry sub-humid (0.5-0.65), sub-humid (0.65-0.75), humid (0.75-1.25), very humid (1.25-2.5) and wet (> 2.5).

pH: The map of soil pH at 1 km resolution was obtained from http://soilgrids.org in tiles. Tiles were merged and reprojected to the EPSG:3035 coordinate system.

Texture: The map of content in coarse fragments (%) was prepared by the Land Resource Management Unit (Institute for Environmental Sustainability) of the Joint Research Centre (JRC) of the European Commission and is available for download at http://esdac.jrc.ec.europa.eu/content/topsoil-physical-properties-europe-based-lucas-topsoil-data [97, 210]. This map was reprojected to the EPSG:3035 coordinate system and aggregated from 500 m to 1 km resolution using the average of the percent values within each 1 km pixel.

Geology: The European map of parent material (PARMADO1) was obtained from the ESDB, prepared by the Land Resource Management Unit (Institute for Environmental Sustainability) of the JRC of the European

Commission and available at the website: http://esdac.jrc.ec.europa.eu/content/european-soil-database-v20-vector-and-attribute-data [210, 211]. The shapefile of geological classes (1 - Consolidated-clastic-sedimentary rocks; 2 - Sedimentary rocks-chemically precipitated, evaporated, or organogenic or biogenic in origin; 3 - Igneous rocks; 4 - Metamorphic rocks; 5 - Unconsolidated deposits-alluvium, weathering residuum and slope deposits; 6 - Unconsolidated glacial deposits/glacial drift; 7 - Eolian deposits and 8 - Organic materials) was rasterized and used to create binary maps representing the absence/presence of each type of parent material.

Land Cover: CORINE Land Cover 2006 raster maps at 250 m resolution were downloaded from the site http://www.eea.europa.eu/data-and-maps. The original CLC classes were reclassified to match with the land cover attribute (LC1) within the LUCAS database: 1 - Artificial surfaces; 2 - Croplands; 3 - Woodlands; 4 - Shrublands; 5 - Grassland; 6 - Bare soils; 7 - Water bodies; 8 - Wetlands (Table 4.1). Binary maps representing the absence/presence of each land cover class were calculated and aggregated to 1 km resolution using the nearest neighbor criteria.

TABLE 4.1: Reclassification of CLC classes.

Grid Code	CLC Code	CLC category	New Class
1	111	Continuous urban fabric	Artificial
2	112	Discontinuous urban fabric	Artificial
3	121	Industrial or commercial units	Artificial
4	122	Road and rail networks and associated land	Artificial
5	123	Port areas	Artificial
6	124	Airports	Artificial
7	131	Mineral extraction sites	Bareland
8	132	Dump sites	Bareland
9	133	Construction sites	Artificial
10	141	Green urban areas	Artificial
11	142	Sport and leisure facilities	Artificial
12	211	Non-irrigated arable land	Cropland
13	212	Permanently irrigated land	Cropland
14	213	Rice fields	Cropland
15	221	Vineyards	Cropland
16	222	Fruit trees and berry plantations	Cropland
17	223	Olive groves	Cropland
18	231	Pastures	Cropland
19	241	Annual crops associated with permanent crops	Cropland
20	242	Complex cultivation patterns	Cropland
21	243	Land principally occupied by agriculture	Cropland
22	244	Agro-forestry areas	Cropland
23	311	Broad-leaved forest	Woodland
24	312	Coniferous forest	Woodland
25	313	Mixed forest	Woodland
26	321	Natural grasslands	Grassland
27	322	Moors and heathland	Shrubland
28	323	Sclerophyllous vegetation	Shrubland
29	324	Transitional woodland-shrub	Woodland
30	331	Beaches, dunes, sands	Bareland
31	332	Bare rocks	Bareland
32	333	Sparsely vegetated areas	Bareland
33	334	Burnt areas	Shrubland
34	335	Glaciers and perpetual snow	Water
35	411	Inland marshes	Wetlands
36	412	Peat bogs	Wetlands
37	421	Salt marshes	Water
38	422	Salines	Wetlands
39	423	Intertidal flats	Wetlands
40	511	Water courses	Water
41	512	Water bodies	Water
42	521	Coastal lagoons	Water
43	522	Estuaries	Water
44	523	Sea and ocean	Water

4.5 Modelling procedures

The statistical steps used in this study are summarized in Figure 4.3. The basic idea here proposed is that the soil geochemical properties are correlated to the environmental conditions existing at each location. As a result, environmental conditions also leave their imprint in the spectroscopic signal of the soil samples. Since organic matter content has a strong influence on the spectroscopic signature, differences in its content will highly determine the values recorded in the infrared spectra. Under this assumption, the spectroscopic information can be used to identify samples where SOC content and the environmental conditions that promote SOC accumulation follow a similar pattern. In this study we aimed to develop a statistical approach to capture the spatial variability of the factors influencing SOC accumulation at European scale.

We identified the number of different groups of soil samples in the LUCAS database by means of their VNIR signature. The first step in our methodology is a PCA analysis over the VNIR spectra to reduce the dimensionality of the data. The scores from PCA were used in a k-means classification algorithm to determine groups of samples with similar spectroscopic features. In a second step, for each group identified we calculated the buffer distances between samples in order to account for the effect of the location of the

different groups. In a third step, these buffer maps were combined with the environmental maps of covariates to derive a number of spatial predictive components by PCA that are then used as covariates to build a RF probability model as in [212], which allows to create a map depicting the group membership of each pixel site in a raster map of Europe. The last step consists in the creation of a RF regression model to determine the relationship between SOC and the environmental covariates within each of the groups of samples previously identified. Each single RF regression equation from each group was applied upon the correspondent area from the membership classification to determine the SOC content and the uncertainty associated to SOC predictions at each European site. We used Ordinary Kriging upon the RF residuals to account for the unexplained SOC variability in the RF model and to map the Standard Error (SE) associated to our model predictions. The resulted map of SOC was validated with a 5-fold cross validation method.

Principal Component Analysis (PCA): PCA is a well-know method used to reduce the dimensionality of the data transforming the variables into new ones called "principal components", which are linear combination of the original variables. PCA decomposes the original database in a matrix of loadings (representing the correlation between each wavenumber and the different components) and a matrix of

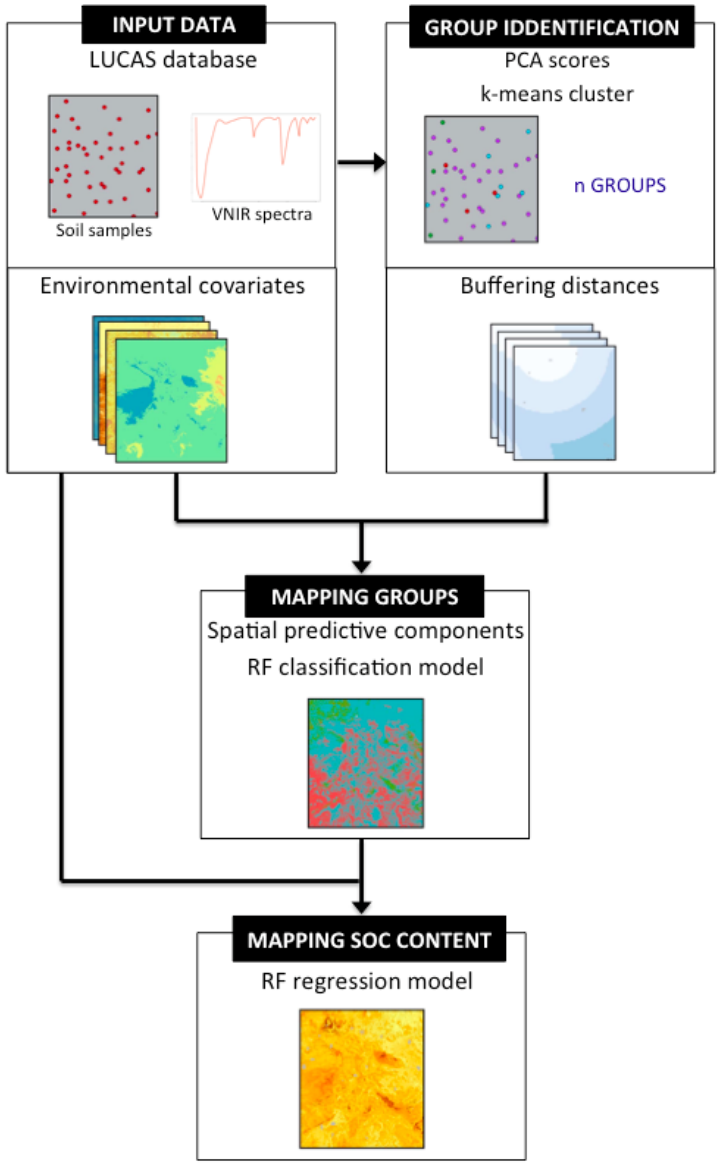


FIGURE 4.3: Statistical framework used for mapping SOC content in European topsoils.

scores representing the correlation of each sample in the same components [213].

k-means clustering: k-means is an unsupervised classification method that creates a subset of n samples $X_{tr} = [x_{trj}]_{(j=1)}^{n}$ from a given set of N samples $X = [x_i]_{(i=1)}^{N}$. The algorithm performs a k-means clustering of X using n clusters, extract the n centroids and calculates the distance of each sample to these centroids. Finally for each centroid allocates in X_{tr} its closest sample found in X [214, 215]. In the elbow method, the Sum of Squared Error (SSE) obtained between different splitting options are calculated and compared to determine the optimal number of groups to split the dataset. This optimal number corresponds to the value where SSE drops, producing an elbow in the plot.

Random Forest (RF): RF is a machine learning method that can be used for regression and classification purposes. It derives from bagging algorithm [169], adding an additional layer of randomness to improve the prediction performance [92, 172]. The RF algorithm generates multiple randomized CART. Each CART is calibrated on a random set of bootstrap samples from the original data and provides a robust estimation for the remaining data, the so-called OOB samples. The best split at each node of the tree is created using a random subset of predictors. The final output of RF is a single prediction that corresponds to the average prediction

from all regression trees. During the process, an estimation of the OOB samples is produced and the Mean Square Error (MSE_{OOB}) is calculated by aggregating the predictions from all trees according to the following expression:

$$MSE_{OOB} = n^{(-1)} \sum_{i=1}^{n} (z_i - \hat{z}_i^{OOB})^2 \tag{4.1}$$

where n is number of trees and $\hat{z}_i^{OOB}$ is the average of all OOB predictions for the z_i observations [92, 172, 173]. The GOF of the model can be directly obtained from the MSE_{OOB} and the total variance of the response variable (var_z) as the proportion of explained variance (var_{ex}) [92]:

$$var_{ex} = 1 - (MSE_{OOB}/var_z) \tag{4.2}$$

RF presents a series of advantages, over other predictive methodologies, such as no overfitting, higher prediction performance, low correlation of the individual trees, low bias, low variance due to averaging over a large number of trees and robust error estimates by using the OOB dataset. In addition, the RF algorithm can handle either continuous or categorical auxiliary variables [72, 92].

4.6 Use of infrared data for mapping

The statistical methodology here applied firstly requires the classification of samples according to the similarity of their VNIR signature. We used PCA to reduce the dimensionality of the original spectroscopic data and to enable the identification of groups of samples. PCA showed that three principal components explained 93% of the total spectral variability. A k-means clustering analysis, upon the scores of these three components, was then used to identify groups of samples with similar spectroscopic signature. The elbow method used within the k-means clustering revealed an optimal number of four clusters of samples within the entire dataset (Figure 4.4). From this step we obtained a georeferenced database of groups of samples classified by their VNIR similarity (Figure 4.7) that was used within a probability RF classification model to map the occurrence of each group in the conterminous Europe (Figure 4.6). This map was obtained from the maximum membership probability calculated as the count of votes of the individual trees in the ensemble (Figure 4.7). The prediction error of the RF classification model was 0.11.

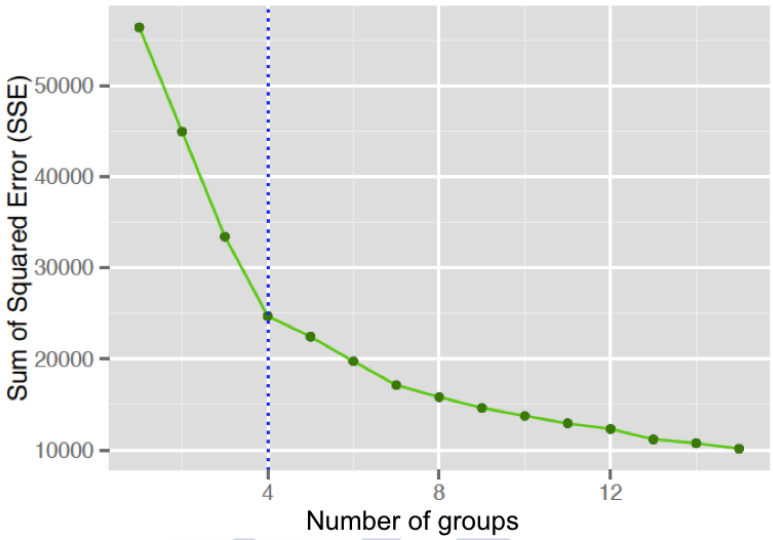


FIGURE 4.4: Elbow method for selecting the optimal number of groups.

4.7 Factors influencing the amount of SOC

Climate, land-cover, deforestation, biomass burning, ploughing, drainage of wetlands and low-input farming or shifting have been identified as environmental factors influencing the content of SOC [25]. Since different combinations of such parameters can lead to similar values of SOC, the creation of a single statistical model, calibrated upon data from an extensive area including a high variety of environmental conditions, can underestimate, at local scale, the effect of some relevant processes masked within the global

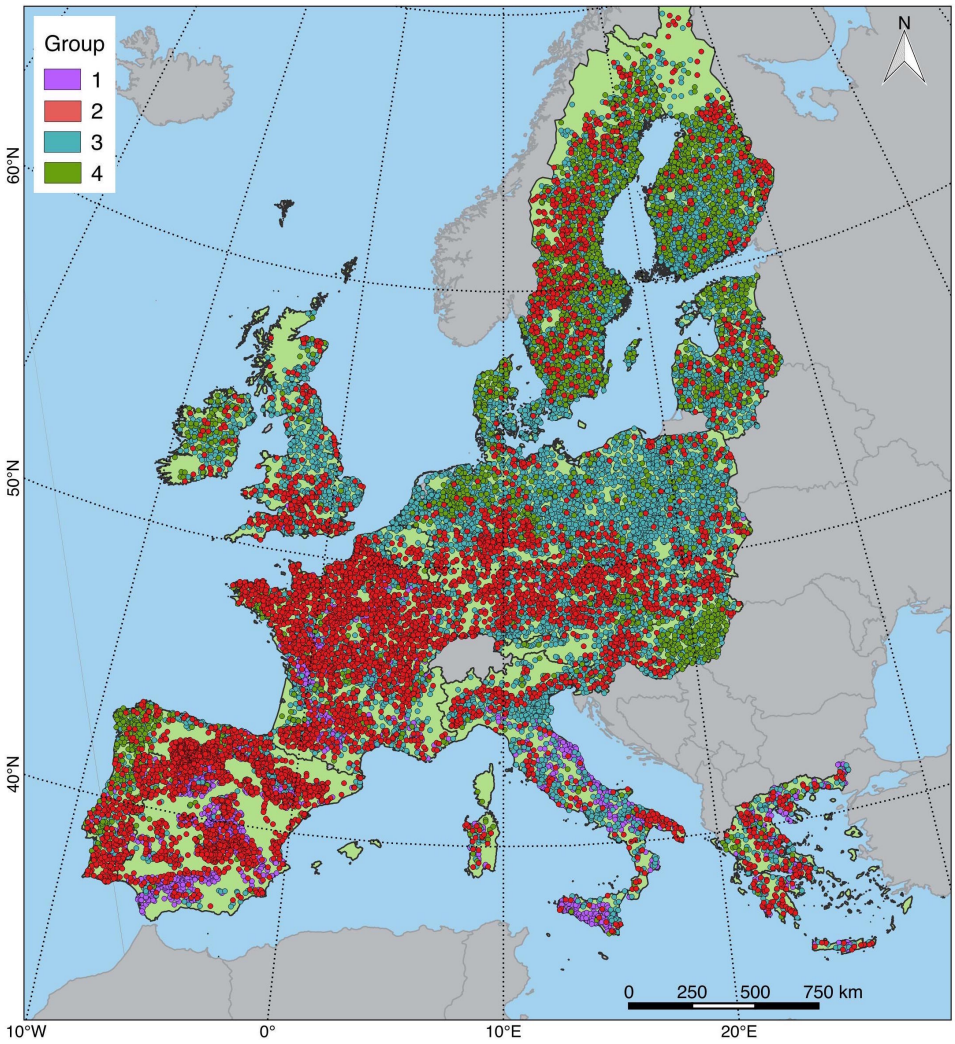


FIGURE 4.5: Map showing the group corresponding to each LUCAS soil sample.

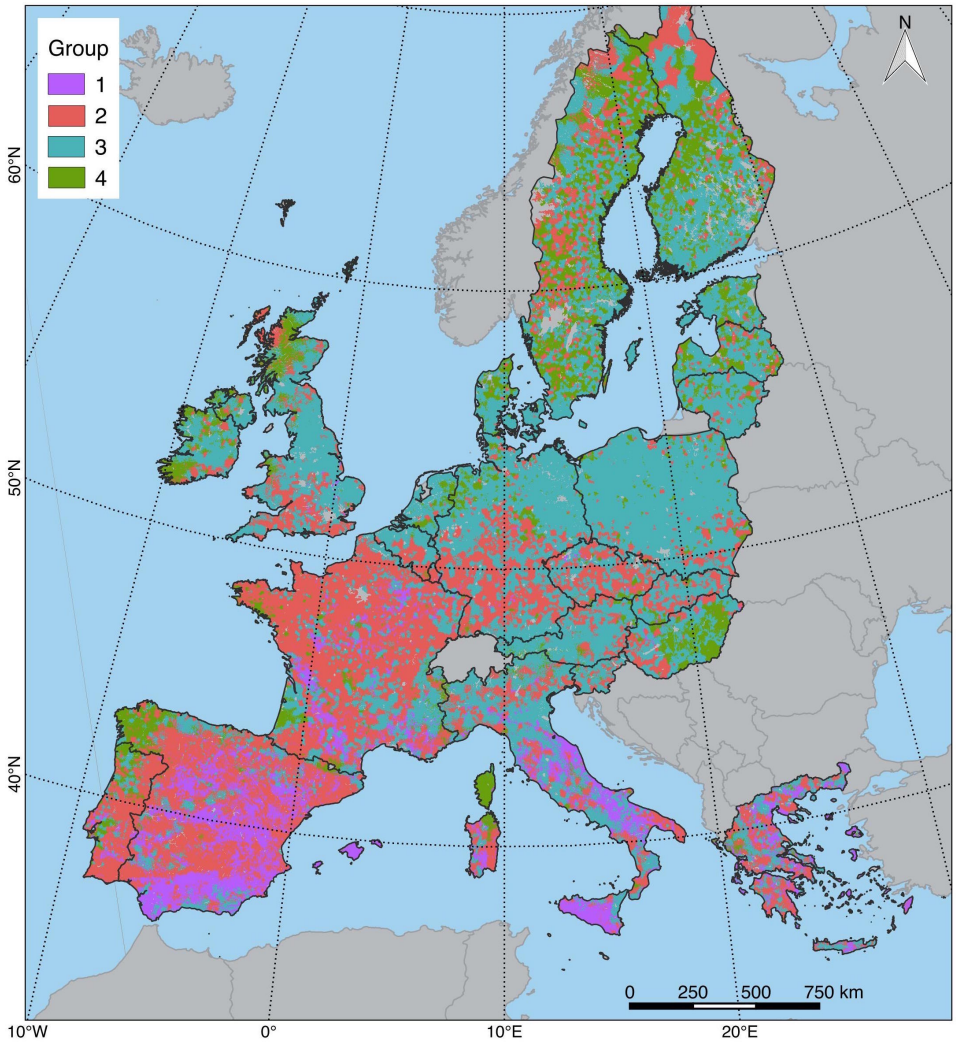


FIGURE 4.6: Map of the spatial distribution of groups in Europe.

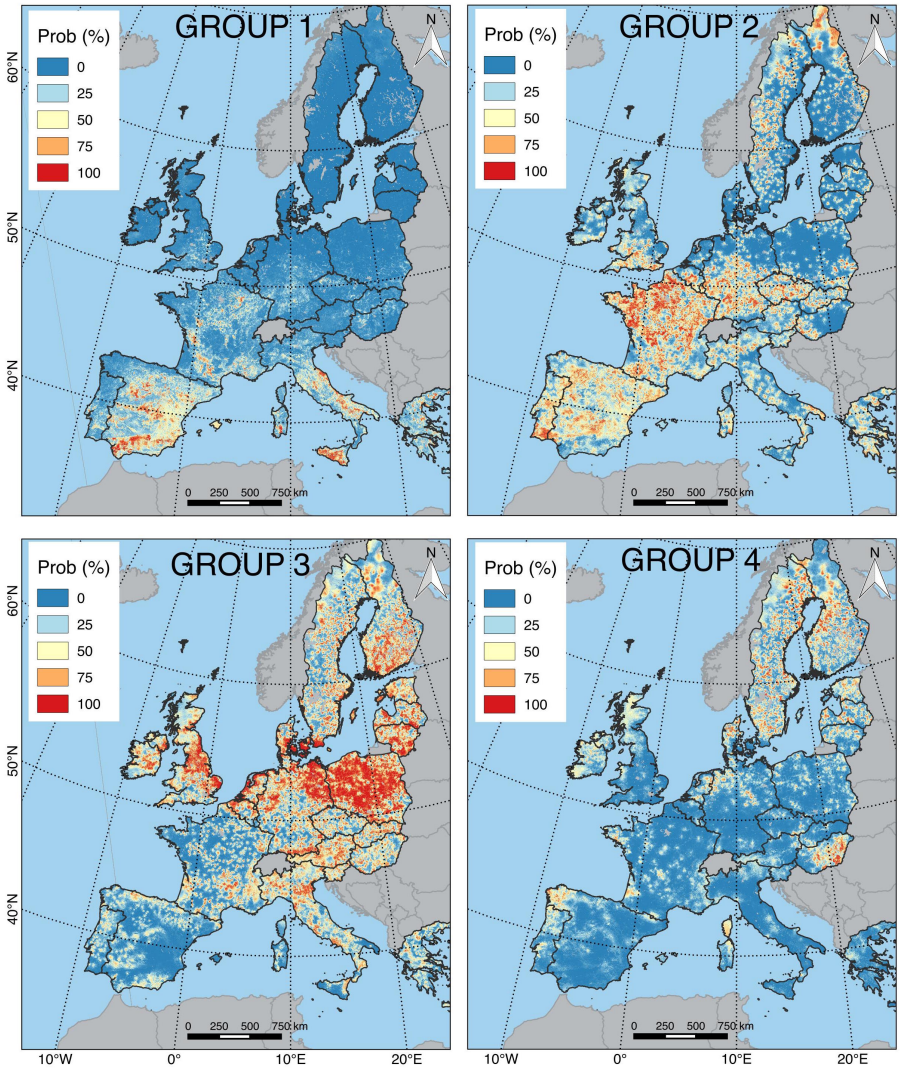


FIGURE 4.7: Maps showing the membership probability for each group obtained from the RF classification model.

model. The solution here proposed is to calibrate different predictive models for regions identified by having similar environmental conditions according to their spectral properties. For each RF model we also analyzed the variable importance to identify the main environmental conditions governing the SOC contents found within each group (Figure 4.9 and 4.10). In this case, the variable importance has been determined by the correspondent increment in the mean square error of predictions when a certain covariate was not considered during the construction of the single regression trees in RF. The characteristics, the main influential environmental parameters and the geographical extent of the models obtained for each group are:

Group 1: The expected distribution of soils in this group, obtained through the RF classification algorithm, indicates that they represent 8% of European area, and are located in central and southern areas of the Iberian Peninsula, Italy, Greece and some areas of France (Figure 4.6). The RF regression model showed low predictive capacity (R^2 = 0.38, RMSE = 0.72, MAE = 0.49). The variable importance indicates that climate (aridity index, mean annual temperature) and parent material (pH) are the most influential factors (Figure 4.9). SOC content is negatively correlated to aridity index [57], a climatic factor limiting the primary production of organic matter [216]. Under this situation, soils from arid areas are expected to have the lowest

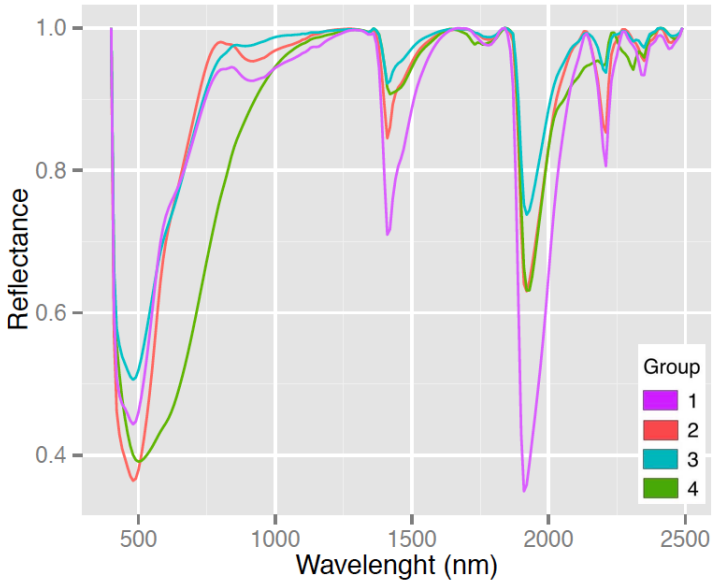


FIGURE 4.8: Mean VNIR spectra of each group.

SOC contents. The second factor is the presence of relatively high annual mean temperatures that, as in the case of aridity, may increase the mineralization rates of organic carbon in soils [217, 218]. The third influential factor identified is soil pH. High pH values have been correlated to low SOC, as they are usually associated to high soil microbial activity [52, 219], which promotes the mineralization of the organic matter. Aridity in soils from southern Europe is the main cause for the low SOC content. In addition, soils in this group are mainly located on large river basins, presenting limited drainage and high contents of Ca^{2+}. Under such conditions, the formation of 2:1 smectitic clays is favored [220, 221]. This

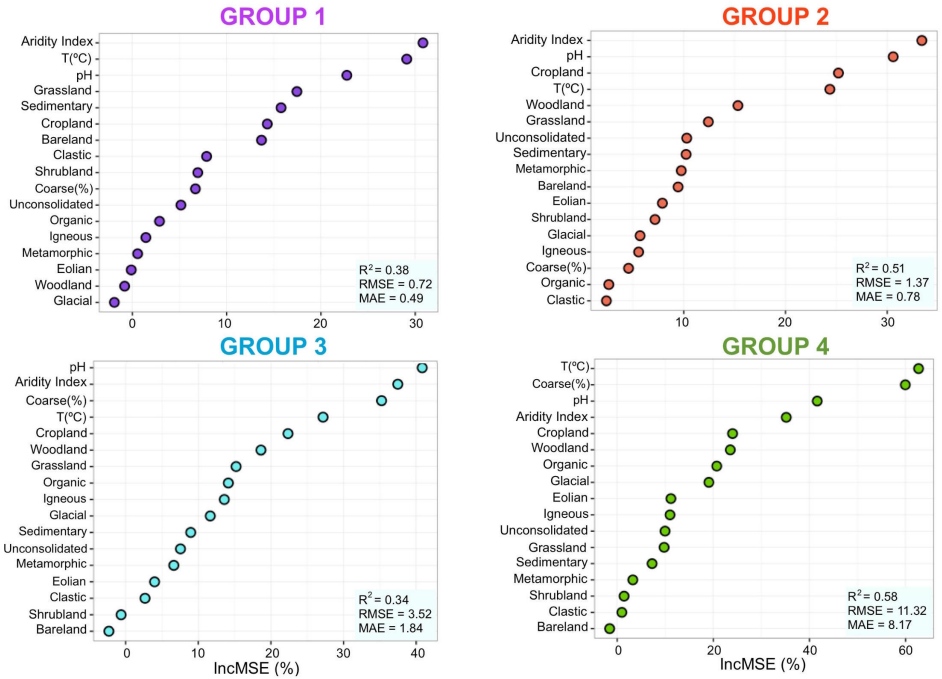


FIGURE 4.9: Variable importance plots from RF analysis.

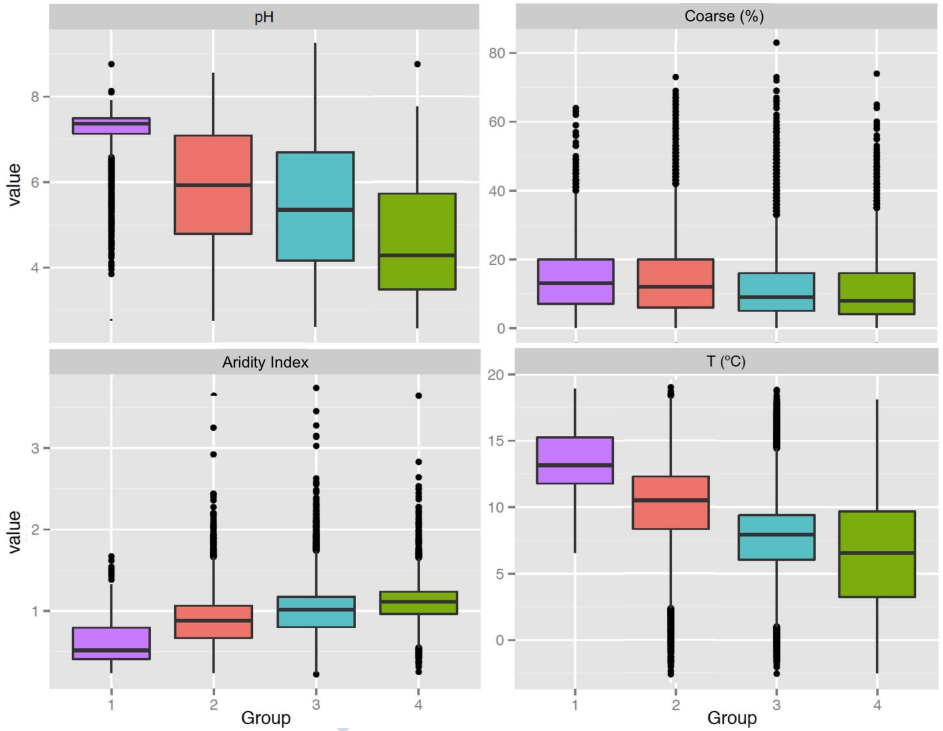


FIGURE 4.10: Boxplots showing distribution of continuous environmental variables in each group.

is consistent with the high bands at 1500, 1900 and 2200 nm [123, 125, 222] in the mean spectra of this group (Figure 4.8). Croplands are the dominant land cover type. Management techniques such as the exportation of organic material and the promotion of organic matter decomposition due to ploughing, are responsible for their low SOC contents [52].

Group 2: The classification model shows that they cover about 33% of the European surface here analyzed, and are mainly located in southern European countries (Portugal, Spain, France, Italy and Greece) although also present in some areas of southern United Kingdom and central Europe (Figure 4.6). The RF analyses showed a moderate predictive capacity (R^2 = 0.51, RMSE = 1.37, MAE = 0.78). This model indicates that SOC content in this area is influenced by climatic conditions (aridity index, and mean annual temperature values), parent material (pH) and land-cover type. Even if the presence of these soils with low carbon content can be attributed to climatic, pedologic and land use factors, the intensity of these factors in the geographic space is not equivalent for the areas classified within this group. Climate is the main factor operating in southern Europe. About 30% of the soils in this group are located in semiarid to sub humid conditions in Spain, Portugal, Italy and Greece. For France, southern United Kingdom and central European countries, the climatic effect is likely to have a lower importance (these areas present regimes from humid, to dry

sub-humid) and the SOC contents are linked to the predominance of croplands within this group, with slightly higher soil pH values (Figures 4.2). The map in Figure 4.6 shows some soils from this group in parts of Sweden and northern Finland. Higher SOC contents would be expected for these areas, due to the accumulation processes associated to cold climates [223]. However, the presence of soils from groups 2, 3 and 4 along Scandinavia (Figure 4.6), indicates that the variation of SOC in this region is not only controlled by climate and that management changes over the past 100 years have probably created imbalances between litter inputs and SOC mineralization [224–226]. In many cases, human activities led to the substitution of natural forests by productive forest and the development of cropland areas [227, 228]. This local effect of land management would explain the low SOC found in samples from this area. Some authors also pointed to low nitrogen deposition rates in ecosystems from high latitudes as a potential cause for the presence of soils with low SOC [225, 229, 230], as low N deposition rates would result in a faster decomposition of soil humus in northern areas of Scandinavia.

Group 3: The classification model indicates it covers 46% of European territory, mainly in northern and central Europe (Scandinavia, Lithuania, Poland, Austria, Germany, NW United Kingdom) and mountainous areas in the Pyrenees, Alps and Apennines (Figure 4.6). The predictive accuracy of

the RF model is the lowest found in this study (R^2 = 0.34, RMSE = 3.52, MAE = 1.84). The variable importance indicates that parent material (pH, coarse fragments) and climatic factors (aridity index and mean annual temperature) are the most influential (Figure 4.9). In central Europe, soils show near neutral pH values and develop on humid and cold climates. Soil acidity and cold mean temperatures limit microbial activity, resulting in low mineralization rates and thus net SOC accumulation. In addition, mineral weathering rates are low, being illite, inherited from the weathering of primary minerals, the likely dominant soil clays. Illite formation is favored by the presence of high concentrations of Al^{3+} and K^{+} typical of soils over basement that show slightly acidic pH [221], as those from the areas represented in this group. In fact, we identified illite in the average spectrum of this group - as the lower intensity peaks on the spectral bands located at 1500, 1900 and 2200 nm (Figure 4.8). These bands indicate the stretching of O-H bonds and bending vibrations of metal-OH bonds from illite [123, 125, 222]. Croplands are the most frequent land-use type in this subset ((Figures 4.12), which are responsible for the lower SOC content found in areas such as Poland.

Group 4: Woodland is the dominant land cover type in the samples from this group, which are often associated to higher SOC contents [191]. This group can be easily identified by the width of the reflectance bands located between

400-1000 nm (Figure 4.8) [123, 231]. The classification model indicates that soils from this group occupy about 13% of the European area, and are mainly located in Scandinavia, north of United Kingdom and Ireland, NW of the Iberian Peninsula, French Landes, eastern areas in Hungary and Corsica island (Figure 4.6). The RF model showed the highest performance among all the models here created (R^2 = 0.58, RMSE = 11.32, MAE = 8.17). The variable importance analysis (Figure 4.9) indicates that climatic factors (temperature and aridity index), coarse fragments, and pH have the higher weights in the predictions. In this case, low temperatures and wet climates are positively related while the abundance of coarse fragments is negatively related to high SOC contents. Low contents in coarse fragments may be associated to soils under environmental conditions of biostasia [232], developing on more or less stable surfaces where the intensity of erosion is low. This favors soil development and, indirectly, the accumulation of SOC. The relative influence of these factors varies spatially. In northern countries, low temperatures are likely to be the dominant factor, while in oceanic areas SOC accumulation is driven by the humid conditions prevailing most of the year (Figure 4.2). For Hungary, samples falling in this group correspond to dry-sub-humid areas with alkaline soil pH, where Chernozem soils dominate, which are characterized by a large accumulation of organic matter [233].

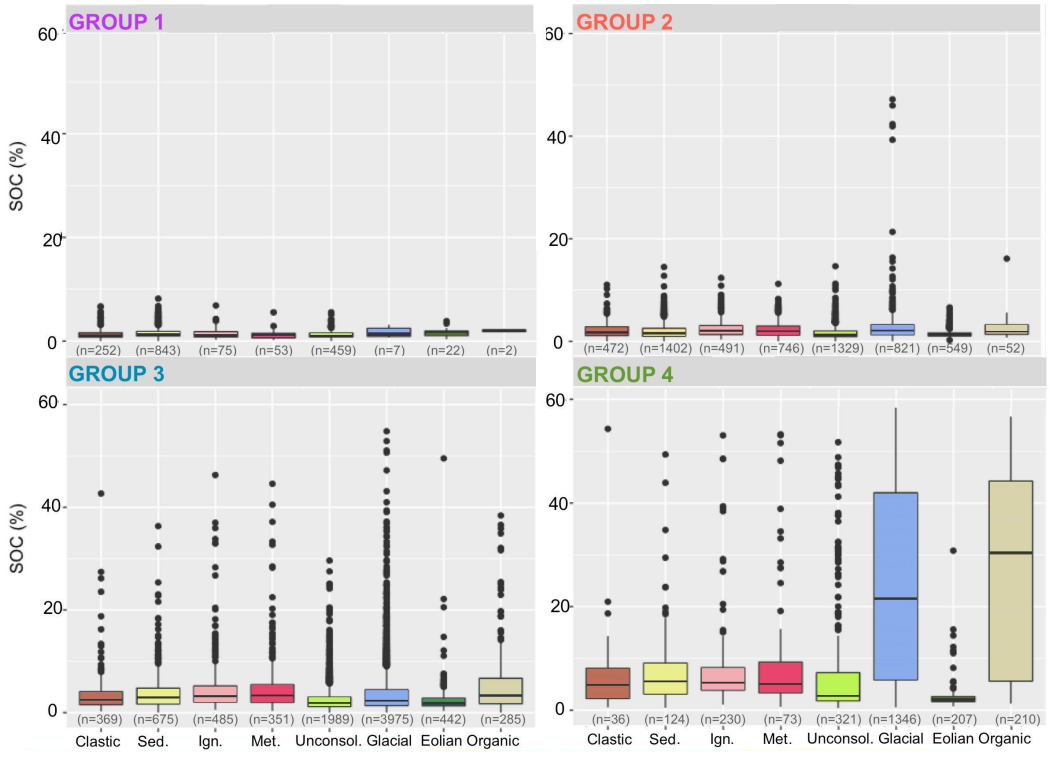


FIGURE 4.11: Boxplot relating SOC content to geology types in samples from Europe.

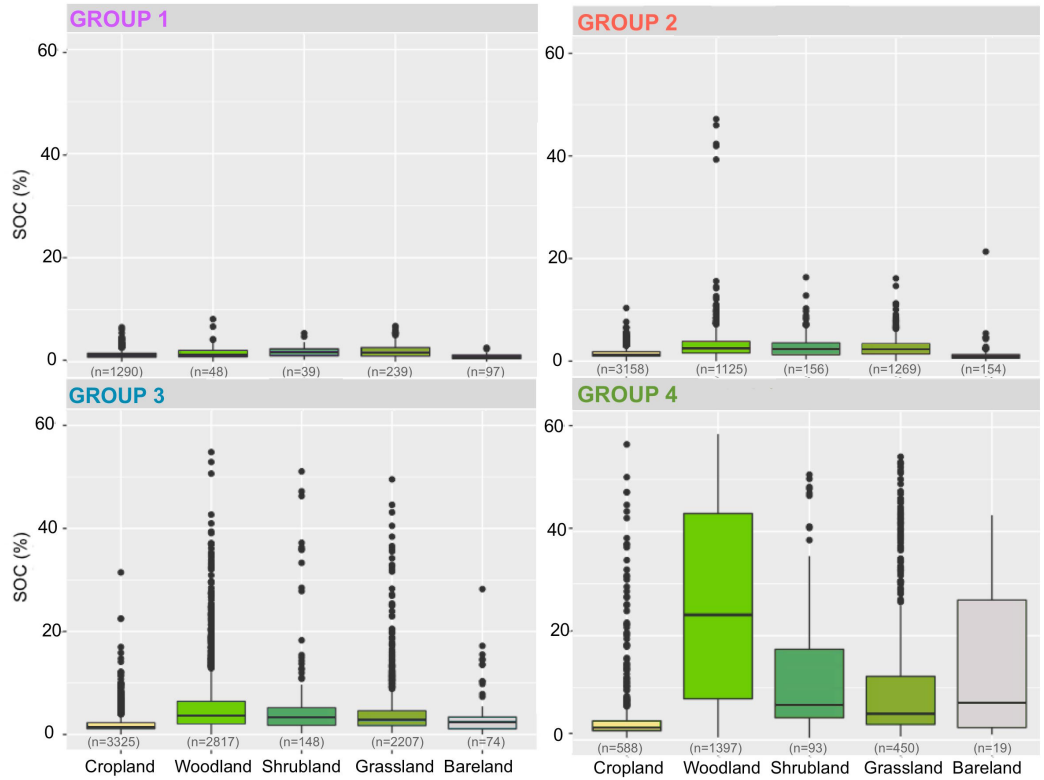


FIGURE 4.12: Boxplot relating SOC content to land cover type in samples from Europe.

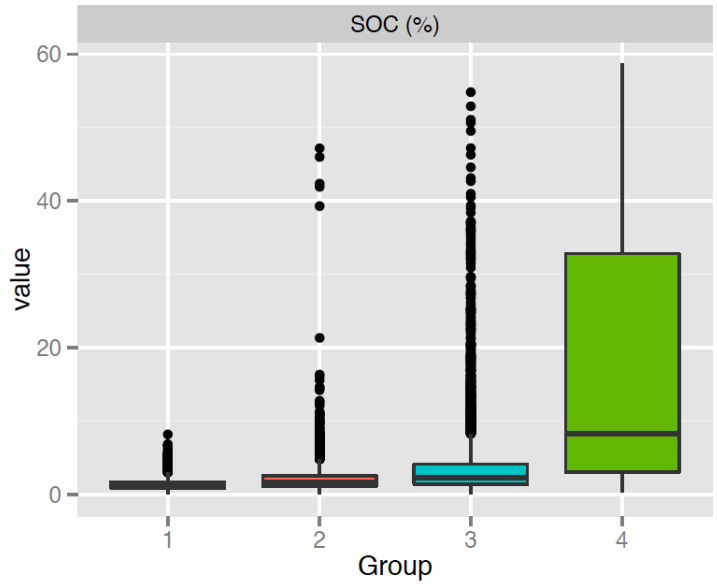


FIGURE 4.13: Boxplot relating SOC content to groups in samples from Europe.

4.8 Spatial distribution of SOC content

The samples from groups 1 and 2 have the lowest SOC content, with a mean value of 1.4% and 2.1% SOC respectively. The samples in group 3 show moderate SOC content values with a mean value of 3.6% SOC. The samples in group 4 have the highest SOC content (up to 32% SOC and with a mean value of about 18% SOC) (Figure 4.13).

The map of SOC content (Figure 4.14) shows that the highest values (>15% SOC) are found in countries from Northern Europe, in high elevation mountain ranges such as the Alps, Pyrenees and Apennines and in the NW of United Kingdom. The lowest values (<1% SOC) are found in arid areas from the south of the Iberian Peninsula and Italy, but also in some agricultural regions of Poland. The results are similar to those found in previous studies [199, 201], which also used the LUCAS database.

The approaches developed by de Brogniez *et al.* [199], Yigini and Panagos [200] and Aksoy *et al.* [201] consider that the statistical relationships between SOC content and the environmental variables explaining its content are spatially stationary, i.e. constant in the geographic space. However, it is clear that the influence each of the different environmental parameters is not constant at pan-European scale. Thus, the use of spatially non-stationary approaches is recommended to account for differences at larger scales.

4.9 Uncertainty of SOC estimations

The map of the SE of the estimates (Figure 4.15) shows that the highest uncertainty is found in areas with high SOC such as the northern Scandinavian countries, areas located at higher altitudes and NW of United Kingdom, NW of the

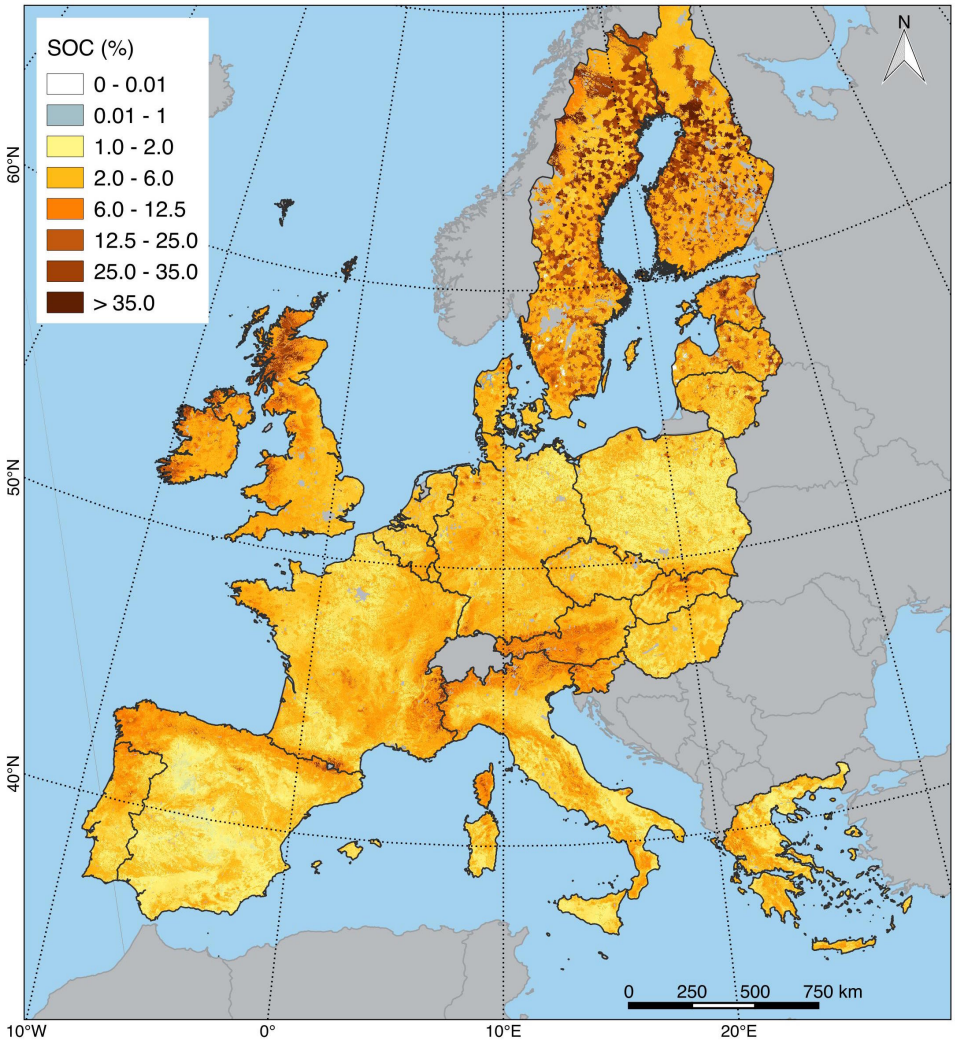


FIGURE 4.14: Map of predicted SOC content in European topsoils.

TABLE 4.2: R-squared values obtained for mapping SOC content by using the LUCAS database.

Approach	**Model**	**R^2**
de Brogniez *et al.*	GAM	0.28
Yigini and Panagos	RK	0.40
Aksoy *et al.*	RK	0.40
This approach	RF group 1	0.38
	RF group 2	0.51
	RF group 3	0.34
	RF group 4	0.58

Iberian Peninsula, SW France, Corsica and central southern Austria.

Table 4.2 shows the R-squared values obtained in these approaches as well as that from the study of Yigini and Panagos [200]. The generalized additive model developed by de Brogniez *et al.* [199] explained 28% of the SOC variability of European soils. The model performed poorly for soils from Scandinavia ($R^2 = 0.09$), and the authors suggested that these results are like due to the use of an unimodal distribution in their methodology. Yigini and Panagos [200] and Aksoy *et al.* [201] improved these results by applying a regression kriging approach, obtaining R-squared values up to 0.40. While Aksoy *et al.* [201] do not present an uncertainty map for their results, Yigini and Panagos [200] also showed a high uncertainty in their predictions for the Scandinavian area.

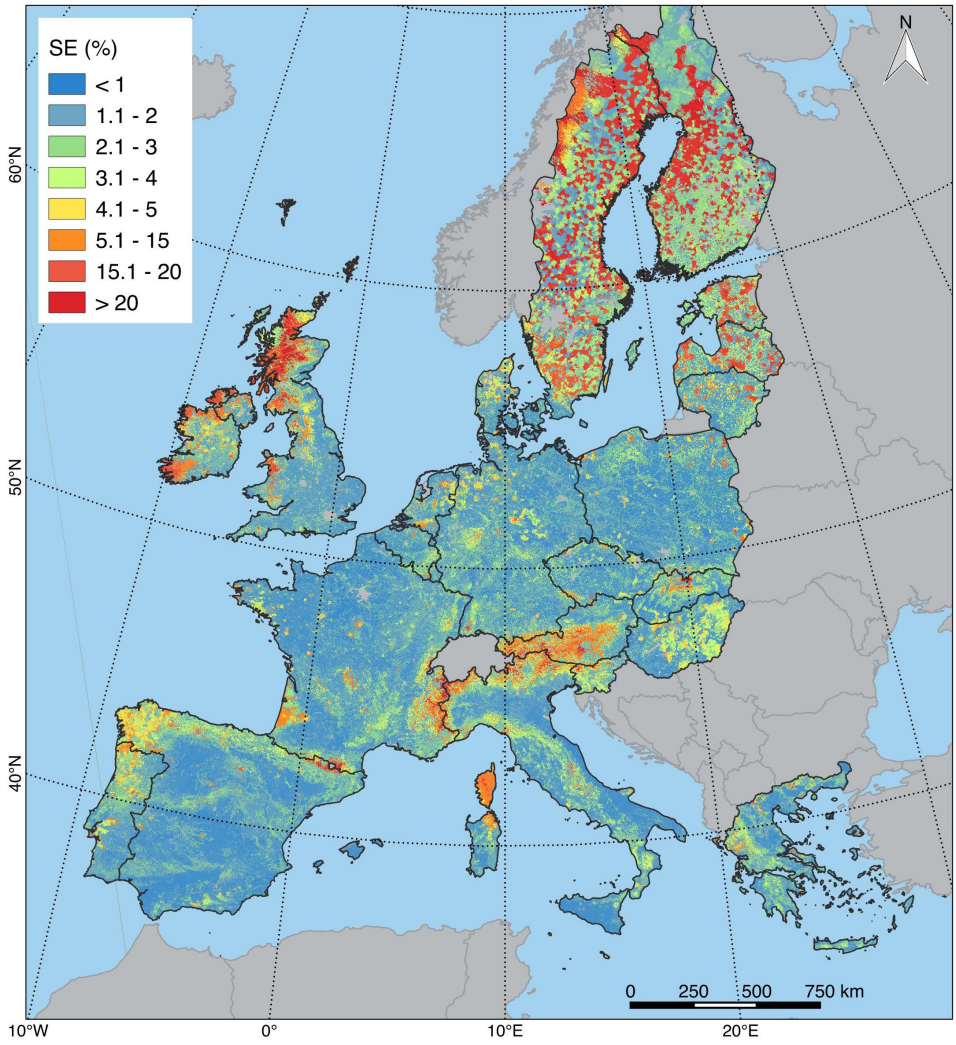


FIGURE 4.15: Map of SE of the SOC predictions in Europe.

Our spatially non-stationary approach improves the results obtained by de Brogniez *et al.* [199] for all the groups here identified. It also provides a better estimation of SOC than those obtained by Yigini and Panagos [200] and Aksoy *et al.* [201] for groups 2 and 4 (Table 4.2). Our model classified the soil samples from Scandinavia into groups 2, 3 and 4. The respective R-squared values obtained were $R^2 = 0.57$, $R^2 = 0.26$ and $R^2 = 0.39$, which indicates that the use of a non-stationary approach improves the model predictions at regional scale. The R-squared values obtained using an aggregation at country level (Figure 4.16) show that our approach improves the prediction accuracy for Denmark, Sweden and Finland in relation to that obtained by de Brogniez *et al.* [199].

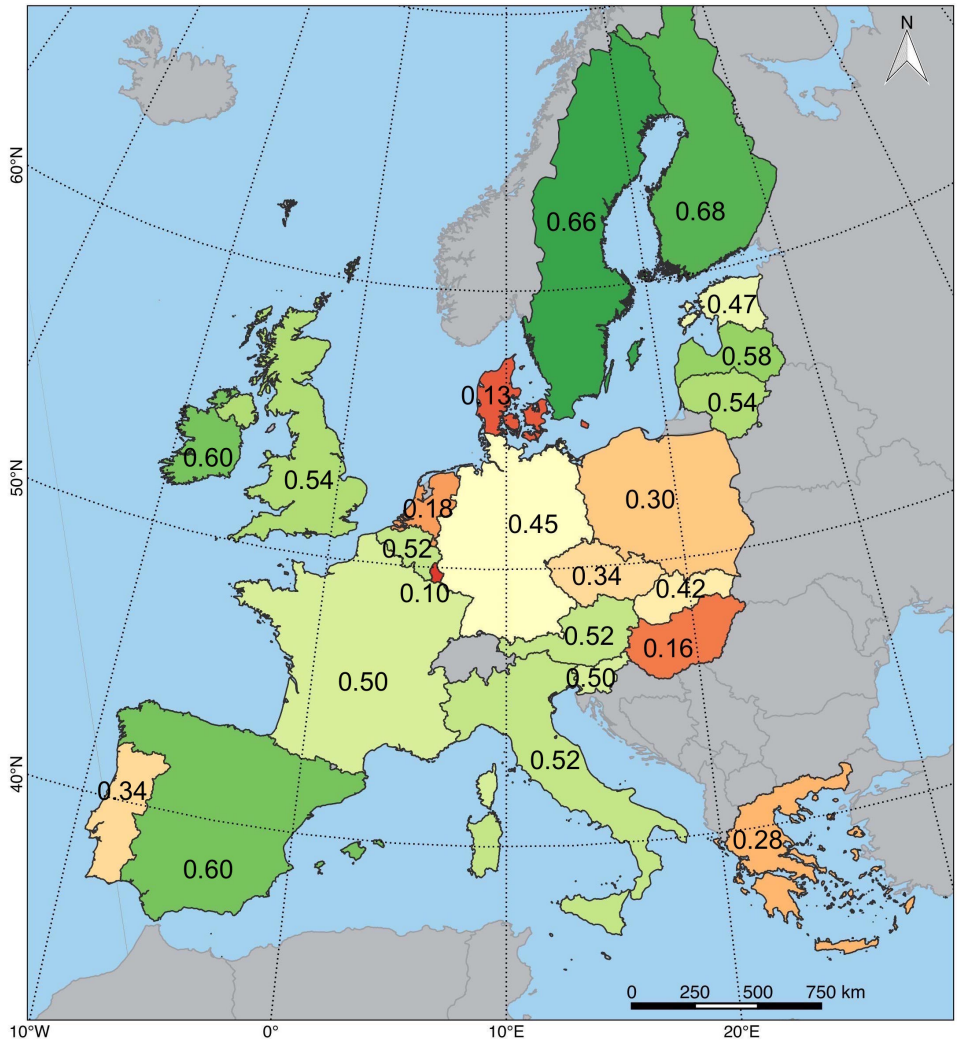


FIGURE 4.16: Map of R-squared values for each country obtained from the predictions of the RF regression models.

Chapter 5

STUDY CASE III: Santa Cruz Island, Galapagos

5.1 Background and study area

There is an increasing concern on how human activities could affect the ecosystems of the Galapagos Islands, a hot-spot for biodiversity recognized worldwide. The Galapagos Islands are a pristine area that constitutes an excellent location to study the impact of human activities on the environment [234]. Climate change was pinpointed as one of the major threats for the ecosystems in this area [235–237]. For the Eastern Pacific region, it was estimated that climate change will produce an intensification of ENSO events during the next decades, an

increase in the sea surface temperature, rainfall rates and sea level and will decrease ocean pH and the intensity of ocean upwelling [238]. The creation of an environmental monitoring system was suggested to detect the potential negative impacts of climate change in Galapagos [237].

Despite the high number of studies related to the ecological relationships between species in these islands, almost no research was conducted on the description of their soils, a basic component of the ecosystems of Galapagos [239–243]. Apart from the data used here, descriptions of soils in the entire archipelago are constrained to a data published in Adelinet el al. [239], White *et al.* [244] and of two soil profiles from San Cristobal island included in the HWSD (http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/). In total, the number of soil profiles with SOC measurements available from these publications is only 2 samples from San Cristobal Island.

Due to state-mandated restrictions on soil collecting in the Galapagos Islands at present, this study is based upon soil samples from the last geo-pedological expedition in 1962, which aimed to compile information about the islands' main soil types and properties [242]. The expedition described and sampled fifty-eight soil profiles across a transect from Academic Bay - southern coast of Santa Cruz Island - to an

altitude of approximately 500 m. Between three and five horizons within each profile were properly sampled, stored and analyzed in the laboratory [245]. Figure 5.1 shows the distribution of the 36 topsoil samples recovered from this old expedition and used in this study. Samples were collected following the bioclimatic belts that exist in the island. Table 5.1 summarizes the main characteristics of these vegetation zones.

5.2 Geochemical data

Total carbon content was measured by combustion using a LECO carbon analyzer CHNS-932 (LECO Corp., St Joseph. MI) on the fine-earth fraction ($\phi < 2$mm) of topsoil samples. The analyzed soils do not contain any source of inorganic carbon, thus the obtained values correspond to SOC content in %.

5.3 Spectroscopic data

Soil samples were finely ground using a Retsch MM 301 Mixer Mill (model 01-462-0201). FTIR-ATR spectra were sampled at 4 cm^{-1} using an Agilent Cary 630 FTIR spectrometer (Agilent Technologies, USA) attached to a diamond crystal

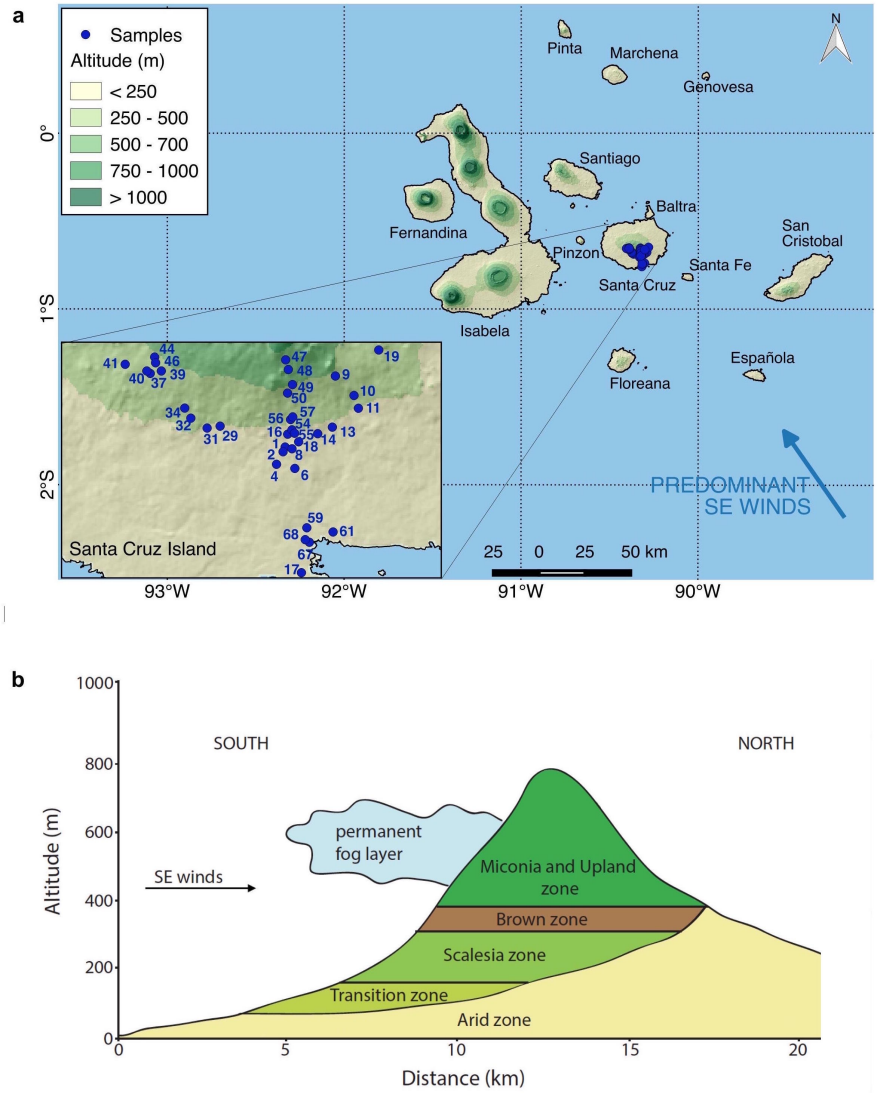


FIGURE 5.1: **Sampling scheme in Santa Cruz Island. a,** Geographic location of the topsoil samples collected by the Belgian geo-pedological mission in 1962. **b,** Zonation of bioclimatic belts in Santa Cruz Island.

TABLE 5.1: Main characteristics of the vegetation zones in Santa Cruz Island as described by Laurelle (1963) [246].

Zone	Alt (m)	Soil type	Parental material	Color	Depth	Vegetation
Arid	100 - 120	Lithosol	Basalts	Red	5 cm	Opuntia echios gigantea and Jasminocereus howellii
Transition	180 - 240	Brown soil with AC profile	Weathering of basaltic rock. Locally mixed deposits of pyroplastic origin	Brown	70 cm	Psidium galapageun. Pisonia floribunda. Piscidia erythrina
Scalesia (Humid)	300 - 400	Sol brun or sol brun lessivé	Weathering products of basalts mixed with pyroclastic material	Reddish hue	1 m	Scalesia
Brown (Humid)	400 - 500	Andosol	Pyroclastic deposits	Reddish brown	1-3 m	Psidium galapageun and Zanthoxylum fagara
Miconia (Humid)	> 500	Andosol	Pyroclastic deposits			Miconia robinsoniana and Pteridium aquilinum

ATR device and a Deuterated Ttriglycine Sulphate (DTGS) detector. Spectra were baseline corrected in order to avoid bias in the spectroscopic signal due to scattering, reflection, temperature, concentration or instrument anomalies [167].

5.4 Environmental variables

A series of GIS-based raster maps that include information on rainfall, at 1 km resolution, were used as environmental covariates to model the spatial distribution SOC stocks over the study area. These maps were downscaled at 90 m resolution using the relationship between rainfall maps and those of topographic parameters.

Rainfall: Simulated precipitation raster maps, at 30 arc-second grid resolution (Figures 5.2 and 5.3), were obtained from the Global Climate Data repository for ecological modeling and GIS V1.4 ([208], http://www.worldclim.org/) for periods 1950-2000, 2041-2060 and 2061-2080. The maps were reprojected to the WGS84/UTM 15S (EPSG: 32715) coordinate system. The precipitation maps representing future conditions correspond to the mean forecast obtained from 10 individual Global Climate Models from the Coupled Model Intercomparison (CMIP5) Project (Table 5.2) and considering four different RCP scenarios. All these maps were downscaled to 90 m

resolution by linear regression using the precipitation values at 30" resolution at sampling locations as the dependent variable and the respective values of altitude and wind effect at 90 m resolution as independent parameters.

TABLE 5.2: CMIP5 Global Climate Models used for derive future SOC contents.

CMIP5 model	**Code**	**Sponsor**
BCC-CSM1-1	bc	Beijing Climate Center, China
CCSM4	cc	Canadian Centre for Climate, Canada
CESM1-CAM5	ce	National Center for Atmospheric Research, USA
GFDL-ESM2G	gd	Geophysical Fluid Dynamics Laboratory, USA
HadGEM2-ES	he	Met Office Hadley Centre, UK
INM-CM4	in	Institute for Numerical Mathematics, Russia
IPSL-CM5A-LR	ip	Institut Pierre-Simon Laplace, France
MIROC-ESM	mr	Japan Agency for Marine-Earth Science and Technology, Japan
MPI-ESM-LR	mp	Max Planck Institute for Meteorology, Germany
NorESM1-M	no	Norwegian Climate Centre, Norway

Topographic parameters: A DEM at 90 m grid resolution (Figure 5.4a) was downloaded from the CGIAR Consortium for Spatial Information at the website: http://srtm.csi.cgiar.org/. The DEM was projected to the WGS84/UTM 15S (EPSG: 32715) coordinate system and used as a topographic template to calculate a map of wind effect (Figure 5.4b). The map of wind effect [247] was obtained using the SAGA-Wind effect module within the Geographic Information System QGIS v.2.12.

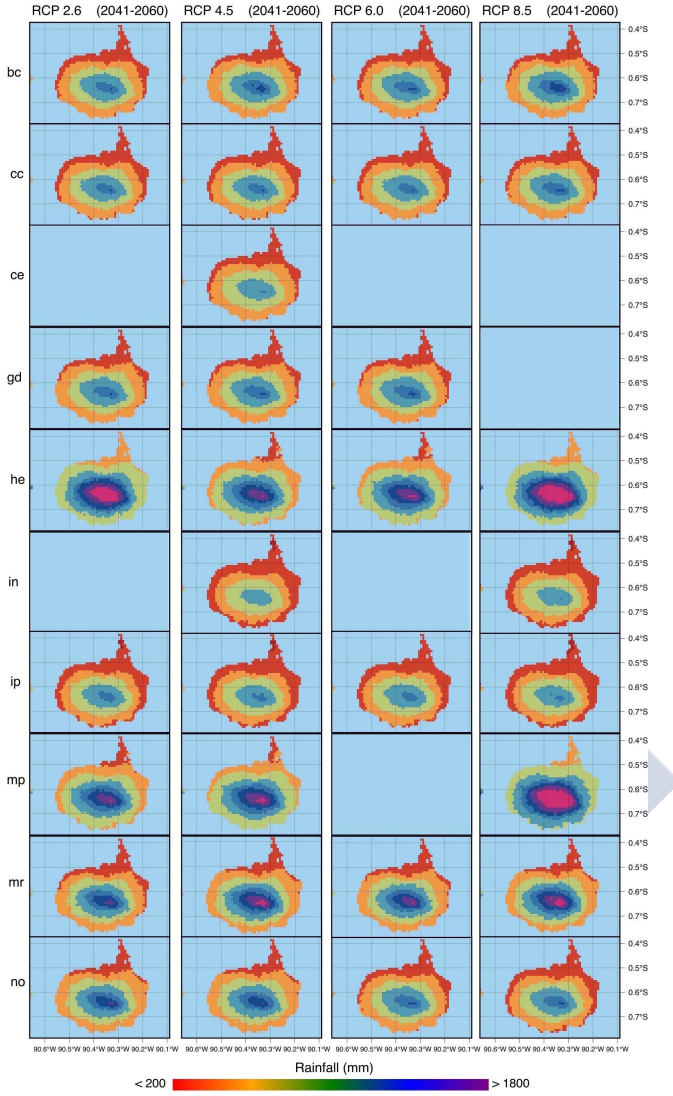


FIGURE 5.2: Estimation of mean annual precipitation for the period 2041-2060 under future climate scenarios according to different models within the CMIP5 project (Listed in Table 5.2). Missing maps correspond to RCP experiments for which models are not available.

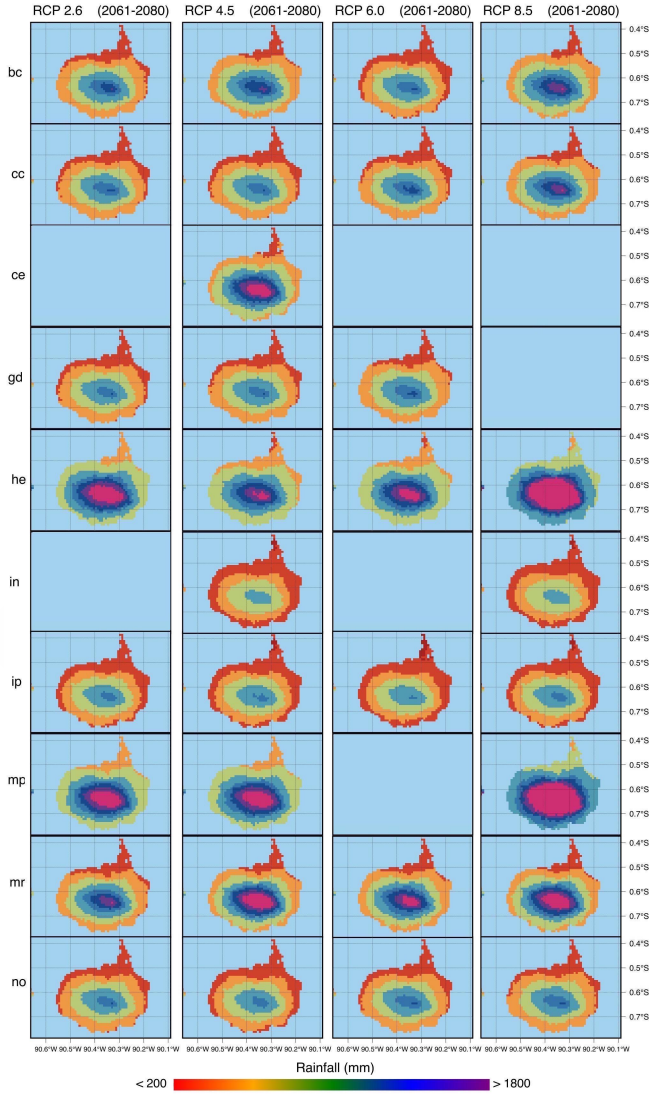


FIGURE 5.3: Estimation of mean annual precipitation for the period 2061-2080 under future climate scenarios according to different models within the CMIP5 project (Listed in Table 5.2). Missing maps correspond to RCP experiments for which models are not available.

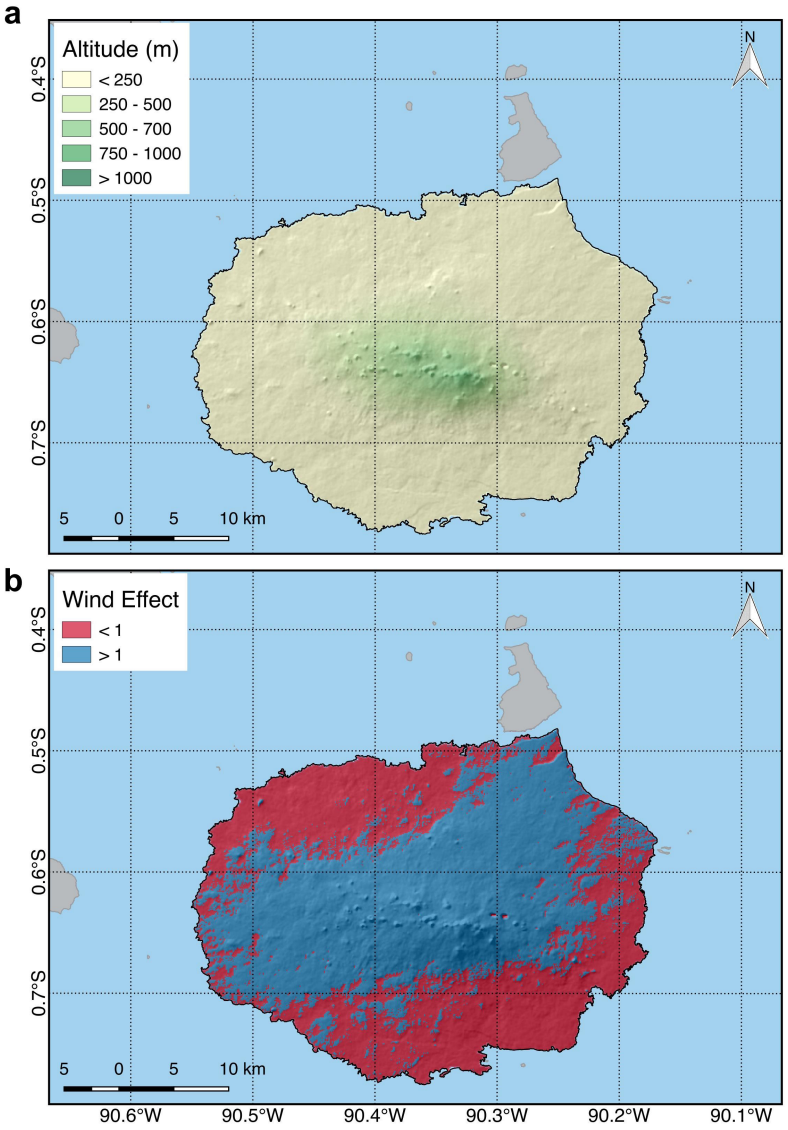


FIGURE 5.4: Environmental variables used for mapping purposes in Santa Cruz Island.In the map of wind effect the values above 1 indicate the areas more exposed to SE winds and viceversa.

5.5 Modelling procedures

Figure 5.5 summarizes the statistical steps used in this study. Firstly, we tested the potential of infrared data to predict SOC contents by means of a PLS regression model. As explained in the previous chapter, the algorithm [88] summarizes complex information from infrared data into few new artificial uncorrelated variables, the so-called latent vectors that summarize the maximum amount of spectroscopic information while maximizing the explained variance of SOC. In a second stage, we map SOC stocks for future and present times using rainfall data as covariate. 30 arc-second grid rainfall rasters ($\simeq$ 920 m) were downscaled to 90 m resolution data by relating rainfall data to elevation data and wind effect at the sampling locations by means of linear regression. The obtained models were then used to generalize the results of the different rainfall scenarios to the whole archipelago at higher spatial resolution. Modeled precipitations for period 1950-2000 at sampling locations and squared-root transformed SOC measurements were related by GWR, a method of spatially non-stationary linear regression.

Geographically Weighted Regression (GWR): GWR is a statistical method that can be used to determine changing relationships in space between the dependent and independent variables [89, 248]. Model predictions at location i (Y_i) are obtained according to the Equation 5.1):

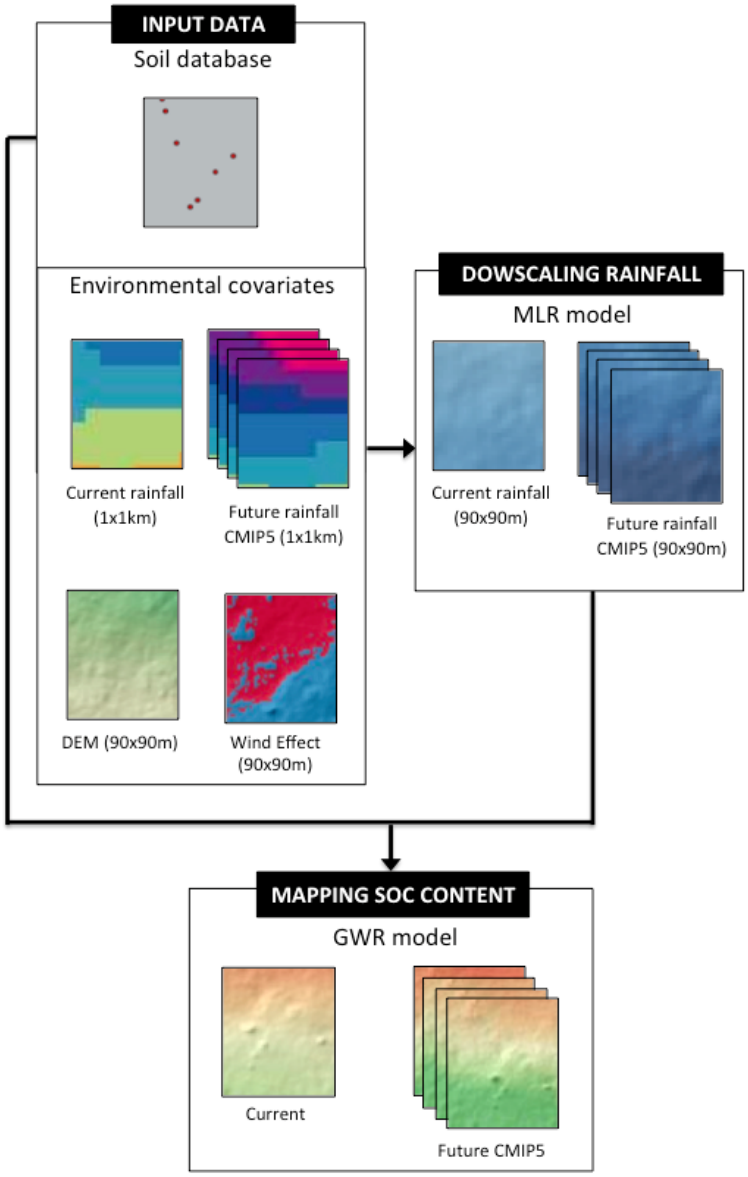


FIGURE 5.5: Statistical framework used for mapping SOC stocks Santa Cruz Island topsoils.

$$Y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) X_{ik} + \varepsilon_i \tag{5.1}$$

where (u_i, v_i) are the spatial coordinates at i, β_0 and β_k are the estimated regression coefficients, X_{ik} are the values of the independent variables at i and ε_i is the residual error. The regression coefficients (β) are determined by means of a weighting function, shown in Equation 5.2):

$$\beta_(u_i, v_i) = (X^T W(u_i, v_i) X)^{(-1)} X^T W(u_i, v_i) Y \tag{5.2}$$

where X^T is the matrix of environmental variables, Y is the response variable and $W(u_i, v_i)$ are weighting factors used to estimate the influence of each observation on the predicted values within its neighbourhood [89, 249]. The GWR model was validated by using Leave-One-Out (LOO) cross validation, due to the limited number of observations in the dataset.

The SE of predictions was used to depict the spatial uncertainty of SOC forecasts obtained by our GWR model. It was calculated through the following equation [89]:

$$SE = \sqrt{Var(\beta(u_i, v_i))} \tag{5.3}$$

where $\beta(u_i, v_i)$ is the regression coefficient of the GWR equation at each location.

The values of SOC content predicted by GWR and soil bulk density was used to calculate SOC stocks. Soil bulk density (g cm^{-3}) was estimated by the Adams equation:

$$\rho_b = \frac{100}{\frac{OM\%}{\rho_{OM}} + \frac{(100-OM\%)}{\rho_M}} \tag{5.4}$$

where ρ_{OM} is the density of organic matter ($\rho_{OM} = 0.224$ g cm^{-3}), ρ_M is the mineral bulk density ($\rho_M = 1.64$ g cm^{-3}) and $OM\% = 1.72\ SOC\%$.

Finally, SOC stocks (Kg m^{-2}) were calculated for the upper 10 cm (Equation 5.5):

$$SOC_{STOCK} = \frac{SOC}{\rho_b} 1000d \tag{5.5}$$

where d is the depth of the soil [191]. We have considered a constant depth of 10 cm that corresponds to the mean topsoil depth in the original field descriptions. Finally, the GWR model was used to estimate future SOC content and SOC stocks using predicted rainfall data from the different CMIP5 models previously mentioned and the differences between the SOC stocks for the period 1950-2000 and future SOC stocks

for all the CMIP5 individual and the ensemble models were calculated (Table 5.3).

5.6 Use of infrared data for mapping

We evaluated the capacity of infrared spectroscopy (FTIR-ATR) to predict SOC content in Santa Cruz Island soils. The PLS model here created indicates that regression on 2 latent vectors can explain most of the variability in SOC in our samples. The model was validated by LOO cross validation showing a relatively good accuracy (R^2 = 0.58, RMSE = 2.56, MAE = 2.18). Infrared spectroscopy can be applied in situ [155] and on a small amount of sample. Thus, the good correlation between infrared data and SOC content suggests that spectroscopic measurements can be used in regular monitoring programs to evaluate SOC stocks and changes in this highly sensitive ecosystem, where human disturbances must be minimized.

5.7 Factors influencing the amount of SOC

Mean normal rainfall (P) for the period 1950-2000 was downscaled from 1 km to 90 m by linear regression using

TABLE 5.3: Variations in the SOC stocks according to the different CMIP5 models under the RCP scenarios here considered.

	Period 2041-2060				**Period 2061-2080**			
CMIP5 code	**RCP 2.6**	**RCP 4.5**	**RCP 6.0**	**RCP 8.5**	**RCP 2.6**	**RCP 4.5**	**RCP 6.0**	**RCP 8.5**
bc	-0.38%	4.72%	0.34%	5.84%	10.29%	15.34%	-0.82%	21.31%
cc	-1.86%	-4.46%	-1.29%	1.69%	1.56%	31.27%	4.25%	15.41%
ce	-	-1.99%	-	-	-	0.35%	-	-
gd	1.84%	2.41%	2.87%	-	-0.91%	0.48%	5.26%	-
he	42.14%	24.29%	27.45%	46.94%	48.22%	31.84%	36.38%	66.07%
in	-	-12.20%	-	-9.94%	-	-14.54%	-	-13.71%
ip	-4.67%	-6.78%	-4.55%	-8.58%	-5.28%	-7.35%	-10.73%	-5.86%
mr	23.38%	30.85%	12.64%	53.40%	44.48%	47.03%	22.14%	73.93%
mp	9.11%	17.75%	-	19.17%	13.54%	31.33%	-	31.33%
no	9.68%	9.38%	1.77%	1.34%	-1.22%	-4.60%	1.45%	2.37%
ensemble	10.70%	6.93%	5.93%	15.47%	14.92%	14.53%	8.98%	28.22%
SD	15.78	14.07	11.01	24.28	21.02	21.05	15.80	32.02

elevation (DEM) and wind direction (WIND) as independent proxies (Equation 5.6). This period corresponds to the period in which soil samples were collected. The model showed a high performance for the period 1950-2000 (R^2= 0.92, RMSE = 55.74, MAE = 42.58).

$$P = 533.94 + 42.88WIND + 0.92DEM \tag{5.6}$$

The same procedure was used to downscaled rainfall scenarios for periods 2041-2060 and 2061-2080 (Tables 5.4 and 5.5).

The GWR model here developed shows a high positive relationship between SOC and rainfall for period 1950-2000, indicating that rainfall is the main factor controlling SOC accumulation in Santa Cruz soils.

The regression coefficients in Equation 5.6 show that both elevation and wind effect are positively correlated to precipitation. Galapagos climate is controlled by the north-south migration of the Inter-Tropical Convergence Zone (ITCZ) that generates the two seasons that characterize the islands climatic conditions. In the dry cold season the ITCZ migrates northwards and the southeast coolest trade winds predominate over the islands producing a condensation above 250 meters altitude that creates a permanent fog layer in the windward areas. These clouds result in vertical and occult

TABLE 5.4: Fitted values obtained for downscaling rainfall data extracted from each CMIP5 model under fourth RCP scenarios for the period 2041-2060. (-) RCP experiment not available.

CMIP5 code	**RCP 2.6**	**RCP 4.5**	**RCP 6.0**	**RCP 8.5**
bc	R^2=0.92 RMSE=60.79 MAE=46.40	R^2=0.92 RMSE=65.04 MAE=49.67	R^2=0.92 RMSE=63.46 MAE=48.37	R^2=0.92 RMSE=68.88 MAE=52.70
cc	R^2=0.92 RMSE=59.42 MAE=45.41	R^2=0.91 RMSE=51.38 MAE=39.26	R^2=0.92 RMSE=59.08 MAE=45.09	R^2=0.92 RMSE=61.67 MAE=47.24
ce	-	R^2=0.92 RMSE=59.48 MAE=45.54	-	-
gd	R^2=0.92 RMSE=59.65 MAE=45.32	R^2=0.92 RMSE=62.58 MAE=47.57	R^2=0.92 RMSE=65.10 MAE=49.55	-
he	R^2=0.91 RMSE=97.02 MAE=73.57	R^2=0.91 RMSE=81.23 MAE=61.61	R^2=0.91 RMSE=82.64 MAE=62.56	R^2=0.91 RMSE=102.33 MAE=77.61
in	-	R^2=0.91 RMSE=51.01 MAE=38.95	-	R^2=0.91 RMSE=54.16 MAE=41.34
ip	R^2=0.91 RMSE=60.19 MAE=45.79	R^2=0.92 RMSE=58.63 MAE=44.67	R^2=0.92 RMSE=58.86 MAE=44.74	R^2=0.92 RMSE=57.13 MAE=43.60
mr	R^2=0.91 RMSE=76.55 MAE=58.53	R^2=0.91 RMSE=80.76 MAE=61.73	R^2=0.92 RMSE=79.93 MAE=60.93	R^2=0.90 RMSE=99.51 MAE=75.95
mp	R^2=0.92 RMSE=76.29 MAE=58.16	R^2=0.92 RMSE=84.96 MAE=64.66	-	R^2=0.92 RMSE=88.11 MAE=67.08
no	R^2=0.92 RMSE=69.90 MAE=53.45	R^2=0.92 RMSE=71.14 MAE=54.23	R^2=0.92 RMSE=62.81 MAE=47.84	R^2=0.92 RMSE=62.56 MAE=47.84
ensemble	R^2=0.92 RMSE=69.96 MAE=53.33	R^2=0.92 RMSE=66.6 MAE=50.79	R^2=0.92 RMSE=67.40 MAE=51.30	R^2=0.91 RMSE=74.27 MAE=56.67

TABLE 5.5: Fitted values obtained for downscaling rainfall data extracted from each CMIP5 model under fourth RCP scenarios for the period 2061-2080. (-) RCP experiment not available.

CMIP5 code	**RCP 2.6**	**RCP 4.5**	**RCP 6.0**	**RCP 8.5**
bc	R^2=0.92 RMSE=69.30 MAE=52.66	R^2=0.91 RMSE=73.03 MAE=55.78	R^2=0.92 RMSE=63.93 MAE=48.77	R^2=0.91 RMSE=79.59 MAE=60.68
cc	R^2=0.92 RMSE=63.32 MAE=48.32	R^2=0.92 RMSE=97.05 MAE=73.49	R^2=0.92 RMSE=65.69 MAE=50.01	R^2=0.92 RMSE=75.24 MAE=57.49
ce	-	R^2=0.92 RMSE=60.86 MAE=46.55	-	-
gd	R^2=0.92 RMSE=55.67 MAE=43.80	R^2=0.92 RMSE=58.49 MAE=44.59	R^2=0.92 RMSE=63.05 MAE=47.89	-
he	R^2=0.90 RMSE=102.58 MAE=77.69	R^2=0.91 RMSE=84.48 MAE=63.97	R^2=0.91 RMSE=90.50 MAE=68.76	R^2=0.91 RMSE=131.76 MAE=99.31
in	-	R^2=0.91 RMSE=48.97 MAE=37.20	-	R^2=0.91 RMSE=49.17 MAE=37.19
ip	R^2=0.92 RMSE=57.94 MAE=44.18	R^2=0.91 RMSE=59.12 MAE=44.98	R^2=0.91 RMSE=57.46 MAE=43.82	R^2=0.91 RMSE=59.38 MAE=45.21
mr	R^2=0.91 RMSE=99.33 MAE=75.74	R^2=0.91 RMSE=96.28 MAE=73.42	R^2=0.92 RMSE=91.19 MAE=69.46	R^2=0.90 RMSE=125.80 MAE=95.84
mp	R^2=0.92 RMSE=81.18 MAE=61.88	R^2=0.92 RMSE=107.47 MAE=81.83	-	R^2=0.92 RMSE=105.77 MAE=80.61
no	R^2=0.92 RMSE=59.20 MAE=45.18	R^2=0.92 RMSE=58.21 MAE=44.49	R^2=0.92 RMSE=61.46 MAE=46.97	R^2=0.92 RMSE=61.81 MAE=47.12
ensemble	R^2=0.91 RMSE=73.80 MAE=56.18	R^2=0.92 RMSE=74.37 MAE=56.63	R^2=0.92 RMSE=70.45 MAE=53.67	R^2=0.91 RMSE=86.00 MAE=65.43

rainfall due to the condensation of fog on the vegetation and on the ground. In the opposite, in the hot season the ITCZ migrates southwards producing a weakening in the trade winds and the prevalence of typical tropical conditions marked by the orographic rainfall that increases with altitude [250–252]. Most of the rain events occur during the Wet Hot Season, when winds are weak, and elevation is the parameter controlling the amount of water entering in the soil by precipitation. However, the dynamics of SOC in this area is also affected by the amount of occult rainfall (water condensation from fogs) entering in the system during the Dry Cold Season. Pryet [253] determined that the occult precipitation in high altitudes of Santa Cruz Island for the year 2010 was 86 ± 50 mm, representing the 22 ± 13% of water entries to the soil system. This estimation is slightly higher than the values obtained from our equation but were calculated only for one year period and not using values for a normal climatic period as we used in this study.

Rainfall simulations were used to indicate the trend of SOC stocks in Santa Cruz expected under the different climate scenarios forecasted by the IPCC. However, a great variability exist among the predictions of the individual CMIP5 climatic models (Figure 5.7c). The predictions for future climate scenarios in this area are highly uncertain. Climate in Galapagos is markedly influenced by both its position in the ITCZ and the incidence of ENSO events.

Paleoclimatic studies on lake sediments in different islands in the Pacific region [250, 251] revealed that the location of ITCZ and the incidence of ENSO varied during the last millennia, producing an alternation of long humid and arid periods. To accurately model the effects of global warming taking into account the periodicity (2-7 years) and intensity of future ENSO phenomenon is still a challenge for climate modellers [254]. Since the beginning of the Industrial Revolution, climate in Galapagos shifted towards wetter climatic conditions [237, 251] and an intensification of ENSO events due to global warming is expected in the near future [254]. An ENSO intensification can modify the capacity of soils to remove atmospheric CO_2 at global scale [6]. During ENSO, the proliferation of invasive species better adapted to extreme climatic conditions and a rise in the mortality of native species in the arid zone were reported [255, 256]. Vegetation in the humid zone presents a higher resilience due to its association to a persistent fog layer that buffers hydric deficits in soils and vegetation. The formation of this layer is mainly a function of Sea Surface Temperature (SST), wind direction and humidity. However, changes in these parameters, such as a weakening in the Walker winds circulation, can lead to the reduction of the fog layer and the modification of the existing vegetation types [237, 257, 258]. In addition, the uncertainty in the definition of the ITCZ and the possible occurrence of a double ITCZ constitutes an

important limitation for the existing models to obtain robust precipitation forecasts in the medium and long term [259, 260] (Figure 5.2 and 5.3).

5.8 Spatial distribution of SOC content

Mean SOC content in samples is 9.9%, with values ranging from 0.7 to 18.1%. The higher SOC contents correspond to the samples located at higher elevations, while minimum values were observed in soils from the arid areas in the coastal region (Figure 5.6).

A GWR model was used to relate SOC content in topsoil samples to the predicted rainfall values at sampling locations. The model showed a high spatial correlation between squared-root transformed SOC measurements and rainfall data. The validation of the model was calculated by LOO cross validation indicating a good model performance (R^2 = 0.66, RMSE = 0.51, MAE = 0.43). The highest SOC content (>16%) was found for soils from the humid highlands, while the lower contents (about 4%) correspond to soils placed in the arid coastal area. SOC content and bulk density were used to calculate SOC stocks in Santa Cruz Island for period 1950-2000. Our results indicate that soils in Santa Cruz,

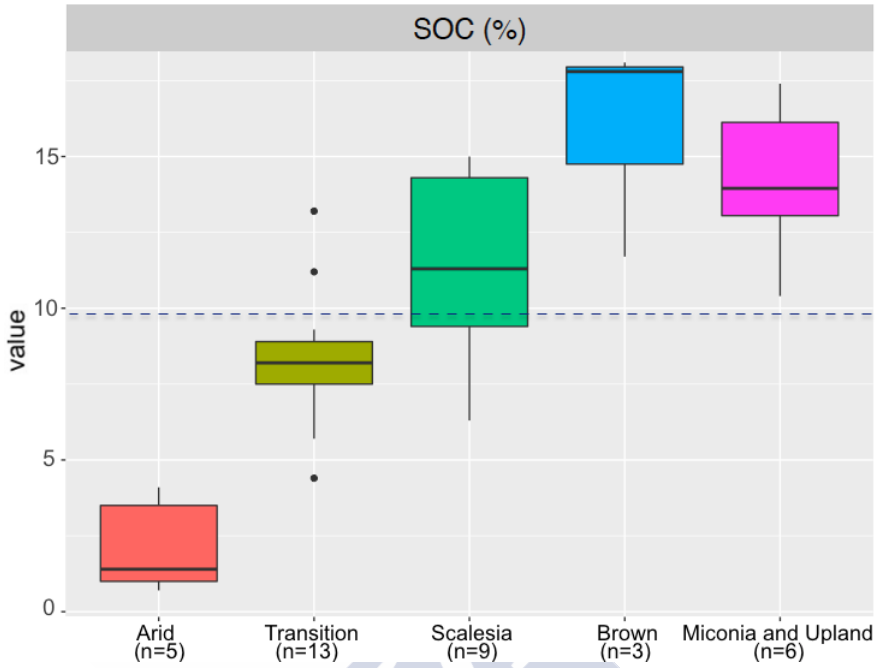


FIGURE 5.6: Boxplot relating SOC content to bioclimatic belts in samples from Santa Cruz Island. The dotted line represents the mean SOC value.

during the sampling period, accumulated about 706 Gg of SOC in the upper 10 cm (Figure 5.7a). All the islands of the archipelago present similar environmental conditions. Vegetated zones in the windward areas were clearly positively correlated to elevation [246, 261, 262]. In leeward slopes, the arid region extends to higher altitudes due to the shadow effect in wet and diffuse precipitation coming from the dominant humid SE trade winds [242] (Figure 5.1). The spatial variation in rainfall rates is a key factor influencing the

differentiation of vegetation and soils in the area [243, 246]. Although the number of soil samples available for the area is limited and mostly located in the windward area of Santa Cruz Island, the same precipitation pattern, highly correlated to altitude and windward direction values, is expected for the remaining islands. In this situation, a generalization of the results of this model to the entire archipelago indicates that soils in Galapagos may accumulate up to 54 Tg of C in the upper 10 cm.

The GWR model was also used to estimate the variation in SOC stocks associated with future changes in rainfall rates as forecasted by the fourth RCP scenarios [263] from the Intergovernmental Panel on Climate Change Fifth Assessment Report (IPCC AR5) (Figure 5.7). Table 5.3 shows the differences of SOC stocks for future times with respect to stock value for the sampling period 1950-2000. We used the rainfall data from 10 climatic models developed by the groups joined to the CMIP5 to calculate SOC stocks and differences. CMIP5 models are usually expressed as an ensemble multi-model calculated from average precipitation of the 10 CMIP5 models. Ensemble models indicate an increase in SOC stocks for all periods and RCP scenarios, with a minimum SOC stock value of 748 Gg in period 2041-2060 under RCP 6.0 and a maximum value of 906 Gg in 2061-2080 under RCP 8.5. This indicates an increase in the SOC stocks of about 5.93% and 28.22% for periods 2041-2060 RCP 6.0 and

2061-2080 RCP 8.5 respectively in relation to the stocks estimated for period 1950-2000 (Figure 5.7).

Our model uses CMIP5 weather forecasts for lag periods of 20 years. In a recent study, Carvalhais *et al.* [56] used the same set of CMIP5 models to analyze the global covariation of carbon turnover with climate, finding that the mean carbon turnover rate for equatorial areas is about 15 years. These authors also identified a strong association between SOC and precipitation indicating that, contrary to what was considered in most of the Earth system models used at present, future climate/carbon-cycle feedbacks will strongly depend on changes in the hydrological cycle. This estimated turnover value for SOC in equatorial areas is slightly lower than the climatic lag periods considered in this study, thus the response in SOC contents due to such climate variations was accounted for our model.

The resulting forecasts of SOC stocks under future climatic scenarios (Figure 5.7b) show an increase of SOC stocks in soils in Santa Cruz. An increase in rainfall rates can also influence the extent of the arid zone, promoting changes in the vegetation cover and further enhancing the accumulation of SOC in the island.

Due to the high correlation between precipitation rates and SOC, we consider that long-term differences in precipitation rates due to the potential effect of climate change could be

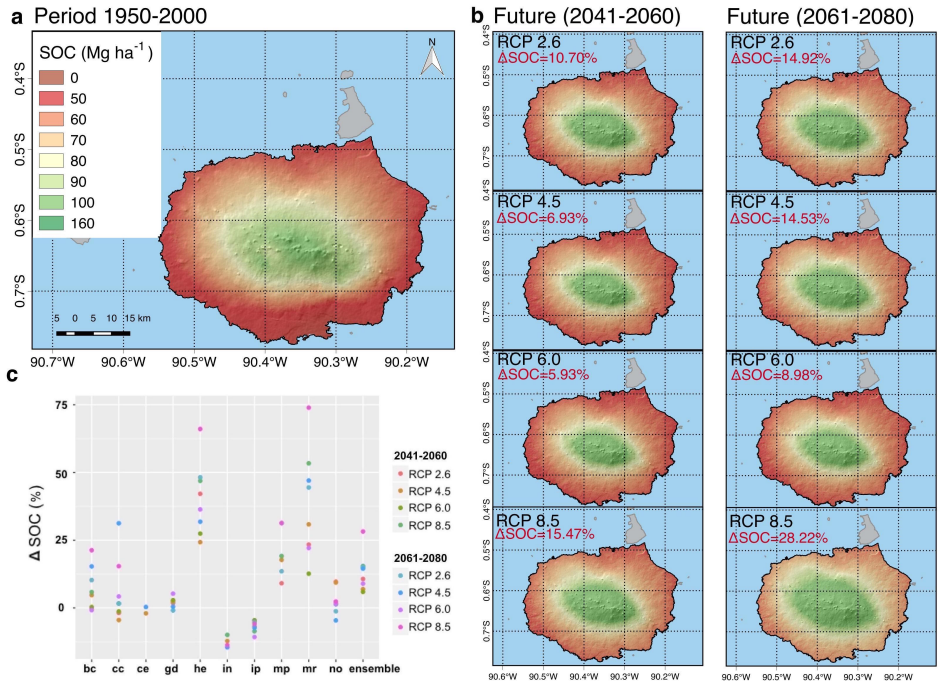


FIGURE 5.7: **Map of predicted SOC stocks in Santa Cruz Island.**
a,b, SOC accumulation in the upper 10 cm of soils (Mg ha^{-1}) for period 1950-2000 and future RCP climate change scenarios. Percentage of SOC variation (ΔSOC) represents the increase in SOC stocks with respect to the values for the period 1950-2000 and the scenarios considered.
c, Variation in SOC stocks between the sampling period and future for each CMIP5 model.

tracked through a monitoring program based in the evolution of the SOC stocks.

5.9 Uncertainty of SOC estimations

A map of SE was calculated to account for the spatial uncertainty of our SOC predictions (Figure 5.8). The results show that high model uncertainty (± 0.37%) is associated with high elevation areas of the island.

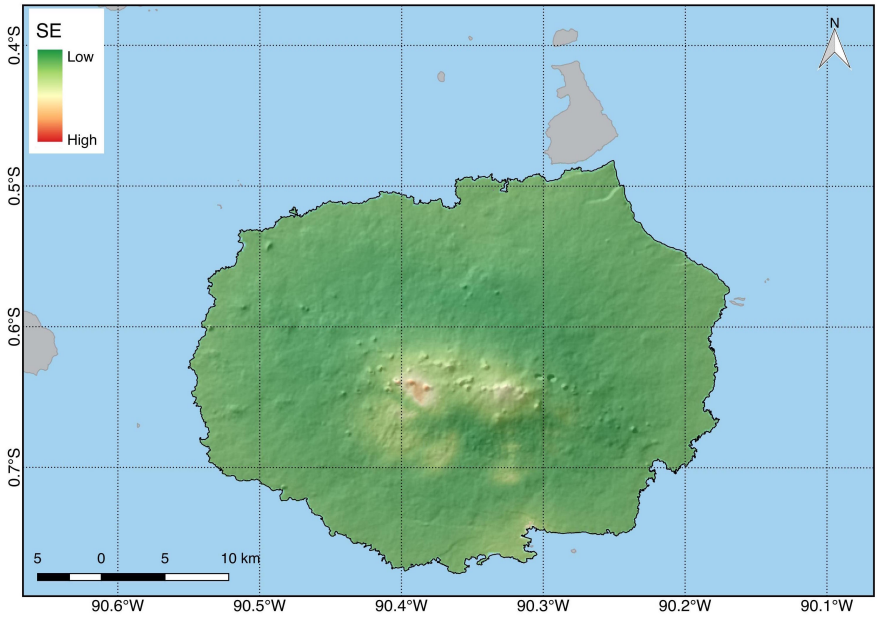


FIGURE 5.8: Map of SE of the SOC predictions in Santa Cruz Island.
(mean = ±0.13% SOC, max = ±0.37% SOC)

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 6

CONCLUSIONS

- The potential of spectroscopic techniques as a tool for mapping soil organic carbon at topsoil level was demonstrated by using samples from three different areas.
- Spectroscopic data was used in two regression approaches showing its capacity for predict soil organic carbon content, minimizing once a time the analytical cost and time efforts.
- Spectroscopic data was used in a classification approach proving the capacity of spectra for aggregating similar soil samples.
- The developed statistical models showed a relative good predictive performance in the majority of cases.

- The factors that promote SOC accumulation differ between the three study cases.

- The statistical framework developed to model the distribution of topsoil organic carbon in Galicia (NW Spain) from infrared data had a high predictive power.
- A random forest model allowed the identification of regions of the spectra more relevant for soil organic carbon predictions in Galicia.
- The availability of soil water is the main parameter influencing the accumulation of organic carbon in Galicia.
- The highest soil organic carbon values found in Galician topsoils correspond to areas located on the mountain ranges and the Atlantic coast.
- Soil organic carbon predictions obtained by infrared data were similar to those obtained by wet chemistry data for Galician topsoils.

- A spatially non-stationary approach to map organic carbon in Europe was successfully created.
- Visible-near infrared spectra let to classify European samples in groups according to their similar spectroscopic features.
- The tagged groups differ in the values of soil pH, content of coarse fragments and climatic features.
- The statistical method developed to map SOC content allows knowing the spatial weight of factors promoting carbon storage along Europe.
- Climate is the main factor influencing carbon accumulation in Europe at continental scale.
- We also identified the processes promoting the accumulation at local level such as the land cover type.
- The highest soil organic carbon content in Europe was found for the highest altitudes in the Alps, Pyrenees and Apennines, in the NE of United Kingdom and in the northern Scandinavian countries.
- This model partially improved the accuracy of previous studies made at European scale using the same geochemical database.

- A Geographically Weighted Regression model permitted to predict the distribution of organic carbon stocks in topsoils from Santa Cruz Island (Galapagos) from legacy data.

- Rainfall is the main driver of soil organic carbon accumulation in soils from Santa Cruz Island.

- The higher carbon concentrations were found at high altitude in the windward slope of Santa Cruz Island, where rainfall rates are greater.

- Climatic forecasts indicate that climate change will produce an increase of soil organic carbon stocks in the Galapagos Islands.

- It was demonstrated the capacity of infrared spectroscopy to make predictions of soil organic carbon content in a quickly manner by using soils from Santa Cruz Island.

- Infrared data could be used as a monitoring system to control the progression of climate change in the Galapagos Islands.

Bibliography

[1] P. Falkowski, "The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System," *Science*, vol. 290, no. 5490, pp. 291–296, 2000.

[2] T. M. Kusky, *Climate change: shifting glaciers, deserts, and climate belts*. New York: Facts on File, 2009.

[3] J. G. Canadell, C. Le Quéré, M. R. Raupach, C. B. Field, E. T. Buitenhuis, P. Ciais, T. J. Conway, N. P. Gillett, R. A. Houghton, and G. Marland, "Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 47, pp. 18866–18870, 2007.

[4] IPCC, *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC, Core Writing Team and Pachauri, R. K. and Mayer, L. ed., 2014.

[5] R. A. Houghton, "Revised estimates of the annual net flux of carbon to the atmosphere from changes in land use and land management 1850-2000," *Tellus B*, vol. 55, no. 2, pp. 378–390, 2003.

[6] P. Cox, R. Betts, C. Jones, S. Spall, and I. Totterdell, "Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model," *Nature*, vol. 408, pp. 184–187, 2000.

[7] S. Doetterl, A. Stevens, J. Six, R. Merckx, K. Van Oost, M. Casanova Pinto, A. Casanova-Katny, C. Muñoz, M. Boudin, E. Zagal Venegas, and P. Boeckx, "Soil carbon storage controlled by interactions between geochemistry and climate," *Nature Geoscience*, vol. 8, no. 10, pp. 780–783, 2015.

[8] S. Solomon, G.-K. Plattner, R. Knutti, and P. Friedlingstein, "Irreversible climate change due to carbon dioxide emissions," *Proceedings of the National Academy of Sciences*, vol. 106, no. 6, pp. 1704–1709, 2009.

[9] A. J. McMichael, R. E. Woodruff, and S. Hales, "Climate change and human health: present and future risks," *The Lancet*, vol. 367, no. 9513, pp. 859–869, 2006.

[10] N. Oreskes, "Beyond The Ivory Tower: The Scientific Consensus on Climate Change," *Science*, vol. 306, no. 5702, pp. 1686–1686, 2004.

[11] C. Le Quéré, R. M. Andrew, J. G. Canadell, S. Sitch, J. I. Korsbakken, G. P. Peters, A. C. Manning, T. A. Boden, P. P. Tans, R. A. Houghton, R. F. Keeling, S. Alin, O. D. Andrews, P. Anthoni, L. Barbero, L. Bopp, F. Chevallier, L. P. Chini, P. Ciais, K. Currie, C. Delire, S. C. Doney, P. Friedlingstein, T. Gkritzalis, I. Harris, J. Hauck, V. Haverd, M. Hoppema, K. Klein Goldewijk, A. K. Jain, E. Kato, A. Körtzinger, P. Landschützer, N. Lefèvre, A. Lenton, S. Lienert, D. Lombardozzi, J. R. Melton, N. Metzl, F. Millero, P. M. S. Monteiro, D. R. Munro, J. E. M. S. Nabel, S.-i. Nakaoka, K. O'Brien, A. Olsen, A. M. Omar, T. Ono, D. Pierrot, B. Poulter, C. Rödenbeck, J. Salisbury, U. Schuster, J. Schwinger, R. Séférian, I. Skjelvan, B. D. Stocker, A. J. Sutton, T. Takahashi,

H. Tian, B. Tilbrook, I. T. van der Laan-Luijkx, G. R. van der Werf, N. Viovy, A. P. Walker, A. J. Wiltshire, and S. Zaehle, "Global Carbon Budget 2016," *Earth System Science Data*, vol. 8, no. 2, pp. 605–649, 2016.

[12] R. Lal, "Beyond COP 21: Potential and challenges of the "4 per Thousand" initiative," *Journal of Soil and Water Conservation*, vol. 71, no. 1, pp. 20A–25A, 2016.

[13] IGBP Terrestrial Carbon Working Group, "The Terrestrial Carbon Cycle: Implications for the Kyoto Protocol," *Science*, vol. 280, no. 5368, pp. 1393–1394, 1998.

[14] C. Breidenich, D. Magraw, A. Rowley, and J. W. Rubin, "The Kyoto Protocol to the United Nations Framework Convention on Climate Change," *The American Journal of International Law*, vol. 92, no. 2, pp. 315–331, 1998.

[15] N. Anger and J. Sathaye, "Reducing Deforestation and Trading Emissions: Economic Implications for the Post-Kyoto Carbon Market," *SSRN Electronic Journal*, 2008.

[16] N. H. Stern, *The economics of climate change: the Stern review*. UK: Cambridge University Press, 2007.

[17] M. Wara, "Is the global carbon market working?," *Nature*, vol. 445, no. 7128, pp. 595–596, 2007.

[18] M. H. Babiker, "Climate change policy, market structure, and carbon leakage," *Journal of International Economics*, vol. 65, no. 2, pp. 421–445, 2005.

[19] A. F. Bouwman, International Soil Reference and Information Centre, Netherlands, Commission of the European Communities, and United Nations Environment Programme, *Soils and the*

greenhouse effect: the present status and future trends concerning the effect of soils and their cover on the fluxes of greenhouse gases, the surface energy balance, and the water balance: proceedings of the International Conference Soils and the Greenhouse Effect. Chichester, New York: Wiley, 1990.

[20] J. P. Scharlemann, E. V. Tanner, R. Hiederer, and V. Kapos, "Global soil carbon: understanding and managing the largest terrestrial carbon pool," *Carbon Management*, vol. 5, no. 1, pp. 81–91, 2014.

[21] M. Köchy, R. Hiederer, and A. Freibauer, "Global distribution of soil organic carbon – Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world," *SOIL*, vol. 1, no. 1, pp. 351–365, 2015.

[22] Y. Yigini, L. Montanarella, and P. Panagos, "European Contribution Towards a Global Assessment of Agricultural Soil Organic Carbon Stocks," *Advances in Agronomy*, vol. 142, pp. 385–410, 2017.

[23] E. G. Jobbágy and R. B. Jackson, "The vertical distribution of Soil Organic Carbon and its relation to climate and vegetation," *Ecological Applications*, vol. 10, no. 2, pp. 423–436, 2000.

[24] H. Tian, C. Lu, J. Yang, K. Banger, D. N. Huntzinger, C. R. Schwalm, A. M. Michalak, R. Cook, P. Ciais, D. Hayes, M. Huang, A. Ito, A. K. Jain, H. Lei, J. Mao, S. Pan, W. M. Post, S. Peng, B. Poulter, W. Ren, D. Ricciuto, K. Schaefer, X. Shi, B. Tao, W. Wang, Y. Wei, Q. Yang, B. Zhang, and N. Zeng, "Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions," *Global Biogeochemical Cycles*, vol. 29, no. 6, pp. 775–792, 2015.

[25] R. Lal, "Soil Carbon Sequestration Impacts on Global Climate Change and Food Security," *Science*, vol. 304, no. 5677, pp. 1623–1627, 2004.

[26] E. Lugato, F. Bampa, P. Panagos, L. Montanarella, and A. Jones, "Potential carbon sequestration of European arable soils estimated by modelling a comprehensive set of management practices," *Global Change Biology*, vol. 20, no. 11, pp. 3557–3567, 2014.

[27] E. Lugato, K. Paustian, P. Panagos, A. Jones, and P. Borrelli, "Quantifying the erosion effect on current carbon budget of European agricultural soils at high spatial resolution," *Global Change Biology*, vol. 22, no. 5, pp. 1976–1984, 2016.

[28] S. M. O'Rourke, D. A. Angers, N. M. Holden, and A. B. McBratney, "Soil organic carbon across scales," *Global Change Biology*, vol. 21, no. 10, pp. 3561–3574, 2015.

[29] K. E. O. Todd-Brown, J. T. Randerson, F. Hopkins, V. Arora, T. Hajima, C. Jones, E. Shevliakova, J. Tjiputra, E. Volodin, T. Wu, Q. Zhang, and S. D. Allison, "Changes in soil organic carbon storage predicted by Earth system models during the 21st century," *Biogeosciences*, vol. 11, no. 8, pp. 2341–2356, 2014.

[30] V. Arora, G. J. Boer, P. Friedlingstein, M. Eby, C. D. Jones, J. R. Christian, G. Bonan, L. Bopp, V. Brovkin, P. Cadule, T. Hajima, T. Ilyina, K. Lindsay, J. F. Tjiputra, and T. Wu, "Carbon–Concentration and Carbon–Climate Feedbacks in CMIP5 Earth System Models," *Journal of Climate*, vol. 26, no. 15, pp. 5289–5314, 2013.

[31] V. O. Polyakov and R. Lal, "Soil erosion and carbon dynamics under simulated rainfall," *Soil Science*, vol. 169, no. 8, pp. 590–599, 2004.

[32] K. R. Olson, M. Al-Kaisi, R. Lal, and L. Cihacek, "Impact of soil erosion on soil organic carbon stocks," *Journal of Soil and Water Conservation*, vol. 71, no. 3, pp. 61A–67A, 2016.

[33] A. Chappell, J. Baldock, and J. Sanderman, "The global significance of omitting soil erosion from soil organic carbon cycling schemes," *Nature Climate Change*, vol. 6, pp. 187–191, 2015.

[34] F. Kirkels, L. Cammeraat, and N. Kuhn, "The fate of soil organic carbon upon erosion, transport and deposition in agricultural landscapes — A review of different concepts," *Geomorphology*, vol. 226, pp. 94–105, 2014.

[35] P. Lagacherie and A. McBratney, *Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping*, vol. 31 of *Digital Soil Mapping An Introductory Perspective*. Elsevier, 2007.

[36] B. Minasny, A. B. McBratney, and R. M. Lark, *Digital Soil Mapping Technologies for Countries with Sparse Data Infrastructures*. Netherlands: Springer, 2008.

[37] P. Lagacherie, A. B. McBratney, M. Voltz, and Global Workshop on Digital Soil Mapping, *Digital soil mapping: an introductory perspective*. Amsterdam: Elsevier, 2007.

[38] J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, and S. Kienast-Brown, *Digital Soil Mapping*. Dordrecht: Springer, 2010.

[39] B. Minasny, B. Malone, and A. B. McBratney, eds., *Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia*. Australia: CRC Press, 2012.

[40] A. McBratney, M. Mendonça Santos, and B. Minasny, "On digital soil mapping," *Geoderma*, vol. 117, no. 1-2, pp. 3–52, 2003.

[41] B. Minasny and A. McBratney, "Digital soil mapping: A brief history and some lessons," *Geoderma*, vol. 264, pp. 301–311, 2016.

[42] J. Schelling, "Soil genesis, soil classification and soil survey," *Geoderma*, vol. 4, no. 3, pp. 165–193, 1970.

[43] *Soil Survey Manual, US. Department of Agriculture Handbook 18. Soil Conversation Service.* Soil Survey Division Staff, 1993.

[44] V. Mulder, *Spectroscopy-supported digital soil mapping.* PhD thesis, Wageningen University, Netherlands, 2013.

[45] R. Tomlinson, *Design considerations for digital soil map systems. 11th Congress of Soil Science.* Canada: ISSS, 1978.

[46] E. C. Brevik, C. Calzolari, B. A. Miller, P. Pereira, C. Kabala, A. Baumgarten, and A. Jordán, "Soil mapping, classification, and pedologic modeling: History and future directions," *Geoderma*, vol. 264, pp. 256–274, 2016.

[47] D. Arrouays, M. G. Grundy, A. E. Hartemink, J. W. Hempel, G. B. Heuvelink, S. Y. Hong, P. Lagacherie, G. Lelyk, A. B. McBratney, N. J. McKenzie, M. Mendonça Santos, B. Minasny, L. Montanarella, I. O. Odeh, P. A. Sanchez, J. A. Thompson, and G.-L. Zhang, "GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties," *Advances in Agronomy*, vol. 125, pp. 93–134, 2014.

[48] P. A. Sanchez, S. Ahamed, F. Carre, A. E. Hartemink, J. Hempel, J. Huising, P. Lagacherie, A. B. McBratney, N. J. McKenzie, M. Mendonça Santos, B. Minasny, L. Montanarella, P. Okoth, C. A. Palm, J. D. Sachs, K. D. Shepherd, T.-G. Vagen, B. Vanlauwe, M. G.

Walsh, L. A. Winowiecki, and G.-L. Zhang, "Digital Soil Map of the World," *Science*, vol. 325, no. 5941, pp. 680–681, 2009.

[49] T. Hengl, J. M. de Jesus, R. A. MacMillan, N. H. Batjes, G. B. M. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J. G. B. Leenaars, M. G. Walsh, and M. R. Gonzalez, "SoilGrids1km — Global Soil Information Based on Automated Mapping," *PLOS ONE*, vol. 9, no. 8, p. e105992, 2014.

[50] T. Hengl, 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, "SoilGrids250m: Global gridded soil information based on machine learning," *PLOS ONE*, vol. 12, no. 2, p. e0169748, 2017.

[51] H. Jenny, *Factors of soil formation: a system of quantitative pedology*. New York: McGraw-Hill, 1941.

[52] R. J. A. Jones, R. Hiederer, E. Rusco, P. J. Loveland, and L. Montanarella, *The map of organic carbon in topsoils in Europe, Version 1.2, September 2003: Explanation of Special Publication Ispra 2004 No.72 (S.P.I.04.72). European Soil Bureau Research Report No.17, EUR 21209 EN, 26pp. and 1 map in ISO B1 format.* Luxembourg: Office for Official Publications of the European Communities, 2004.

[53] A. Bot and J. Benites, *The importance of soil organic matter: key to drought-resistant soil and sustained food production*. No. 80 in FAO soils bulletin, Rome: Food and Agriculture Organization of the United Nations, 2005.

[54] C. Tarnocai, J. G. Canadell, E. A. G. Schuur, P. Kuhry, G. Mazhitova, and S. Zimov, "Soil organic carbon pools in the northern circumpolar permafrost region," *Global Biogeochemical Cycles*, vol. 23, no. 2, pp. 1–11, 2009.

[55] M. Köchy, A. Don, M. K. van der Molen, and A. Freibauer, "Global distribution of soil organic carbon – Part 2: Certainty of changes related to land use and climate," *SOIL*, vol. 1, no. 1, pp. 367–380, 2015.

[56] N. Carvalhais, M. Forkel, M. Khomik, J. Bellarby, M. Jung, M. Migliavacca, M. Iu, S. Saatchi, M. Santoro, M. Thurner, U. Weber, B. Ahrens, C. Beer, A. Cescatti, J. T. Randerson, and M. Reichstein, "Global covariation of carbon turnover times with climate in terrestrial ecosystems," *Nature*, vol. 514, no. 7521, pp. 213–217, 2014.

[57] M. Delgado-Baquerizo, F. T. Maestre, A. Gallardo, M. A. Bowker, M. D. Wallenstein, J. L. Quero, V. Ochoa, B. Gozalo, M. García-Gómez, S. Soliveres, P. García-Palacios, M. Berdugo, E. Valencia, C. Escolar, T. Arredondo, C. Barraza-Zepeda, D. Bran, J. A. Carreira, M. Chaieb, A. A. Conceição, M. Derak, D. J. Eldridge, A. Escudero, C. I. Espinosa, J. Gaitán, M. G. Gatica, S. Gómez-González, E. Guzman, J. R. Gutiérrez, A. Florentino, E. Hepper, R. M. Hernández, E. Huber-Sannwald, M. Jankju, J. Liu, R. L. Mau, M. Miriti, J. Monerris, K. Naseri, Z. Noumi, V. Polo, A. Prina, E. Pucheta, E. Ramírez, D. A. Ramírez-Collantes, R. Romão, M. Tighe, D. Torres, C. Torres-Díaz, E. D. Ungar, J. Val, W. Wamiti, D. Wang, and E. Zaady, "Decoupling of soil nutrient cycles as a function of aridity in global drylands," *Nature*, vol. 502, no. 7473, pp. 672–676, 2013.

[58] O. Hararuk, M. J. Smith, and Y. Luo, "Microbial models with data-driven parameters predict stronger soil carbon responses to climate change," *Global Change Biology*, vol. 21, no. 6, pp. 2439–2453, 2015.

[59] Z. Zhu, S. Piao, R. B. Myneni, M. Huang, Z. Zeng, J. G. Canadell, P. Ciais, S. Sitch, P. Friedlingstein, A. Arneth, C. Cao, L. Cheng, E. Kato, C. Koven, Y. Li, X. Lian, Y. Liu, R. Liu, J. Mao, Y. Pan, S. Peng, J. Peñuelas, B. Poulter, T. A. M. Pugh, B. D. Stocker, N. Viovy, X. Wang, Y. Wang, Z. Xiao, H. Yang, S. Zaehle, and N. Zeng, "Greening of the Earth and its drivers," *Nature Climate Change*, 2016.

[60] A. E. Hartemink, A. B. McBratney, and M. Mendonça Santos, *Digital soil mapping with limited data*. London: Springer, 2008.

[61] L. B. Guo and R. M. Gifford, "Soil carbon stocks and land use change: a meta analysis," *Global Change Biology*, vol. 8, no. 4, pp. 345–360, 2002.

[62] M. W. I. Schmidt, M. S. Torn, S. Abiven, T. Dittmar, G. Guggenberger, I. A. Janssens, M. Kleber, I. Kögel-Knabner, J. Lehmann, D. A. C. Manning, P. Nannipieri, D. P. Rasse, S. Weiner, and S. E. Trumbore, "Persistence of soil organic matter as an ecosystem property," *Nature*, vol. 478, no. 7367, pp. 49–56, 2011.

[63] R. Prasad and J. F. Power, *Soil fertility management for sustainable agriculture*. New York: CRC Press, 1997.

[64] P. Scull, J. Franklin, O. A. Chadwick, and D. McArthur, "Predictive soil mapping: a review," *Progress in Physical Geography*, vol. 27, no. 2, pp. 171–197, 2003.

[65] B. Brockett, *An interdisciplinary approach to mapping soil carbon*. PhD thesis, Lancaster University, UK, 2016.

[66] S. Kumar, R. Lal, and D. Liu, "A geographically weighted regression kriging approach for mapping soil organic carbon stock," *Geoderma*, vol. 189-190, pp. 627–634, 2012.

[67] S. Kumar and R. Lal, "Mapping the organic carbon stocks of surface soils using local spatial interpolator," *Journal of Environmental Monitoring*, vol. 13, no. 11, pp. 3128–3135, 2011.

[68] G. C. Simbahan, A. Dobermann, P. Goovaerts, J. Ping, and M. L. Haddix, "Fine-resolution mapping of soil organic carbon based on multivariate secondary data," *Geoderma*, vol. 132, no. 3-4, pp. 471–489, 2006.

[69] F. López-Granados, M. Jurado-Expósito, J. Peña-Barragán, and L. García-Torres, "Using geostatistical and remote sensing approaches for mapping soil properties," *European Journal of Agronomy*, vol. 23, no. 3, pp. 279–289, 2005.

[70] K. Phachomphon, P. Dlamini, and V. Chaplot, "Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables," *Geoderma*, vol. 155, no. 3-4, pp. 372–380, 2010.

[71] M. Martin, T. Orton, E. Lacarce, J. Meersmans, N. Saby, J. Paroissien, C. Jolivet, L. Boulonne, and D. Arrouays, "Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale," *Geoderma*, vol. 223-225, pp. 97–107, 2014.

[72] R. Grimm, T. Behrens, M. Märker, and H. Elsenbeer, "Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis," *Geoderma*, vol. 146, no. 1-2, pp. 102–113, 2008.

[73] B. Henderson, E. Bui, C. Moran, and D. Simon, "Australia-wide predictions of soil properties using decision trees," *Geoderma*, vol. 124, no. 3-4, pp. 383–398, 2005.

[74] U. Franko, B. Oelschlägel, and S. Schenk, "Simulation of temperature-, water- and nitrogen dynamics using the model CANDY," *Ecological Modelling*, vol. 81, no. 1-3, pp. 213–222, 1995.

[75] B. M. Petersen, J. Berntsen, S. Hansen, and L. S. Jensen, "CN-SIM—a model for the turnover of soil organic matter. I. Long-term carbon and radiocarbon development," *Soil Biology and Biochemistry*, vol. 37, no. 2, pp. 359–374, 2005.

[76] W. J. Parton, D. S. Schimel, C. V. Cole, and D. S. Ojima, "Analysis of Factors Controlling Soil Organic Matter Levels in Great Plains Grasslands1," *Soil Science Society of America Journal*, vol. 51, no. 5, pp. 1173–1179, 1987.

[77] T. Mueller, L. S. Jensen, S. Hansen, and N. E. Nielsen, *Simulating soil carbon and nitrogen dynamics with the soil-plant-atmosphere system model DAISY*. Berlin: Springer, 1996.

[78] C. Li, S. Frolking, and R. Harriss, "Modeling carbon biogeochemistry in agricultural soils," *Global Biogeochemical Cycles*, vol. 8, no. 3, pp. 237–254, 1994.

[79] W. J. Williams, "The Erosion-Productivity Impact Calculator (EPIC) Model: A Case History," *Philosophical Transactions: Biological Sciences*, vol. 329, no. 1255, pp. 421–428, 1990.

[80] O. Andrén and T. Kätterer, "ICBM: The Introductory Carbon Balance Model for exploration of soil carbon balances," *Ecological Applications*, vol. 7, no. 4, pp. 1226–1236, 1997.

[81] O. Chertov, A. Komarov, M. Nadporozhskaya, S. Bykhovets, and S. Zudin, "ROMUL — a model of forest soil organic matter dynamics as a substantial tool for forest ecosystem modeling," *Ecological Modelling*, vol. 138, no. 1-3, pp. 289–308, 2001.

[82] K. Coleman and D. S. Jenkinson, *RothC-26.3 - A Model for the turnover of carbon in soil*. Berlin: Springer, 1996.

[83] T. Hengl, G. B. Heuvelink, and D. G. Rossiter, "About regression-kriging: From equations to case studies," *Computers & Geosciences*, vol. 33, no. 10, pp. 1301–1315, 2007.

[84] B. Minasny and A. B. McBratney, "Spatial prediction of soil properties using EBLUP with the Matérn covariance function," *Geoderma*, vol. 140, no. 4, pp. 324–336, 2007.

[85] R. M. Lark, B. R. Cullis, and S. J. Welham, "On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML," *European Journal of Soil Science*, vol. 57, no. 6, pp. 787–799, 2006.

[86] A. B. McBratney, I. O. Odeh, T. F. Bishop, M. S. Dunbar, and T. M. Shatar, "An overview of pedometric techniques for use in soil survey," *Geoderma*, vol. 97, no. 3-4, pp. 293–327, 2000.

[87] G. M. Laslett, A. B. McBratney, P. J. Pahl, and M. F. Hutchinson, "Comparison of several spatial prediction methods for soil pH," *Journal of Soil Science*, vol. 38, no. 2, pp. 325–341, 1987.

[88] H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS Regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2010.

[89] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: the analysis of spatially varying relationships*. USA: Wiley, 2002.

[90] R. Bivand, E. J. Pebesma, and V. Gómez-Rubio, *Applied spatial data analysis with R.* Use R!, New York: Springer, 2008.

[91] G. De'ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178–3192, 2000.

[92] M. Wiesmeier, F. Barthold, B. Blank, and I. Kögel-Knabner, "Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem," *Plant and Soil*, vol. 340, no. 1-2, pp. 7–24, 2011.

[93] M. R. Pahlavan Rad, N. Toomanian, F. Khormali, C. W. Brungard, C. B. Komaki, and P. Bogaert, "Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran," *Geoderma*, vol. 232-234, pp. 97–106, 2014.

[94] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[95] T. Behrens, H. Förster, T. Scholten, U. Steinrücken, E.-D. Spies, and M. Goldschmitt, "Digital soil mapping using artificial neural networks," *Journal of Plant Nutrition and Soil Science*, vol. 168, no. 1, pp. 21–33, 2005.

[96] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning.* UK: Addison-Wesley Pub. Co, 1989.

[97] C. Ballabio, P. Panagos, and L. Monatanarella, "Mapping topsoil physical properties at European scale using the LUCAS database," *Geoderma*, vol. 261, pp. 110–123, 2016.

[98] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[99] K. Taalab, R. Corstanje, J. Zawadzka, T. Mayr, M. Whelan, J. Hannam, and R. Creamer, "On the application of Bayesian Networks in Digital Soil Mapping," *Geoderma*, vol. 259-260, pp. 134–148, 2015.

[100] B. Heung, H. C. Ho, J. Zhang, A. Knudby, C. E. Bulmer, and M. G. Schmidt, "An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping," *Geoderma*, vol. 265, pp. 62–77, 2016.

[101] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?," *SIGKDD Explorations*, vol. 2, no. 2, pp. 1–13, 2000.

[102] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 1–25, 1995.

[103] E. N. Bui, "Soil survey as a knowledge system," *Geoderma*, vol. 120, no. 1-2, pp. 17–26, 2004.

[104] W. Wielemaker, S. de Bruin, G. Epema, and A. Veldkamp, "Significance and application of the multi-hierarchical landsystem in soil mapping," *CATENA*, vol. 43, no. 1, pp. 15–34, 2001.

[105] P. E. Hart, R. O. Duda, and M. T. Einaudi, "PROSPECTOR—A computer-based consultation system for mineral exploration," *Journal of the International Association for Mathematical Geology*, vol. 10, no. 5, pp. 589–610, 1978.

[106] A. K. Skidmore, P. J. Ryan, W. Dawes, D. Short, and E. O'Loughlin, "Use of an expert system to map forest soils from a geographical information system," *International journal of geographical information systems*, vol. 5, no. 4, pp. 431–445, 1991.

[107] S. Banwart, "Save our soils," *Nature*, vol. 474, no. 7350, pp. 151–152, 2011.

[108] C. Guerrero, R. A. Viscarra Rossel, and A. M. Mouazen, "Special issue 'Diffuse reflectance spectroscopy in soil science and land resource assessment'," *Geoderma*, vol. 158, no. 1-2, pp. 1–2, 2010.

[109] R. A. Viscarra Rossel, C. R. Lobsey, C. Sharman, P. Flick, and G. McLachlan, "Novel Proximal Sensing for Monitoring Soil Organic C Stocks and Condition," *Environmental Science & Technology*, vol. 51, no. 10, pp. 5630–5641, 2017.

[110] R. Mirzaeitalarposhti, M. S. Demyan, F. Rasche, G. Cadisch, and T. Müller, "Mid-infrared spectroscopy to support regional-scale digital soil mapping on selected croplands of South-West Germany," *CATENA*, vol. 149, pp. 283–293, 2017.

[111] A. Ångström, "The Albedo of Various Surfaces of Ground," *Geografiska Annaler*, vol. 7, pp. 323–342, 1925.

[112] A. M. O'Neal, "The effect of moisture on the color of certain Iowa soils," *American Soil Survey Association Bulletin*, vol. B8, pp. 158–174, 1927.

[113] W. T. Carter, "Color Analysis of Soils with Spectrophotometer," *Soil Science Society of America Journal*, vol. B12, no. 2001, p. 169, 1931.

[114] J. B. Hester, "The Relation of Soil Texture and Color to the Organic Matter Content," *Soil Science Society of America Journal*, vol. 3, no. C, p. 112, 1939.

[115] A. H. Al-Abbas, P. H. Swain, and M. F. Baumgardner, "Relating Organic Matter and Clay Content to the Multispectral Radiance of Soils," *Soil Science*, vol. 114, no. 6, pp. 477–485, 1972.

[116] S. Kristof, M. F. Baumgardner, and C. Johannsen, "Spectral Mapping of Soil Organic Matter," *LARS Technical Reports*, no. 26, 1973.

[117] H. L. Mathews, R. L. Cunningham, and G. W. Petersen, "Spectral Reflectance of Selected Pennsylvania Soils," *Soil Science Society of America Journal*, vol. 37, no. 3, pp. 421–424, 1973.

[118] E. R. Stoner and M. F. Baumgardner, "Characteristic Variations in Reflectance of Surface Soils," *Soil Science Society of America Journal*, vol. 45, no. 6, p. 1161, 1981.

[119] R. Linker, *Application of FTIR Spectroscopy to Agricultural Soils Analysis*. InTech, 2011.

[120] R. C. Dalal and R. J. Henry, "Simultaneous Determination of Moisture, Organic Carbon, and Total Nitrogen by Near Infrared Reflectance Spectrophotometry," *Soil Science Society of America Journal*, vol. 50, no. 1, pp. 120–123, 1986.

[121] L. Janik and J. Skjemstad, "Characterization and analysis of soils using mid-infrared partial least-squares.2. Correlations with some laboratory data," *Australian Journal of Soil Research*, vol. 33, no. 4, pp. 637–650, 1995.

[122] R. A. Viscarra Rossel, D. Walvoort, A. McBratney, L. Janik, and J. Skjemstad, "Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties," *Geoderma*, vol. 131, no. 1-2, pp. 59–75, 2006.

[123] R. A. Viscarra Rossel, T. Behrens, E. Ben-Dor, D. Brown, J. Demattê, K. Shepherd, Z. Shi, B. Stenberg, A. Stevens, V. Adamchuk, H. Aïchi, B. Barthès, H. Bartholomeus, A. Bayer, M. Bernoux, K. Böttcher, L. Brodský, C. Du, A. Chappell, Y. Fouad, V. Genot, C. Gomez, S. Grunwald, A. Gubler, C. Guerrero, C. Hedley, M. Knadel, H. Morrás, M. Nocita, L. Ramirez-Lopez, P. Roudier, E. R. Campos, P. Sanborn,

V. Sellitto, K. Sudduth, B. Rawlins, C. Walter, L. Winowiecki, S. Hong, and W. Ji, "A global spectral library to characterize the world's soil," *Earth-Science Reviews*, vol. 155, pp. 198–230, 2016.

[124] J. M. Soriano-Disla, L. J. Janik, R. A. Viscarra Rossel, L. M. Macdonald, and M. J. McLaughlin, "The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties," *Applied Spectroscopy Reviews*, vol. 49, no. 2, pp. 139–186, 2014.

[125] B. Stenberg, R. A. Viscarra Rossel, A. M. Mouazen, and J. Wetterlind, "Visible and Near Infrared Spectroscopy in Soil Science," *Advances in Agronomy*, vol. 107, pp. 163–215, 2010.

[126] T. Shi, Y. Chen, Y. Liu, and G. Wu, "Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals," *Journal of Hazardous Materials*, vol. 265, pp. 166–176, 2014.

[127] A. Horta, B. Malone, U. Stockmann, B. Minasny, T. Bishop, A. McBratney, R. Pallasser, and L. Pozza, "Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: A prospective review," *Geoderma*, vol. 241-242, pp. 180–209, 2015.

[128] R. A. Viscarra Rossel, "The Soil Spectroscopy Group and the development of a global soil spectral library," *NIR news*, vol. 20, no. 4, pp. 14–15, 2009.

[129] D.-W. Sun, *Infrared spectroscopy for food quality analysis and control*. Amsterdam: Elsevier, 2009.

[130] B. Stuart, *Infrared spectroscopy: fundamentals and applications*. Analytical techniques in the sciences, UK: Wiley, 2004.

[131] R. A. Viscarra Rossel, A. B. McBratney, and B. Minasny, eds., *Proximal soil sensing*. Progress in soil science, New York: Springer, 2010.

[132] T. Chen, Q. Chang, J. Clevers, and L. Kooistra, "Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy," *Environmental Pollution*, vol. 206, pp. 217–226, 2015.

[133] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[134] J. Huang, S. Romero-Torres, and M. Moshgbar, "Practical Considerations in Data Pre-treatment for NIR and Raman Spectroscopy," *American Pharmaceutical Review*, 2010.

[135] B. C. Smith, *Fundamentals of Fourier transform infrared spectroscopy*. New York: CRC Press, 2011.

[136] A. Gholizadeh, L. Borůvka, M. Saberioon, J. Kozák, R. Vašát, and K. Němeček, "Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features," *Soil and Water Research*, vol. 10, no. 4, pp. 218–227, 2016.

[137] R. N. Clark and T. L. Roush, "Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications," *Journal of Geophysical Research: Solid Earth*, vol. 89, no. B7, pp. 6329–6340, 1984.

[138] R. F. Kokaly, D. G. Despain, R. N. Clark, and K. Livo, "Mapping vegetation in Yellowstone National Park using spectral feature analysis of AVIRIS data," *Remote Sensing of Environment*, vol. 84, no. 3, pp. 437–456, 2003.

[139] P. Geladi, D. MacDougall, and H. Martens, "Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat," *Applied Spectroscopy*, vol. 39, no. 3, pp. 491–500, 1985.

[140] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[141] S. Wold, H. Antti, F. Lindgren, and J. Öhman, "Orthogonal signal correction of near-infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 44, no. 1-2, pp. 175–185, 1998.

[142] Y. Ozaki, W. F. McClure, and A. A. Christy, *Near-infrared spectroscopy in food science and technology.* New Jersey: Wiley, 2007.

[143] M. Dhanoa, S. Lister, R. Sanderson, and R. Barnes, "The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra," *Journal of Near Infrared Spectroscopy*, vol. 2, no. 1, pp. 43–47, 1994.

[144] R. A. Storey and I. Ymen, *Solid state characterization of pharmaceuticals.* UK: Wiley, 2011.

[145] R. A. Viscarra Rossel and A. McBratney, *Diffuse Reflectance Spectroscopy as a Tool for Digital Soil Mapping.* Netherlands: Springer, 2008.

[146] R. A. Viscarra Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no. 1-2, pp. 46–54, 2010.

[147] B. E. Madari, J. B. Reeves, P. L. Machado, C. M. Guimarães, E. Torres, and G. W. McCarty, "Mid- and near-infrared

spectroscopic assessment of soil compositional parameters and structural indices in two Ferralsols," *Geoderma*, vol. 136, no. 1-2, pp. 245–259, 2006.

[148] H. Yang and A. M., *Vis/Near- and Mid- Infrared Spectroscopy for Predicting Soil N and C at a Farm Scale*. InTech, 2012.

[149] J. A. Demattê, R. C. Campos, M. C. Alves, P. R. Fiorio, and M. R. Nanni, "Visible–NIR reflectance: a new approach on soil evaluation," *Geoderma*, vol. 121, no. 1-2, pp. 95–112, 2004.

[150] E. Ben-Dor and A. Banin, "Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties," *Soil Science Society of America Journal*, vol. 59, no. 2, pp. 364–372, 1995.

[151] P. H. Fidêncio, R. J. Poppi, J. C. de Andrade, and H. Cantarella, "Determination of organic matter in soil using near-infrared spectroscopy and partial least squares regression," *Communications in Soil Science and Plant Analysis*, vol. 33, no. 9-10, pp. 1607–1615, 2002.

[152] K. D. Shepherd and M. G. Walsh, "Development of Reflectance Spectral Libraries for Characterization of Soil Properties," *Soil Science Society of America Journal*, vol. 66, no. 3, pp. 988–998, 2002.

[153] K. W. Daniel, N. K. Tripathi, and K. Honda, "Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand)," *Australian Journal of Soil Research*, vol. 41, no. 1, pp. 47–59, 2003.

[154] D. J. Brown, K. D. Shepherd, M. G. Walsh, M. Dewayne Mays, and T. G. Reinsch, "Global soil characterization with VNIR diffuse

reflectance spectroscopy," *Geoderma*, vol. 132, no. 3-4, pp. 273–290, 2006.

[155] V. Bellon-Maurel and A. McBratney, "Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives," *Soil Biology and Biochemistry*, vol. 43, no. 7, pp. 1398–1410, 2011.

[156] J. B. Reeves, "Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done?," *Geoderma*, vol. 158, no. 1-2, pp. 3–14, 2010.

[157] L. Rodríguez-Lado and A. Martínez-Cortizas, "Modelling and mapping organic carbon content of topsoils in an Atlantic area of southwestern Europe (Galicia, NW-Spain)," *Geoderma*, vol. 245-246, pp. 65–73, 2015.

[158] L. Rodríguez-Lado, *Análisis y Cartografía de las cargas críticas de acidez y eutrofización de suelos*. PhD thesis, Universidade de Santiago de Compostela, Santiago de Compostela, 2004.

[159] A. Martínez-Cortizas, F. Castillo, and A. Pérez-Alberti, "Factores que influyen en la precipitación y el balance de agua en Galicia," *Boletín de la A.G.E.*, vol. 18, pp. 79–96, 1994.

[160] T. Taboada and C. Garcia, "Smectite formation produced by weathering in a coarse granite saprolite in Galicia (NW Spain)," *CATENA*, vol. 35, no. 2-4, pp. 281–290, 1999.

[161] S. Rivas-Martínez, *Global Bioclimatics (Clasificación bioclimática de la Tierra)*. Madrid: Phytosociological Research Center, 2004.

[162] S. Bará and G. Toval, *Calidad de estación del Pinus pinaster Ait. en Galicia*. Madrid: Instituto Nacional de Investigaciones Agrarias (INIA), 1983.

[163] F. Díaz and F. Gil, *Capacidad productiva de los suelos de Galicia*. Santiago de Compostela: Universidade de Santiago de Compostela, 1984.

[164] F. Guitián, T. Carballas, and M. Muñoz-Taboada, *Suelos naturales de la provincia de Lugo*. Santiago de Compostela: Consejo Superior de Investigaciones Científicas, 1982.

[165] F. Macías and R. Calvo, *Los suelos de la provincia de La Coruña*. Santiago de Compostela: Diputación de la Coruña, 1992.

[166] A. Walkley and A. Black, "An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method," *Soil Science*, vol. 27, pp. 29–37, 1934.

[167] P. R. Griffiths and J. A. De Haseth, *Fourier transform infrared spectrometry*. No. v. 171 in Chemical analysis, New Jersey: Wiley, 2007.

[168] R. Saphire, Y. Freund, P. Barlett, and W. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, pp. 1651–1689, 1998.

[169] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[170] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random Forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[171] M. Verleysen, *Learning high-dimensional data*. Netherlands: IOS Press, 2003.

[172] A. Liaw and M. Weiner, "Classification and regression by randomForest," *R News*, vol. 2, pp. 18–22, 2002.

[173] K. Were, D. T. Bui, Ø. B. Dick, and B. R. Singh, "A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape," *Ecological Indicators*, vol. 52, pp. 394–403, 2015.

[174] L. Breiman, *Random forests – random features. Technical Report 567*. CA: Statistics Departament UC, 1999.

[175] G. Baffi, E. Martin, and A. Morris, "Non-linear projection to latent structures revisited: the quadratic PLS algorithm," *Computers & Chemical Engineering*, vol. 23, no. 3, pp. 395–411, 1999.

[176] T. Naes and B.-H. Mevik, "Understanding the collinearity problem in regression and discriminant analysis," *Journal of Chemometrics*, vol. 15, no. 4, pp. 413–426, 2001.

[177] R. C. Team, *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2013.

[178] C. Loiseau, R. J. Harrigan, C. Bichet, R. Julliard, S. Garnier, Á. Z. Lendvai, O. Chastel, and G. Sorci, "Predictions of avian Plasmodium expansion under climate change," *Scientific Reports*, vol. 3, no. 1126, pp. 1–6, 2013.

[179] R. Ellerbrock and M. Kaiser, "Stability and composition of different soluble soil organic matter fractions–evidence from ^{13}C and FTIR signatures," *Geoderma*, vol. 128, no. 1-2, pp. 28–37, 2005.

[180] J. A. Pedersen, M. A. Simpson, J. G. Bockheim, and K. Kumar, "Characterization of soil organic carbon in drained thaw-lake basins of Arctic Alaska using NMR and FTIR photoacoustic spectroscopy," *Organic Geochemistry*, vol. 42, no. 8, pp. 947–954, 2011.

[181] I. Simkovic, P. Dlapa, S. H. Doerr, J. Mataix-Solera, and V. Sasinkova, "Thermal destruction of soil water repellency and associated changes to soil organic matter as observed by FTIR spectroscopy," *CATENA*, vol. 74, no. 3, pp. 205–211, 2008.

[182] M. Vohland, M. Ludwig, S. Thiele-Bruhn, and B. Ludwig, "Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection," *Geoderma*, vol. 223-225, pp. 88–96, 2014.

[183] E. Garcia-Rodeja, B. M. Silva, and F. Macías, "Andosols developed from non-volcanic materials in Galicia, NW Spain," *Journal of Soil Science*, vol. 38, no. 4, pp. 573–591, 1987.

[184] E. Garcia-Rodeja, T. Taboada, A. Martínez-Cortizas, B. Silva, and C. García, *Soils with "andic" properties developed from non-volcanic materials. Genesis and implications in soil classification.* Volcanic Soil Resources in Europe. COST Action 622 final meeting, Iceland: Agriculture Research Institute. Reykavík, 2004.

[185] F. Macías, M. Puga, and F. Guitián, "Caracteres ándicos en suelos sobre gabros de Galicia," *Anales de edafología y agrobiología*, vol. 37, pp. 187–203, 1978.

[186] J. R. Verde, M. Camps-Arbestain, and F. Macías, "Efecto de las prácticas agrícolas sobre la estabilidad de los complejos organoalumínicos en suelos ándicos de Galicia," *Edafología*, vol. 11, pp. 319–328, 2004.

[187] A. Freibauer, M. D. Rounsevell, P. Smith, and J. Verhagen, "Carbon sequestration in the agricultural soils of Europe," *Geoderma*, vol. 122, no. 1, pp. 1–23, 2004.

[188] I. A. Janssens, "Europe's Terrestrial Biosphere Absorbs 7 to 12% of European Anthropogenic CO_2 Emissions," *Science*, vol. 300, no. 5625, pp. 1538–1542, 2003.

[189] D. R. Cameron, M. Van Oijen, C. Werner, K. Butterbach-Bahl, R. Grote, E. Haas, G. B. M. Heuvelink, R. Kiese, J. Kros, M. Kuhnert, A. Leip, G. J. Reinds, H. I. Reuter, M. J. Schelhaas, W. De Vries, and J. Yeluripati, "Environmental change impacts on the C- and N-cycle of European forests: a model comparison study," *Biogeosciences*, vol. 10, no. 3, pp. 1751–1773, 2013.

[190] W. H. Schlesinger, "Carbon Sequestration in Soils," *Science*, vol. 284, no. 5423, pp. 2095–2095, 1999.

[191] W. M. Post and K. C. Kwon, "Soil carbon sequestration and land-use change: processes and potential," *Global Change Biology*, vol. 6, no. 3, pp. 317–327, 2000.

[192] P. Smith, "Carbon sequestration in croplands: the potential in Europe and the global context," *European Journal of Agronomy*, vol. 20, no. 3, pp. 229–236, 2004.

[193] G.-J. Nabuurs, M. Lindner, P. J. Verkerk, K. Gunia, P. Deda, R. Michalak, and G. Grassi, "First signs of carbon sink saturation in European forest biomass," *Nature Climate Change*, vol. 3, no. 9, pp. 792–796, 2013.

[194] K. Naudts, Y. Chen, M. J. McGrath, J. Ryder, A. Valade, J. Otto, and S. Luyssaert, "Europe's forest management did not mitigate climate warming," *Science*, vol. 351, no. 6273, pp. 597–600, 2016.

[195] R. Valentini, G. Matteucci, A. J. Dolman, E.-D. Schulze, C. Rebmann, E. J. Moors, A. Granier, P. Gross, N. O. Jensen, K. Pilegaard, A. Lindroth, A. Grelle, C. Bernhofer, T. Grünwald,

M. Aubinet, R. Ceulemans, A. S. Kowalski, T. Vesala, ü. Rannik, P. Berbigier, D. Loustau, J. Guamundsson, H. Thorgeirsson, A. Ibrom, K. Morgenstern, R. Clement, J. Moncrieff, L. Montagnani, S. Minerbi, and P. G. Jarvis, "Respiration as the main determinant of carbon balance in European forests," *Nature*, vol. 404, no. 6780, pp. 861–865, 2000.

[196] R. J. A. Jones, R. Hiederer, E. Rusco, and L. Montanarella, "Estimating organic carbon in the soils of Europe for policy support," *European Journal of Soil Science*, vol. 56, no. 5, pp. 655–671, 2005.

[197] G. Tóth, A. Jones, L. Montanarella, European Commission, Joint Research Centre, and Institute for Environment and Sustainability, *LUCAS topsoil survey methodology, data and results.* Luxembourg: Publications Office, 2013.

[198] G. Tóth, A. Jones, and L. Montanarella, "The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union," *Environmental Monitoring and Assessment*, vol. 185, no. 9, pp. 7409–7425, 2013.

[199] D. de Brogniez, C. Ballabio, A. Stevens, R. J. A. Jones, L. Montanarella, and B. van Wesemael, "A map of the topsoil organic carbon content of Europe generated by a generalized additive model: Soil organic carbon content at pan-European level," *European Journal of Soil Science*, vol. 66, no. 1, pp. 121–134, 2015.

[200] Y. Yigini and P. Panagos, "Assessment of soil organic carbon stocks under future climate and land cover changes in Europe," *Science of The Total Environment*, vol. 557-558, pp. 838–850, 2016.

[201] E. Aksoy, Y. Yigini, and L. Montanarella, "Combining Soil Databases for Topsoil Organic Carbon Mapping in Europe," *PLOS ONE*, vol. 11, no. 3, p. e0152098, 2016.

[202] M. Bell and F. Worrall, "Estimating a region's soil organic carbon baseline: The undervalued role of land-management," *Geoderma*, vol. 152, no. 1-2, pp. 74–84, 2009.

[203] M. P. Martin, M. Wattenbach, P. Smith, J. Meersmans, C. Jolivet, L. Boulonne, and D. Arrouays, "Spatial distribution of soil organic carbon stocks in France," *Biogeosciences*, vol. 8, no. 5, pp. 1053–1065, 2011.

[204] J. Meersmans, F. De Ridder, F. Canters, S. De Baets, and M. Van Molle, "A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium)," *Geoderma*, vol. 143, no. 1-2, pp. 1–13, 2008.

[205] P. Panagos, C. Ballabio, Y. Yigini, and M. B. Dunbar, "Estimating the soil organic carbon content for European NUTS2 regions based on LUCAS data collection," *Science of The Total Environment*, vol. 442, pp. 235–246, 2013.

[206] K. Adhikari, A. E. Hartemink, B. Minasny, R. Bou Kheir, M. B. Greve, and M. H. Greve, "Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark," *PLOS ONE*, vol. 9, no. 8, p. e105519, 2014.

[207] J. Rodríguez Martín, J. Álvaro-Fuentes, J. Gonzalo, C. Gil, J. Ramos-Miras, J. Grau Corbí, and R. Boluda, "Assessment of the soil organic carbon stock in Spain," *Geoderma*, vol. 264, pp. 117–125, 2016.

[208] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, "Very high resolution interpolated climate surfaces for global land areas," *International Journal of Climatology*, vol. 25, no. 15, pp. 1965–1978, 2005.

[209] R. J. Zomer and International Water Management Institute, *Carbon, land and water a global analysis of the hydrologic dimensions of climate change mitigation through afforestation/reforestation.* Sri Lanka: International Water Management Institute, 2006.

[210] P. Panagos, M. Van Liedekerke, A. Jones, and L. Montanarella, "European Soil Data Centre: Response to European policy support and public data requirements," *Land Use Policy*, vol. 29, no. 2, pp. 329–338, 2012.

[211] M. Van Liedekerke, A. Jones, and P. Panagos, *The European Soil Database distribution version 2.0.* European Commission and the European Soil Bureau Network, CD-ROM, 2004.

[212] J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, and A. Ziegler, "Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines," *Methods of Information in Medicine*, vol. 51, no. 1, pp. 74–81, 2011.

[213] I. T. Jolliffe, *Principal Component Analysis.* New York: Springer, 1986.

[214] T. Næs, *A user-friendly guide to multivariate calibration and classification.* UK: NIR Publications, 2004.

[215] A. Stevens and L. Ramirez-Lopez, "An introduction to the prospectr package," *https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf*, 2013.

[216] F. T. Maestre, M. Delgado-Baquerizo, T. C. Jeffries, D. J. Eldridge, V. Ochoa, B. Gozalo, J. L. Quero, M. García-Gómez, A. Gallardo, W. Ulrich, M. A. Bowker, T. Arredondo, C. Barraza-Zepeda, D. Bran, A. Florentino, J. Gaitán, J. R. Gutiérrez, E. Huber-Sannwald, M. Jankju, R. L. Mau, M. Miriti, K. Naseri, A. Ospina,

I. Stavi, D. Wang, N. N. Woods, X. Yuan, E. Zaady, and B. K. Singh, "Increasing aridity reduces soil microbial diversity and abundance in global drylands," *Proceedings of the National Academy of Sciences*, p. 201516684, 2015.

[217] S. E. Trumbore, O. A. Chadwick, and R. Amundson, "Rapid Exchange Between Soil Carbon and Atmospheric Carbon Dioxide Driven by Temperature Change," *Science*, vol. 272, no. 5260, pp. 393–396, 1996.

[218] W. Knorr, I. C. Prentice, J. I. House, and E. A. Holland, "Long-term sensitivity of soil carbon turnover to warming," *Nature*, vol. 433, no. 7023, pp. 298–301, 2005.

[219] R. I. Griffiths, B. C. Thomson, P. Plassart, H. S. Gweon, D. Stone, R. E. Creamer, P. Lemanceau, and M. J. Bailey, "Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets," *Applied Soil Ecology*, vol. 97, pp. 61–68, 2016.

[220] G. Borchardt, *Montmorillonite and other smectite minerals*. Minerals in soil environments, Soil Science Society of America, 1977.

[221] W. A. Deer, R. A. Howie, and J. Zussman, *An introduction to the rock-forming minerals*. UK: Pearson/Prentice Hall, 2009.

[222] A. Stevens, M. Nocita, G. Tóth, L. Montanarella, and B. van Wesemael, "Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy," *PLOS ONE*, vol. 8, no. 6, p. e66409, 2013.

[223] O. Andrén, T. Kätterer, T. Karlsson, and J. Eriksson, "Soil C balances in Swedish agricultural soils 1990–2004, with preliminary projections," *Nutrient Cycling in Agroecosystems*, vol. 81, no. 2, pp. 129–144, 2008.

[224] G. I. Ågren, R. Hyvönen, and T. Nilsson, "Are Swedish forest soils sinks or sources for CO_2—model analyses based on forest inventory data," *Biogeochemistry*, vol. 89, no. 1, pp. 139–149, 2008.

[225] D. Berggren Kleja, M. Olsson, M. Svensson, and P.-E. Jansson, *Soil C dynamics in Swedish forest soils - gradients from south to north.* Greenhouse-gas budget of soils under changing climate and land use (BurnOut). COST 639, 2010.

[226] C. A. Ortiz, J. Liski, A. I. Gärdenäs, A. Lehtonen, M. Lundblad, J. Stendahl, G. I. Ågren, and E. Karltun, "Soil organic carbon stock changes in Swedish forest soils—A comparison of uncertainties and their sources through a national inventory and two simulation models," *Ecological Modelling*, vol. 251, pp. 221–231, 2013.

[227] M. J. Metzger, R. G. H. Bunce, R. H. G. Jongman, C. A. Mücher, and J. W. Watkins, "A climatic stratification of the environment of Europe: A climatic stratification of the European environment," *Global Ecology and Biogeography*, vol. 14, no. 6, pp. 549–563, 2005.

[228] R. H. G. Jongman, R. G. H. Bunce, M. J. Metzger, C. A. Mücher, D. C. Howard, and V. L. Mateus, "Objectives and Applications of a Statistical Environmental Stratification of Europe," *Landscape Ecology*, vol. 21, no. 3, pp. 409–419, 2006.

[229] M. Olsson, M. Erlandsson, L. Lundin, T. Nilsson, Å. Nilsson, and J. Stendahl, "Organic carbon stocks in Swedish Podzol soils in relation to soil hydrology and other site characteristics," *Silva Fennica*, vol. 43, no. 2, 2009.

[230] J. Heikkinen, E. Ketoja, V. Nuutinen, and K. Regina, "Declining trend of carbon in Finnish cropland soils in 1974-2009," *Global Change Biology*, vol. 19, no. 5, pp. 1456–1469, 2013.

[231] M. Nocita, A. Stevens, G. Toth, P. Panagos, B. van Wesemael, and L. Montanarella, "Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach," *Soil Biology and Biochemistry*, vol. 68, pp. 337–347, 2014.

[232] H. Erhart, *La genèse des sols en tant que phénomène géologique. Esquisse d'une théorie géologique et géochimique. Biostasie et rhéxistasie.* Paris: Masson, 1956.

[233] N. János and R. Tamás, *Soil Cultivation and Land Use.* Hungary: University of Debrecen, Service Sciences Methodology Centre, 2013.

[234] J. Huxley, *Charles Darwin: Galápagos and After*, vol. The Galápagos: Proceedings of the Symposia of the Galápagos Internation Scientific Project. CA: University of California Press, 1966.

[235] N. d'Ozouville, G. DiCarlo, F. Ortiz, F. De Koning, H. Scott, and E. Pidgeon, "Galápagos in the face of climate change: considerations for biodiversity and associated human well-being," *Galápagos Report 2009-2010*, pp. 170–176, 2009.

[236] C. Ebbesmeyer, *1976 Step in the Pacific Climate: Forty Environmental Changes Between 1968-1975 and 1977-1984.* CA: Seventh Annual Pacific Climate (PACLIM) Workshop, 1991.

[237] I. Larrea, *Climate Change Vulnerability Assessment of the Galápagos Islands.* USA: WWF and Conservation International, 2011.

[238] J. P. Sachs and N. Ladd, "Climate and oceanography of the Galápagos in the 21st century: expected changes and research needs," *Galapagos Research*, vol. 67, pp. 50–54, 2010.

[239] M. Adelinet, J. Fortin, N. d'Ozouville, and S. Violette, "The relationship between hydrodynamic properties and weathering of soils derived from volcanic rocks, Galapagos Islands (Ecuador)," *Environmental Geology*, vol. 56, no. 1, pp. 45–58, 2008.

[240] H. Eswaran, "A contribution to the study of soil formation on Isla Santa Cruz, Galápagos.," *Pedologie*, vol. 23, pp. 100–122, 1973.

[241] J. Laurelle and G. Stoops, "Minor elements in Galapagos soils," *Pedologie*, vol. 17, pp. 232–258, 1967.

[242] G. Stoops, "Soils and Paleosoils of the Galápagos Islands: What We Know and What We Don't Know, A Meta-Analysis," *Pacific Science*, vol. 68, no. 1, pp. 1–17, 2014.

[243] T. Taboada, L. Rodríguez-Lado, C. Ferro-Vázquez, G. Stoops, and A. Martínez-Cortizas, "Chemical weathering in the volcanic soils of Isla Santa Cruz (Galápagos Islands, Ecuador)," *Geoderma*, vol. 261, pp. 160–168, 2016.

[244] W. M. White, A. R. McBirney, and R. A. Duncan, "Petrology and geochemistry of the Galápagos Islands: Portrait of a pathological mantle plume," *Journal of Geophysical Research*, vol. 98, no. B11, pp. 19533–19563, 1993.

[245] G. Stoops and P. De Paepe, "Vijftig jaar geleden. Een Belgische geopedologische zending naar de Galápagoseilanden (mei - oktober 1962)," *Mededelingen Zittingen Koninklijke Academie voor Overzeese Wetenschappen*, vol. 59, no. 2-4, pp. 323–343, 2013.

[246] J. Laurelle, "Exploration Geo-Pedologique de L'Ile Santa Cruz," *Noticias de Galápagos*, vol. 1, pp. 11–13, 1963.

[247] J. Böhner and O. Antonić, "Chapter 8 Land-Surface Parameters Specific to Topo-Climatology," *Developments in Soil Science*, vol. 33, pp. 195–226, 2009.

[248] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," *Geographical Analysis*, vol. 28, no. 4, pp. 281–298, 2010.

[249] C. Zhang, Y. Tang, X. Xu, and G. Kiely, "Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland," *Applied Geochemistry*, vol. 26, no. 7, pp. 1239–1248, 2011.

[250] P. A. Colinvaux, "Climate and the Galapagos Islands," *Nature*, vol. 240, no. 5375, pp. 17–20, 1972.

[251] J. P. Sachs, D. Sachse, R. H. Smittenberg, Z. Zhang, D. S. Battisti, and S. Golubic, "Southward movement of the Pacific intertropical convergence zone AD 1400–1850," *Nature Geoscience*, vol. 2, no. 7, pp. 519–525, 2009.

[252] M. Trueman and N. d'Ozouville, "Characterizing the Galapagos Terrestrial Climate in the face of global climate change," *Noticias de Galápagos*, vol. 67, pp. 26–37, 2010.

[253] Pyret, *Hydrogeology of volcanic islands:a case-study in the Galapagos Archipelago (Ecuador)*. PhD thesis, Université Paris 6 Pierre et Marie Curie, Paris, 2011.

[254] M. Collins, S.-I. An, W. Cai, A. Ganachaud, E. Guilyardi, F.-F. Jin, M. Jochum, M. Lengaigne, S. Power, A. Timmermann, G. Vecchi, and A. Wittenberg, "The impact of global warming on the tropical Pacific Ocean and El Niño," *Nature Geoscience*, vol. 3, no. 6, pp. 391–397, 2010.

[255] O. Hamann, *The El Niño influence on the Galápagos vegetation*. Ecuador: Charles Darwin Foundation, 1985.

[256] A. Tye and I. Aldaz, "Effects of the 1997-98 El Niño event on the vegetation of Galapagos," *Noticias de Galápagos*, vol. 60, pp. 22–24, 1999.

[257] P. N. DiNezio, A. C. Clement, G. A. Vecchi, B. J. Soden, B. P. Kirtman, and S.-K. Lee, "Climate Response of the Equatorial Pacific to Global Warming," *Journal of Climate*, vol. 22, no. 18, pp. 4873–4892, 2009.

[258] G. A. Vecchi, B. J. Soden, A. T. Wittenberg, I. M. Held, A. Leetmaa, and M. J. Harrison, "Weakening of tropical Pacific atmospheric circulation due to anthropogenic forcing," *Nature*, vol. 441, no. 7089, pp. 73–76, 2006.

[259] H. Bellenger, E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard, "ENSO representation in climate models: from CMIP3 to CMIP5," *Climate Dynamics*, vol. 42, no. 7-8, pp. 1999–2018, 2014.

[260] B. Oueslati and G. Bellon, "The double ITCZ bias in CMIP5 models: interaction between SST, large-scale circulation and precipitation," *Climate Dynamics*, vol. 44, no. 3-4, pp. 585–607, 2015.

[261] J. Laurelle, *Study of a Soil Sequence on Indefatigable Island*. CA: University of California Press, 1966.

[262] A. Stewart, *A botanical survey of the Galapagos Islands*. CA: California Academy of Sciences, 1911.

[263] D. P. van Vuuren, J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G. C. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, T. Masui, M. Meinshausen, N. Nakicenovic, S. J. Smith, and S. K. Rose, "The representative concentration pathways: an overview," *Climatic Change*, vol. 109, no. 1-2, pp. 5–31, 2011.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

APPENDICES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Publications

Refereed research papers:

- **M. Rial**, A. Martínez Cortizas, T. Taboada, and L. Rodríguez-Lado, "Soil organic carbon stocks in Santa Cruz Island, Galapagos, under different climate change scenarios," CATENA, vol. 156C, pp. 74-81, 2017. (doi: 10.1016/j.catena.2017.03.020).

- **M. Rial**, A. Martínez Cortizas, and L. Rodríguez-Lado, "Mapping soil organic carbon content using spectroscopic and environmental data: A case study in acidic soils from NW Spain," Science of Total Environment, vol. 529, pp. 26-35, 2016. (doi: 10.1016/j.scitotenv.2015.08.088).

Papers in refereed conference proceedings:

- L. Rodríguez-Lado, **M. Rial**, T. Taboada, and A. Martínez Cortizas, "A pedotransfer function to map soil bulk density from limited data," Procedia Environmental Sciences, vol. 27, pp. 45-48, 2015. (doi: 10.1016/j.proenv.2015.07.112).

- **M. Rial**, A. Martínez Cortizas, and L. Rodríguez-Lado, "A novel approach to map soil organic carbon content using spectroscopic and environmental data," Procedia Environmental Sciences, vol. 27, pp. 49-52, 2015. (doi: 10.1016/j.proenv.2015.07.113).

- **M. Rial**, A. Martínez Cortizas, and L. Rodríguez-Lado, "Mapping soil organic carbon content using spectroscopic data," in 8th European Congress on Regional Geoscientific Cartography Information Systems. Geological 3D Modelling and Soils: functions and threats. Proceedings, pp. 160, Barcelona: Institut Cartogràfic i Geològic de Catalunya, 2015. (ISBN: 978-84-393-9292-7).

Working papers under revision or review:

- **M. Rial**, A. Martínez Cortizas, and L. Rodríguez-Lado, "Understanding the spatial distribution of factors controlling topsoil organic carbon content in European soils".

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Summary in Spanish

Antecedentes

Uno de los grandes retos de la ciencia del suelo es comprender el papel del suelo en el cambio climático global. Las actividades antropogénicas, surgidas alrededor de la industria del carbono y de los combustibles fósiles, han provocado un incremento en la concentración atmosférica de gases de efecto invernadero como el dióxido de carbono o el metano. El aumento de la concentración de estos gases en la atmósfera acelera el cambio climático de una forma drástica, provocando de esta manera el calentamiento del planeta debido al efecto invernadero. El cambio climático está asociado a ciertos sucesos que pondrían en riesgo la vida en la Tierra tal y como la conocemos hoy en día. Entre algunos de estos sucesos destacan el aumento de las temperaturas, cambios en los patrones climáticos, en las corrientes atmosféricas y oceánicas

o el incremento del nivel del mar, entre otros. Esta situación ha llevado a los organismos internacionales a tomar medidas para detener el avance del cambio climático, o en todo caso a mitigar sus efectos.

El suelo es uno de los componentes básicos en el ciclo del carbono, con lo cual tiene un papel crucial en las políticas destinadas a disminuir las concentraciones de los gases de efecto invernadero. Concretamente, la comunidad científica y los responsables políticos han expresado la necesidad de disponer de información precisa, actualizada y geo-referenciada sobre la distribución del carbono orgánico del suelo. El carbono orgánico representa la mayor reserva de carbono terrestre, siendo uno de los componentes más relevantes en el ciclo del carbono y en el cambio climático. Para evaluar los lugares en los cuales los suelos son más vulnerables a las pérdidas de carbono o cuales son los usos de la tierra que pueden promover su secuestro y mitigar el incremento de las concentraciones de gases de efecto invernadero, es necesario disponer de un buen conocimiento acerca de su distribución espacial. Esta necesidad ha quedado patente en la última cumbre del clima, que tuvo lugar en París en el 2015 y en la cual se presentó la iniciativa "4 per mill". Según esta iniciativa, los stocks globales de carbono orgánico deberían incrementarse un 0.4% anualmente para compensar las emisiones de gases de efecto invernadero. Con lo cual, es importante obtener información espacial sobre la

distribución del carbono orgánico y evaluar correctamente las zonas más adecuadas para establecer medidas políticas, como cambios en los usos del suelo, que consigan incrementar los stocks de carbono de manera efectiva.

Objetivos

En la presente tesis se pretende evaluar la capacidad de la cartografía digital, combinada con la espectroscopía infrarroja, para mostrar la distribución espacial del carbono orgánico a escala regional y continental. Además, también se pretende evaluar la capacidad de los modelos de cartografía digital para predecir los stocks de carbono esperables en un futuro cercano bajo las condiciones predichas en distintos escenarios climáticos. Para todo ello, se han utilizado muestras de horizontes superficiales de suelos recogidas en tres áreas de estudio: Galicia, Europa y las Islas Galápagos.

Fundamentos teóricos

La cartografía digital consiste en la creación de un sistema de información geográfica que muestra la distribución espacial de

ciertas propiedades del suelo como, por ejemplo, el carbono orgánico, la capacidad de intercambio catiónico, el pH, la textura de los suelos, etc. Las muestras de suelo recogidas en una determinada área, se analizan en el laboratorio para obtener valores de las propiedades que queremos cartografiar. Una vez hecho esto, se utiliza una función matemática, llamada función espacial predictiva, para determinar la relación entre dichas propiedades y otras variables ambientales de las cuales ya existen mapas y cuyos valores son extraídos para los puntos de muestreo. Esta función matemática busca una relación entre la propiedad de estudio y las variables ambientales auxiliares. Finalmente, esta relación se utiliza para extrapolar y obtener la distribución espacial de la propiedad de interés a lo largo de un área geográfica.

El primer paso en la cartografía digital es la identificación de las variables que más influyen en la explicación de una propiedad concreta del suelo. Es decir, aquellas que permiten un mejor ajuste de la función espacial predictiva. De forma general, la ecuación de McBratney identifica 7 factores como los más influyentes en la formación del suelo. Es de esperar que estos factores también contribuyan significativamente en las dinámicas del resto de propiedades del suelo. De acuerdo con esta ecuación, las propiedades que queremos modelar serían función de otras propiedades del suelo previamente medidas, de las condiciones climáticas, de los organismos, de los atributos topográficos, del material de partida sobre el que

se desarrolla el suelo, del tiempo y de la localización geográfica. Evidentemente, la importancia de estas variables son especificas del área de estudio y de la propiedad que se pretende determinar y, de hecho, algunas de ellas pueden carecer de importancia en determinados casos.

Con el boom de la era informática, los estudios en el campo de la cartografía digital se han incrementado significativamente, demostrando el enorme potencial de la cartografía digital para inferir la distribución espacial del carbono orgánico con un esquema de muestreo limitado. Los continuos avances en la ciencia computacional han propiciado el uso de nuevas funciones matemáticas que permiten obtener mejores ajustes en los modelos empleados para obtener la distribución espacial de distintas propiedades del suelo. Las funciones o algoritmos matemáticos que se emplean en la cartografía digital varían desde algoritmos geo-estadísticos simples, como la regresión linear multiple, kriging, la regresión de mínimos cuadrados o la regresión ponderada geográficamente, hasta algoritmos mucho más complejos, como las máquinas de aprendizaje.

La principal limitación de la cartografía digital, en vista a su posible uso para monitorizar el cambio en las propiedades del suelo con el paso de los años, es el procesado de las muestras y la obtención de datos analíticos en el laboratorio. La espectroscopía infrarroja es una técnica con un enorme

potencial para complementar a la cartografía digital debido a sus numerosas ventajas, como por ejemplo, la posibilidad de realizar medidas in-situ, su rapidez, bajo precio y la posibilidad de hacer análisis no destructivos que suponen una alternativa poderosa frente a los métodos químicos tradicionales de combustión u oxidación química. Durante las últimas décadas, las técnicas espectroscópicas se han empleado para predecir diferentes propiedades de los suelos; sin embargo, su uso como complemento a la cartografía digital todavía no ha sido ampliamente explorado. Entre algunos de los ejemplos en los que se ha utilizado la espectroscopía infrarroja se encuentran el estudio de diversas propiedades físicas, químicas y biológicas como el contenido en agua, el tamaño de partícula, el contenido de carbono orgánico, el contenido de materia orgánica, la capacidad de intercambio catiónico, los macronutrientes, micronutrientes, la mineralogía del suelo, la conductividad eléctrica, el pH, la concentración de ciertos contaminantes o la biomasa microbiana. Al igual que sucede con la cartografía digital, se emplea un algoritmo matemático para buscar la relación entre el espectro y la propiedad de estudio.

Cartografía digital del carbono orgánico en Galicia, NO España

En el primer caso de estudio presentado en este trabajo se desarrolla una metodología estadística para estimar el contenido de carbono orgánico en Galicia. Este estudio supone una continuación de un estudio previo, llevado a cabo en nuestro grupo de investigación, en el cual se logró obtener la distribución espacial del carbono orgánico en nuestra comunidad autónoma. En el presente estudio se pretende evaluar la capacidad de la espectroscopía infrarroja para obtener dicha distribución espacial.

En primer lugar, se determina el contenido de carbono orgánico en 221 muestras de epipedones de suelo utilizando el método Walkley-Black y se obtiene el espectro en el rango infrarrojo medio (FTIR-ATR) del mismo conjunto de muestras. Como variables auxiliares se emplean diversos mapas que incluyen el tipo de geología, los usos del suelo y mapas que muestran las condiciones climáticas propias de esta región.

El proceso estadístico se divide en varias partes:

i) El algoritmo Random Forest (RF) se utiliza para establecer la relación que existe entre las concentraciones de

carbono orgánico y el espectro FTIR-ATR. Con esto conseguimos identificar cuales son las bandas de absorbancia que explican la mayor variabilidad del carbono. Los resultados indican que la banda a 1697 cm^{-1} explica la mayor parte de la variabilidad. El buen ajuste obtenido en el modelo matemático demuestra que los datos espectroscópicos pueden usarse para obtener la concentración del carbono orgánico, minimizando los costes analíticos asociados a la técnica Walkley-Black. Los datos bibliográficos indican que la banda a 1697 cm^{-1} está asociada a la vibración del enlace carbonilo de aldehídos, cetonas y ácidos carboxílicos presentes en compuestos del suelo relacionados con altos contenidos de materia orgánica.

ii) Una regresión de mínimos cuadrados parciales (PLS) permite modelar la distribución espacial de la banda espectroscópica, empleando como covariables una serie de mapas de la distribución de los parámetros ambientales citados anteriormente (clima, uso de la tierra y geología).

iii) El mapa obtenido en el paso anterior, que muestra la distribución de la señal espectroscópica, se utiliza para cartografiar la distribución espacial del carbono orgánico a lo largo de la región gallega. Para ello, se crea un modelo de regresión lineal que relaciona las concentraciones de carbono orgánico a partir de los valores espectroscópicos a 1697 cm^{-1}.

El mapa del contenido de carbono se obtiene mediante sustitución directa en la ecuación de regresión.

Los coeficientes de correlación entre el carbono orgánico y las covariables ambientales indican que la acumulación de carbono está influenciada principalmente por una alta precipitación y la presencia y disponibilidad de agua en el suelo. Otros factores como la naturaleza del material geológico o los usos del suelo, juegan un papel minoritario en la acumulación y degradación del carbono orgánico.

Los buenos parámetros obtenidos en los distintos algoritmos demuestran que el método propuesto es capaz de estimar y obtener la distribución espacial del carbono orgánico a partir de los datos espectroscópicos. Para evaluar el error asociado a las estimaciones, hemos comparado el mapa de carbono obtenido en esta aproximación con el obtenido anteriormente por el grupo de investigación, y en el cual no se utilizaban datos espectroscópicos. A pesar de que los valores de validación obtenidos en la aproximación anterior son ligeramente mejores, las diferencias entre los dos mapas son bajas en la mayor parte del territorio, demostrando la capacidad de las técnicas espectroscópicas para cartografiar de forma efectiva el contenido de carbono orgánico en Galicia. El método estadístico desarrollado podría aplicarse para estimar y cartografiar el contenido de carbono orgánico en regiones con condiciones climáticas similares.

Cartografía digital del carbono orgánico en Europa

En vista del buen ajuste obtenido a escala regional, en el segundo caso de estudio se desarrolla un método estadístico, basado en medidas espectroscópicas en el rango visible-infrarrojo cercano (VNIR), para cartografiar y monitorizar el contenido de carbono orgánico y la distribución espacial de los factores ambientales que controlan su almacenamiento a escala europea. Como variables auxiliares se utilizan mapas que muestran la distribución del pH, los tipos de geología, los usos del suelo, la temperatura y el índice de aridez y el contenido en fragmentos gruesos.

Se utiliza la base de datos Land Use/Cover Area frame statistical Survey (LUCAS), disponible para su descarga libre a través de la web del Joint Research Centre. Esta base de datos contiene los valores de ciertas propiedades medidas mediante métodos estándar certificados. Entre ellas se encuentra la textura, el pH, el contenido carbono orgánico, el contenido en carbonatos, el contenido de fosforo, nitrógeno total, el contenido de potasio, la capacidad de intercambio catiónico y el espectro VNIR de casi 20000 muestras de suelos recogidas en 25 países europeos.

Los pasos estadísticos utilizados para cartografiar el carbono orgánico y los factores que influyen en su acumulación son los siguientes:

i) En primer lugar se realiza la identificación de grupos de muestras con una señal espectroscópica similar. La metodología estadística propuesta se basa en que las propiedades geoquímicas en las muestras de suelo son determinadas por las condiciones ambientales existentes en cada lugar. Estos factores ambientales dejan su huella en la señal espectroscópica de las muestras. El contenido de materia orgánica también tiene una fuerte contribución a la señal espectroscópica. Por lo tanto, la acumulación o degradación de carbono orgánico en las muestras con un espectro similar se atribuye a procesos similares. Para identificar los grupos de muestras a partir de los datos espectroscópicos se realiza un análisis de componentes principales (PCA) para reducir la dimensión de los datos mediante la creación de nuevas variables, o componentes principales, que son combinación lineal de las variables originales. Una vez hecho esto, las puntuaciones del PCA, que representan los datos espectroscópicos en el nuevo espacio de variables, se utilizan para realizar un análisis de clasificación (k-means) que consiste en la agrupación de las muestras en función de las similitudes entre sus espectros. Utilizando el método "elbow" se identifican cuatro grupos de muestras.

ii) Una vez identificados los grupos, para cada uno de ellos se calculan las distancias "buffer" entre las muestras. De esta forma se tiene en cuenta el efecto de la localización espacial de las muestras pertenecientes a los diferentes grupos. Estos mapas "buffer" se combinan con los mapas de las propiedades ambientales para calcular los componentes predictivos espaciales utilizando PCA. Estos componentes son utilizados como covariables para construir un modelo de clasificación Random Forest (RF) probabilístico. Este modelo permite crear un mapa que muestra la distribución espacial de los grupos en Europa.

iii) El tercer paso de la metodología estadística desarrollada consiste en la creación de un modelo de regresión RF para cada uno de los grupos de muestras identificados previamente, cuya finalidad es determinar la relación entre el contenido de carbono orgánico del suelo y las variables ambientales. Cada una de las cuatro ecuaciones de regresión (una por grupo) se aplica al área correspondiente en el mapa de grupos, que ha sido calculado en el paso previo, con el fin de obtener la distribución espacial del contenido de carbono en Europa. Además, se obtiene el mapa del error estándar asociado a estas predicciones mediante kriging de los residuales de los modelos de regresión.

Los modelos de regresión RF mostraron un ajuste relativamente bueno para dos de los grupos de muestras identificados ($R^2 = 0.58$ y $R^2 = 0.51$), mientras que el ajuste

fue peor para los grupos restantes ($R^2 = 0.38$ y $R^2 = 0.34$). El mapa que muestra la distribución espacial del contenido de carbono orgánico en el suelo sigue el patrón encontrado en estudios previos.

El método desarrollado también permite definir áreas con características ambientales similares, ayudando así a una mejor comprensión la distribución espacial de las condiciones ambientales que influyen en el contenido de carbono orgánico. Analizando el espectro medio de cada grupo, los gráficos de importancia extraídos de los modelos de regresión RF, la distribución de las variables ambientales entre las muestras pertenecientes a cada grupo y la distribución espacial de los grupos, se determinan cuales son las variables ambientales más influyentes en la acumulación de carbono orgánico en cada zona geográfica. Los resultados muestran que existe una clara tendencia de los factores que promueven la acumulación del carbono orgánico, siendo el pH, los fragmentos gruesos y el índice de aridez los principales reguladores del contenido de carbono en la mayoría del territorio europeo. Las condiciones climáticas son el principal factor que influye en la acumulación/degradación del carbono orgánico a escala continental, pero también otros parámetros como el tipo de uso del suelo influyen en la cantidad de carbono a escalas locales. Una comprensión profunda de los factores que influyen en el contenido de carbono orgánico en Europa es crucial para implementar futuras decisiones dirigidas a

mejorar la capacidad de secuestro de carbono atmosférico de los suelos europeos.

Cartografía digital del carbono orgánico en la isla Santa Cruz, Galápagos

En el último caso de estudio se cartografía el contenido de carbono orgánico almacenado en los suelos de la isla Santa Cruz, Galápagos. Existe una creciente preocupación sobre como las actividades humanas podrían afectar los ecosistemas de las islas Galápagos, reconocidas mundialmente por su enorme biodiversidad. Concretamente, se ha sugerido la creación de un sistema de monitorización ambiental para detectar los impactos negativos que podría tener el cambio climático sobre los ecosistemas de las islas. En este sentido, nuestro estudio pretende evaluar la capacidad de la espectroscopía infrarroja como método de monitorización y cuales serán los contenidos de carbono orgánico en los próximos años estimados a partir de simulaciones de las condiciones climáticas esperadas bajo distintos escenarios de cambio climático. Estos escenarios van desde un aumento

moderado hasta un aumento acusado en la concentración de gases de efecto invernadero a nivel atmosférico.

A pesar del gran número de estudios relacionados con las relaciones ecológicas entre especies en estas islas, prácticamente no se han realizado investigaciones sobre sus suelos, un componente básico de los ecosistemas de dichas islas. Debido a las restricciones impuestas a la hora de realizar campañas de muestreo en la actualidad en las islas Galápagos inherentes a su estado de protección, en este estudio se utilizan muestras recogidas en la última expedición geo-pedológica. Ésta tuvo lugar en 1962 por investigadores de una misión franco-belga, liderada por la Universidad de Gante y su finalidad era recopilar información detallada sobre la diversidad del suelo en el archipiélago.

Se utilizan un total de 36 muestras de epipedones de suelo recogidas en la isla Santa Cruz para cuantificar la cantidad de carbono orgánico y se registra el espectro infrarrojo (FTIR-ATR) de dichas muestras. Debido a que disponemos de pruebas evidentes de la influencia de la precipitación en la formación y el desarrollo de los suelos de la isla, esta variable climática se utiliza como variable auxiliar para cartografiar el carbono orgánico. Debido a la baja resolución de los mapas de precipitación obtenidos se utiliza un modelo de elevación digital y un mapa que muestra la distribución de vientos en las islas para cambiar la resolución de los mapas de

precipitación de 1 km a 90 m. Para ello se utiliza un modelo lineal que busca la relación entre los mapas de estas dos variables y el de precipitación.

Se emplea el algoritmo Geographically Weighted Regression (GWR) para cartografiar la cantidad de carbono orgánico en las isla Santa Cruz. Los resultados indican que los suelos de la isla almacenan aproximadamente 706 Gg de carbono en los 10 cm superiores. El buen ajuste de la función indica que la acumulación de carbono está principalmente impulsada por factores climáticos. Éstos, a su vez, están altamente influenciados por la altitud y la dirección de los vientos predominantes. De hecho, los resultados indican que las concentraciones más altas de carbono aparecen en áreas situadas a mayor altitud en las laderas de barlovento, donde las tasas de precipitación son mayores. Aunque el número de muestras disponibles está limitado a la ladera de barlovento de la Isla de Santa Cruz, se espera el mismo patrón de precipitaciones para el resto de las islas del archipiélago. Según esto, una generalización de este modelo indica que los suelos de Galápagos acumulan 54 Tg de carbono en los 10 cm superiores.

Los datos de precipitación predichos para los periodos 2041-2060 y 2061-2080, se usan para calcular cual sería la cantidad de carbono en las isla Santa Cruz bajo diferentes escenarios climáticos. Tal y como predicen la mayoría de los

modelos climáticos, los futuros escenarios del cambio climático supondrán un aumento en la cantidad de lluvia que, a su vez, estará asociado a un incremento general de la cantidad de carbono orgánico y que, probablemente, modificará la composición de las especies vegetales existentes en los diferentes estratos bioclimáticos de las islas. La incertidumbre en las predicciones climáticas, junto con el tamaño de muestras recogidas, que es relativamente pequeño, puede limitar la capacidad de nuestro modelo para predecir las reservas de carbono. Sin embargo, este estudio constituye el primero que consigue cartografiar y evaluar el carbono orgánico total almacenado en los suelos en estas islas.

Por último, se utiliza un modelo de regresión por mínimos cuadrados parciales (PLS) para demostrar que el contenido carbono orgánico de la isla Santa Cruz puede obtenerse fácilmente a partir de medidas espectroscópicas. Con lo cual, un programa de monitorización, basado en análisis espectroscópicos, podría utilizarse en el futuro para determinar las variaciones temporales de los stocks de carbono de manera rápida, rentable y, a la vez, minimizando las perturbaciones humanas en las islas Galápagos.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA