Novel applications of spectroscopy to characterize soil variation

Mario Fajardo P.

A thesis submitted in fulfilment of requirements for the degree of

Doctor of Philosophy

Faculty of Agriculture and Environment

University of Sydney

2016

Certificate of Originality

The text of this thesis contains no material which has been accepted as part of the requirements for any other degree or diploma in any university or any material previously published or written, unless due reference to the material has been made.

Mario Fajardo P.

Summary

The first modern soil science studies involving the measurement of soil properties such as those made by Professor Von Liebig, focused on the quantification of pure mineral elements present in the soil matrix. As the first half of the 20th century passed, the first works in the area of soil spectroscopy took place and the number of measurable soil attributes went from less than ten to many thousands. As an inmediate consequence, the use of multivariate methodologies and computer-based analysis became indispensable. Today, we are witnessing a moment in history where the amount of soil information available probably exceeds our analytic capabilities; therefore the need for transforming this raw data into useful information for final users is a continuous challenge.

This thesis embodies a collection of novel studies related to the use of multivariate information provided by spectroscopic tools such as Visible and Near Infrared (Vis-NIR) spectrometers to represent soil variation. The thesis general structure is organized to follow the increasing levels of soil complexity, starting from the characterization of soil aggregates and the identification of soil colloids, to the recognition of soil horizons and their boundaries in the soil profile, to finally the depiction of soil type's distribution in the landscape.

It is shown, as the complexity of the soil target attribute increases (soil colloids < soil aggregates < soil horizons < soil types), the use of the multivariate information needs to be adapted as well, starting from recognition of individual soil properties (identification of individual soil colloids represented in a single spectrum in Chapter 2) to the use of the

full variation contained in the spectroscopic information (creation of multidimensional indices in Chapter 5).

Briefly, Chapter 1 is written as a rationale, presenting the need for up-to-date methodologies for making effective use of the increasing amount of soil information produced worldwide. It gives examples of the existing types of soil information and suggests the use of soil spectroscopy and multivariate analyses techniques as candidates for the study of key aspects of soil at different scales and levels of complexity.

Accordingly, starting from the micro-scale and the analysis of individual soil properties, Chapter 2 presents a newly developed methodology for the measure of soil aggregate stability and the further use of spectroscopic information to predict its values. The new methodology makes use of an image recognition algorithm that allows measurement of the slaking of soil aggregates in water over time. The resulting increase of area in time was fitted to a function and three "slaking" coefficients were obtained. Coefficient *a,* was related to the highest possible area increase due to slaking in time. Coefficient *b* was associated with the initial time of slaking and finally coefficient *c* was linked to the rate of area increase in time. It was found that these coefficients were related with known soil properties, for example, coefficient *a* was positively related with exchangeable sodium (cmol(+) kg^{-1}) and negatively related to exchangeable calcium–exchangeable magnesium ratio (Ca Mg^{-1}). On the other hand, coefficients *b* and *c* were related with percentage of total carbon (%TC) and percentage of total nitrogen (%N).

The second part of this chapter presents the use of different types of spectroscopic instruments, namely Vis-NIR and Middle Infrared (MIR) both portable and bench spectrometers, for the prediction of the slaking coefficient *a.* The regions of the electromagnetic spectra identified as important, were associated with Fe oxides and Fe bearing minerals, organic carbon and presence of kaolinite bearing materials. This chapter also presents the possibility of predicting slaking values using Vis-NIR and MIR information, showing that if an appropriate calibration dataset is used, the predictions are moderately acceptable ($R^2 = 0.6$).

Following the same line of research, in terms of the prediction of individual soil properties by using Vis-NIR spectroscopy, Chapter 3 demonstrates the use of selected state-of-the-art spectroscopic modelling techniques and large spectral libraries. The example presented in this chapter, involves the use of two large spectral libraries containing soil samples from the conterminous United States for the prediction of %TC in samples of a certain region of U.S. or "local samples". The results of this chapter put special emphasis on the large amount of variability that Vis-NIR information could contain and on the importance of a properly selected set of soil samples to create Vis-NIR models. Even though all the modelling approaches were successful ($R^2 > 0.9$), the important effect of similar composition regarding their geographic similarity which significantly affected the final bias of the models was noticeable (0.8 vs 0.02 %TC).

As the complexity of the target soil feature increases (soil colloids < soil aggregates < soil horizons), it is observed that rather than using multivariate information to predict separate soil properties, this information can be directly used to describe soil variation. In this sense Chapter 4 details the development of a new method for the identification of soil horizons and their boundaries using fuzzy clustering of Vis-NIR spectra. The results of this chapter showed that the newly identified horizons (spectrally derived horizons), bear a resemblance to those described by pedologists.

Chapter 5 follows the same essential concepts of Chapter 4 in terms of using the raw information contained in the Vis-NIR region of the spectra to characterize the soil profile rather than using individual soil attributes. This chapter expands into a new way of measuring the diversity of soils in the landscape. It presents two new indices for measuring soil functional diversity or “Functional Pedodiversity” inspired by previous studies in Functional Ecology where individuals are characterized in a multidimensional space composed by several continuous properties. The new “Functional pedodiversity indices” are then calculated using information from the whole Vis-NIR spectra and then are compared with the conventional approaches or “Taxonomic pedodiversity indices” which use previously classified soil taxonomic orders. The results of this chapter show the close connection between soil attribute variation and soil taxonomic units. This chapter offers a new perspective on the measurement of pedodiversity and represents the first study that assesses pedodiversity by spectral means.

Finally Chapter 6 discusses the main findings of this thesis and foresees issues, challenges and opportunities in the area of spectroscopy and multivariate soil data analysis.

Acknowledgements

This is the hardest part that I have written in this thesis so far[1]. It is difficult to give proper recognition to all the people that made this work possible (some of them without knowing it).

If someone asks me for the first person that I need to acknowledge, that one will be Alex. To be honest, before coming to Sydney I have literally no idea the extent of his work and how much I was about to learn here. Thanks Alex for the countless times that you told me: "Ok that's good, but what if we do ... <*Insert something incredibly difficult to do here*>". I think the best ideas presented here came from those questions, and as I told you several times "*That's why you are the boss*". Thank you for all the help and the patience, really.

Thank you Brett for the unconditional help and friendship, as I have said many times (and also the boss) *"You are the most intelligent person that I know"*. To be able to make some serious science while being able to have a normal conversation is a rare talent that only some people have.

Thank you Budi for the incredible capacity of solving every (very difficult) problem that I had during my thesis, you are always two or three steps beyond, and still you had the time to go back and lend some help when needed.

[1] By "so far" I am not referring to the Summary of course … despite this is the first part, these are the last few lines that I will write in this thesis.

Thanks Damien for all the ideas that you gave me, and the countless times that we shared a good laugh! The office wouldn't be "the office" without you, thanks for making every day's work, more than just sitting in front of a screen, thanks[2]!

Thanks Pinto for all the *"mornings"* that we have shared! To be honest, you should be first in the acknowledgements, but *"the bosses are the bosses"*. Thanks for every little moment that we have spent, you are by far the most important person for me here in Oz.

Thanks José and Nico for being such good friends all these years, you made Australia feel like Chile literally, and thanks to all the other Chileans, if I named you all, it will be two more pages (lol…the future generations may not understand this expression).

Thanks to the Aussies (and not that Aussies) Phil, Seb, Brendan, Uta, Kanika, Maryam, Kanko (Yes you are considered an Aussie here why not, it's my thesis), Patito, Liana, Eduardo Jamezonez, Ish, Nathan, Stacey, J! (you are the best), Olivier, Jessica, Ali, Odeh (Chief), Tom, Floris, Willem, Yolima, Paola, Mana, Felipe, and many, many, many others. Especial mention to Ivan, Neil and Lori, the stories that we "survived" in the field were just incredible! Thank you all who helped somehow to make this possible!

Thanks to my parents Mario and Taly and my sisters Ivonne and Pauli, I love you all, and I'll be there soon to say LLEGUÉ!

It's not just about doing a PhD, it's about *to be able to do it!*

[2] And thanks for watering my desk's plant everyday …without you she would be dead for sure!

Publications

The following Chapters are published or under revision as research articles in scientific journals:

Chapter 2, *part1*.

Fajardo, M., McBratney, A., Field, D., Minasny, B., 2016. Soil slaking assessment using image recognition. Soil and Tillage Research 163,119-129.

Chapter 4.

Fajardo, M., McBratney, A., Whelan, B., 2016. Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles. Geoderma 263, 244-253.

Chapter 5.

Fajardo, M., McBratney, A., Minasny, B., 2016. Measuring functional pedodiversity using spectroscopic information. Catena. *In revision*

Table of Contents

Certificate of Originality ii

Summary iii

Acknowledgements vii

Publications ix

Table of Contents x

List of figures xvii

List of tables xxii

Chapter 1 Rationale 1

1.1 Soil information and soil variation 2

1.2 Information that we have 4

1.2.1 Soil quantitative information 4

1.2.2 Proximal soil sensing and soil spectroscopy 5

1.3 Information that we need 8

1.4 How can we get that information? 11

1.4.1 Soil colloids 11

1.4.2 Soil aggregation 13

1.4.3 Soil morphological horizons 15

1.4.4 Pedodiversity 17

1.5 Thesis objectives 19

Chapter 2 Soil Aggregate Stability 24

Part 1: Soil slaking assessment using image recognition 24

2.1 Summary 25

2.2 Introduction 26

2.3 Materials and Methods 28

2.3.1 Dataset 28

2.3.2 Sample preparation and complementary analyses 29

2.3.3 Digital single lens reflex (DSLR) camera images 30

2.3.4 Image processing and segmentation procedure 33

2.3.5 Statistical analyses 36

2.4 Results and Discussion 37

2.4.1 Segmentation performance 37

2.4.2 Soil disaggregation dynamics 39

2.4.3 Relationships with soil properties 41

2.4.4 Relationships with land-use 44

2.5 Discussion 46

2.6 Conclusions 46

Part 2: Spectral information related to soil structure 51

2.7 Summary 52

2.8 Introduction 54

2.9 Materials and Methods 56

2.9.1 Datasets 56

2.9.2 Sample preparation and soil analyses 56

2.9.3 Spectral treatments and statistical analyses 57

2.10 Results and discussion 60

2.10.1 Measured soil properties 60

2.10.2 Spectral information 60

2.10.3 Spectral model performance 62

2.10.4 Model usage 63

2.10.4.1 Vis-NIR part of the spectra 63

2.10.4.2 Vis-NIR + MIR part of the spectra 65

2.10.5 Two-step modelling 67

2.11 Uncertainty of predictive models and general discussion of the chapter 70

2.12 Conclusions 73

Chapter 3 Soil spectral databases 78

Use of large Vis-NIR spectral libraries for the prediction of local soil properties: A continental scale example 78

3.1 Summary 79

3.2 Introduction 81

3.2.1 Spectral libraries 81

3.2.2 Case scenario: RaCA dataset 83

3.3 Materials and methods 85

3.3.1 Datasets 85

3.3.2 Area of study 85

3.3.3 Spectral pre-treatments 88

3.3.4 Spectral modelling methodologies 88

3.3.4.1 Legacy approach 89

3.3.4.2 Legacy in-hull approach 89

3.3.4.3 Legacy in-hull plus spike approach 90

3.3.4.4 Legacy Latin hypercube sample within convex hull approach 91

3.3.4.5 Legacy Latin hypercube within convex hull sample plus a spike approach 91

3.3.4.6 RaCA model approach 92

3.3.4.7 RaCA in-hull model approach 92

3.3.4.8 RaCA local model approach 92

3.3.4.9 RaCA local in-hull model approach 93

3.3.4.10 Model developed with only information from a spike 93

3.4 Results and Discussion 94

3.4.1 TC in the datasets 94

3.4.2 Spectral similarity of the datasets 95

3.4.3 Overall model performance 96

3.4.4 Legacy approaches 97

3.4.5 RaCA approaches 99

3.4.6 RaCA local approaches 100

3.4.7 Spectral variability considerations and discussion 102

3.5 Conclusions 105

Chapter 4 Soil Morphology 109

Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles 109

4.1 Summary 110

4.2 Introduction 111

4.3 Materials and Methods 112

4.3.1 Study area 112

4.3.2 Sampling Design 113

4.3.3 Processing of Vis-NIR spectra for SPD hor detection 114

4.3.4 Fuzzy clustering 115

4.3.5 Classification performance assessment 116

4.3.6 Digital gradients for horizon boundary detection 117

4.3.7 Spectrally derived horizon identification performance 117

4.4 Results and discussion 118

4.4.1 Smoothing by wavelength to capture spectral features 118

4.4.2 Relations between horizons and Vis-NIR values 120

4.4.3 Clustering the dataset: global classes in the soil core 120

4.4.4 SPD *hor* detection 124

4.4.5 Overall performance 127

4.5 Final considerations 129

4.6 Conclusions 130

Chapter 5 Pedodiversity 133

Use of spectral information and new indices to quantify pedodiversity 133

5.1 Summary 134

5.2 Introduction 135

5.3 Materials and Methods 138

5.3.1 Dataset 138

5.3.2 Pedodiversity measurements 140

5.3.2.1 In-hull diversity 140

5.3.2.2 Eigenvalue diversity 141

5.3.2.3 Conventional diversity indices 142

5.3.3 Area–pedodiversity relationships 144

5.3.3.1 Point scale pedodiversity 144

5.3.3.1.1 Soil surface, 0 to10 cm 144

5.3.3.1.2 Soil profile, 0 to 100 cm 144

5.3.3.2 Large scale pedodiversity 144

5.4 Results and Discussion 145

5.4.1 Point scale pedodiversity: Soil surface 145

5.4.2 Point scale pedodiversity: Profile 148

5.4.2.1 Within-sample variation 148

5.4.2.2 Between-samples variation 150

5.4.3 Large-scale pedodiversity: Topsoil 151

5.4.4 Large scale pedodiversity: Profile 157

5.4.5 Large scale pedodiversity: Conventional Indices 160

5.4.6 Functional and taxonomic pedodiversity 162

5.5 Conclusions 163

Chapter 6 General discussion, conclusions and future work 168

6.1 Studying soil complexity 169

6.2 Electromagnetic spectrum 170

6.3 Getting the big picture 171

6.4 General conclusions 174

6.5 Opportunities and future work 176

6.5.1 Practical significance of the Slaking Index 176

6.5.2 Spectral information related to soil slaking 177

6.5.3 Large spectral databases 177

6.5.4 Soil morphological horizons 178

6.5.5 Pedodiversity 179

Appendices 182

List of figures

Figure 1-1 Percent reflectance vs. wavelength of incident radiation at various moisture contents (moisture contents indicated directly above each curve). *Extracted from* Bowers and Hanks (1965). 7

Figure 1-2 Theoretic hierarchical organization of pedological studies. 10

Figure 1-3 X-Ray diffraction patterns. *Extracted from* Chipera and Bish (2001). 12

Figure 1-4 Diagnostic absorptions of (a) smectite (1912 nm), (b) kaolinite (2165 nm) and (c) illite (2345 nm). The plots show the spectrum of each of the reference clay minerals (thicker gray line) and the spectra of the 5th, 6th, 50th, 84th and 95th percentiles of the soil samples (thinner black lines). *Extracted from* Viscarra Rossel (2011). 13

Figure 1-5 Pseudo three-dimensional pore structure of two soil samples at different wetting-drying cycles obtained from X-ray micro-tomography analysis. *Extracted from* Ma et al. (2015). 14

Figure 1-6 Field horizonation and calculated differences between layers for a pedon in Louisiana: (a) differences of clay, (b) differences of laboratory analysis, (c) elemental differences of portable X-ray fluorescence spectrometry (PXRF) readings on bulk density samples, (d) elemental differences of PXRF readings under field conditions, (e) elemental differences of PXRF readings on samples under laboratory conditions, and (f) elemental differences of PXRF readings on monolith. Extracted from Weindorf et al. (2012). 16

Figure 2-1 Area of study, NSW, Australia. 29

Figure 2-2 Aggregates and camera setup. 30

Figure 2-3 Aggregates before water immersion; *Note the high brightness of the picture is due to high exposure settings.* 31

Figure 2-4 Aggregates after 5 seconds of water immersion. 32

Figure 2-5 Red band of soil aggregates image. 33

Figure 2-6 Green band of soil aggregates image. 34

Figure 2-7 Blue band of soil aggregates image. 34

Figure 2-8 Detail of image processing of a single aggregate. 35

Figure 2-9 Detail of a two aggregates image segmentation process, different colours represent different labels. 36

Figure 2-10 Area (pixels) of 5 aggregates as a function of time. 38

Figure 2-11 Slaking index of soil aggregates with time of a single soil sample. The dots represent the mean value, and the grey area represents the envelope within ± 1standard deviation. 38

Figure 2-12 Dynamics of contrasting disaggregation patterns and the respective Gompertz function model. 40

Figure 2-13 Biplot of SI values and measured properties on all the samples. 43

Figure 2-14 The relationship between Total C content (%TC) and coefficient *a.* 44

Figure 2-15 Boxplot of coefficient *a* values by land use. Dots correspond to values outside 95% of observations. 45

Figure 2-16 Boxplot of coefficient *b* values by land use. Dots correspond to values outside 95% of observations. 45

Figure 2-17 Boxplot of coefficient *c* values by land use. Dots correspond to values outside 95% of observations. 45

Figure 2-18. Vis-NIR spectra convex hulls for three datasets. 61

Figure 2-19. Portable MIR spectra convex hulls for three datasets. 61

Figure 2-20. Lab MIR spectra convex hulls for three datasets. 62

Figure 2-21 Wavelengths usage of Vis-NIR models. 63

Figure 2-22 Main wavenumbers of MIR-NIR models (NIR part of portable MIR). 64

Figure 2-23 Main wavenumbers of portable MIR models. 65

Figure 2-24 Main wavenumbers of laboratory MIR models. 66

Figure 2-25 Observed vs. predicted values of Two-step model procedure in two validation datasets. 68

Figure 3-1 Effect of different parameters on absorbance at 1915 nm, *extracted from* Stevens et al. (2013b). 82

Figure 3-2 RaCA LUGR regions of RaCA project, *extracted from* Soil Survey Staff (2013). 86

Figure 3-3 Dominant U.S. Soil taxonomy soil orders, *extracted from* National soil survey center staff (2013). 86

Figure 3-4 Physiographic map of Texas, *extracted from* Wermund *(1996).* 87

Figure 3-5 Legacy approach scheme. A is Legacy dataset and B is Validation dataset. 89

Figure 3-6 Legacy in-hull approach scheme. A is Legacy dataset and B is Validation dataset, section in red corresponds as the in-hull dataset used as training. 90

Figure 3-7 Spiked legacy dataset within RaCA convex hull scheme. A is Legacy dataset and B is Validation dataset, area filled in red correspond to in-hull dataset and circles in green to RaCA spike. 90

Figure 3-8 Latin hypercube sample within RaCA convex hull scheme. A is Legacy dataset and B is Validation dataset, circles filled in red correspond to the Latin hypercube in-hull dataset used as training. 91

Figure 3-9 Latin hypercube spike of RaCA plus legacy scheme. A is Legacy dataset and B is Validation dataset, circles filled in green correspond to the latin hypercube spike dataset added to the legacy training. 92

Figure 3-10 Kernel density distribution for Total Carbon % values of datasets. 94

Figure 3-11 Convex hull of first the 2 Principal components of the used datasets. 95

Figure 3-12 Observed vs predicted values (%TC) of model using spike as training. Continuous line represents 1:1 observed - predicted ratio and dashed line represents linear regression of predictions. 98

Figure 3-13 RaCA dataset (Light yellow asterisks), RaCA in hull (red circles), RaCA local (black circles) and RaCA local in hull (red over black circles) sampling locations. 99

Figure 3-14 Observed vs predicted values of model using RaCA training. Continuous line represents 1:1 observed - predicted ratio and dashed line represents linear regression of predictions. 100

Figure 3-15 Bias values for all the approaches. Boxes represent the values of twenty iterations 101

Figure 3-16 TC% for RaCA local, RaCA local in hull and validation dataset. 101

Figure 3-17 Example of change in convex hull in legacy in hull dataset after addition of a 20 samples spike. *Note that legacy in hull dataset (in red) had almost 12,000 samples.* 103

Figure 3-18 Logical scheme for use of reflectance spectral libraries in a risk-based approach to prediction of soil functional attributes. Extracted from (Shepherd and Walsh, 2002). 104

Figure 4-1 Lower Hunter valley study area and sample locations. 113

Figure 4-2 Spectral variation of intact, processed cores and smoothing procedure. 119

Figure 4-3 *left* and *center*, Relations between horizons origin and associated spectra; *right* showing A1 horizons in the ellipse. 120

Figure 4-4 Chosen number of fuzzy classes. 121

Figure 4-5 Effect of membership exponent in overall confusion index values. 121

Figure 4-6. Distribution of classes by type of horizon. 122

Figure 4-7 Fuzzy classes centroids. *Left*, Spectra of Centroids; *Right*, Euclidean distance between centroids. 123

Figure 4-8 Detailed spectra of centroids selected regions. 124

Figure 4-9 Digital gradients in a soil core. The value of the DG comparison window λ was set to 4 cm. Boundaries distinctness is relative to the thickness of the boundary. 125

Figure 4-10 Distribution of ICC values for 56 described soil cores (after outlier removal) of traditional horizons designation and *SPD hor.* 128

Figure 5-1 Vis-NIR soil scans by core. *Left*: First 60cm of 1m core, *Right*: Schema of Vis-NIR observations. 139

Figure 5-2 Schema of area-pedodiversity calculations, *N represents the number of sites considered.* 145

Figure 5-3 Three sites example of convex hulls created using a score matrix with spectra as input (First 10 cm observations). The three selected samples are separated by approximately 50km each. 146

Figure 5-4 Three sites example of convex hulls created using PC scores of five properties (First 10 cm observations). The three selected samples are separated by approximately 50km each. 146

Figure 5-5 Convex hull created using PC scores of spectra as input for one site (Profile). 149

Figure 5-6 Convex hull created using a score matrix with properties as input for one site (Profile). 149

Figure 5-7 Convex hulls created using with the first 2 PC scores of the spectra for three sites (Profile). The three selected samples are separated by approximately 50km each. 150

Figure 5-8 An example of convex hulls for 3 sites created using PC scores of five soil properties (Profile). The three selected samples are separated by approximately 50km each. 151

Figure 5-9 Mean annual temperature map of NSW. Source: (ABARES et al., 2014). 152

Figure 5-10 Elevation map of NSW. Source: (ABARES et al., 2014). 153

Figure 5-11 Australian soil classification soil types of NSW. Source: (ABARES et al., 2014). 153

Figure 5-12 HULLdiv index calculated using Vis-NIR spectra at different scales across the two transects. 154

Figure 5-13 HULLdiv index calculated using 5 soil properties at different scales for the two transects. 155

Figure 5-14 EIGENdiv index calculated using Vis-NIR spectra at different scales for the two transects. 156

Figure 5-15 EIGENdiv index calculated using predicted properties at different scales for the two transects. 156

Figure 5-16 HULLdiv index calculated using Vis-NIR spectra at different scales for the two transects (Profile). 158

Figure 5-17 HULLdiv index calculated using soil properties at different scales across the two transects (Profile). 158

Figure 5-18 EIGENdiv index calculated using Vis-NIR spectra at different scales for the two transects (Profile). 159

Figure 5-19 EIGENdiv index calculated using predicted properties at different scales for the two transects (Profile). 160

Figure 5-20 Shannon index calculated using Australian Soil Classification orders for two transects and land uses. 161

Figure 5-21 Simpson index calculated using Australian Soil Classification orders for two transects and landuse. 161

Figure 5-22 Functional vs Taxonomic pedodiversity. Note that values were standardized in order to show the results in the same scale. 162

List of tables

Table 2-1 DSLR camera settings. 32

Table 2-2 Measured soil properties of the dataset. 42

Table 2-3 Spectral pre-treatments. *S-G filt1 and S-G filt2: Second order Savitsky-golay filter with none and first derivative output respectively; Strip5 and Strip10: Selection of every 5th and 10th wavelength respectively; Trim: Elimination of the extreme wavelengths prone to high noise to signal ratio; SNV: Standard normal variate transformation; ChBLC: Convex hull baseline correction; Det: Detrending algorithm as specified in Stevens and Ramirez-Lopez (2013).* 58

Table 2-4 Spectral models performance on independent and external dataset. 62

Table 2-5 Performance parameters of Two-step model procedure in two validation datasets. 68

Table 2-6 Time-cost-uncertainty of SI methodologies. The time used in the image recognition analysis considered the use of five digital cameras working in parallel. Cost was calculated assuming an average salary of $AUD 21 per hour for a laboratory assistant. 71

Table 3-1 Different model approaches performance for the mean predictions of 20 iterations ordered by lower to higher RMSE values. *LH*: Latin hypercube sample. 96

Table 4-1 Example of class percentages in one soil profile. 126

Table 5-1 Example of pedodiversity indices for three selected samples using the first 2 PC scores of the spectra and soil properties. 148

Chapter 1 *Rationale*

1.1 Soil information and soil variation

During the last decade we have seen an increase in the global awareness of the importance of the soil resource, noted by The Food and Agriculture Organization of the United Nations (FAO) as “the medium for preservation and advancement of life on earth” (Omuto et al., 2013) and also by United Nations (UN) in its resolution A/RES/68/232 adopted by the general assembly as “The foundation for agricultural development, essential ecosystem functions and food security and hence key to sustaining life on Earth” (United Nations, 2014). These statements seem to agree with the increasing number of publications related to soil science and the renewed interest in the specialty, as Hartemink and McBratney (2008) state, we are witnessing a “Soil science renaissance”, where high quality and up-to-date information is becoming a necessary asset. In order to fully understand this statement, it is necessary to first define what is understood by soil information.

In a few words, soil information can be considered as the qualitative or quantitative representation of the soil condition in a determined area or volume. Now, the assessment of this condition would not be necessary if soil was totally homogenous, but it is not. Since soil is one of the most variable natural resources on earth we indeed require soil information to describe its variation in space and time.

As an example, soil condition in an area of interest can be assessed using a number of methodologies, depending on the aspect that needs to be described e.g., content of Nitrogen in a determined location is commonly assessed by Kjeldahl method (Bremner,

1960) and is expressed as a numerical result in a continuous scale (e.g., mg N g soil^{-1}) which allows comparison with other locations and thus efficiently describes its state.

Following the same example, to describe how nitrogen content varies in space, it will be required to assess it with the same methodology in another place, and ideally at many different locations. If the previous exercise is performed at regular intervals we can describe systematically its variation in space using the variogram (Matheron, 1965), or the average variance[3] of any two points within a finite area separated by a given distance and direction, and possibly predicting with this the nitrogen content at unsampled locations.

The soil can be characterized by multiple attributes e.g., colour, clay content, number of horizons in a pedon, etc. All of these attributes contributing with a certain amount of information that helps in characterizing it as a whole. However, when trying to describe different aspects of soil, we find that not all the information that is produced comes in the same format. As an example, a common way to describe the soil structure in the field is by visually describing the grade of soil aggregates e.g., massive or structureless, moderate, strong, etc. Yet, in laboratory conditions it is possible to describe it in another more precise but time consuming way e.g., mean weight diameter (Van Bavel, 1950).

Each of the previously mentioned methods has its own benefits and limitations, so if our intention is to describe the soil variation in the best possible way, first it is needed to determine what kind of soil information can be considered strategically important to efficiently assess the soil condition. From this statement we can formulate two simple

[3] Referred to its statistical meaning, defined as the average of the squared differences from the mean.

questions: what kind of information do we have and what kind of information do we need.

1.2 Information that we have

1.2.1 Soil quantitative information

In answering the first question, we immediately can make a clear distinction in what has been probably one of the paradigm shifts in soil science of the last 50 years, the creation of a formal term for the use of quantitative methodologies applied to soil science i.e., Pedometrics.

Interestingly, the use of the first recognized pedometric technique was the variogram, which came from the same field of science (Geology), from which many years ago soil science was formed in the Russia of the 19th century. Accordingly, the first area of soil science which saw a big influence of this new paradigm was the one of modern geostatistics applied to soil science and the study of its spatial autocorrelation (Mercer and Hall, 1911). From that time until now, the use of geostatistical methods applied to the mapping and prediction of soil attributes in the landscape or Digital Soil Mapping (DSM) has been the most prolific of all. However this success was not only due to the high expertise of the DSM community, but to a conjunction of factors which led to its success namely, the production of vast amounts of data of different sources (mainly remote Sensing e.g., aerial photography or hyperspectral imagery), the use of global navigation satellite systems (e.g. GPS), and to the creation of new methodologies to make good use of all this information.

Some examples of good use of remote sensing information are the already mentioned variogram, the use of an empirical approach of Jenny's soil forming factors (i.e., The *scorpan-SSPFe* model (McBratney et al., 2003)) to efficiently predict soil properties in the landscape, and the integration of deterministic and stochastic modelling like regression kriging procedures (Odeh et al., 1995), among others. However as Webster (1994) explained when he was told of the term Pedometrics, the original meaning was thought as a whole quantitative approach and especially relevant to observations made in the field, which take us to another source of soil information, this is Proximal soil sensing.

1.2.2 Proximal soil sensing and soil spectroscopy

Historically, and within its very core, soil science has been linked with the needs of society, and just like economic cycles (Kondratiev, 1925), it observes cycles of high and low demand. The previous observed "peak" occurred after the Second World War as commented upon by Tinker (1985). Today, as some authors have noticed (So and Lal, 2002; Hartemink and McBratney, 2008; Churchman, 2010) a new "peak" has started, but this time the trigger is not only hunger alleviation, but a whole set of society needs, such as, inter alia, economic growth, maintaining biodiversity, improving water availability, reducing desertification and land degradation and also to provide food security (United Nations, 2014). In other words, a very large number of issues drives a reason to start working on fast, scalable solutions to provide soil information in an effective and efficient way to final users and policy makers (Bouma, 2010; Bouma and McBratney, 2013).

As mentioned in section 1.2.1 an effective and also efficient way to provide soil information has been through the increasing use of remote sensing technologies resulting in a change of perspective of soil research from farm scale to a global extent (Jensen, 2009). In the same line of thinking, as the development of GPS and satellite observation is to remote sensing, the use of soil spectroscopy has been pivotal to proximal soil sensing.

The term spectroscopy groups all the related research involving the direct or indirect use of the different parts of the electromagnetic spectrum for describing soil variation (*See* Fig 1 *in* McBratney et al. (2003)). Likewise, the use of spectroscopy is not exclusive to field observations neither to soil science. A little bit of history tells us that the first recognized works in the field of spectroscopy date to 1648 in the book written by Marcus Marci of Kronland (1595 - 1667) entitled "*Thaumantias liber de arcu coelesti deque colorum apparentium natura…* " and the first researcher regarded to apply the word "*spectrum*" as it is used today was Isaac Newton, even though he was apparently influenced by the previously used term "*iris*" coined by Robert Boyle as reviewed by Burns (1987).

In the case of soil science, the first formal work using spectroscopy techniques as we see today was most probably the one of Bowers and Hanks (1965)[4], where they observed the influence of moisture, organic matter and particle size on the absorbance of the wavelengths ranging from 500 to 2500 nm (Figure 1-1).

[4] There are other previous works like the one of Carter (1931), however his work was not conclusive

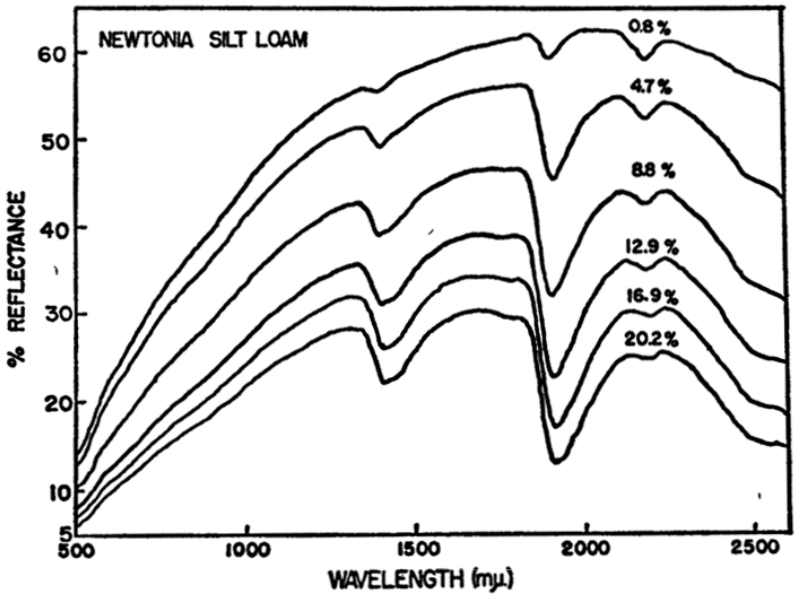


Figure 1-1 Percent reflectance vs. wavelength of incident radiation at various moisture contents (moisture contents indicated directly above each curve). *Extracted from* Bowers and Hanks (1965).

Later, further research using multivariate techniques (Krishnan et al., 1980; Dalal and Henry, 1986; Ben-Dor and Banin, 1995) led to what soil spectroscopy now mainly focuses upon, an alternative to traditional methods of wet chemistry to measure common soil properties e.g. organic carbon, clay percentage, pH, etc. (Nocita et al., 2015).

One of the most useful attributes of soil spectral information resides in its high dimensionality, allowing the prediction of multiple attributes at the same time from a single observation or spectrum (Ben-Dor and Banin, 1995). From all the regions of the electromagnetic spectrum currently under study, the one that has received more attention is that which includes the visible, middle and near infrared regions of the electromagnetic spectrum or those wavelengths ranging from 400 to 25.000 nm.

Chabrillat et al. (2013) described the state of the art methods in the previously mentioned area of soil spectroscopy, highlighting its use in prediction of carbon content (McDowell et al., 2012), the effect of raindrop energy on the water infiltration in relation to soil surface spectral reflectance (Goldshleger et al., 2012) and to the detection of total petroleum hydrocarbons in the soil (Schwartz et al., 2012). Naturally, and mainly because of its simplicity and speed in obtaining the measurements, the current spectroscopy studies have been rapidly moved to field conditions (Hedley et al., 2014; Ackerson et al., 2015).

Accordingly, and as noticed by Hartemink and Minasny (2014), the use of spectroscopic tools in the field has revolutionized the way soil scientists are perceiving and describing the soil profile, and it has occupied a place in a whole sub-discipline recently recognized as one of the newest IUSS working groups named “Digital soil morphometrics”. This reflects the continuous renewal of the soil science “toolbox”, by using new methods for describing the variation of the strategically important attributes of soil, which leads us to our second question, what information do we need?

1.3 Information that we need

Even though historically soil science has being divided into sub-disciplines i.e., pedology, soil physics, chemistry, biology, mineralogy and fertility among others, in a review of the key elements that define soil studies, Churchman (2010) identifies three aspects that can be considered as exclusive to the study of soils as a whole. These are i) the formation and properties of soil horizons, ii) the occurrence and properties of aggregates in soil and iii) the occurrence and behaviour of soil colloids.

These three soil science areas that may appear unconnected share one key element, they are the result of the interactions of multiple soil attributes and their study requires the analysis of many factors acting at the same time. Even more, they seem to relay on each other as Churchman (2010) also noticed: *"...Further, it is likely that improvements in descriptions of the aggregates could in turn help to contribute to improvements in characterisations of soil profiles and their constituent horizons, even if the most complete description of these latter is most likely to be given at a larger scale than that of the aggregates..."*

From here, it would be possible to generalize this idea and imply that the study of soil involves the study of interactions of multiple attributes at not only different scales but different "organization levels" as understood by Wimsatt (1994), where *"each level of organization is organized by part-whole relations, in which wholes at one level function as parts in the next (and all the higher) levels"*.

Following this line of thinking it can be hypothesized that, if we have an idea of the content of "idealized" soil colloids, let's say kaolinite and goethite in a soil sample, it will be possible to look for relations between those attributes, and with this, describing the process of aggregation. Further and as mentioned before, if the information related to those colloids and the interaction between them resulting in soil aggregates is identified, our understanding about soil horizons could be improved, since now we will have an idea of the mineralogical composition of soils and also of the process of aggregation which is known for direct influence with other more complex physical properties such as hydraulic conductivity, which in the end will determine soil horizonation processes.

And from here we can continue, if we consider that soil horizons are a prerequisite for classifying soil types e.g. soil orders, then all the information of the previous levels will govern the taxonomic output. A final step in this pyramid-like structure can be highlighted, this is the distribution of soil types in the landscape or the concept known as "Pedodiversity", briefly described as the systematic study of the richness of different soil types and their relative distribution in the landscape (McBratney, 1992; Huston, 1994; Ibáñez and Bockheim, 2013).

Based on the previous and under the assumption that pedological studies possess "levels" of organization, a simple hierarchical structure can be schematized in (**Figure 1-2**).

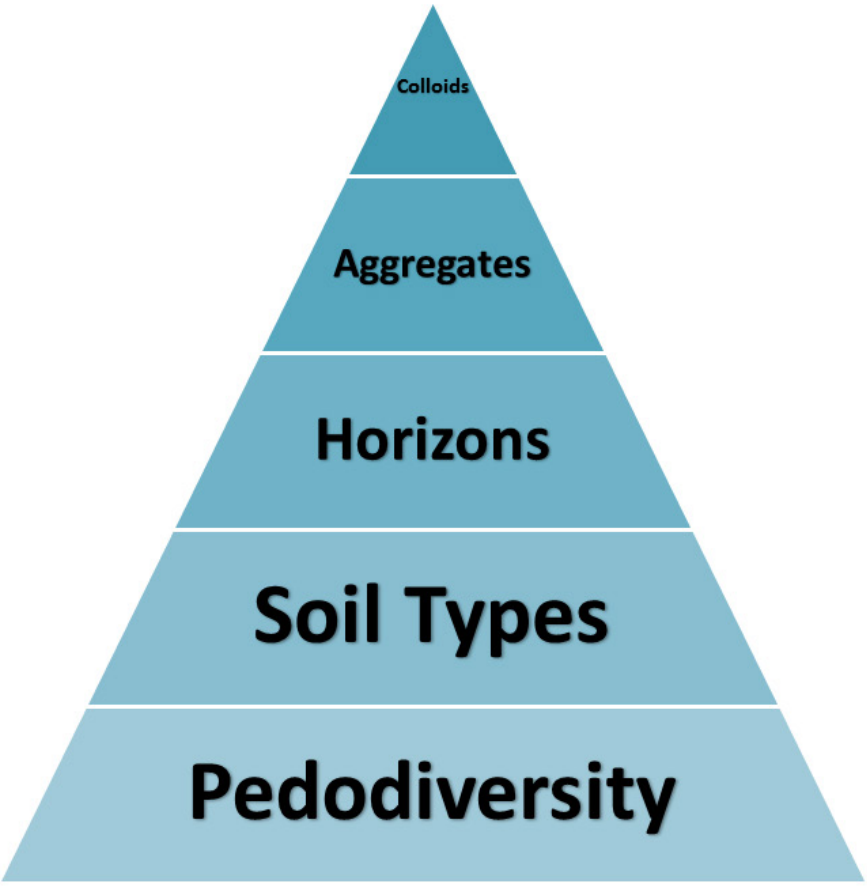


Figure 1-2 Theoretic hierarchical organization of pedological studies.

1.4 How can we get that information?

It has been showed that pedological studies are in general terms connected by a hierarchical thread, and each level in this structure relies on the preceding. From the previous it is possible to see that by obtaining information from each level, we are partially acquiring knowledge for the next level and so on, however, how can we relate the information from one level with the next one? To do this, the information needs to be expressed in the same format. As it was commented in earlier sections, an efficient way to express soil information is on a continuous scale i.e., quantitative information.

As commented earlier, soil spectroscopy can provide in fine detail, information related to the physico-chemical composition of soils in a continuous scale. Even though there are a number of studies that use a spectroscopic approach to assess the soil condition at the previously defined levels of soil organization (**Figure 1-2**), as seen in the next sections, the use of spectroscopic tools in most of the before mentioned areas is still considered as experimental.

1.4.1 Soil colloids

The area of soil colloids is most probably the one that has more research using spectroscopic related methodologies. The reason is because of the implicit need for computer based analyses to measure clay minerals or organic colloids due to their small size. Even more, the first studies involving formal spectroscopy methodologies applied to soil science were related with the occurrence of soil colloids detected by X-Ray diffraction, in the landscape and in depth (Wilson and Cradwick, 1972; Herbillon et al., 1981; Chipera and Bish, 2001).

The use of spectroscopy in these studies involved the detection of “peaks” in the spectra by analysing pure minerals and later detecting those peaks in another sample to finally asses its composition (Figure 1-3).

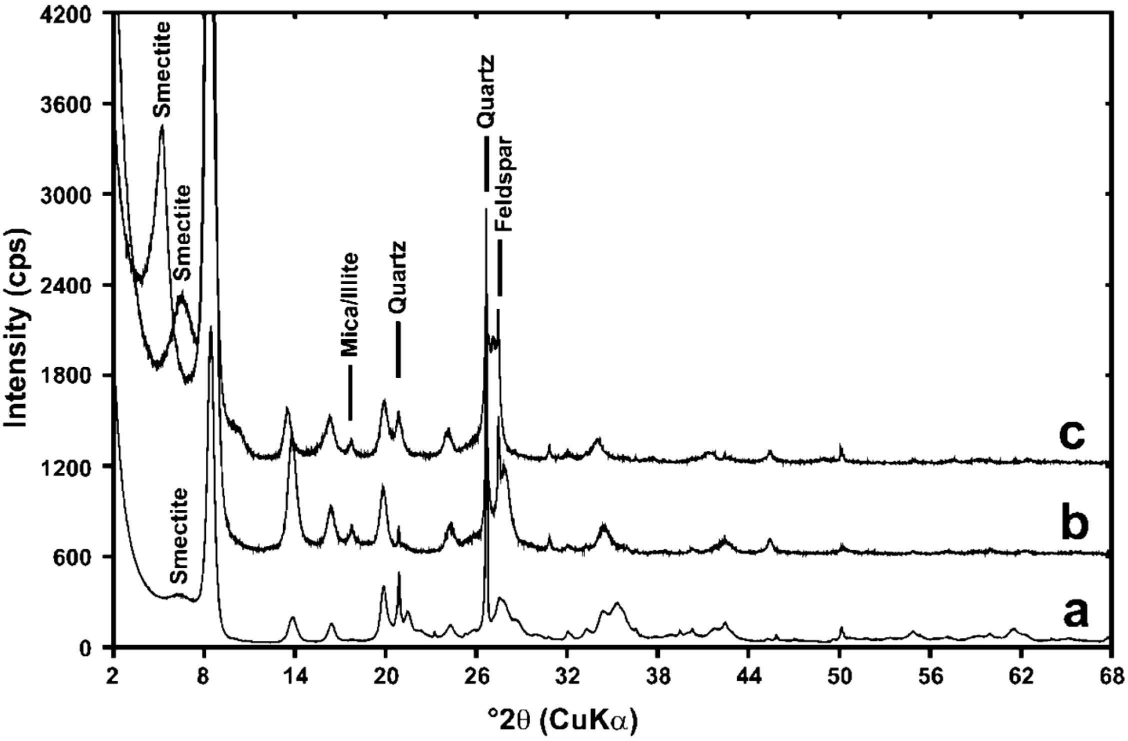


Figure 1-3 X-Ray diffraction patterns. *Extracted from* Chipera and Bish (2001).

The techniques in these former studies evolved with the application of current technologies (proximal and remote sensing) into continental studies like the one of Viscarra Rossel (2011) published as a map of the occurrence of clay minerals in Australia also using spectroscopic techniques. The recognition of idealized soil colloids in this study did not changed considerably from previous works, as the protocol for identifying colloids and their relative abundance, consisted in measuring the differences between reference clay minerals and soil samples reflectance values at selected wavelengths (Figure 1-4).

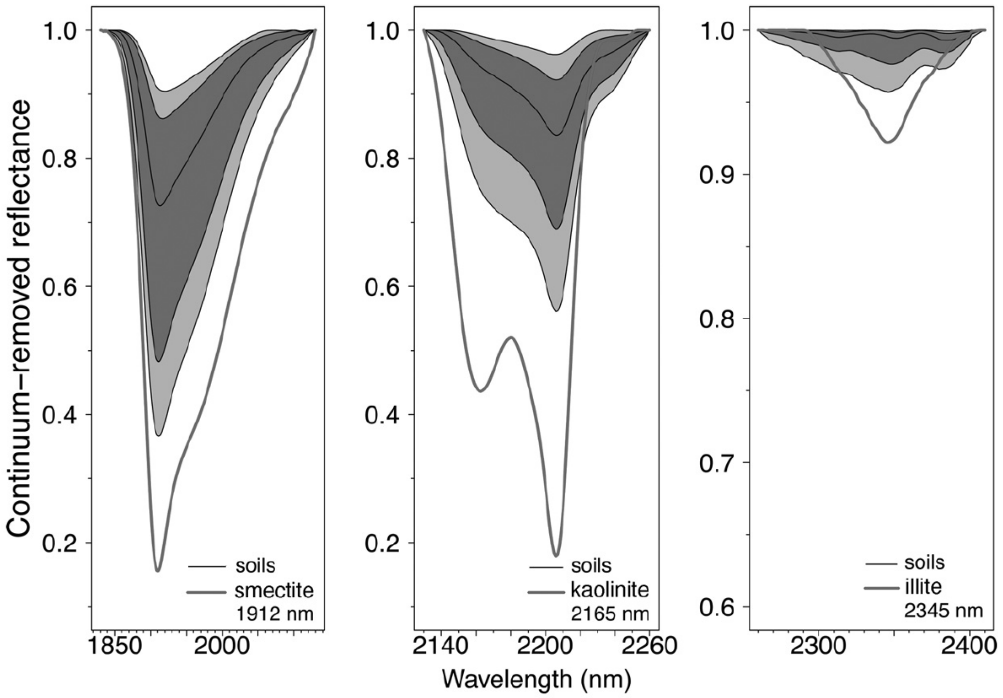


Figure 1-4 Diagnostic absorptions of (a) smectite (1912 nm), (b) kaolinite (2165 nm) and (c) illite (2345 nm). The plots show the spectrum of each of the reference clay minerals (thicker gray line) and the spectra of the 5th, 6th, 50th, 84th and 95th percentiles of the soil samples (thinner black lines). *Extracted from* Viscarra Rossel (2011).

1.4.2 Soil aggregation

As the complexity of the soil attribute increases, paradoxically the use of new spectroscopy techniques appears to be even less common. In the case of soil aggregation studies, only a few make use of new approaches, for example those involving soil micro aggregation and X-ray tomography (Moran et al., 1989; Ma et al., 2015). These types of methodologies use the absorbance values of a sample after the interaction with a light source within the X-ray part of the electromagnetic spectrum (~ 10 nm – 10 pm) to create images or “slices” of a soil aggregate repeatedly. This technique allows the creation of a pseudo-tridimensional image, where each image or “slice” is then analysed with an image segmentation algorithm to extract the areas and resultant volume of the soil porous system (Figure 1-5).

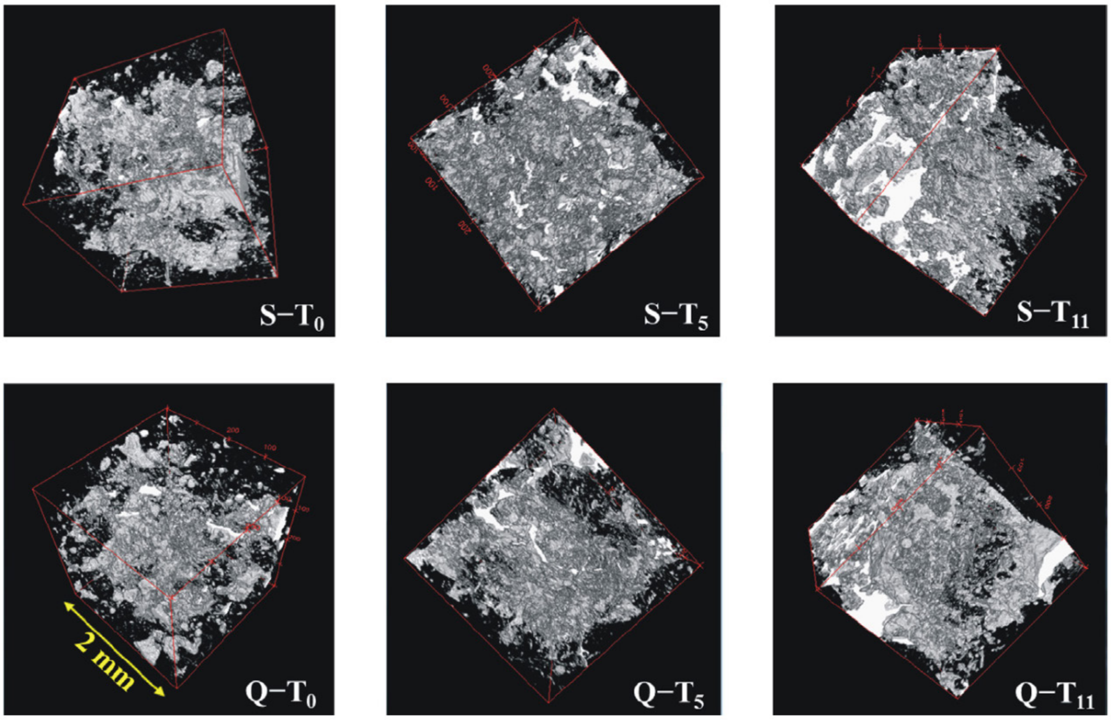


Figure 1-5 Pseudo three-dimensional pore structure of two soil samples at different wetting-drying cycles obtained from X-ray micro-tomography analysis. *Extracted from* Ma et al. (2015).

Despite the existence of more sophisticated methodologies and mainly because of their inherent cost, the commonly used methods for the characterization of the soil aggregation (mainly macroaggregation ~ >250 µm) date to the first part of the 20th century (Yoder, 1936). Just recently some works have explored the possibility of using Vis-NIR spectroscopy as indicators of soil aggregate stability (Cañasveras et al., 2010; Askari et al., 2015). These studies make use of the soil absorbance values within the Vis-NIR and MIR regions of the spectra and multivariate analyses like partial least squares regression, to predict soil aggregation coefficients measured by conventional methods i.e., (Yoder, 1936).

1.4.3 Soil morphological horizons

Soil morphological horizons and their identification in the profile takes place in the very centre of any pedological study, however, there are only a few works that apply new methodologies, other than the traditional methods of pedon morphological descriptions for their assessment e.g., Schoeneberger (2002) or Isbell (2002).

They can be grouped based on the aspect of soil horizonation in study, namely a) the systematic use of soil sensing methods for the allocation of different soil types in the landscape based on their horizonation and b) the identification of soil horizons in a pedon using proximal sensing techniques.

Some noteworthy examples for the first case are Chaplot et al. (2001) mapping hydromorphic horizons with a radio magnetotelluric resistivity instrument. They used the known relation between horizons of different compositions and their respective electric resistivity, and tested a radio magnotelluric-resistivity instrument to identify poorly drained horizons in the landscape.

Another example is the work of Odgers et al. (2011a) where they investigated the use of continuous soil “layers” created by using the properties predicted from their MIR spectra as analogues for the commonly described soil morphological horizons but with a continuous approach, where each observation can belong to a soil layer and at the same time, since the layers creation is based on a continuous scale, a relative distance to another layer can be calculated without the implicit bias produced by human-made soil descriptions.

For the second case, i.e., the identification of soil horizons in a pedon, Weindorf et al. (2012) studied the possibility of assessing horizonation with a portable X-ray device (Figure 1-6). They found a close relation between changes in concentrations of elements quantified using X-ray fluorescence with depth and the boundaries of soil morphological horizons.

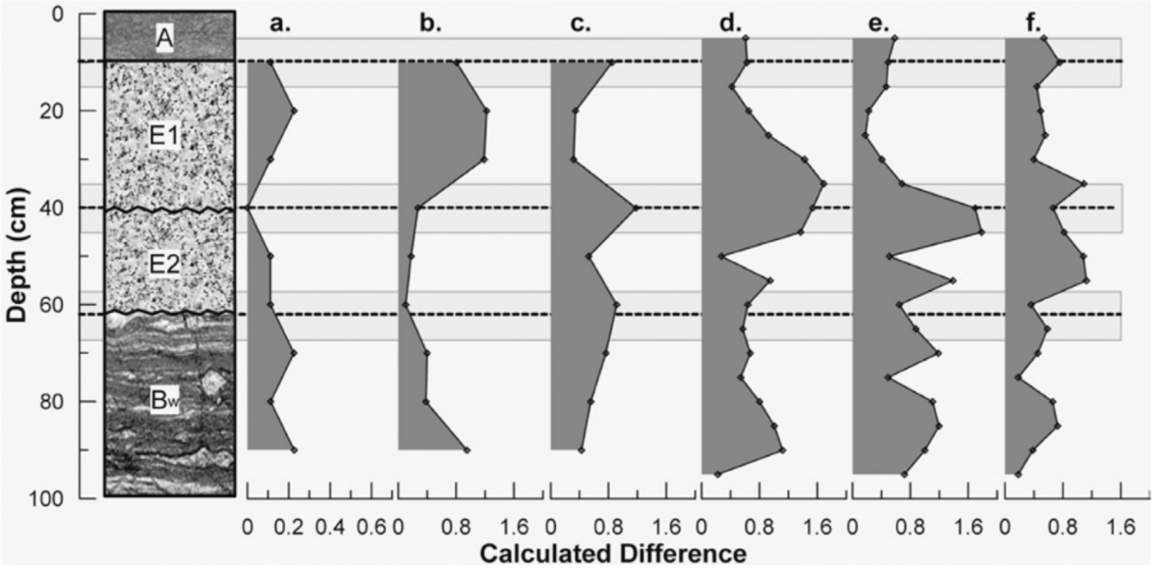


Figure 1-6 Field horizonation and calculated differences between layers for a pedon in Louisiana: (a) differences of clay, (b) differences of laboratory analysis, (c) elemental differences of portable X-ray fluorescence spectrometry (PXRF) readings on bulk density samples, (d) elemental differences of PXRF readings under field conditions, (e) elemental differences of PXRF readings on samples under laboratory conditions, and (f) elemental differences of PXRF readings on monolith. Extracted from Weindorf et al. (2012).

Similar works are provided by Ben-Dor et al. (2008) and Demattê et al. (2012), where they observed that commonly described soil morphological horizons had a distinctive "spectral signature" in the Vis-NIR range of the electromagnetic spectrum highlighting its possible use in discrimination and further classification.

1.4.4 Pedodiversity

So far it can be seen that there are studies that have already attempted to describe the variation observed on the different levels of soil organization by spectral means, however the complexity in the analysis has to increase as the scale and degree of organization increases e.g., soil colloids to soil horizons.

It has been shown that colloids relate with specific regions of the electromagnetic spectra and can be efficiently identified by systematically detecting peaks in the spectrum. Further, aggregates require a more sophisticated analysis and involve for example, multivariate calibrations which make use of several peaks and structural attributes that could be derived using spectral means. Soil horizons instead, require not only a multivariate approach like those studies that have attempted to predict soil aggregation parameters e.g., (Cañasveras et al., 2010), but the use of the multivariate information contained in a single observation plus the variation of that multivariate information over depth e.g., multiple samples analysed with a Vis-NIR spectra like those studies of Ben-Dor et al. (2008) or Odgers et al. (2011a).

This phenomenon can imply that a higher hierarchy level can be also analysed by spectral means following the same schema, which is by using the multivariate information of a single sample i.e., Vis-NIR spectrum from one scan; plus the variation from these measurements with depth i.e., soil horizon and finally, plus the variation in the landscape. Despite this, pedodiversity studies are the ones that have experienced the lowest use of spectroscopic techniques.

Based on the latest publications related with pedodiversity, it is worthy to note that they still depend on the discretization of the soil information i.e., Soil classes (McBratney and Minasny, 2007; Ibáñez and Bockheim, 2013; Costantini and L'Abate, 2015; Kooch et al., 2015). This may be one important reason for the lack of use of soil sensing instrumentation like spectroscopy, which intrinsically provides continuous information.

Toomanian and Esfandiarpoor (2010), in a review of the challenges in the area of pedodiversity, observed the connection that soil property variation and pedodiversity have, and the importance of considering multiple soil attributes simultaneously in order to describe the pedodiversity of soils in an efficient way.

Considering the great value of spectroscopic techniques as an input to multivariate soil information, it is easy to anticipate the great potential for its use in pedodiversity studies, however there is no literature related with the use of spectroscopy applied to pedodiversity so far.

Having briefly summarized the great value of the soil information that can be obtained by using spectroscopic techniques and also understanding the inherent hierarchic organization of soil, starting from the colloidal fraction to the organization of different soil types in the landscape, it is possible to hypothesize that the soil variation contained at different levels of organization i.e., soil colloids to pedodiversity, can be effectively described by spectral means.

Finally, the objectives of the present thesis can be summarized as follows:

1.5 Thesis objectives

- To develop and test a methodology to measure soil aggregate stability and to explore the possibility of predicting its values with Visible, Near and Middle Infrared spectroscopy.

- To test the use of large spectral libraries and state-of-the-art spectroscopy techniques for the prediction of soil properties at a local scale.

- To develop and test a numerical approach based on the use of Visible and Near Infrared spectroscopy for the identification of morphological horizons in soil profiles.

- To develop and test new pedodiversity indices based on continuous information and to use Visible and near infrared spectroscopy information for assessing the new pedodiversity indices at different scales.

Chapter references

Ackerson, J.P., Demattê, J.A.M., Morgan, C.L.S., 2015. Predicting clay content on field-moist intact tropical soils using a dried, ground VisNIR library with external parameter orthogonalization. Geoderma 259-260, 196-204.

Askari, M.S., Cui, J., O'Rourke, S.M., Holden, N.M., 2015. Evaluation of soil structural quality using VIS–NIR spectra. Soil and Tillage Research 146, 108-117.

Ben-Dor, E., Banin, A., 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Science Society of America Journal 59(2), 364-372.

Ben-Dor, E., Heller, D., Chudnovsky, A., 2008. A novel method of classifying soil profiles in the field using optical means. Soil Science Society of America Journal 72(4), 1113-1123.

Bouma, J., 2010. Implications of the Knowledge Paradox for Soil Science. 106, 143-171.

Bouma, J., McBratney, A., 2013. Framing soils as an actor when dealing with wicked environmental problems. Geoderma 200-201, 130-139.

Bowers, S.A., Hanks, R.J., 1965. REFLECTION OF RADIANT ENERGY FROM SOILS. Soil science 100(2), 130-130.

Bremner, J., 1960. Determination of nitrogen in soil by the Kjeldahl method. The Journal of Agricultural Science 55(01), 11-33.

Burns, D.T., 1987. Aspects of the Development of Colorimetric Analysis and Quantitative Molecular Spectroscopy in the Ultraviolet-Visible Region, Advances in Standards and Methodology in Spectrophotometry. Elsevier Amsterdam.

Cañasveras, J.C., Barrón, V., del Campillo, M.C., Torrent, J., Gómez, J.A., 2010. Estimation of aggregate stability indices in Mediterranean soils by diffuse reflectance spectroscopy. Geoderma 158(1-2), 78-84.

Carter, W., 1931. Color analysis of soils with spectrophotometer. Soil Science Society of America Journal 12(2001), 169-170.

Chabrillat, S., Ben-Dor, E., Rossel, R.A.V., Demattê, J.A.M., 2013. Quantitative Soil Spectroscopy. Applied and Environmental Soil Science 2013, 1-3.

Chaplot, V., Walter, C., Curmi, P., Hollier-Larousse, A., 2001. Mapping field-scale hydromorphic horizons using Radio-MT electrical resistivity. Geoderma 102(1–2), 61-74.

Chipera, S.J., Bish, D.L., 2001. Baseline studies of the clay minerals society source clays: powder X-ray diffraction analyses. Clays and Clay Minerals 49(5), 398-409.

Churchman, G.J., 2010. The philosophical status of soil science. Geoderma 157(3-4), 214-221.

Costantini, E.A.C., L'Abate, G., 2015. Beyond the concept of dominant soil: Preserving pedodiversity in upscaling soil maps. Geoderma.

Dalal, R.C., Henry, R.J., 1986. Simultaneous Determination of Moisture, Organic Carbon, and Total Nitrogen by Near Infrared Reflectance Spectrophotometry. Soil Science Society of America Journal 50(1), 120-123.

Demattê, J.A.M., Terra, F.d.S., Quartaroli, C.F., 2012. Spectral behavior of some modal soil profiles from São Paulo State, Brazil. Bragantia 71, 413-423.

Goldshleger, N., Chudnovsky, A., Ben-Dor, E., 2012. Using Reflectance Spectroscopy and Artificial Neural Network to Assess Water Infiltration Rate into the Soil Profile. Applied and Environmental Soil Science 2012, 1-9.

Hartemink, A.E., McBratney, A., 2008. A soil science renaissance. Geoderma 148(2), 123-129.

Hartemink, A.E., Minasny, B., 2014. Towards digital soil morphometrics. Geoderma 230, 305-317.

Hedley, C., Roudier, P., Maddi, L., 2014. VNIR Soil Spectroscopy for Field Soil Analysis. Communications in Soil Science and Plant Analysis 46(sup1), 104-121.

Herbillon, A., Frankart, R., Vielvoye, L., 1981. An occurrence of interstratified kaolinite-smectite minerals in a red-black soil toposequence. Clay Minerals 16(2), 195-201.

Huston, M.A., 1994. Biological diversity: the coexistence of species on changing landscapes. Cambridge University Press.

Ibáñez, J.J., Bockheim, J.G., 2013. Pedodiversity. CRC Press.

Isbell, R.F., 2002. The Australian soil classification / R.F. Isbell. Australian soil and land survey handbook ; vol. 4. CSIRO Publishing, Collingwood, Vic. :.

Jensen, J.R., 2009. Remote sensing of the environment: An earth resource perspective 2/e. Pearson Education India.

Kondratiev, N.D., 1925. The major economic cycles. Voprosy Konjunktury 1(1), 28-79.

Kooch, Y., Hosseini, S.M., Scharenbroch, B.C., Hojjati, S.M., Mohammadi, J., 2015. Pedodiversity in the Caspian forests of Iran. Geoderma Regional 5, 4-14.

Krishnan, P., Alexander, J.D., Butler, B.J., Hummel, J.W., 1980. Reflectance Technique for Predicting Soil Organic Matter. Soil Science Society of America Journal 44(6), 1282-1285.

Ma, R., Cai, C., Li, Z., Wang, J., Xiao, T., Peng, G., Yang, W., 2015. Evaluation of soil aggregate microstructure and stability under wetting and drying cycles in two Ultisols using synchrotron-based X-ray micro-computed tomography. Soil and Tillage Research 149, 1-11.

Matheron, G., 1965. Les variables régionalisées et leur estimation une application de la théorie des fonctions aléatoires aux sciences de la nature. Masson, Paris.

McBratney, A., 1992. On variation, uncertainty and informatics in environmental soil management. Soil Research 30(6), 913-935.

McBratney, A., Minasny, B., 2007. On measuring pedodiversity. Geoderma 141(1–2), 149-154.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117(1-2), 3-52.

McDowell, M.L., Bruland, G.L., Deenik, J.L., Grunwald, S., 2012. Effects of Subsetting by Carbon Content, Soil Order, and Spectral Classification on Prediction of Soil Total Carbon with Diffuse Reflectance Spectroscopy. Applied and Environmental Soil Science 2012, 14.

Mercer, W., Hall, A., 1911. The experimental error of field trials. The Journal of Agricultural Science 4(02), 107-132.

Moran, C.J., McBratney, A.B., Koppi, A.J., 1989. A Rapid Method for Analysis of Soil Macropore Structure. I. Specimen Preparation and Digital Binary Image Production. Soil Science Society of America Journal 53(3).

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthes, B., Dor, E.B., Brown, D., Clairotte, M., Csorba, A., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67(3–4), 215-226.

Odgers, N.P., McBratney, A.B., Minasny, B., 2011a. Bottom-up digital soil mapping. I. Soil layer classes. Geoderma 163(1–2), 38-44.

Omuto, C., Nachtergaele, F., Rojas, R.V., 2013. State of the Art Report on Global and regional Soil Information: Where are we? Where to go? Food and Agriculture Organization of the United Nations.

Schoeneberger, P.J., 2002. Field book for describing and sampling soils, Version 3.0. Government Printing Office.

Schwartz, G., Ben-Dor, E., Eshel, G., 2012. Quantitative Analysis of Total Petroleum Hydrocarbons in Soils: Comparison between Reflectance Spectroscopy and Solvent Extraction by 3 Certified Laboratories. Applied and Environmental Soil Science 2012, 11.

So, H.B., Lal, R., 2002. Encyclopedia of Soil Science. Taylor & Francis.

Tinker, P., 1985. Soil science in a changing world. Journal of Soil Science 36(1), 1-8.

Toomanian, N., Esfandiarpoor, I., 2010. Challenges of pedodiversity in soil science. Eurasian Soil Science 43(13), 1486-1502.

United Nations, 2014. Resolution A/RES/68/232 adopted by the General Assembly on the report of the Second Committee (A/68/444) UN.

Van Bavel, C., 1950. Mean weight-diameter of soil aggregates as a statistical index of aggregation. Soil Science Society of America Journal 14(C), 20-23.

Viscarra Rossel, R.A., 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. Journal of Geophysical Research 116(F4).

Webster, R., 1994. The development of pedometrics. Geoderma 62(1), 1-15.

Weindorf, D.C., Zhu, Y., Haggard, B., Lofton, J., Chakraborty, S., Bakr, N., Zhang, W., Weindorf, W.C., Legoria, M., 2012. Enhanced Pedon Horizonation Using Portable X-ray Fluorescence Spectrometry. Soil Sci. Soc. Am. J. 76(2), 522-531.

Wilson, M.J., Cradwick, P.D., 1972. Occurrence of interstratified kaolinite-montmorillonite in some Scottish soils. Clay Minerals 9(435-437).

Wimsatt, W.C., 1994. The ontology of complex systems: levels of organization, perspectives, and causal thickets. Canadian Journal of Philosophy 24(sup1), 207-274.

Yoder, R.E., 1936. A direct method of aggregate analysis of soils and a study of the physical nature of erosion losses. Agronomy Journal 28(5), 337-351.

Chapter 2 *Soil Aggregate Stability*

Part 1: Soil slaking assessment using image recognition

"Mater artium necessitas"

Chapter 2, Part 1 published as:

Fajardo, M., McBratney, A.B., Field, D.J., Minasny, B., 2016. Soil slaking assessment using image recognition. Soil and Tillage Research. 163,119-129.

2.1 Summary

We have developed a new methodology for the assessment of soil slaking in fast wetting conditions. We applied an image recognition algorithm to a set of digital images of soil aggregates immersed in water taken at regular intervals. The kinetics of the slaking process were captured by measuring the projected area change over time. The methodology was tested in a dataset covering a great part of the agro-ecological variability of New South Wales (NSW), Australia. An empirical model which captures the rapid and slow slaking process was fitted to the data and three new slaking coefficients (*a*, *b* and *c*) were obtained and related to selected soil properties and land-use. The coefficient *a*, equivalent with the maximum slaking potential of the samples, was linearly related to exchangeable sodium, pH, clay percentage, calcium/magnesium and total carbon/nitrogen, and non-linearly related with total carbon. The coefficients *b* and *c*, equivalent with the initial slaking and the rate of change respectively, were found to be linearly related to nitrogen and total carbon. The coefficient *a*, was significantly lower in the natural sites reflecting a higher aggregate stability in those soils. The methodology is fast, inexpensive and simple; furthermore, it provides a new perspective in soil aggregate stability experiments, since it considers the slaking dynamics during the entire disaggregation process.

2.2 Introduction

An aggregate is a group of primary particles that cohere to each other and their stability is a function of the binding forces between them that withstand after an applied disruptive event (Kemper and Rosenau, 1986). As reviewed by many authors e.g., (Tisdall and Oades, 1982; Kemper and Rosenau, 1986; Field et al., 1997; Amézketa, 1999; Dıaz-Zorita et al., 2002; Bronick and Lal, 2005), aggregate stability is an important soil physical property which influences a wide range of biological and chemical processes in natural and agricultural environments. The stability of aggregates in water is an essential property for maintaining soil productivity by helping in minimizing soil erosion, environmental pollution and further degradation.

Notwithstanding its importance, many soil studies do not consider the analysis of this physical property, due to the time needed for the analysis, the instrumentation specificity used in each observation and the lack of accessible (in terms of cost and/or availability) private services to outsource its evaluation.

There are many methods that can be used for measuring soil stability in water, namely a) end-over-end shaking methods (Middleton, 1930; Quirk, 1950), b) wet sieving techniques (Yoder, 1936; Perfect et al., 1992), c) raindrop impact or rainfall simulation techniques (Young, 1984), d) immersion methods (Loveday and Pyle, 1973; Field et al., 1997) and e) ultrasonic dispersion (Edwards and Bremner, 1967).

Among these previously mentioned, the usually preferred method to measure and present the results of aggregate stability in water, is in the form of a Mean Weight Diameter

(MWD) after wet sieving of soil aggregates (Van Bavel, 1950; Youker and McGuinness, 1957; Henin et al., 1958; Kemper and Rosenau, 1986; Le Bissonnais, 1996). This method is simple and intuitive, the basic idea behind it is that resultant bigger aggregates after sieving are related with greater stability (Nimmo and Perkins, 2002), however it requires specialized equipment e.g., wet sieving apparatus, and the time required in the analysis is a limiting factor in processing a large dataset, due to the need of post-processing i.e., oven-drying and weighting of resultant soil fractions.

On the other hand, a more practical and faster approach developed specifically for in-field assessment of sodicity problems, is the immersion method presented by Field et al. (1997) called ASWAT (Aggregate Stability in Water). This method uses a qualitative scale, or ASWAT scores, to rank the degree of slaking/dispersion of a sample, with 0 equivalent to non-disruption and 16 to full dispersion.

Inspired by the simplicity of this method in terms of reduced sample preparation and fast results, we foresaw the possibility to automate it by using an image recognition algorithm and by doing this, providing a continuous index which reflects the slaking area increase after the immersion of soil aggregates in water. There is extensive literature related to the applicability of image recognition procedures applied to soil structure studies and it has been shown that they are especially useful in terms of accuracy, speed and replicability e.g., (O'Callaghan and Loveday, 1973; Ringrose-Voase and Bullock, 1984; Moran et al., 1989; Vogel, 1997; Pagliai et al., 2004).

This newly developed methodology measures the projected area of soil aggregates while disaggregating in water over time, using a digital camera coupled with an image recognition algorithm. The benefits of this new method reside in its simplicity, objectivity, continuous scale and its ability to visualize the dynamics of disaggregation.

As a part of major project which is exploring the relationships among soil properties and microbial communities at a large scale (State of NSW, Australia), the need for a fast and simple methodology for assessing soil structure stability on a continuous scale was a key element in relating soil biological and physical properties, therefore, we developed and employed the methodology on the before mentioned dataset which covers a great part of the agro-ecological variability in NSW, Australia.

The aim of the study is a) to describe the method development and performance under a considerable range of soil conditions and b) to explore its relations with basic chemical and physical properties in order to characterize their influences over soil aggregate stability.

2.3 Materials and Methods

2.3.1 Dataset

The area of study is located in New South Wales (NSW), Australia, following a survey of two transects. The first transect encompasses 27 paired sampled sites in native vegetation and nearby agricultural sites separated by a distance of approximately 50 km (see Appendix 1) at two depths, 0 to 5 and 5 to 10 cm. The transect covers a total length of 900 km from North to South following the same precipitation regime (550 mm isohyet) from the border with the state of Queensland in the north, to the border with the state of Victoria in the south (Figure 2-1).

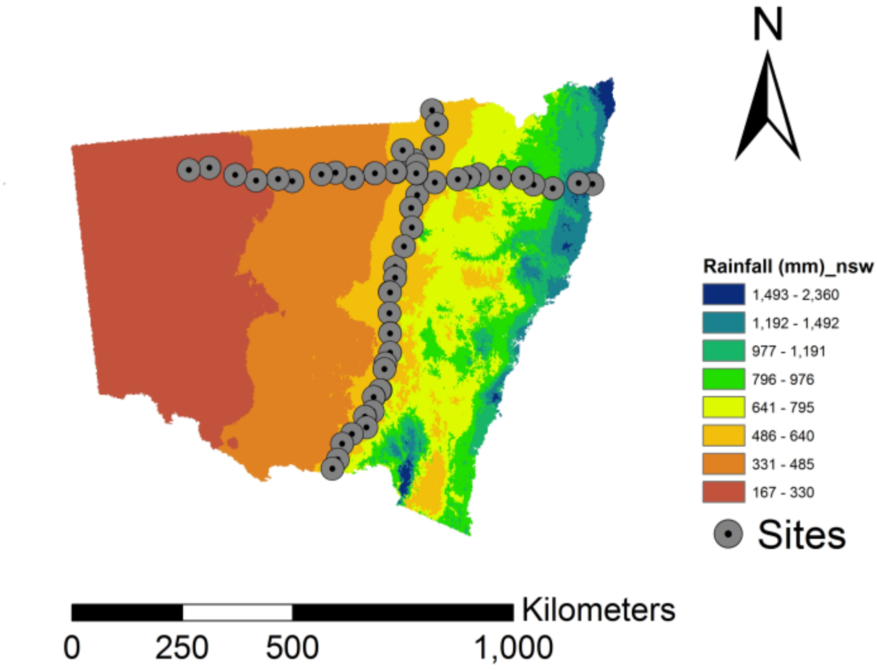


Figure 2-1 Area of study, NSW, Australia.

The second transect comprises 22 paired sites with the same sampling protocol in a 930 km long East-West transect perpendicular to the North-South transect from the city of Coffs Harbour on the east coast, to the town of Wanaaring in western NSW (rainfall gradient from >1500mm to <300mm) (Figure 2-1).

2.3.2 Sample preparation and complementary analyses

Basic soil chemical analyses were performed on all samples: Total Carbon content (%TC) and Nitrogen content (%N) by dry combustion (LECO instrument, CSBP laboratory Ltd., Western Australia), Extractable Phosphorus (P) and Potassium (K) expressed in mg kg^{-1} (Colwell, 1963), Ammonium (NH_4) and Nitrate (NO_3) expressed in mg kg^{-1} (Rayment and Higginson, 1992), Exchangeable cations (Ca, Mg, Na, K, Al) expressed in cmol(+) kg^{-1} (Rayment and Higginson, 1992), Cation Exchange Capacity (CEC) expressed in cmol(+) kg^{-1} (Rayment and Higginson, 1992), Electric conductivity expressed in dS m^{-1}, pH in water and particle size analysis expressed in % using the hydrometer method (Bouyoucos, 1962), all the analyses excepting particle size analysis were performed by CSBP Laboratory Ltd., Western Australia.

2.3.3 Digital single lens reflex (DSLR) camera images

Soil aggregates were air-dried for three days and a set of 5 aggregates ranging between 3 to 10 mm diameter from each site were placed in an empty Petri dish over a white workbench where a first picture was taken (Figure 2-2). The aggregates were removed and the Petri dish was filled with water followed by the immersion of the five soil aggregates at the same time.

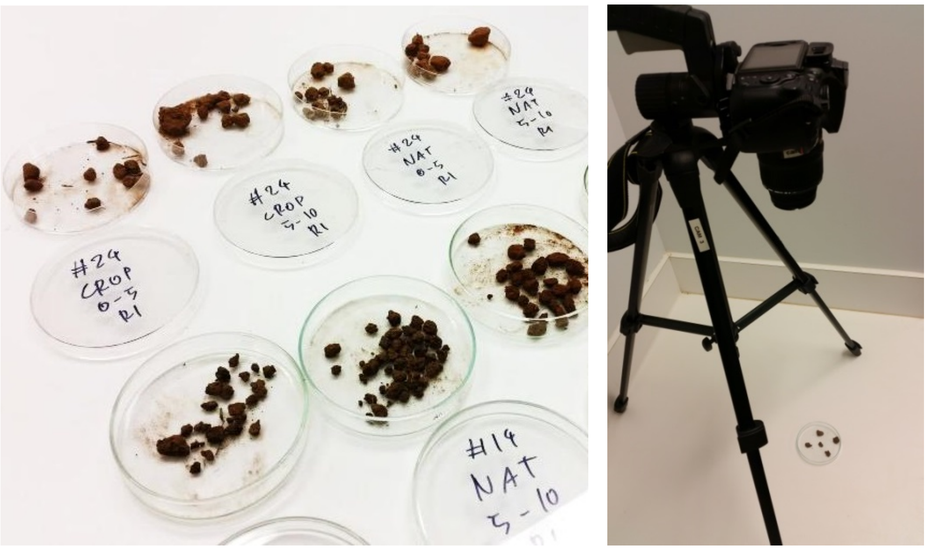


Figure 2-2 Aggregates and camera setup.

A Digital single lens reflex (DSLR) camera was set at a vertical distance of approximately 30 cm. The study was performed in a laboratory with halogen lights and no extra illumination was added e.g., camera flash. A set of pictures was taken every second from the first contact with water until 120 seconds elapsed, then a set of 12 pictures was taken every 20 seconds and finally 12 more pictures were taken at an interval of 10 minutes until 2 hours approx. (117 min) making a total of 145 pictures per sample (Figure 2-3 and Figure 2-4).

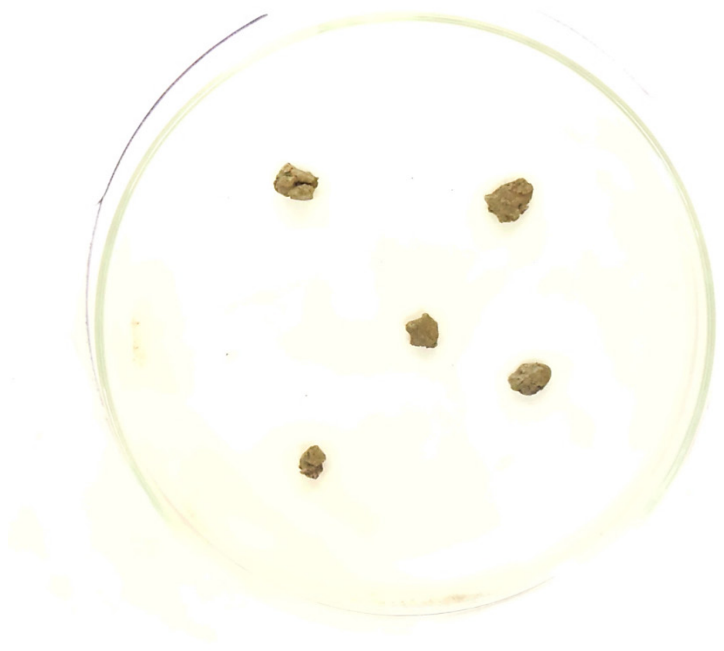


Figure 2-3 Aggregates before water immersion; *Note the high brightness of the picture is due to high exposure settings.*

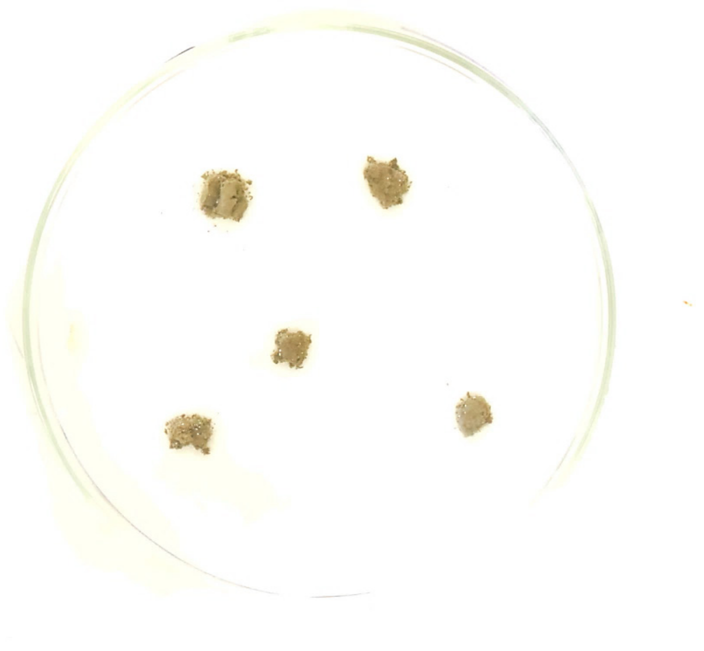


Figure 2-4 Aggregates after 5 seconds of water immersion.

Each picture was set to a high exposure level in order to reduce image noise (e.g., eventual floating particles in water or soil aggregates shade). The camera model and settings are specified in Table 2-1.

Table 2-1 DSLR camera settings.

Setting	***Value***
Camera	*model NIKON D5300*
Dimensions	*2992 x 2000 pixels*
ISO-speed	*640*
Exposure bias	*+4 step*
Focal length	*46 mm*
Exposure time 1/250 sec.	*1/250 sec.*
Dots per inch (DPI)	*300*
Bit depth	*24*
F-stop	*f/8*

2.3.4 Image processing and segmentation procedure

An example with the image recognition R code is presented in Appendix 2 . Each of the 145 images was imported into R (R Core Team, 2013) using function "*readImage"* of the R package *EBImage* (Pau et al., 2014). The resulting object is a 2 slots S4 object, the first one an array of dim1 x dim2 x *RGB* bands where dim1 and 2 are the dimensions of the picture and *RGB* bands are the respective values ranging from 0 to 1 for each pixel in a coordinate $RGB_{dim1,\mathrm{dim2}}$. The second slot contains the respective "*colormode"* for internal use of the *EBImage* package.

Out of the three colour bands (Figure 2-5, Figure 2-6 and Figure 2-7), the blue band was selected for the next process of segmentation and identification of each aggregate after testing the segmentation.

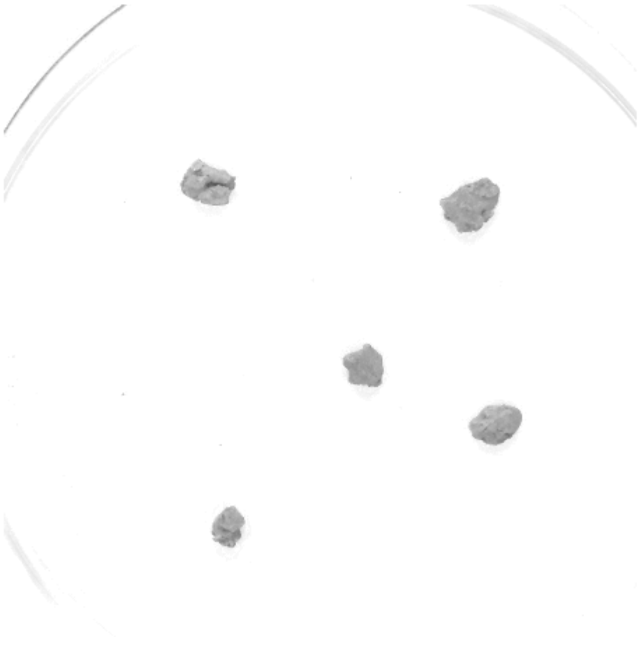


Figure 2-5 Red band of soil aggregates image.

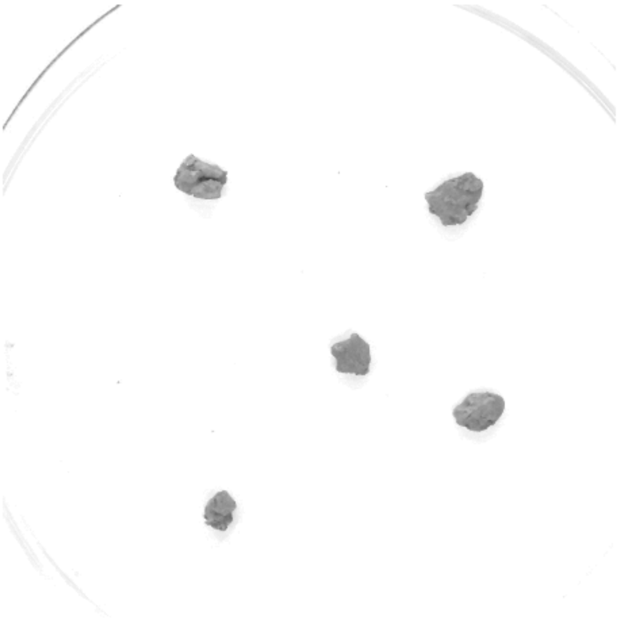


Figure 2-6 Green band of soil aggregates image.

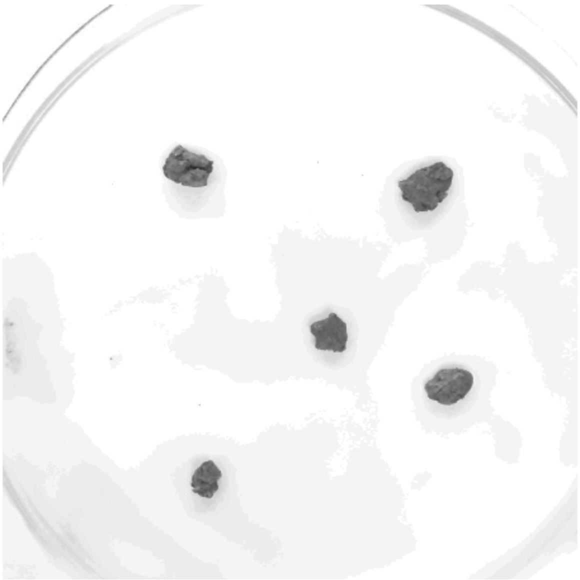


Figure 2-7 Blue band of soil aggregates image.

A binary matrix was created where a value of 1 was assigned to blue band values higher than 0.9 and 0 to the rest; this threshold was established empirically based on visual inspection of soil aggregates discrimination from the background after testing the segmentation procedure for red, green and blue bands with different thresholds. It is important to note that other bands and threshold values can also be used. The band and threshold used in this study were set, based on the variability of our dataset and it may be affected by the colour of the soil aggregates, in our case the Munsell colours ranged from Black (N 1/0) to Greyish red (10R 4/2) with most of the samples with Brown colours with Hue values between 10YR to 5YR, Chromas of 2 to 4 and Values of 4 to 5.

Finally, a two-step image filter was applied using the function "*closing*" which involves a dilation followed by an erosion filter, removing the eventual background values inside the aggregate, usually produced by air bubbles in the sample (Figure 2-8).

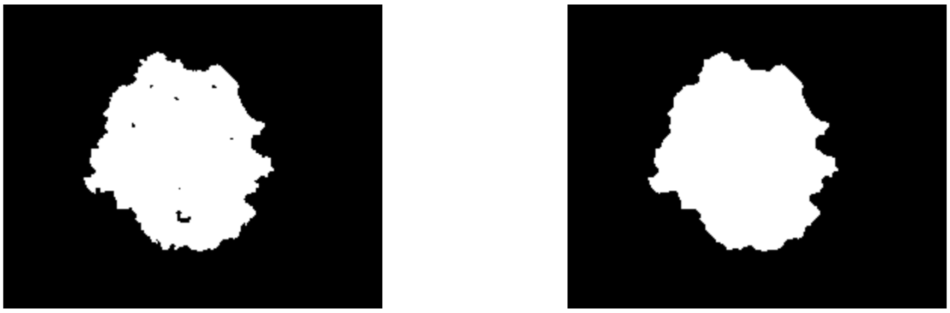


Figure 2-8 Detail of image processing of a single aggregate.

Once the binary matrices were produced, the function "*bwlabel*" was employed which applies a segmentation algorithm implemented by Pau et al. (2014) and assigns an identification number to each group of pixels with same adjacent values, in this way, every group of pixels was labelled with a single number. We selected those 5 labelled objects which contain the maximum amount of pixels (i.e., 5 soil aggregates). In this way, every other smaller labelled object was assigned to a value of 0 i.e. background (Figure 2-9).

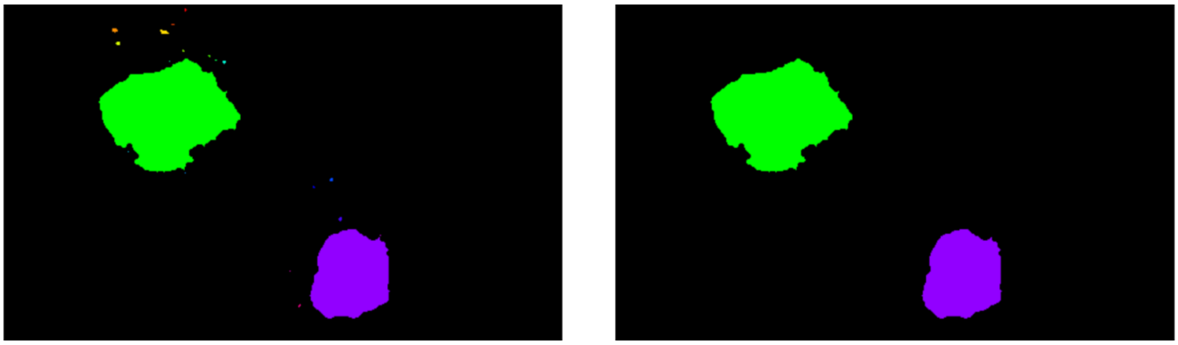


Figure 2-9 Detail of a two aggregates image segmentation process, different colours represent different labels.

Finally, the function "*computeFeatures.shape*" was applied to calculate the area, perimeter, mean radius, standard deviation of the mean radius, the maximum and minimum radius of the 5 labelled objects (in pixels).

2.3.5 Statistical analyses

Using the area values for each of the five aggregates at time 0 (initial size) and the areas for the subsequent time frames, a Slaking Index (SI) was calculated as follows (Equation 1):

$$SI_t = \frac{A_t - A_{t0}}{A_{t0}} \tag{1}$$

where A_t equals the projected area at time *t*, using this approach SI_{t0} will be always 0, and SI_t = 1, means that the aggregate area at *t* is twice as big as at t_0. The *SI* value for each site was calculated as the arithmetic mean of the five aggregate *SI.*

Of the 196 Site-depth-system combinations, 148 were analysed i.e. 37 out of 49 sites. The reason for not measuring the index at the remaining 12 sites was either a lack of stable aggregates (e.g., too sandy) or the hydrophobicity of aggregates due to high organic matter content.

2.4 Results and Discussion

2.4.1 Segmentation performance

The segmentation and shape analyses were automated for each of the 145 images for each soil sample (a total of 148 samples). The computer processing time per sample was approximately 5 to 6 minutes using an 8 cores computer (i7-2600 CPU @ 3.40 GHz) with 5 to 6 GB of physical memory used in the process depending on the size and shapes of the aggregates. It is important to mention that the resolution of the images impacts directly in the total time of analysis, and can be a limiting factor in images processing (see Table 2-1). After the image recognition, it was observed that the disaggregation process was very homogeneous within samples i.e., the behaviour of 5 different aggregates was similar within a sample, even with different sized soil aggregates (Figure 2-10). This was reflected in the low standard deviation of *SI* values in relation with traditional soil aggregate stability methods (Pulido Moncada et al., 2013) (Figure 2-11).

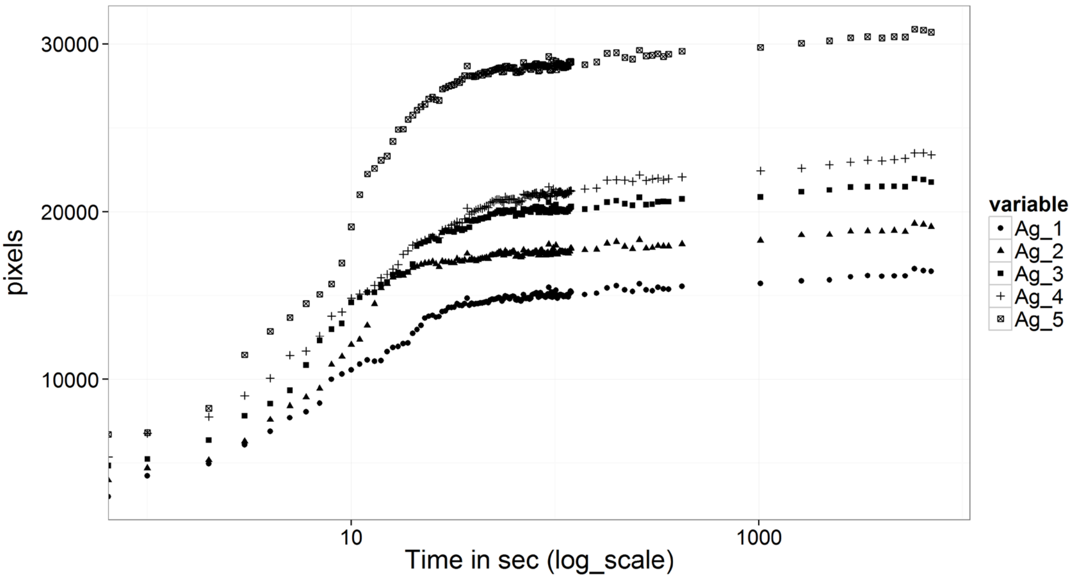


Figure 2-10 Area (pixels) of 5 aggregates as a function of time.

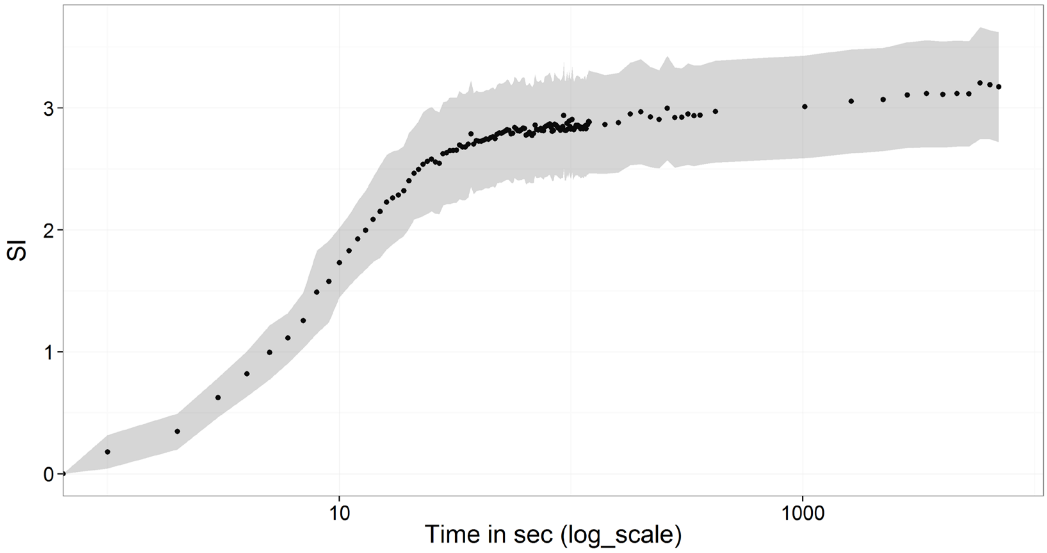


Figure 2-11 Slaking index of soil aggregates with time of a single soil sample. The dots represent the mean value, and the grey area represents the envelope within ± 1standard deviation.

2.4.2 Soil disaggregation dynamics

One of the advantages of the method was the ability to observe the dynamics of the disaggregation process (Figure 2-11). As observed by Yoder (1936) the initial process of trapped air release was evident during the first 20 seconds, contributing significantly to the final *SI* values. This same pattern was identified in all the samples, whereby a first rapid increase in area was followed by an almost linear increment in *SI* values.

The few studies that have modelled the dynamics of the disaggregation process as a function of time have used an exponential function (Russell and Feng, 1947; Zanini et al., 1998; Braudeau and Mohtar, 2006). At first examination, the dynamics seemed to follow this exponential behaviour; however the exponential function did not model the dynamics satisfactorily, as we observed 2 types of kinetics with time, one faster at the beginning and one slower at longer times. Due to this, a model with 3 parameters (*a*, *b* and *c*) was employed based on the theoretical hierarchical organization of aggregates (Oades and Waters, 1991) and in the observed disaggregation process, as follows (Equation 2) :

$$\boldsymbol{Si_t = \; a\, e^{\left(-b\, e^{\left(-c\, log(t)\right)}\right)}} \tag{2}$$

where the SI value measured on a log scale time is equal to a log-scaled Gompertz function (Gompertz, 1825), with an initial fast area increase due to bigger aggregates imbibing water, followed by a slower area increase due to medium and smaller aggregates disaggregation. The parameter *b* represents the initial time of fast slaking and *a* the asymptotic *SI* value or maximum possible slaking, which is conditioned ultimately by the smaller aggregates, elemental particles and the conditions of the medium e.g., electrical conductivity of water. The parameter *c* is the increase rate, which is highly correlated with parameter *b* (Figure 2-12 and Figure 2-13).

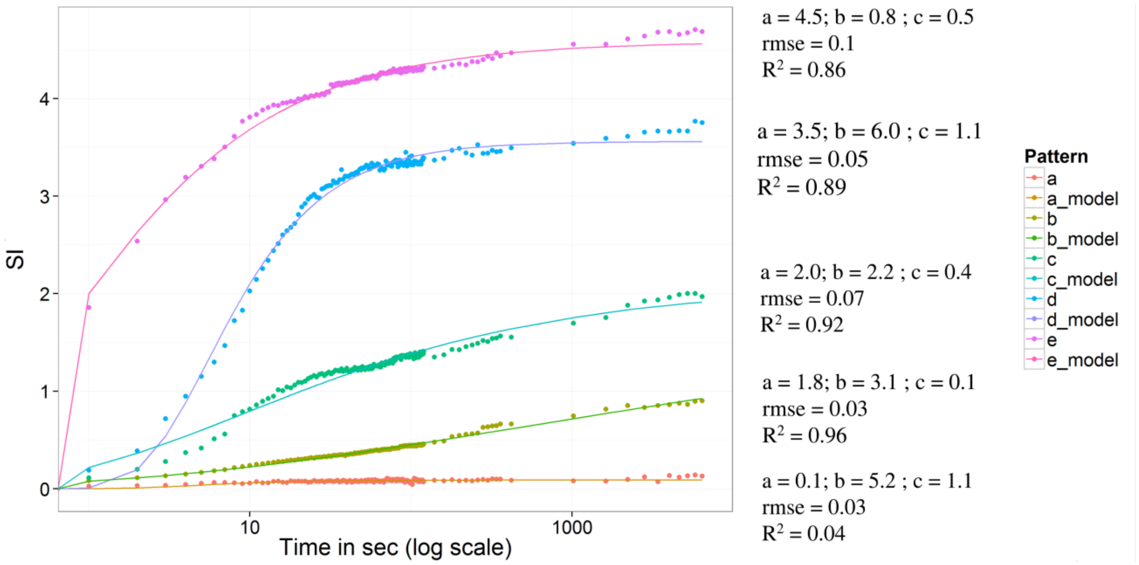


Figure 2-12 Dynamics of contrasting disaggregation patterns and the respective Gompertz function model.

The advantage of a Gompertz function over an exponential function approach is that it models two different processes i.e., rapid and gradual disaggregation. The function was fitted to all the samples based on the non-linear least-squares algorithm. It was observed that low slaking samples had low R^2 values, this was mainly due to the known influence of the range of values over the R^2 parameter (Davies and Fearn, 2006). Nevertheless, it was observed that the Gompertz function fitted all data quite well. The fitted Gompertz parameters were then related to the measured soil properties as described in the following section.

2.4.3 Relationships with soil properties

Le Bissonnais (1996) presented the factors that must be considered by any soil aggregate stability test. These factors are the ones usually occurring in natural processes such as rain events or soil saturation-drying cycles. They are breakdown caused by air release during initial wetting, by differential swelling, by raindrops impact and by physicochemical dispersion due to osmotic stress. This new methodology allowed clear identification of the initial fast disaggregation due to the release of entrapped-air, and it was observed that this behaviour was more significant in some samples while in others it did not occur or was too small to be measured.

In order to identify the relationship between the *SI* parameters and the measured soil physico-chemical properties (Table 2-2), a principal component analysis was performed (Figure 2-13).

Table 2-2 Measured soil properties of the dataset.

Property	Min	1st Quartile.	Median	Mean	3rd Quartile.	Max
NH4 (mg kg^{-1})	*0.1*	*2.0*	*4.0*	*5.9*	*7.0*	*53.0*
NO$_3$ (mg kg^{-1})	*1.0*	*3.0*	*11.0*	*22.4*	*24.5*	*202.0*
P (mg kg^{-1})	*3.0*	*15.5*	*30.0*	*39.0*	*48.5*	*273.0*
K (mg kg^{-1})	*74.0*	*310.0*	*450.0*	*475.1*	*572.5*	*1232.0*
EC (dS m^{-1})	*0.0*	*0.0*	*0.1*	*0.1*	*0.1*	*0.5*
pH H$_2$O	*4.8*	*6.0*	*6.6*	*6.7*	*7.5*	*8.6*
Exc. Al (cmol(+) kg^{-1})	*0.0*	*0.0*	*0.1*	*0.1*	*0.1*	*1.8*
Exc. Ca (cmol(+) kg^{-1})	*0.3*	*5.0*	*9.5*	*11.1*	*17.0*	*30.6*
Exc. Mg (cmol(+) kg^{-1})	*0.3*	*1.5*	*2.9*	*4.2*	*6.8*	*13.1*
Exc. K (cmol(+) kg^{-1})	*0.2*	*0.7*	*1.1*	*1.2*	*1.4*	*3.0*
Exc. Na (cmol(+) kg^{-1})	*0.0*	*0.0*	*0.2*	*0.4*	*0.6*	*3.9*
TC (%)	*0.1*	*1.0*	*1.5*	*2.2*	*2.5*	*17.8*
N (%)	*0.0*	*0.1*	*0.1*	*0.2*	*0.2*	*1.2*
CEC (cmol(+) kg^{-1})	*1.7*	*7.4*	*13.5*	*17.0*	*27.5*	*38.3*
Clay (%)	*1.7*	*13.2*	*20.0*	*26.6*	*43.8*	*65.1*

As shown in Figure 2-13, the coefficient *a,* or the maximum theoretical *SI* value was highly (positively) correlated with pH and exchangeable sodium, equating with poorly aggregated soils. Clay percentage was positively related with the coefficient *a*, this may imply a relation between clay content and higher micro-aggregation, considering micro aggregates as the resultant particles of the slaking process and so with the maximum area increase or coefficient *a* (See Table 2 *in* Le Bissonnais (1996)). The Ca/Mg ratio was negatively correlated with coefficient *a*, known for its influence in aggregate stabilization (Rengasamy et al., 1986) and also negatively related to C/N which is associated with microbial activity as hinted by Tisdall and Oades (1982) in terms of lower potential oxidation rates by microorganisms i.e., the more the soil aggregate stability the higher the C/N relation.

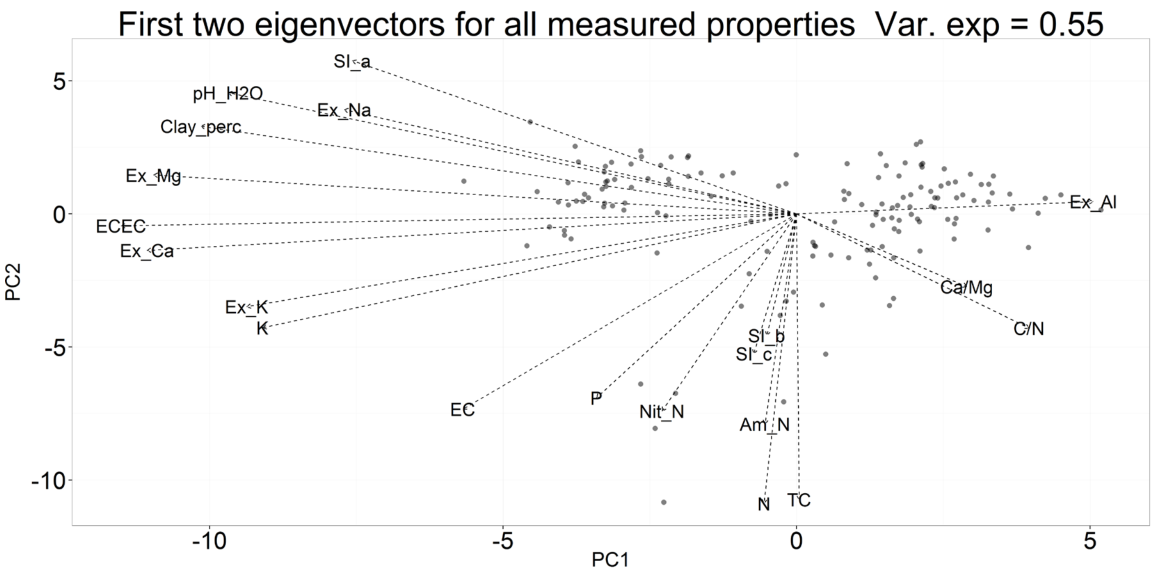


Figure 2-13 Biplot of SI values and measured properties on all the samples.

Both coefficient *b* or the displacement along the x-axis, which could be interpreted as the initial time of fast slaking, and coefficient *c* interpreted as the rate of disaggregation were highly correlated with N and TC. Oades and Waters (1991) observed that macroaggregates (> 250 µm) are usually held together by roots and hyphae, which may explain the direct relation between these two elements (N and TC) and the delay in the rupture of macro aggregates and their rate of disaggregation i.e., coefficients *b* and *c*.

An interesting result was the different correlation between coefficients *a* and *b* with TC and C/N respectively, reflecting the importance of carbon as a binding agent and protecting against immediate disruptive forces (Coefficient *b*) and the final effect (Coefficient *a*) of high C/N values related with a lower potential oxidation rate by microbes (Tisdall and Oades, 1982); Further research will be conducted into these specific matters, as microbial community information will be also considered adding an extra level of complexity.

Finally, despite the lack of linear correlation noticed in the principal component analysis, we plotted the relation between the coefficient *a* and %TC, finding that the slaking index coefficient *a* decreases exponentially as the %TC increases (Figure 2-14). This relation was also observed by Kemper and Rosenau (1986) were they found a logarithmic relation between Organic Carbon and MWD in North American soils.

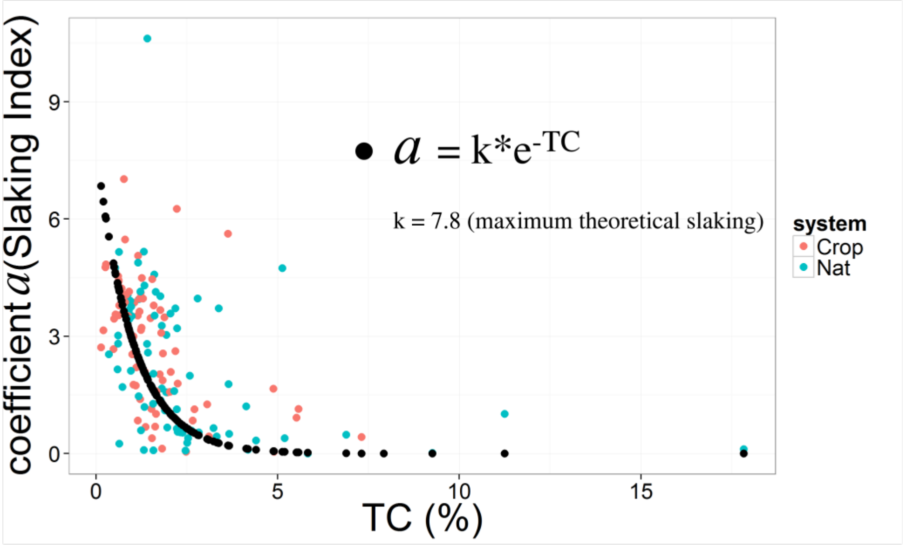


Figure 2-14 The relationship between Total C content (%TC) and coefficient *a*.

2.4.4 Relationships with land-use

During the last 80 years, several authors have revised the influence of land-use on different aggregate stability measures (Yoder, 1936; Tisdall and Oades, 1982; Saygin et al., 2012), highlighting the detrimental effect of arable cropping on aggregation stability. This new methodology detected a similar effect, which was reflected in the values of the coefficient *a*, with significant differences (p-value< 0.05), however coefficient *b* and *c* did not have such differences (Figure 2-15, Figure 2-16 and Figure 2-17).

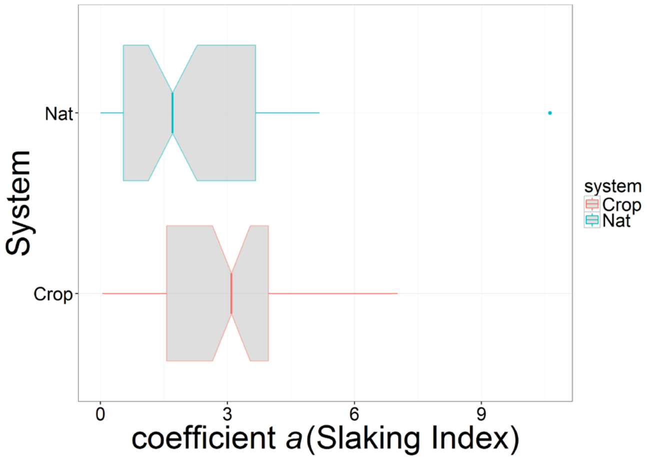


Figure 2-15 Boxplot of coefficient *a* values by land use. Dots correspond to values outside 95% of observations.

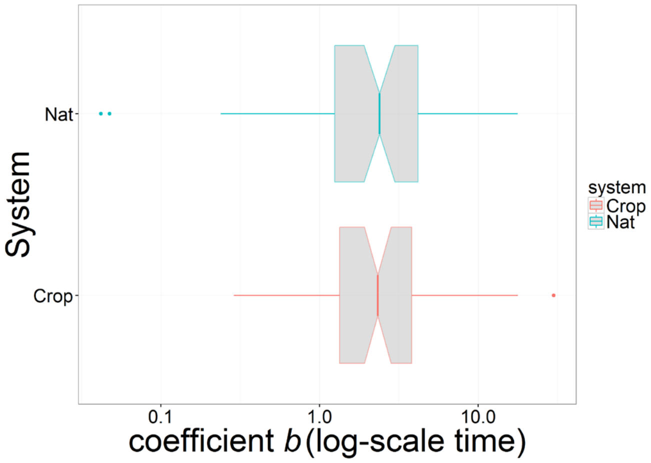


Figure 2-16 Boxplot of coefficient *b* values by land use. Dots correspond to values outside 95% of observations.

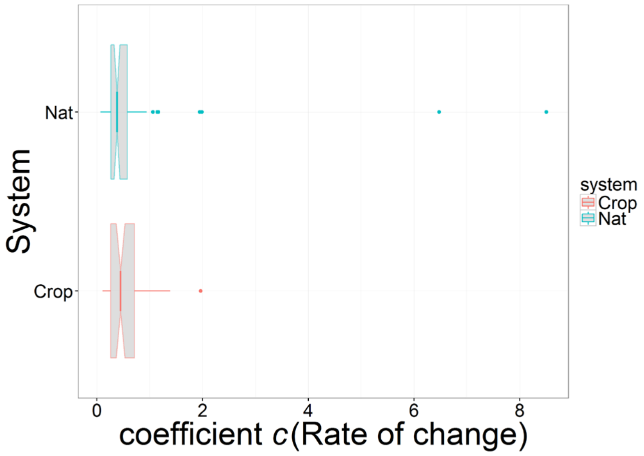


Figure 2-17 Boxplot of coefficient *c* values by land use. Dots correspond to values outside 95% of observations.

2.5 Discussion

Having presented the methodology, some important points are worth highlighting: First, the simplicity of the method, once the images are acquired (the process can also be automated), the area detection is calculated in a matter of minutes and the results are readily available. Further, due to the excellent fit of a model, it is possible to elucidate the dynamics behind the disaggregation process. Second, its scalability, having observed that the final *SI* values are by themselves a good indicator of aggregation stability, it is possible to use just the initial and last value of the curve to calculate an index comparable with the MWD method but without its measurement complexity.

Finally, this new information which describes the entire disaggregation process with time, gives new evidence for the different mechanisms behind the slaking behaviour of soil aggregates highlighting the importance of TC and N in the initial part of the disaggregation process (coefficients *b* and *c*) and of pH, Exc. Na, Ca/Mg and C/N in the final slaking values (coefficient *a*).

2.6 Conclusions

A new methodology for assessing soil aggregate stability has been presented. The method is fast, inexpensive and simple and provides valuable information in terms of slaking patterns in soil aggregates under fast wetting treatments. We observed that the entire slaking process was successfully fitted with an empirical model, resulting in new coefficients which can be used to compare with other soil properties or between treatments. Further research will focus on different interactions between the disaggregation process and other factors e.g., microbial communities.

Chapter 2, first part references

Amézketa, E., 1999. Soil Aggregate Stability: A Review. Journal of Sustainable Agriculture 14(2-3), 83-151.

Bouyoucos, G.J., 1962. Hydrometer method improved for making particle size analyses of soils. Agronomy Journal 54(5), 464-465.

Braudeau, E., Mohtar, R.H., 2006. Modeling the Swelling Curve for Packed Soil Aggregates Using the Pedostructure Concept. Soil Science Society of America Journal 70(2), 494.

Bronick, C.J., Lal, R., 2005. Soil structure and management: a review. Geoderma 124(1-2), 3-22.

Colwell, J., 1963. The estimation of the phosphorus fertilizer requirements of wheat in southern New South Wales by soil analysis. Animal Production Science 3(10), 190-197.

Davies, A.M.C., Fearn, T., 2006. Back to basics: Calibration statistics. Spectroscopy Europe 18(2), 31.

Dıaz-Zorita, M., Perfect, E., Grove, J., 2002. Disruptive methods for assessing soil structure. Soil and Tillage Research 64(1), 3-22.

Edwards, A.P., Bremner, J.M., 1967. Dispersion of soil particles by sonic vibration. Journal of Soil Science 18(1), 47-63.

Field, D.J., McKenzie, D.C., Koppi, A.J., 1997. Development of an improved Vertisol stability test for SOILpak. Soil Research 35(4), 843-852.

Gompertz, B., 1825. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. Philosophical Transactions of the Royal Society of London 115(ArticleType: research-article / Full publication date: 1825 /), 513-583.

Henin, S., Monnier, G., Combeau, A., 1958. Méthode pour l'étude de la stabilité structurale des sols. Ann. Agron 9, 73-92.

Kemper, W., Rosenau, R., 1986. Aggregate Stability and Size Distribution. Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods (methodsofsoilan1), 425-442.

Le Bissonnais, Y., 1996. Aggregate stability and assessment of soil crustability and erodibility: I. Theory and methodology. European Journal of Soil Science 47(4), 425-437.

Loveday, J., Pyle, J., 1973. The Emerson dispersion test and its relationship to hydraulic conductivity.

Middleton, H.E., 1930. Properties of soils which influence soil erosion. US Dept. of Agriculture.

Moran, C.J., McBratney, A.B., Koppi, A.J., 1989. A Rapid Method for Analysis of Soil Macropore Structure. I. Specimen Preparation and Digital Binary Image Production. Soil Science Society of America Journal 53(3).

Nimmo, J.R., Perkins, K.S., 2002. Aggregate stability and size distribution. In: J.H.a.T. Dane, G.C. (Ed.), Methods of soil analysis, Part 4--Physical methods: Soil Science Society of America Book Series No. 5. Soil Science Society of America, Madison, Wisconsin, pp. 317-328.

O'Callaghan, J.F., Loveday, J., 1973. Quantitative measurement of soil cracking patterns. Pattern Recognition 5(2), 83-98.

Oades, J., Waters, A., 1991. Aggregate hierarchy in soils. Soil Research 29(6), 815-828.

Pagliai, M., Vignozzi, N., Pellegrini, S., 2004. Soil structure and the effect of management practices. Soil and Tillage Research 79(2), 131-143.

Pau, G., Oles, A., Smith, M., Sklyar, O., Huber, W., 2014. EBImage: Image processing toolbox for R. R package version 4.4.0.

Perfect, E., Rasiah, V., Kay, B.D., 1992. Fractal Dimensions of Soil Aggregate-size Distributions Calculated by Number and Mass. Soil Science Society of America Journal 56(5).

Pulido Moncada, M., Gabriels, D., Cornelis, W., Lobo, D., 2013. Comparing Aggregate Stability Tests for Soil Physical Quality Indicators. Land Degradation & Development, n/a-n/a.

Quirk, J., 1950. The measurement of stability of soil micro-aggregates in water. Australian Journal of Agricultural Research 1(3), 276-284.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rayment, G., Higginson, F.R., 1992. Australian laboratory handbook of soil and water chemical methods. Inkata Press Pty Ltd.

Rengasamy, P., Greene, R., Ford, G., 1986. Influence of magnesium on aggregate stability in sodic red-brown earths. Soil Research 24(2), 229-237.

Ringrose-Voase, A.J., Bullock, P., 1984. The automatic recognition and measurement of soil pore types by image analysis and computer programs. Journal of Soil Science 35(4), 673-684.

Russell, M.B., Feng, C.L., 1947. Characterization of the stability of soil aggregates. Soil science 63(4), 299-304.

Saygin, S.D., Cornelis, W.M., Erpul, G., Gabriels, D., 2012. Comparison of different aggregate stability approaches for loamy sand soils. Applied Soil Ecology 54, 1-6.

Tisdall, J.M., Oades, J.M., 1982. Organic matter and water-stable aggregates in soils. Journal of Soil Science 33(2), 141-163.

Van Bavel, C., 1950. Mean weight-diameter of soil aggregates as a statistical index of aggregation. Soil Science Society of America Journal 14(C), 20-23.

Vogel, H.J., 1997. Morphological determination of pore connectivity as a function of pore size using serial sections. European Journal of Soil Science 48(3), 365-377.

Yoder, R.E., 1936. A direct method of aggregate analysis of soils and a study of the physical nature of erosion losses. Agronomy Journal 28(5), 337-351.

Youker, R., McGuinness, J., 1957. A SHORT METHOD OF OBTAINING MEAN WEIGHT-DIAMETER VALUES OF AGGREGATE ANALYSES OF SOILS. Soil science 83(4), 291-294.

Young, R., 1984. A method of measuring aggregate stability under waterdrop impact. Transactions of the ASAE-American Society of Agricultural Engineers (USA).

Zanini, E., Bonifacio, E., Albertson, J.D., Nielsen, D.R., 1998. Topsoil aggregate breakdown under water-saturated conditions. Soil science 163(4), 288-298.

Part 2: Spectral information related to soil structure

2.7 Summary

Different modelling approaches using soil reflectance values from the Visible, Near (Vis-NIR) and Middle infrared (MIR) regions of the electromagnetic spectrum were employed to predict Slaking index coefficient *a* (*SIa*). The first approach considered raw information from the spectrum and the second considered a pedotransfer function (PTF), where common soil properties e.g. pH, where first predicted from raw spectra to be used later for the prediction of *SIa*. Finally the performance of each method is compared.

Two datasets were used, the first dataset was the same dataset presented in the first part of this chapter, while the second considered external samples with contrasting spectral signature i.e. different Vis-NIR-MIR reflectance values. It was observed that model accuracy in the first dataset increased as different regions of the spectra were added, while in the second dataset, the predictability was always poor. In the first dataset, the results showed a low predictability (R^2 0.3 to 0.36) using only the Vis-NIR region of the spectra, where only Soil Organic Carbon (SOC) and Fe oxides spectral regions were identified as main predictors (~ 2000 – 2100 nm and ~ 700 – 900 nm). As MIR regions were added, the modelling performance improved (R^2 0.45 to 0.60) where SOC and kaolinite related regions were identified as important predictors (4700 – 4900 cm^{-1}, 3500 – 3700 cm^{-1} or ~ 2000 – 2100 nm and ~ 2700-2850 nm respectively).

The modelling approach considering soil properties predicted from MIR spectra and a PTF considering exchangeable calcium, exchangeable magnesium, pH, clay and cation exchange capacity, had a slightly inferior performance compared with the best spectral models (R^2 0.54).

Since the uncertainty related with the spectroscopic models in the first dataset was relatively low (1.1 slaking index units) it was concluded that is possible to use the current spectroscopic models as a prospecting tool for predicting *SIa*, however the spectral similarity of the samples needs to be verified first, due to the low predictability in samples with a different spectral signature.

2.8 Introduction

The conventional methods for the evaluation of soil aggregate stability require considerable time, work and are also expensive to outsource to commercial laboratories. The first part of this chapter presented a new methodology for assessing soil slaking that proved to be faster and less expensive; furthermore it provided three new indices that are related with key processes of slaking of soil aggregates.

It was shown that the new indices of soil slaking are closely related with known soil properties; therefore it should be possible to create predictive models by considering those existing relations. As commented in the rationale, different authors have used the benefits of Vis-NIR and MIR spectroscopy for the prediction of soil properties since has shown to contain comprehensive information about the soil composition (Ben-Dor and Banin, 1995; Stenberg et al., 2010b; Soriano-Disla et al., 2014). Its use requires minimum sample preparation and when employed with multivariate analytical methods, a number of soil properties can be simultaneously assessed from a single spectrum. Soil properties like total and organic carbon or clay percentage can be determined from a single scan, as well as physicochemical parameters such as pH, conductivity or redox potential (Armenta and de la Guardia, 2014).

Nevertheless, only a few studies have attempted to predict more complex soil properties which depend on a combination of factors, for example Minasny et al. (2008) explored the possibility of predicting physical properties dependant on pore structure like bulk density and hydraulic conductivity, finding that they should not be predicted directly, since spectroscopic methods only provides information of the surface composition. If a prediction of a volume dependant property is intended, a two stepped prediction needs to be performed. First, predict basic properties e.g., organic carbon, and second use PTF (Bouma, 1989), to finally predict the more complex property. Despite this, Minasny et al. (2008) foresaw a possible direct use in predicting other complex properties like Atterberg limits (Casagrande, 1932) or swelling behaviour.

As the Slaking index (*SI*) is a complex property not completely understood and highly dependent on chemical and physical properties as shown in the previous part of this chapter, it is possible to postulate that rather than being a function of the porous structure as reported by previous studies (Hallett et al., 1995; Ma et al., 2015) it could to be the opposite, and the soil porous structure may be a function of soil aggregate stability.

The aims of this work are first, to assess the predictability of *SI* coefficient *a* (*SIa*) directly from Vis-NIR and MIR spectra and to identify the specific spectral features related with the slaking process. Second, to test the performance of a two stepped prediction (Minasny et al., 2008) of the *SIa* by using Vis-NIR and MIR spectroscopy plus a PTF and finally to compare the proposed methods performance.

2.9 Materials and Methods

2.9.1 Datasets

Two datasets were used, the first one is the same dataset presented in part 1 of Chapter 2. The second dataset comprised samples from a Vertosol (Isbell, 2002), taken from 6 sites in a cereal cropping field in northern NSW. This dataset was used as an external validation. The second dataset was chosen to explore the influence of soil properties which could be related to a significantly different observed cereal yields and they formed part of a different study. From that dataset, a total of 9 soil cores were extracted in the site and each soil core was sliced into 0-30, 30-90 and greater than 90 cm depth increments.

2.9.2 Sample preparation and soil analyses

Basic chemical analyses were performed on both datasets. The analyses considered in the first dataset were Extractable Phosphorus (P) and Potassium (K) expressed in mg kg^{-1} (Colwell, 1963), Ammonium (NH_4) and Nitrate (NO_3) expressed in mg kg^{-1} (Rayment and Higginson, 1992), % Total Carbon (TC) by dry combustion (LECO instrument, CSBP laboratory Ltd., Western Australia), electric conductivity (EC) expressed in dS m^{-1}, pH level ($CaCl_2$), pH level (H_2O), Exchangeable cations (Ca, Mg, Na, K, Al) expressed in cmol(+) kg^{-1} (Rayment and Higginson, 1992), Cation Exchange Capacity (CEC) expressed in cmol(+) kg^{-1} (Rayment and Higginson, 1992) and % clay (Bouyoucos, 1962). For the second dataset clay percentage was analysed using MIR spectrometry, and organic carbon was measured instead of total carbon. All the analyses excepting particle size were performed by CSBP Laboratory Ltd., Western Australia.

A set of aggregates between 0.3 to 1.1 cm of diameter per sample (from the two datasets) were selected, air dried and the *SIa* of five of them was measured following Fajardo et al. (2016b) methodology (*see* Chapter 2, part 1). The rest of each sample's soil aggregates were ground and passed through a 2 mm sieve for portable the MIR and Vis-NIR measurements, finally a subsample of the previously sieved soil was ground and passed through a 100 µm sieve for laboratory MIR measurement.

MIR spectra in the laboratory, was measured by a FT-IR Spectrometer TENSOR 37 with a HTS-XT Microplate Reader (400-4000 cm^{-1}) Bruker®. Also, MIR was measured by a portable spectrometer Argilent 4100 ExoScan FTIR (400-6000 cm^{-1}) Argilent Technologies® and finally, Vis-NIR spectrum was measured with an ASP 350-2500 (4000-25000 cm^{-1}) AgriSpec, with a Spectralon® panel as absolute white.

2.9.3 Spectral treatments and statistical analyses

The first dataset was used to create *SIa* spectral models, with a training dataset of 80% of total samples selected using a Latin hypercube algorithm on the first 5 principal components of the spectra (Minasny and McBratney, 2006; Roudier, 2011), leaving the remaining 20% as an independent validation dataset. The second dataset was used as an external validation dataset and it was not included in the training samples.

A Cubist algorithm with 5 committees was used to create the spectral models (Quinlan, 1992; Kuhn et al., 2014). In order to optimize the predictions, different spectral pre-processing treatments were tested as described in Table 2-3.

Table 2-3 Spectral pre-treatments. *S-G filt1 and S-G filt2: Second order Savitsky-golay filter with none and first derivative output respectively; Strip5 and Strip10: Selection of every 5th and 10th wavelength respectively; Trim: Elimination of the extreme wavelengths prone to high noise to signal ratio; SNV: Standard normal variate transformation; ChBLC: Convex hull baseline correction; Det: Detrending algorithm as specified in Stevens and Ramirez-Lopez (2013).*

Spectral range	***Treatment 1***	***Treatment 2***	***Treatment 3***	***Treatment 4***
MIR lab	*S-G filt1 Strip10 SNV*	*S-G filt2 Strip10 SNV*	*S-G fil1 Strip10 ChBLC*	*S-G filt1 Strip10 Det*
MIR portable	*S-G filt1 Strip10 SNV*	*S-G filt2 Strip10 SNV*	*S-G filt1 Strip10 ChBLC*	*S-G filt1 Strip10 Det*
NIR part of Portable MIR	*S-G filt1 Strip5 SNV*	*S-G filt2 Strip5 SNV*	*S-G filt1 Strip5 ChBLC*	
Visible NIR	*S-G filt1 Trim SNV*	*S-G filt2 Trim SNV*	*S-G filt1 Trim ChBLC*	

For each treatment 100 different models were created using a 95% random sample (with resample) of the training dataset, using the mean of those 100 predictions to calculate the performance of the models against the validation set and the external validation dataset.

Out of the 100 modelling iterations for each treatment, kernel density plots were created with the usage of all the models in terms of the wavelengths and wavenumbers considered as important predictors, in order to identify recognizable spectrum zones that may be related with slaking behaviour in soil samples.

A PTF for predicting *SIa* was created with a multiple linear regression (R Core Team, 2013), using the chemical properties with the highest MIR predictability from the training dataset as input. The resulting PTF was used in a two-step approach (Minasny et al., 2008), using first spectral models to predict the parameters of the PTF and with this the *SIa*, from now on named as the “two-step model”.

The performance parameters for spectral, PTF and two-step models were calculated for the independent validation and the external validation datasets. The calculated coefficients were: Coefficient of determination (R^2), Root mean square error (RMSE), and bias i.e. *bias = mean predicted values – mean observed* values. All the analyses were implemented in R (R Core Team, 2013).

2.10 Results and discussion

2.10.1 Measured soil properties

Appendix 3 shows the summary of all the measured properties for training, validation and external validation datasets. It was noticeable that the training dataset covered most of the variability of both validation datasets. However, properties that have proven to be influential over *SIa,* such as pH, Exc. Mg., Exc. K. and Exc. Na. had significantly higher maxima in the external validation; consequentially *SIa* values were higher in the external validation dataset.

2.10.2 Spectral information

The three different types of measured spectra i.e. Vis-NIR and MIR, for the three datasets (training, independent validation and external validation) were projected into principal component space previous Standard Normal Variate transformation (SNV) (Barnes et al., 1989), where it is possible to see that the independent validation dataset was well covered by the training dataset. But as mentioned before, the external validation dataset appeared different, now in terms of its spectral signature, especially in the MIR range of the spectra (Figure 2-18, Figure 2-19 and Figure 2-20).

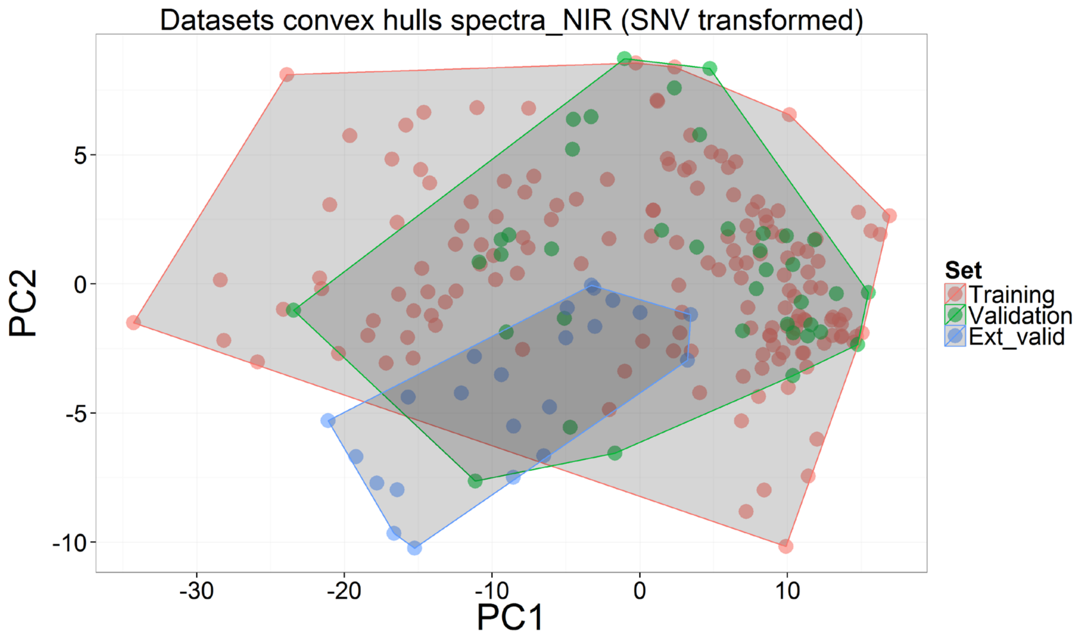


Figure 2-18. Vis-NIR spectra convex hulls for three datasets.

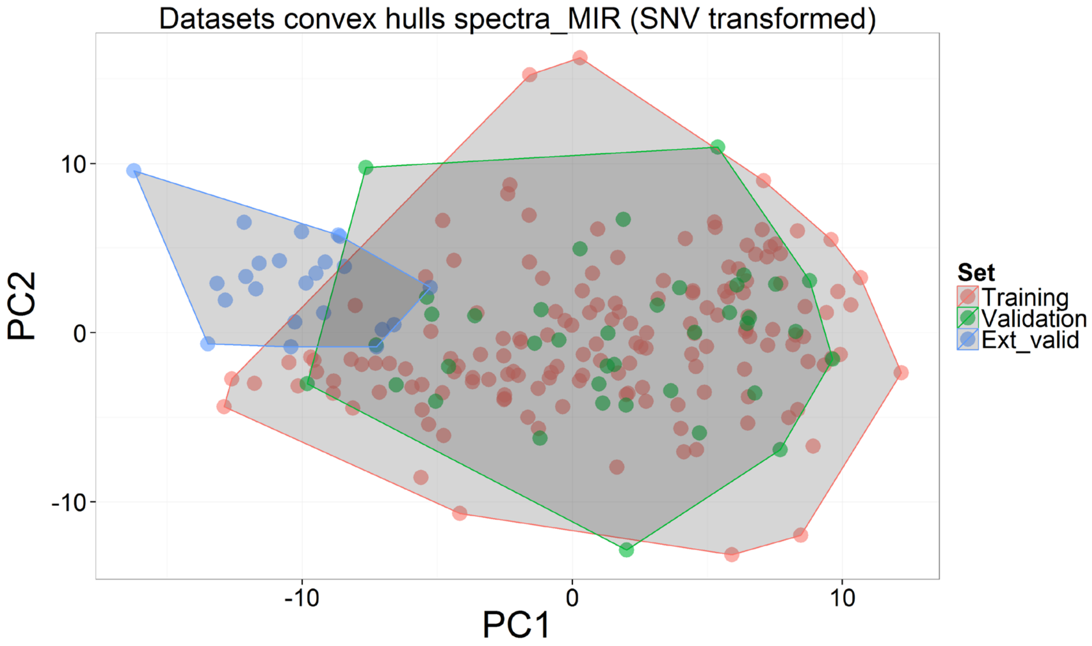


Figure 2-19. Portable MIR spectra convex hulls for three datasets.

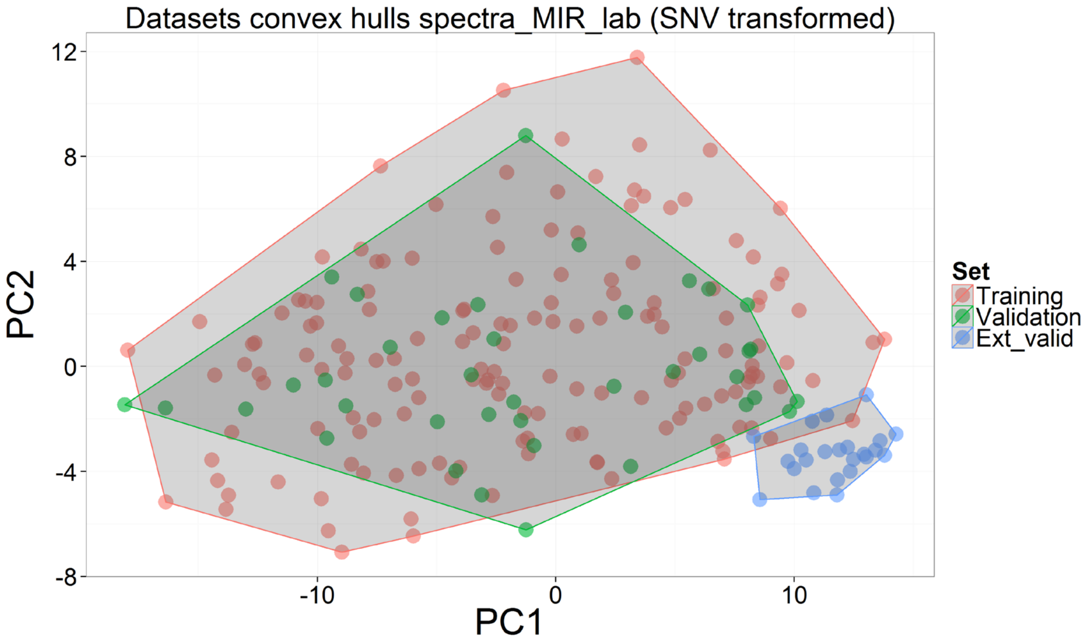


Figure 2-20. Lab MIR spectra convex hulls for three datasets.

2.10.3 Spectral model performance

Table 2-4 shows the performance of the spectral models against the independent and external validation datasets.

Table 2-4 Spectral models performance on independent and external dataset.

MODEL	*Independent*			*External*		
	R^2	*RMSE*	*BIAS*	*R^2*	*RMSE*	*BIAS*
MIR_lab_T1	*0.51*	*1.12*	*0.11*	*0.05*	*3.35*	*-0.82*
MIR_lab_T2	*0.49*	*1.16*	*0.17*	*0.00*	*3.53*	*-1.63*
MIR_lab_T3	*0.45*	*1.19*	*0.13*	*0.00*	*4.07*	*-2.60*
MIR_lab_T4	*0.57*	*1.06*	*0.17*	*0.07*	*3.40*	*-1.13*
MIR_portable_T1	*0.51*	*1.14*	*0.07*	*0.00*	*3.58*	*-1.84*
MIR_portable_T2	*0.54*	*1.13*	*0.04*	*0.07*	*3.68*	*-2.14*
MIR_ portable_T3	*0.60*	*1.04*	*0.05*	*0.00*	*3.73*	*-2.10*
MIR_ portable_T4	*0.52*	*1.14*	*0.03*	*0.00*	*3.71*	*-1.99*
MIR_NIR_T1	*0.48*	*1.18*	*0.05*	*0.00*	*3.76*	*-2.17*
MIR_NIR_T2	*0.42*	*1.24*	*0.06*	*0.01*	*3.94*	*-2.39*
MIR_NIR_T3	*0.50*	*1.14*	*0.07*	*0.00*	*3.52*	*-1.70*
NIR_T1	*0.30*	*1.34*	*0.11*	*0.02*	*3.77*	*-2.09*
NIR_T2	*0.35*	*1.30*	*0.17*	*0.00*	*3.82*	*-2.27*
NIR_T3	*0.36*	*1.29*	*0.19*	*0.14*	*3.41*	*-1.65*

In general terms, model performance on the independent dataset was similar to that observed by Sarathjith et al. (2014) whom predicted Geometric mean weight (GMD) parameters by spectral means (Vis-NIR) in Indian soils. Despite the successful modelling on independent samples, the models did not predict in the external dataset, observing a large underestimation of the *SIa* reflected in a high RMSE and a negative bias consistent with the lower mean *SIa* values of the training datasets compared with the external validation dataset.

2.10.4 Model usage

2.10.4.1 Vis-NIR part of the spectra

Kernel density plots with the models usage were created i.e., main wavelengths in the case of Vis-NIR, or wavenumbers in the case of MIR instruments, for determining which parts of the Vis-NIR spectra, in terms of physicochemical composition were the most important when predicting *SIa* (Figure 2-21 to Figure 2-22).

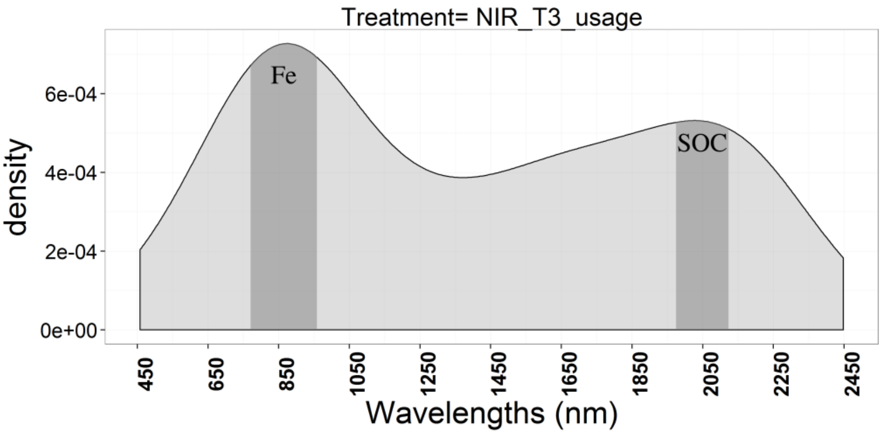


Figure 2-21 Wavelengths usage of Vis-NIR models.

Figure 2-21 shows the main wavelengths used in 100 different Vis-NIR models predicting *SIa* values. It can be seen that the dominant spectral regions are in the range of 700 – 900 nm corresponding to Fe-bearing minerals e.g. hematite, as observed by Sarathjith et al. (2014) and in the range of 2000 – 2100 nm which has been identified as an important zone for predicting soil organic carbon (SOC), as reviewed by Stenberg et al. (2010a).

These results are consistent with those of Duiker et al. (2003) where a close relation between MWD values and Fe oxides was found. The same authors reported that Schahabi and Schwertmann (1970) and Rhoton et al. (1998) obtained similar results.

The same peaks related to SOC, were recognized as important in the NIR part of the portable MIR instrument (Figure 2-22) in the range of 4700 – 4900 cm^{-1} (~ 2000 - 2100 nm) and in the range of 4100 to 4200 cm^{-1} (~ 2380 – 2440 nm) as reviewed by Soriano-Disla et al. (2013); Soriano-Disla et al. (2014).

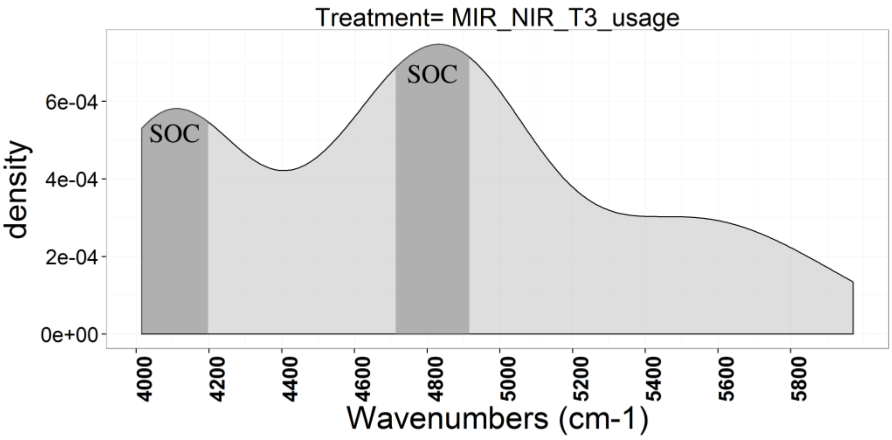


Figure 2-22 Main wavenumbers of MIR-NIR models (NIR part of portable MIR).

Gomez et al. (2013) studied the possibility of predicting MWD indices from Vis-NIR spectra in Mediterranean soils. They found similar results when using “spectrotransfert” functions (spectral models in our case), however they stated that no specific spectral response was related to MWD indices. Our study results showed a clear distinction in the selection of specific Vis-NIR spectral features.

2.10.4.2 Vis-NIR + MIR part of the spectra

As with the portable MIR models (NIR + MIR), the region between wavenumbers 3500 – 3700 cm^{-1} (~ 2700 – 2850 nm) related with kaolinite (Soriano-Disla et al., 2014), was used as a predictor, and also the region of the 4700 cm^{-1} (~ 2120 nm) or SOC related peaks as with previous models (Figure 2-23). These findings are consistent with the results presented by Oades and Waters (1991) on a study based on transmission electron microscopy thin sections of soil micro aggregates (TEM). They proposed an idealised clay aggregate structure which contained as the elemental structure, small (1 to 2 µm) assemblages of kaolinite which where presumably held together by oxides.

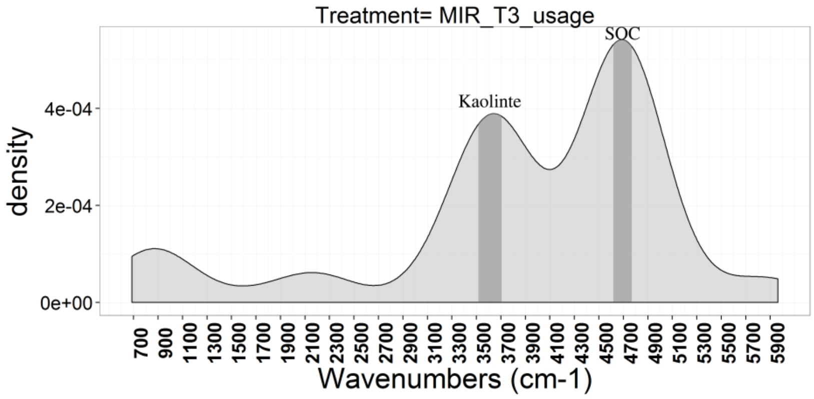


Figure 2-23 Main wavenumbers of portable MIR models.

The before mentioned results were corroborated by the usage of the MIR (laboratory) models, where the wavelengths between 3500 – 3700 cm^{-1} (~ 2700 – 2850 nm) related to kaolinite had the highest importance Figure 2-24.

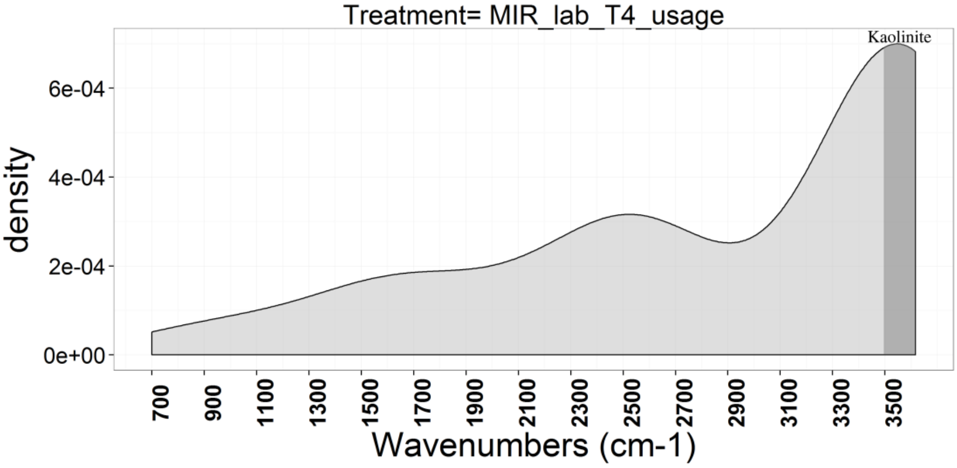


Figure 2-24 Main wavenumbers of laboratory MIR models.

Cañasveras et al. (2010) found good predictability of MWD using Vis-NIR and MIR spectroscopy models and their results showed similarity with this study in terms of the importance of Fe bearing minerals in the NIR region e.g., hematite, however this study showed also a higher importance given to kaolinite in the MIR region.

A closer look to Cañasveras et al. (2010) work shows that their samples had also a high amount of Ca carbonates (see Table 1 *in* Cañasveras et al. (2010)), consequentially several peaks related to calcites were found as the most important spectral features in the MIR region (*see* Figure 5 b in Cañasveras et al. (2010). In this regard, even though regions between 3500-3700 cm^{-1} (~ 2700 – 2850 nm or kaolinite-related) were categorized as important in their study, they seemed to have a secondary role in the aggregation process.

2.10.5 Two-step modelling

Equation 3 shows the PTF created using measured properties from the training dataset as specified in section 2.9.3.

$$SIa = -2.1103 - 0.5039Exc.Ca - 0.3472Exc.Mg + 0.5937pH + 0.0421Clay + 0.4186CEC \tag{3}$$

After predicting each of the required parameters[1] for the PTF function on each of the validation datasets, it was found that the performance in the validation dataset was slightly inferior compared to the direct method with Vis-NIR-MIR spectra (Figure 2-25), most probably because the absence of those properties known to affect directly to the *SI* in the PTF e.g. Fe oxides, or simply due to more complex relations captured in the full spectra and not in separate properties.

[1] R^2 values for Ex.Ca., Exc.Mg., pH, Clay and CEC spectral models were 0.69, 0.73, 0.61, 0.62 and 0.72 respectively in independent dataset.

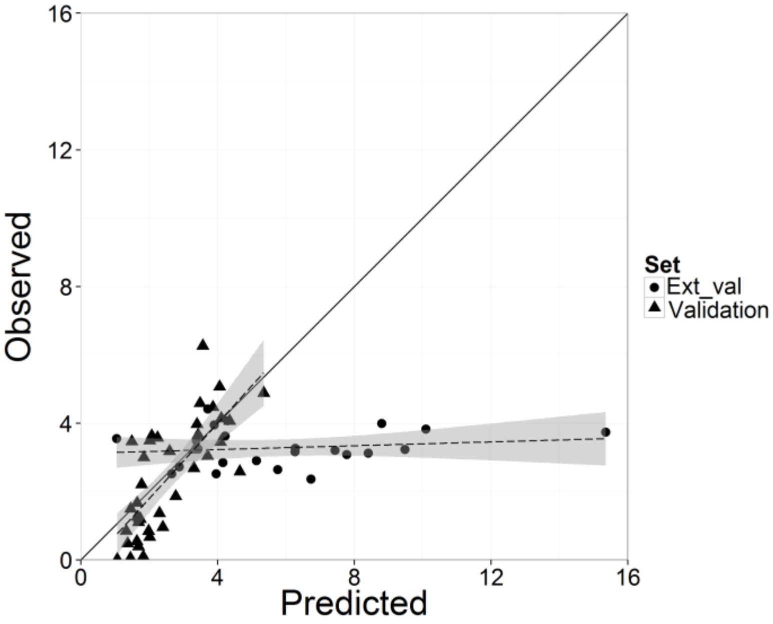


Figure 2-25 Observed vs. predicted values of Two-step model procedure in two validation datasets.

Table 2-5 Performance parameters of Two-step model procedure in two validation datasets.

MODEL	*Independent*			*External*		
	R^2	*RMSE*	*BIAS*	*R^2*	*RMSE*	*BIAS*
Two steps model	*0.54*	*1.18*	*-0.14*	*0.00*	*4.00*	*-2.59*

The use of two-step models for predicting MWD was also employed by Gomez et al. (2013). They found that, while it is a good approach, the success will depend on the ability of spectral information to predict soil properties highly correlated with MWD.

The results of this study showed that as information from a wider range of the electromagnetic spectrum was added (MIR part of the spectra), the predictability increased, due to the addition of spectral information related to other components as observed by previous studies in spectroscopy (Viscarra Rossel et al., 2006). Therefore, if a PTF is used for predicting *SIa,* a first assessment of the specific spectral features considered as important through data mining of spectral models seems to be a useful approach (Viscarra Rossel and Behrens, 2010).

Finally, the performance in the external dataset was again very poor, reflecting the need of a more representative training dataset or a better generalization in the modelling process, these issues will be revised in the next chapter.

2.11 Uncertainty of predictive models and general discussion of the chapter

So far, different approaches for the calculation of the *SIa* have been presented, namely the use of image recognition, and models developed with Vis-NIR and MIR information. It has been shown that the first methodology presented is fast and accurate, based on the standard deviation of the *SIa* values of a set of 5 soil aggregates (*see* Figure 2-11).

Notwithstanding these positive results, the created models using Vis-NIR and MIR information (even though their predictions were not accurate in an external dataset), could represent in the future an alternative for characterizing *SIa*, considering a scenario where a successful model is created (adequate training dataset). For this reason it would desirable to have a notion of the performance in terms of time and cost of the spectroscopic models vs the image recognition methodology.

In order to compare the methodologies, it is possible to use for example, their uncertainty and cost/time used in the analysis. First the uncertainty related to the image recognition algorithm can be represented by the standard deviation (σ) of the calculated *SIa* of 5 soil aggregates in a sample. For the case of the spectroscopic models, an analogue is the root mean squared error (RMSE) of the predictions of *SIa* values per sample, since both are expressed in the units of the target property.

Second, the time employed on the sample preparation plus the time used in the analysis can be measured equally for all the methodologies as well as the associated cost e.g., time used by an operator per 100 samples. Table 2-6 shows the time used for the measurement of the *SIa* per 100 samples, the cost associated to the operator and the respective uncertainty calculated for 100 randomly selected samples.

Table 2-6 Time-cost-uncertainty of SI methodologies. The time used in the image recognition analysis considered the use of five digital cameras working in parallel. Cost was calculated assuming an average salary of $AUD 21 per hour for a laboratory assistant.

Method	***Time in sample preparation (hours per 100 samples)***	***Time used in analysis (hours per 100 samples)***	***Cost per 100 samples ($AUD)***	***Uncertainty***
Image recognition	*1*	*47*	*1,008*	*0.60*
Vis-NIR	*12*	*2*	*294*	*1.13*
MIR laboratory	*24*	*2*	*546*	*1.06*
MIR portable	*12*	*2*	*294*	*1.10*

In the previous exercise, the time of analysis involved the time spent from the beginning of the experiment until obtaining the final result, leaving the time spent for air drying as constant for the four methods.

It can be seen from the table that even though the image recognition algorithm is almost twice more precise, the cost is considerably higher with the current design i.e., 5 digital cameras running in parallel. Also, the uncertainty values of the spectroscopic models can be considered acceptable since the uncertainty is close to 1 SI unit and *SIa* values range commonly from 0 to 10.

On the other hand, the image recognition methodology is a newly developed method, and still has a great space for improvement. Since the images are recognized by a computer and each soil aggregate can be measured separately, it would be possible for example to analyse more samples with just one camera (data not presented showed the possibility to analyse up to 10 soil aggregates within 1 image), and with this reducing the time and cost in the analysis to at least half of the current price.

Despite of the previous, and based on the current methodologies performance, it would be possible to use spectroscopic models as a prospecting tool if no extra information is available, since once the models are created, the time required in the analysis is still less. It is important to highlight however, that if a prediction is intended, the created models need to cover the variability of the samples to predict, as commented in section 2.10.3, where it was shown that the current models cannot predict in samples with a different composition i.e., different spectral signature.

Finally, an important factor that needs to be assessed (if the slaking index method is adopted by the soil science community) is the practical meaning of the index. So far, the values obtained by the methodology have demonstrated to have a direct relation with key soil properties. However, the establishment of thresholds of SI values over for example, agricultural productivity and/or soil erosion, is a task that at needs to be evaluated in order to justify a possible reduction of the operational uncertainty in any method to be utilized in the future.

2.12 Conclusions

The importance of Fe oxides or Fe-bearing clays was found a key component in slaking behaviour, similar results were found by Schahabi and Schwertmann (1970), Rhoton et al. (1998) and Duiker et al. (2003) in aggregate stability measured by wet sieving. It was observed an important influence of wavenumbers ranging from 3500 – 3700 cm^{-1} (~ 2700-2850 nm) related with kaolinite presence over the *SIa* values. The performance of the models increased accordingly when adding this information, reflected in more accurate models.

It was found that a two-step approach can be successfully implemented; nevertheless special attention should be put in the PTF creation, as complex properties (like *SIa*), are dependant of a combination of factors reflected in the better predictability when using the whole spectra. Additionally, the assessment of specific spectral features through data mining seems to be an effective method to identify key properties for the creation of PTFs.

The spectroscopic models uncertainty was relatively low (1.1 *SIa* units) considering the range of values that *SIa* can have more than 10 *SIa* units , therefore it is possible to use the current models as a prospection tool if the samples to be predicted have similar composition and are spectrally similar.

The fact that the external validation was not successfully predicted is an important issue when trying to calculate *SIa* by spectral means. There are several techniques for dealing with spectral libraries incompatibility, namely global spectral databases (Brown et al., 2006) or spiking procedures (Guerrero et al., 2010), however, their use will be analysed in Chapter 3.

Further work needs to assess for example, the agronomical value of the Slaking Index, in order to establish practical thresholds and have a real notion of the implications of the high or low uncertainty inherent to the methodology.

Chapter 2, second part references

Armenta, S., de la Guardia, M., 2014. Vibrational spectroscopy in soil and sediment analysis. Trends in Environmental Analytical Chemistry 2, 43-52.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Appl. Spectrosc. 43(5), 772-777.

Ben-Dor, E., Banin, A., 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Science Society of America Journal 59(2), 364-372.

Bouma, J., 1989. Using soil survey data for quantitative land evaluation, Advances in soil science. Springer, pp. 177-213.

Bouyoucos, G.J., 1962. Hydrometer method improved for making particle size analyses of soils. Agronomy Journal 54(5), 464-465.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132(3-4), 273-290.

Cañasveras, J.C., Barrón, V., del Campillo, M.C., Torrent, J., Gómez, J.A., 2010. Estimation of aggregate stability indices in Mediterranean soils by diffuse reflectance spectroscopy. Geoderma 158(1-2), 78-84.

Casagrande, A., 1932. Research on the Atterberg limits of soils. Public roads 13(8), 121-136.

Colwell, J., 1963. The estimation of the phosphorus fertilizer requirements of wheat in southern New South Wales by soil analysis. Animal Production Science 3(10), 190-197.

Duiker, S.W., Rhoton, F.E., Torrent, J., Smeck, N.E., Lal, R., 2003. Iron (Hydr)Oxide Crystallinity Effects on Soil Aggregation. Soil Science Society of America Journal 67(2).

Fajardo, M., McBratney, A.B., Field, D., Minasny, B., 2016b. Soil slaking assessment using image recognition. Soil and Tillage Research In revision.

Gomez, C., Le Bissonnais, Y., Annabi, M., Bahri, H., Raclot, D., 2013. Laboratory Vis–NIR spectroscopy as an alternative method for estimating the soil aggregate stability indexes of Mediterranean soils. Geoderma 209-210, 86-97.

Guerrero, C., Zornoza, R., Gómez, I., Mataix-Beneyto, J., 2010. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. Geoderma 158(1–2), 66-77.

Hallett, P., Dexter, A., Seville, J., 1995. Identification of pre-existing cracks on soil fracture surfaces using dye. Soil and Tillage Research 33(3), 163-184.

Isbell, R.F., 2002. The Australian soil classification / R.F. Isbell. Australian soil and land survey handbook ; vol. 4. CSIRO Publishing, Collingwood, Vic. :.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2014. Cubist: Rule- and Instance-Based Regression Modeling.

Ma, R., Cai, C., Li, Z., Wang, J., Xiao, T., Peng, G., Yang, W., 2015. Evaluation of soil aggregate microstructure and stability under wetting and drying cycles in two Ultisols using synchrotron-based X-ray micro-computed tomography. Soil and Tillage Research 149, 1-11.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32(9), 1378-1388.

Minasny, B., McBratney, A.B., Tranter, G., Murphy, B.W., 2008. Using soil knowledge for the evaluation of mid-infrared diffuse reflectance spectroscopy for predicting soil physical and mechanical properties. European Journal of Soil Science 59(5), 960-971.

Oades, J., Waters, A., 1991. Aggregate hierarchy in soils. Soil Research 29(6), 815-828.

Quinlan, J.R., 1992. Learning with continuous classes, 5th Australian joint conference on artificial intelligence. Singapore, pp. 343-348.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rayment, G., Higginson, F.R., 1992. Australian laboratory handbook of soil and water chemical methods. Inkata Press Pty Ltd.

Rhoton, F., Römkens, M., Lindbo, D., 1998. Iron oxides erodibility interactions for soils of the Memphis catena. Soil Science Society of America Journal 62(6), 1693-1703.

Roudier, P., 2011. clhs: a R package for conditioned Latin hypercube sampling.

Sarathjith, M.C., Das, B.S., Vasava, H.B., Mohanty, B., Sahadevan, A.S., Wani, S.P., Sahrawat, K.L., 2014. Diffuse Reflectance Spectroscopic Approach for the Characterization of Soil Aggregate Size Distribution. Soil Science Society of America Journal 78(2), 369.

Schahabi, S., Schwertmann, U., 1970. Der Einfluß von synthetischen Eisenoxiden auf die Aggregation zweier Lößbodenhorizonte. Zeitschrift für Pflanzenernährung und Bodenkunde 125(3), 193-204.

Soriano-Disla, J.M., Janik, L., McLaughlin, M.J., Forrester, S., Kirby, J., Reimann, C., 2013. The use of diffuse reflectance mid-infrared spectroscopy for the prediction of the concentration of chemical elements estimated by X-ray fluorescence in agricultural and grazing European soils. Applied Geochemistry 29, 135-143.

Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M.J., 2014. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. Applied Spectroscopy Reviews 49(2), 139-186.

Stenberg, B., Rossel, R.A.V., Mouazen, A.M., Wetterlind, J., 2010a. Chapter five-visible and near infrared spectroscopy in soil science. Advances in agronomy 107, 163-215.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010b. Visible and Near Infrared Spectroscopy in Soil Science. 107, 163-215.

Stevens, A., Ramirez-Lopez, L., 2013. prospectr package. R package version 0.1.3.

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158(1-2), 46-54.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131(1-2), 59-75.

Chapter 3 *Soil spectral databases*

Use of large Vis-NIR spectral libraries for the prediction of local soil properties: A continental scale example

3.1 Summary

Two large spectral libraries from the United States with a total of 7,781 (Rapid Assessment of U.S. Soil Carbon (RaCA) Training) and 19,837 (Legacy dataset) valid samples were used to predict % of Total Carbon (TC) in a local area (state wide) of the U.S. using Visible and Near Infrared (Vis-NIR) spectroscopy. The target area was used as a validation dataset and considered samples from South Central U.S. within the area identified as 'LUGR' region 9 (acronym was designated after Land Use-Land Cover plus Soil Group combination, *see* Soil Survey Staff (2013)) defined in the RaCA project. The aims of this work were to assess the influence of spectral variability in predictive modelling by using a set of state-of-the-art spectroscopic strategies for the prediction of %TC and different spectral libraries, namely the use of i) Legacy libraries, ii) Legacy libraries in spectral hull , iii) Spiked in hull legacy libraries, iv) Latin hypercube legacy libraries, v) Spiked latin hypercube legacy libraries, vi) RaCA libraries, vii) RaCA in spectral hull libraries, viii) RaCA local libraries , ix) RaCA in hull local libraries and x) only spike libraries.

All the approaches excepting the one using only samples from a spike, produced excellent results, comparable with previous works i.e., (Sequeira et al., 2014; Viscarra Rossel et al., 2016). A closer look to the models showed that the factors that most influenced the results where the geographic location, the number of samples in the training dataset, the covered range in the target variable (%TC) and the similarity in terms of spectral characteristics between training and validation datasets. Legacy and Legacy in-hull approaches had the best R^2 of 0.95 and RMSE of 3.7% respectively. RaCA library models were slightly less accurate due to the unbalanced sampling design (R^2 of 0.94 and bias of -0.72) showing

the importance of a well distributed training dataset. RaCA local libraries produced the least biased models (0.02) with a slightly inferior R^2 (0.92) showing the importance of good geographic coverage. The effect of spiking produced an improvement in terms of R^2, RMSE and bias; however a closer look at the model developed with only the spike samples, showed an addition of positive bias to the final predictions.

3.2 Introduction

A major concern when developing any kind of predictive model is its ability to produce an accurate result in an external sample, in other words, to be able to generalize. A frequently used rule of thumb in predictive modelling tells us that a higher model complexity will most probably result in lower predictability (overtfitting), as Occam's razor states *"Nuncquam ponenda et pluralitas sin necesitate"*. However as the target variable increases its complexity, there is an implicit trade-off, as the need for a more specific and complex model will usually produce a more accurate result (Domingos, 1998).

The main problem for soil properties modelling is that, as has been described earlier in this thesis, soil is in fact a very complex entity, and if the intention is to predict the many properties using hypervariate information like Vis-NIR spectra, suddenly to generalize becomes a big problem. A second key element is, as it was observed in Chapter 2, the spectral similarity of the training samples compared with any external validation dataset.

3.2.1 Spectral libraries

Considering that the performance of a predictive model is closely related with the spectral variability in the training dataset, the creation of large spectral libraries covering an important range of soils has being a common strategy used by several authors and organizations (Shepherd and Walsh, 2002; Stevens et al., 2006; Brown, 2007; Viscarra Rossel, 2009; Viscarra Rossel, 2011; Knadel et al., 2012; Viscarra Rossel et al., 2016).

Despite the effectiveness of this method, as Stevens et al. (2013b) reviews, there are important drawbacks in the creation and use of large spectral libraries, since they are usually compiled following different sampling protocols and additional variability introduced by methodological factors is introduced in the dataset e.g., temperature at the moment of the observation (Figure 3-1), changing reference protocols and different nature of the soil spectra in terms of its origin and composition among many others.

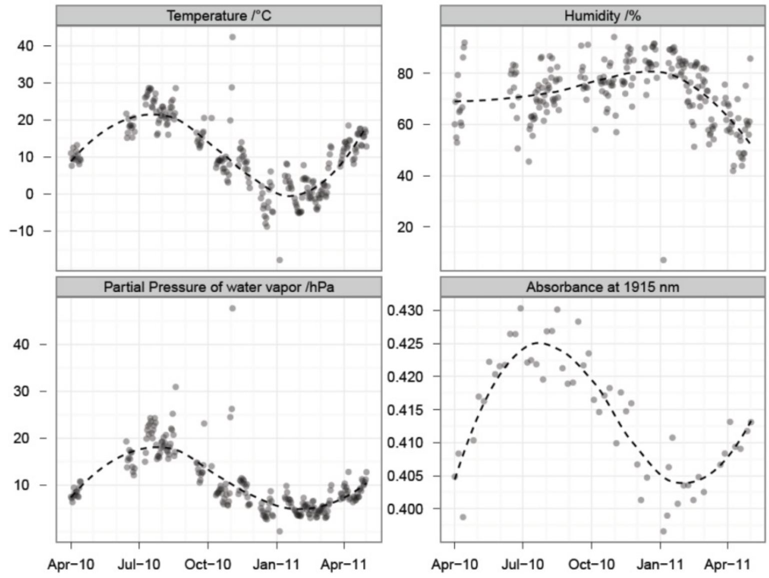


Figure 3-1 Effect of different parameters on absorbance at 1915 nm, *extracted from* Stevens et al. (2013b).

As a result, the use of large spectral libraries is not a simple task and involves a deeper analysis of the factors that may influence the final predictions. Based on the challenging creation of successful predictive models by using spectral libraries, many other alternatives have been devised, all of them trying to deliver accurate predictions of soil properties in soil samples with different composition and from different geographic locations.

Some of the latest modelling approaches involve the stratification of datasets by external metadata e.g., location, depth, land-use or type of spectra (Araújo et al., 2014; Viscarra

Rossel et al., 2016). Others, include the creation of mixed models with spectra plus extra predictors or "auxiliary predictors" like measured soil properties (Stevens et al., 2013a). And lately some of them including or "spiking" a comparatively small representative subsample of the external dataset into the training sample in order to add the variability of the prediction dataset in the model (Guerrero et al., 2010; Guerrero et al., 2016).

3.2.2 Case scenario: RaCA dataset

In 2010, the United States Department of Agriculture (USDA) and the National Resources Conservation Service (NRCS) started an initiative named Rapid Assessment of U.S. Soil Carbon or RaCA, which intended ultimately to estimate the amount and distribution of carbon stocks by means of Vis-NIR information from 144,833 samples taken up to a depth of 100 cm and strategically distributed in the conterminous U.S. (Soil Survey Staff, 2013).

In addition to the previous, the NRCS also has stored an increasing amount of soil samples from different dates and locations in the U.S. All of these samples have also being carefully archived and Vis-NIR scanned, forming an even bigger but more heterogeneous dataset. This type of dataset is commonly called a "Legacy" dataset.

As commented by Viscarra Rossel et al. (2016), there is a need for more research on how to optimally use large and heterogeneous soil spectral databases for local predictions. Also, and as it was discussed in the rationale, a key characteristic of soil spectral information is its transportability, in terms of predicting soil properties in new samples. Therefore inspired by the need of more research related to the use of large scale spectral databases and with the aim of elucidating the role of soil spectral similarities in the modelling process, this chapter will assess the performance of the use of large spectral libraries and state-of-the-art spectral modelling techniques for the prediction of percentage of total carbon (%TC) in soil samples from a particular Region of U.S. (Local samples). The target site covers part of the states of Texas, New Mexico, Oklahoma and Louisiana.

The objectives of this chapter are first to test how different modelling techniques perform with large libraries, both with homogeneous sampling and analysis protocol (RaCA dataset) and a more heterogeneous one (Legacy dataset) and second, to observe how the spectral variability and geographic location affects the spectral modelling procedure.

3.3 Materials and methods

3.3.1 Datasets

The first dataset was generated by the NRCS and several entities under the National Cooperative Soil Survey Data (NCSS) scheme with a total of 19,837 archived samples from different locations, depths and dates in the conterminous U.S. scanned in NRCS laboratories. The dataset considers Vis-NIR spectra and %TC and it will be referred to the "Legacy dataset".

The second dataset was generated by the NRCS and university contributors in the NCSS for the RaCA project as specified in Soil Survey Staff (2013) with a total of 7,781 valid observations with spectra and %TC, referred to as the "RaCA dataset".

3.3.2 Area of study

In order to test the methodologies in a local (state wide) area, LUGR region number 9 was selected from the RaCA dataset (Figure 3-2, Figure 3-3 and Figure 3-4). As a reference, the target region comprises a succession of Aridisols, Mollisols, Alfisols and Vertisol from north-west to south-east in logical relation with its physiographic subdivisions namely High plains, Plateaus (e.g., Edwards plateau) and Coastal prairies which respectively range between 2300 to 0 m.a.s.l. (Wermund, 1996; National soil survey center staff, 2013; Soil Survey Staff, 2013).

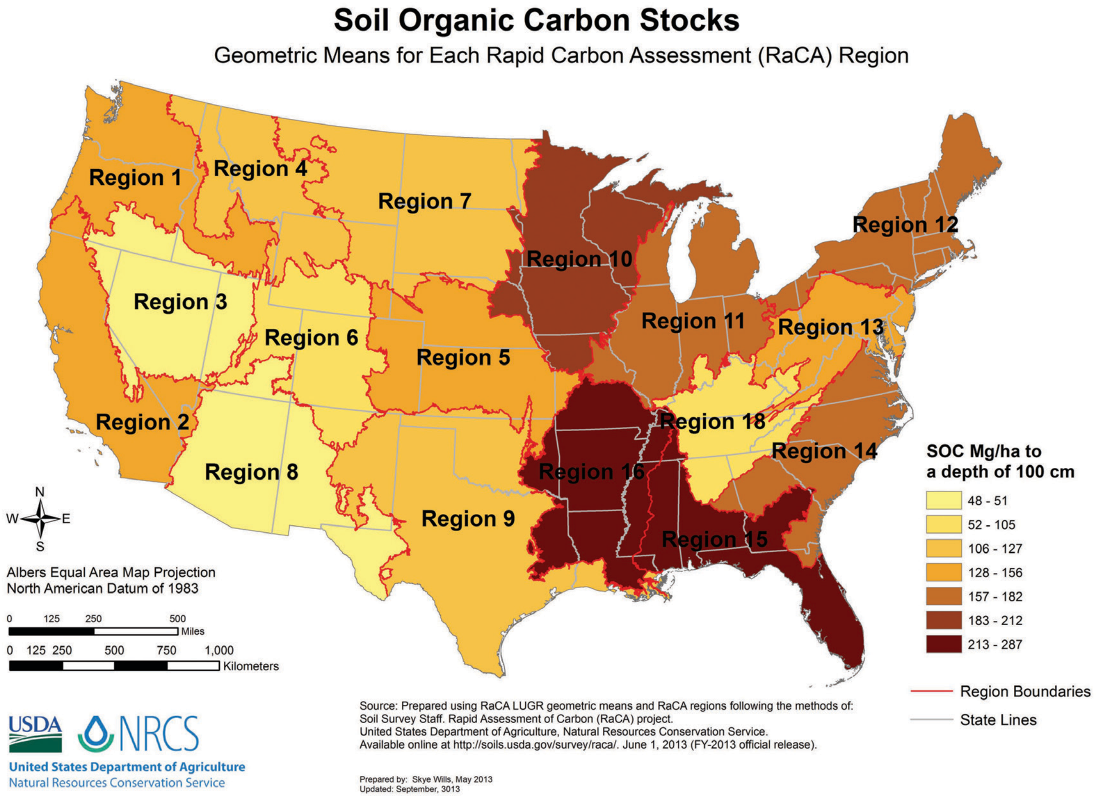

Soil Organic Carbon Stocks
Geometric Means for Each Rapid Carbon Assessment (RaCA) Region
Region 1
Region 2
Region 3
Region 4
Region 5
Region 6
Region 7
Region 8
Region 9
Region 10
Region 11
Region 12
Region 13
Region 14
Region 15
Region 16
Region 18
N
W
E
S
Albers Equal Area Map Projection
North American Datum of 1983
0 125 250 500 Miles
0 125 250 500 750 1,000 Kilometers
SOC Mg/ha to a depth of 100 cm
48 - 51
52 - 105
106 - 127
128 - 156
157 - 182
183 - 212
213 - 287
Region Boundaries
State Lines
USDA NRCS
United States Department of Agriculture
Natural Resources Conservation Service
Source: Prepared using RaCA LUGR geometric means and RaCA regions following the methods of:
Soil Survey Staff. Rapid Assessment of Carbon (RaCA) project.
United States Department of Agriculture, Natural Resources Conservation Service.
Available online at http://soils.usda.gov/survey/raca/. June 1, 2013 (FY-2013 official release).
Prepared by: Skye Wills, May 2013
Updated: September, 2013

Figure 3-2 RaCA LUGR regions of RaCA project, *extracted from* Soil Survey Staff (2013).

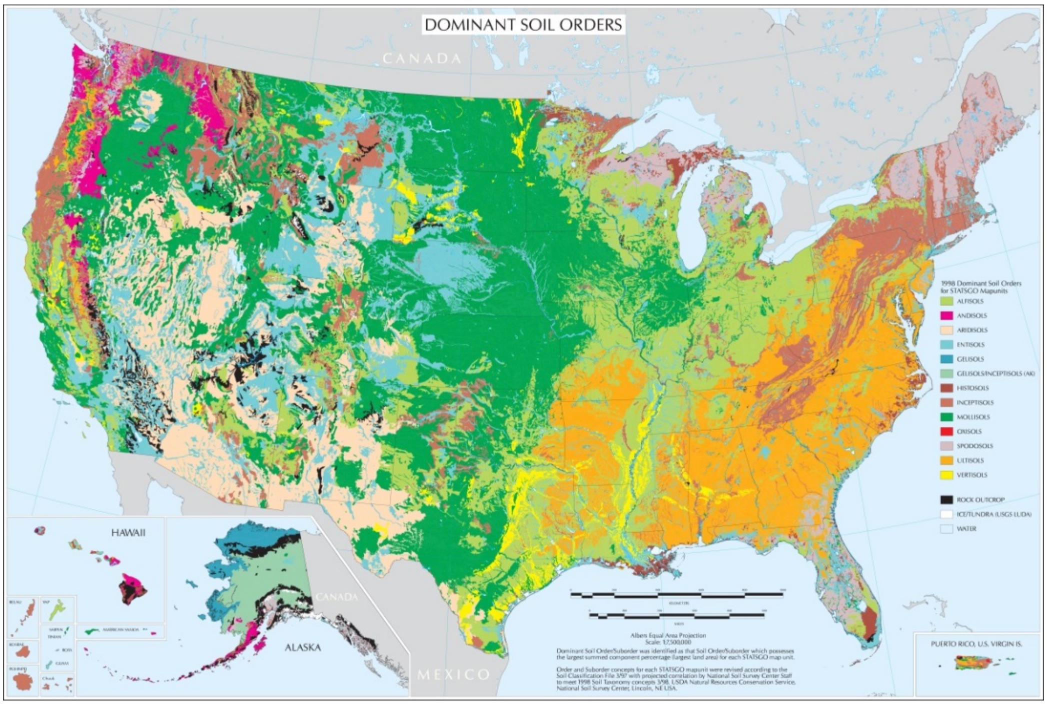
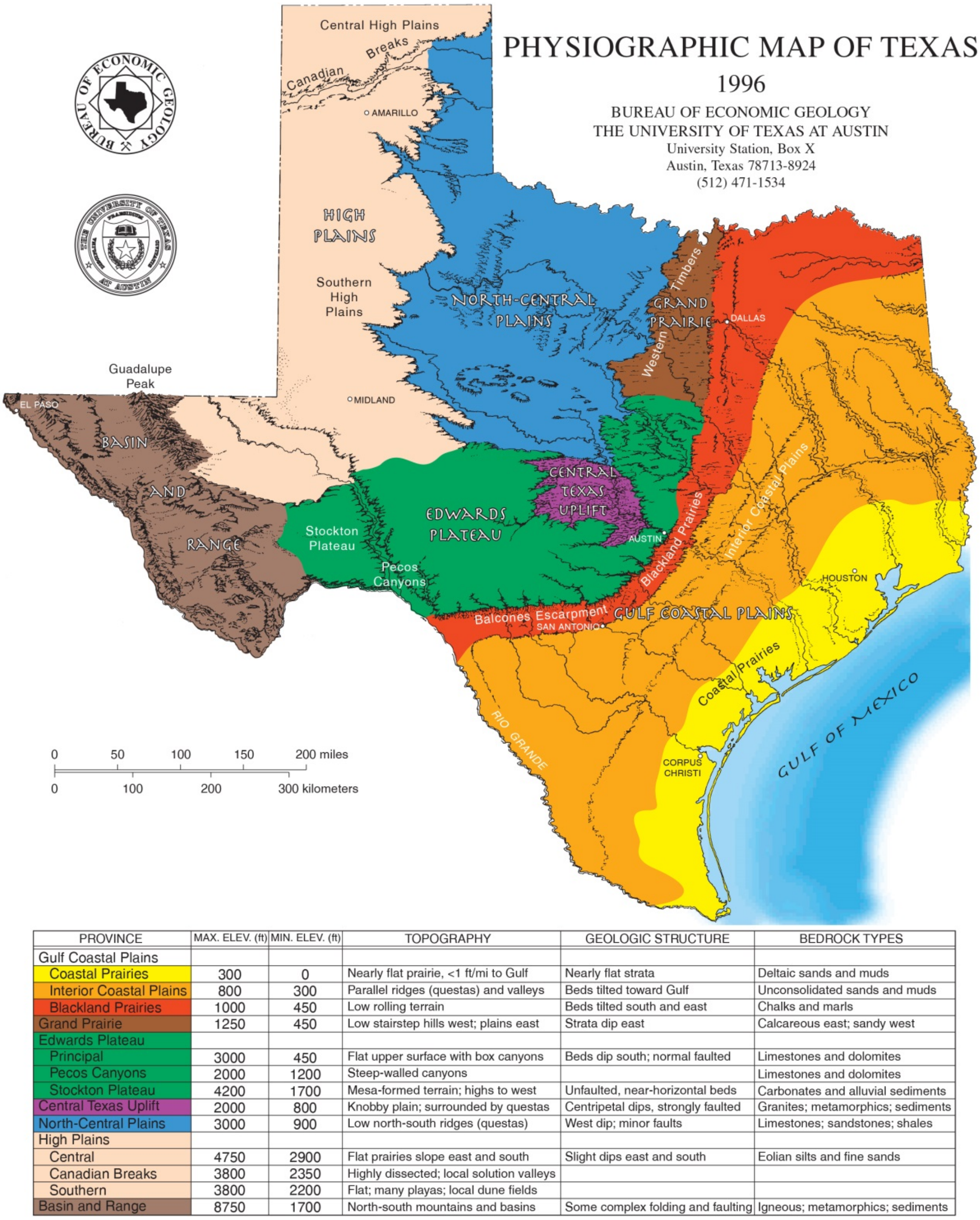

DOMINANT SOIL ORDERS
CANADA
1998 Dominant Soil Orders for STATSGO Mapunits
ALFISOLS
ANDISOLS
ARIDISOLS
ENTISOLS
GELISOLS
GELISOLS/INCEPTISOLS (AK)
HISTOSOLS
INCEPTISOLS
MOLLISOLS
OXISOLS
SPODOSOLS
ULTISOLS
VERTISOLS
ROCK OUTCROP
ICE/TUNDRA (USGS LUDA)
WATER
HAWAII
ALASKA
CANADA
MEXICO
PUERTO RICO, U.S. VIRGIN IS.
Albers Equal Area Projection
Scale: 1:7,500,000
Dominant Soil Order/Suborder was identified as that Soil Order/Suborder which possesses the largest summed component percentage (largest land area) for each STATSGO map unit.
Order and Suborder concepts for each STATSGO mapunit were revised according to the Soil Classification File 3/97 with projected correlation by National Soil Survey Center Staff to meet 1998 Soil Taxonomy concepts 3/98. USDA Natural Resources Conservation Service. National Soil Survey Center, Lincoln, NE USA.

Figure 3-3 Dominant U.S. Soil taxonomy soil orders, *extracted from* National soil survey center staff (2013).



PROVINCE	MAX. ELEV. (ft)	MIN. ELEV. (ft)	TOPOGRAPHY	GEOLOGIC STRUCTURE	BEDROCK TYPES
Gulf Coastal Plains					
Coastal Prairies	300	0	Nearly flat prairie, <1 ft/mi to Gulf	Nearly flat strata	Deltaic sands and muds
Interior Coastal Plains	800	300	Parallel ridges (questas) and valleys	Beds tilted toward Gulf	Unconsolidated sands and muds
Blackland Prairies	1000	450	Low rolling terrain	Beds tilted south and east	Chalks and marls
Grand Prairie	1250	450	Low stairstep hills west; plains east	Strata dip east	Calcareous east; sandy west
Edwards Plateau					
Principal	3000	450	Flat upper surface with box canyons	Beds dip south; normal faulted	Limestones and dolomites
Pecos Canyons	2000	1200	Steep-walled canyons		Limestones and dolomites
Stockton Plateau	4200	1700	Mesa-formed terrain; highs to west	Unfaulted, near-horizontal beds	Carbonates and alluvial sediments
Central Texas Uplift	2000	800	Knobby plain; surrounded by questas	Centripetal dips, strongly faulted	Granites; metamorphics; sediments
North-Central Plains	3000	900	Low north-south ridges (questas)	West dip; minor faults	Limestones; sandstones; shales
High Plains					
Central	4750	2900	Flat prairies slope east and south	Slight dips east and south	Eolian silts and fine sands
Canadian Breaks	3800	2350	Highly dissected; local solution valleys		
Southern	3800	2200	Flat; many playas; local dune fields		
Basin and Range	8750	1700	North-south mountains and basins	Some complex folding and faulting	Igneous; metamorphics; sediments

Figure 3-4 Physiographic map of Texas, *extracted from* Wermund *(1996)*.

3.3.3 Spectral pre-treatments

All spectra undertook a Savitzky-Golay Filter (Savitzky and Golay, 1964) with a 2nd order polynomial and a window size of 11 observations, every 5th wavelength between 500 and 2,450 nm, followed by a Standard Normal Variate (SNV) transformation (Barnes et al., 1989). The outliers of the Legacy dataset were removed with a 5 component PLS model and a chi-square interval with a value of 0.9 (Filzmoser et al., 2005).

3.3.4 Spectral modelling methodologies

As commented in the introduction, a set of different modelling approaches was used for the prediction of target samples in the selected area of study. First, in order to compare all the different methods, a 75% Latin hypercube sample (Minasny and McBratney, 2006; Roudier, 2011) from the RaCA dataset samples within the geographic extent of the target site (274 samples in LUGR 9) was selected i.e., 205 samples, and it was called “RaCA local training dataset”, and the remaining 25% was left aside as an independent validation dataset (69 samples) and called “RaCA validation dataset” which was used for testing the performance of all the modelling approaches.

A Cubist algorithm with 5 committees was used to create the spectral models (Kuhn et al., 2014) and with the purpose of providing a range of predictions (uncertainty), 20 iterations were performed for each of the modelling approaches.

Each of the iterations was made using a 90% random sample from the respective dataset, and by doing this, each iteration will produce a different model, hence a range of predictions.

3.3.4.1 Legacy approach

The first approach involved the use of a totally external dataset for the training of the predictive models, in this case the legacy dataset, with a total of 17,854 samples was used as training (Figure 3-5).

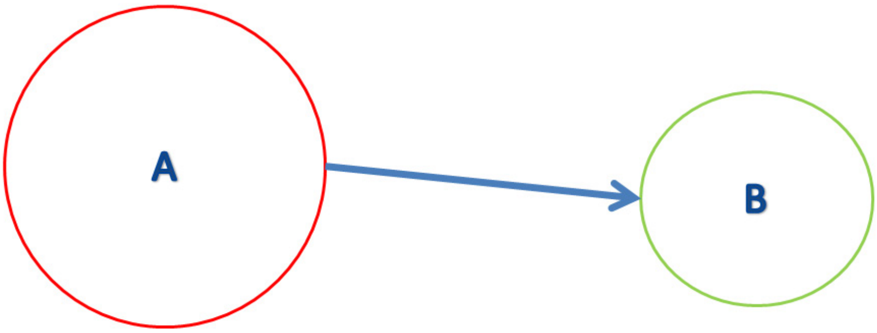


Figure 3-5 Legacy approach scheme. A is Legacy dataset and B is Validation dataset.

3.3.4.2 Legacy in-hull approach

The second approach involved the use of only legacy samples within the RaCA validation dataset convex hull spectral space as a training dataset (i.e., convex hull of two first principal components of the projected datasets), with a total of 11,857 samples. It is important to note that more than half of the samples of the legacy dataset were inside the convex hull of the RaCA validation dataset, reflecting the high spectral diversity of the selected validation area (Figure 3-6).

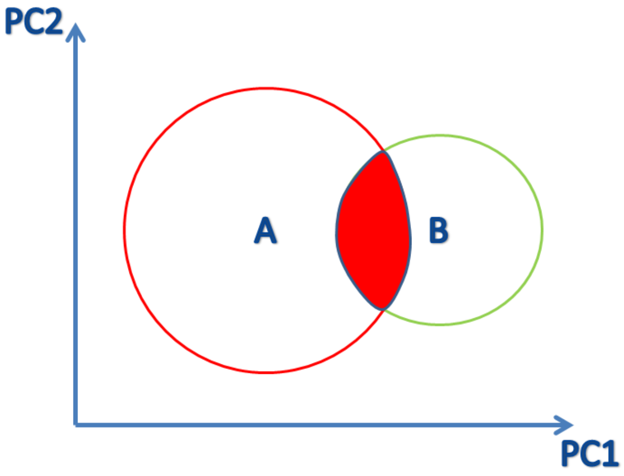


Figure 3-6 Legacy in-hull approach scheme. A is Legacy dataset and B is Validation dataset, section in red corresponds as the in-hull dataset used as training.

3.3.4.3 Legacy in-hull plus spike approach

The third approach involved the use of legacy samples within the RaCA validation dataset spectral convex hull plus a spike of size equal to 10% Latin hypercube sample of RaCA training dataset within the geographical extent of LUGR 9, making a total of 11,877 samples (spike of 20 samples from LUGR 9) (Figure 3-7).

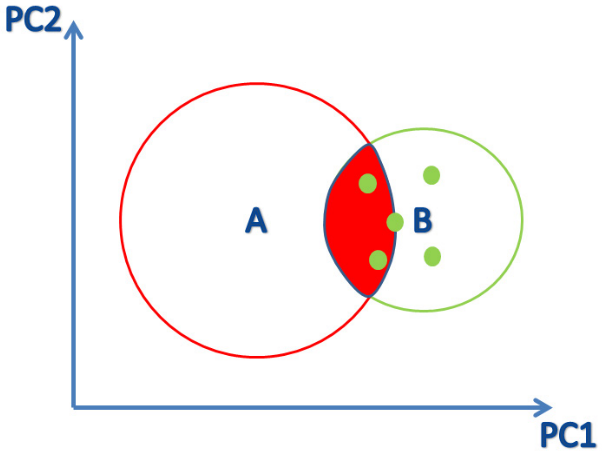


Figure 3-7 Spiked legacy dataset within RaCA convex hull scheme. A is Legacy dataset and B is Validation dataset, area filled in red correspond to in-hull dataset and circles in green to RaCA spike.

3.3.4.4 Legacy Latin hypercube sample within convex hull approach

In order to test how a smaller and more balanced set of samples performed; a 10% Latin hypercube sample of the legacy dataset within the convex hull of the RaCA validation dataset was used with a total of 1,187 samples as presented in Figure 3-8.

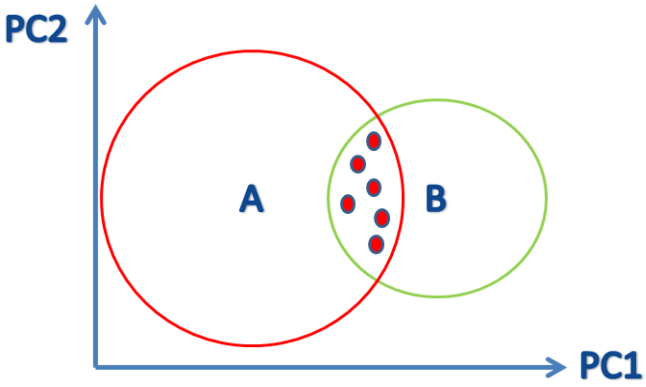


Figure 3-8 Latin hypercube sample within RaCA convex hull scheme. A is Legacy dataset and B is Validation dataset, circles filled in red correspond to the Latin hypercube in-hull dataset used as training.

3.3.4.5 Legacy Latin hypercube within convex hull sample plus a spike approach

The spike sample presented in section 3.3.4.3 was added to the previous dataset, with a total of 1,207 samples. The reason for using a smaller sample from the legacy dataset was to increase the relation spike/legacy as suggested by Guerrero et al. (2010).

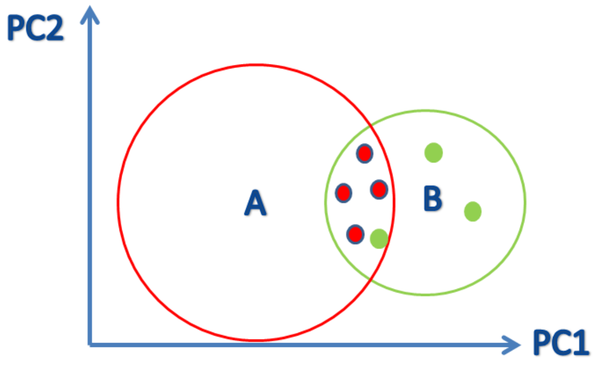


Figure 3-9 Latin hypercube spike of RaCA plus legacy scheme. A is Legacy dataset and B is Validation dataset, circles filled in green correspond to the latin hypercube spike dataset added to the legacy training.

3.3.4.6 RaCA model approach

With the purpose of testing how the RaCA dataset performed as a Global library by itself, RaCA training dataset (with a total of 6,941 samples) was used to predict the target samples.

3.3.4.7 RaCA in-hull model approach

The seventh approach, included RaCA training within the RaCA validation datasets spectral space i.e., convex hull, and this dataset contained 5,807 samples.

3.3.4.8 RaCA local model approach

With the purpose of testing how the RaCA dataset performed as a geographically local library, RaCA training samples in LUGR 9 (a total of 205 samples) were used to predict the RaCA validation dataset.

3.3.4.9 RaCA local in-hull model approach

The ninth approach used RaCA training data within the target sites geographical extent and also spectral space i.e., within LUGR 9 and also convex hull, with a total of 163 samples.

3.3.4.10 Model developed with only information from a spike

Finally, a model using just the samples from the spike was built i.e., 20 samples from the RaCA training dataset within LUGR 9 and within the target sites convex hull. This approach was used to observe the effect of the spike by itself in the overall model performance.

3.4 Results and Discussion

3.4.1 TC in the datasets

An unusually highly left skewed distribution of TC values in the RaCA training dataset was observed. The explanation for this uncommon distribution resided in the low predictability of previously made Vis-NIR models (NRCS team) in organic samples, reason why the sampling effort was biased towards organic samples, with the final purpose to have real values in those samples, accordingly, the validation dataset was highly left skewed as well Figure 3-10.

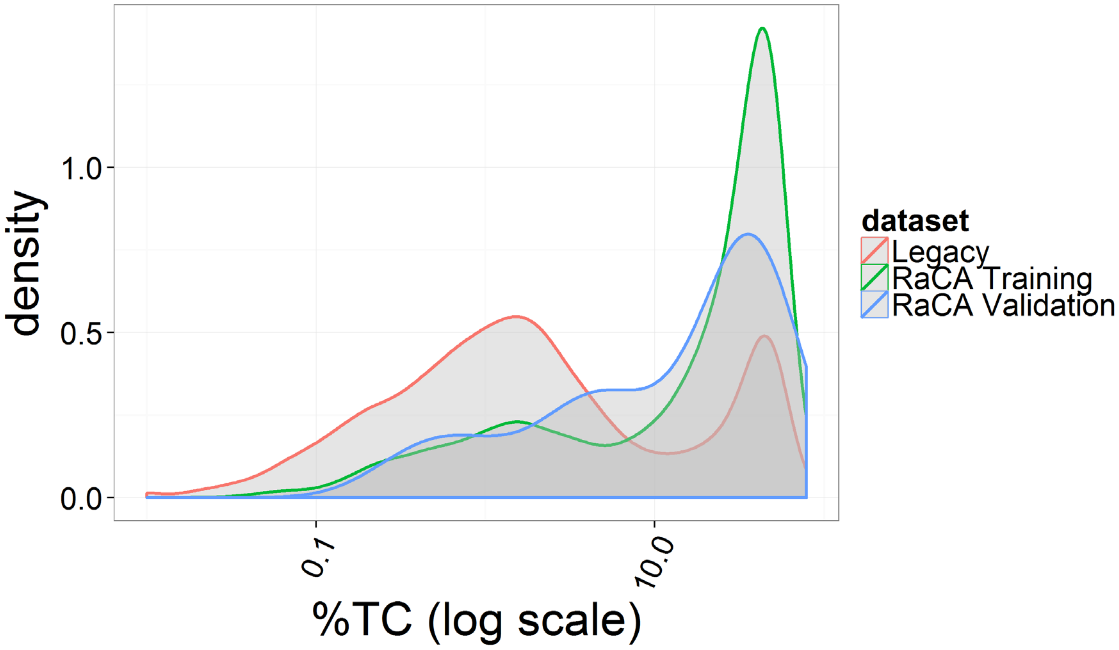


Figure 3-10 Kernel density distribution for Total Carbon % values of datasets.

In respect to the legacy dataset, a double peaked distribution was observed, due to different sampling campaigns. It is common to find mixed distributions in large spectral libraries (Viscarra Rossel et al., 2016). The first pattern responded to a commonly observed distribution with a median centred in lower values (between 1 to 5% of TC) and the second peak (centred in 40 to 70% of TC) related with a sampling campaign expected to cover a bigger proportion of organic horizons; since the metadata information of the legacy dataset was limited only to a few variables, it was technically impossible to accurately determine the reason for this double distribution.

3.4.2 Spectral similarity of the datasets

Figure 3-11 shows the three convex hull regions of the legacy, training and validation datasets, where both legacy and training covered the variability (in terms of the first two principal components explanation i.e. 77%) of the validation dataset. It was also visible that the skewness of the RaCA training dataset was evident even in the spectral space compared with the legacy dataset.

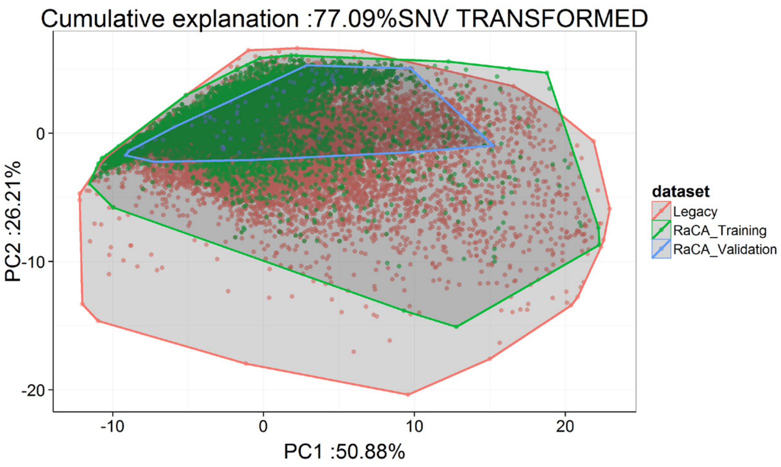


Figure 3-11 Convex hull of first the 2 Principal components of the used datasets.

3.4.3 Overall model performance

Table 3-1 shows three performance parameters i.e., R^2, RSME and bias for all the treatments ordered by their respective RMSE values. In general terms, all the models were accurate and behave as expected from other large scale studies (Stevens et al., 2013b; Viscarra Rossel et al., 2016). It was also evident the effect of spectral similarity i.e., convex hull, where in most models showed improved predictions in terms of R^2 and bias values, when the samples were spectrally closer.

Table 3-1 Different model approaches performance for the mean predictions of 20 iterations ordered by lower to higher RMSE values. *LH*: Latin hypercube sample.

Model	***R2***	***RMSE***	***BIAS***
Spiked legacy in hull	*0.96*	*3.69*	*-0.38*
Legacy	*0.96*	*3.73*	*-0.54*
Legacy in hull	*0.95*	*3.76*	*-0.41*
RaCA in hull	*0.94*	*4.24*	*-0.72*
Spiked LH legacy in hull	*0.94*	*4.44*	*-0.59*
LH legacy in hull	*0.94*	*4.46*	*-0.78*
RaCA	*0.93*	*4.55*	*-0.83*
RaCA Local	*0.92*	*5.02*	*0.02*
RaCA Local in hull	*0.90*	*5.52*	*-0.06*
Spike	*0.79*	*9.46*	*2.90*

The benefits of spectral similarity have been reviewed by several authors and such studies have been critical for the development of accurate spectral distance methodologies (Chang, 1999; Islam et al., 2005; Van der Meer, 2006; Ramirez-Lopez et al., 2013).

With respect of the previous, the latest modelling approaches involving a large spectral variability e.g.,Viscarra Rossel et al. (2016) have revised the improvement in different soil properties predictions from Vis-NIR, after a pre-classification of samples based on the variability explained on their principal components.

3.4.4 Legacy approaches

Among all the models, the ones involving the Legacy dataset (coincidently with the largest amount of samples) were the ones that produced the best R^2 and RMSE values (0.9 and 3.7 respectively). The previously mentioned effect of spectral similarity reduced the bias, consequently Legacy in hull models bias was of 0.41 vs 0.54 in the Legacy models, and the effect of spiking reduced both bias and RMSE values.

Despite the improvement in performance after spiking, a closer look at the model created with only samples from the spike (20 samples from LUGR 9 and in RaCA validation convex hull) showed a strong positive bias, which possibly caused a “masked” improvement in the results by adding samples that caused consistent over predictions (Figure 3-12 and Table 3-1).

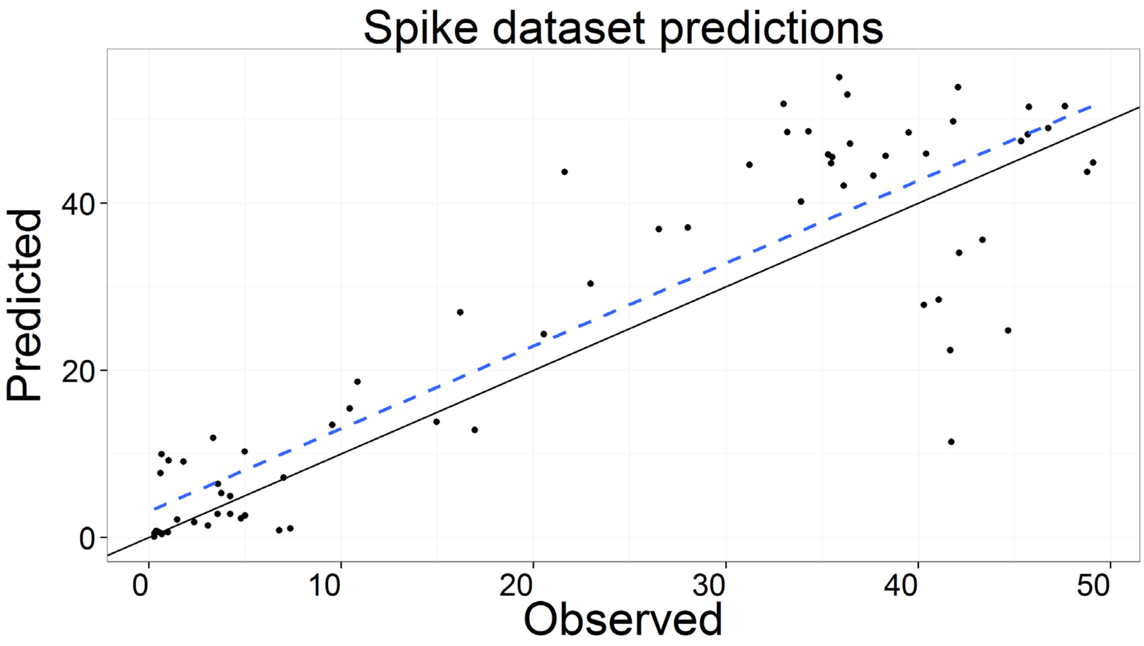


Figure 3-12 Observed vs predicted values (%TC) of model using spike as training. Continuous line represents 1:1 observed - predicted ratio and dashed line represents linear regression of predictions.

In respect of how to select an appropriate Vis-NIR spike, Guerrero et al. (2014) recommended the use of samples evenly distributed in the spectral space, which is the reason why a Latin hypercube algorithm was used for selecting the spike in this study (Minasny and McBratney, 2006).

Nevertheless, due to the skewness of the validation dataset, the Latin hypercube spike sample used here, inherited this behaviour, consequentially producing over-estimations on %TC. It is clear though, how the addition of a minimal amount of samples can produce such a considerable impact in the final result, a fact that makes the use of spikes a delicate task if the prediction of total stocks in large datasets is intended or if the total variability of the target site is not well represented.

In relation to the influence of the legacy samples geographic location, it must be said that a large amount of observations were not geo-referenced, which is why a geographical analysis was not performed to the legacy dataset, in order to avoid misinterpretations.

3.4.5 RaCA approaches

The performance of RaCA as a "Global library" was slightly inferior to the Legacy approaches with higher RMSE and negative bias values. From Figure 3-13 it is possible to see that the bias values in RaCA and RaCA in hull models was related to the high geographical dispersion in both datasets.

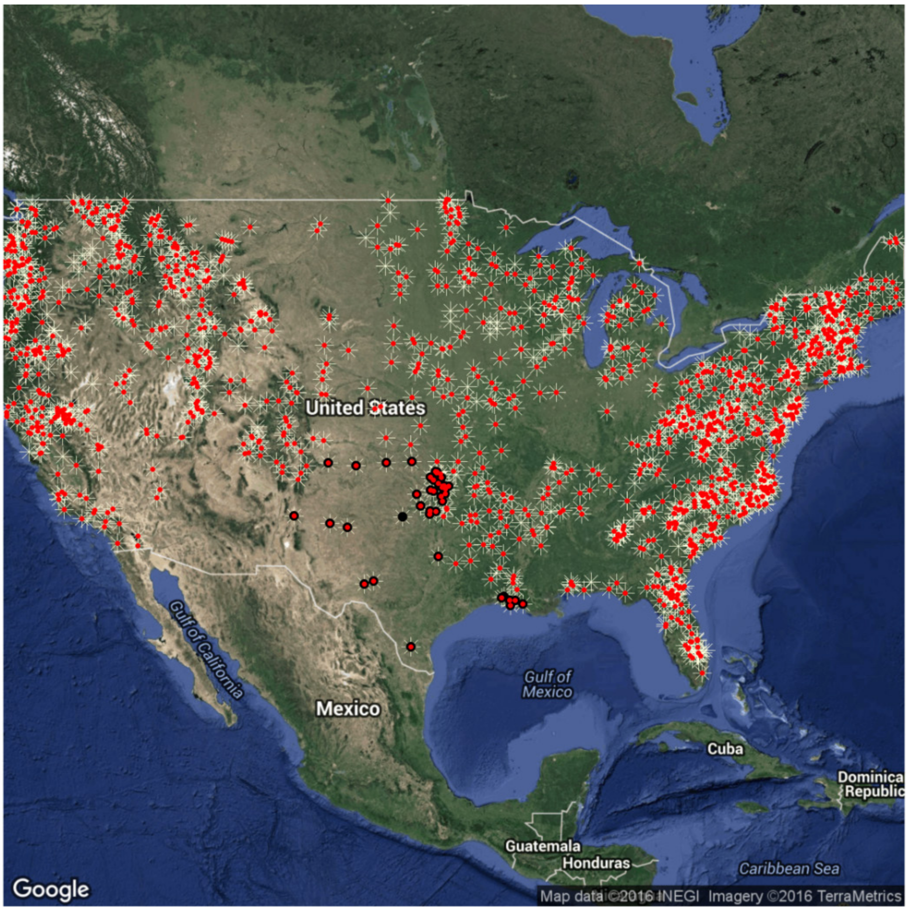


Figure 3-13 RaCA dataset (Light yellow asterisks), RaCA in hull (red circles), RaCA local (black circles) and RaCA local in hull (red over black circles) sampling locations.

The bias values were related to a slight under prediction between 5 to 20% and a bigger one in values higher than 40% TC (Figure 3-14). Nevertheless the models performed reasonably good and comparable to other global approaches e.g., (Guerrero et al., 2010; Guerrero et al., 2016) and to the legacy approaches presented in the previous section.

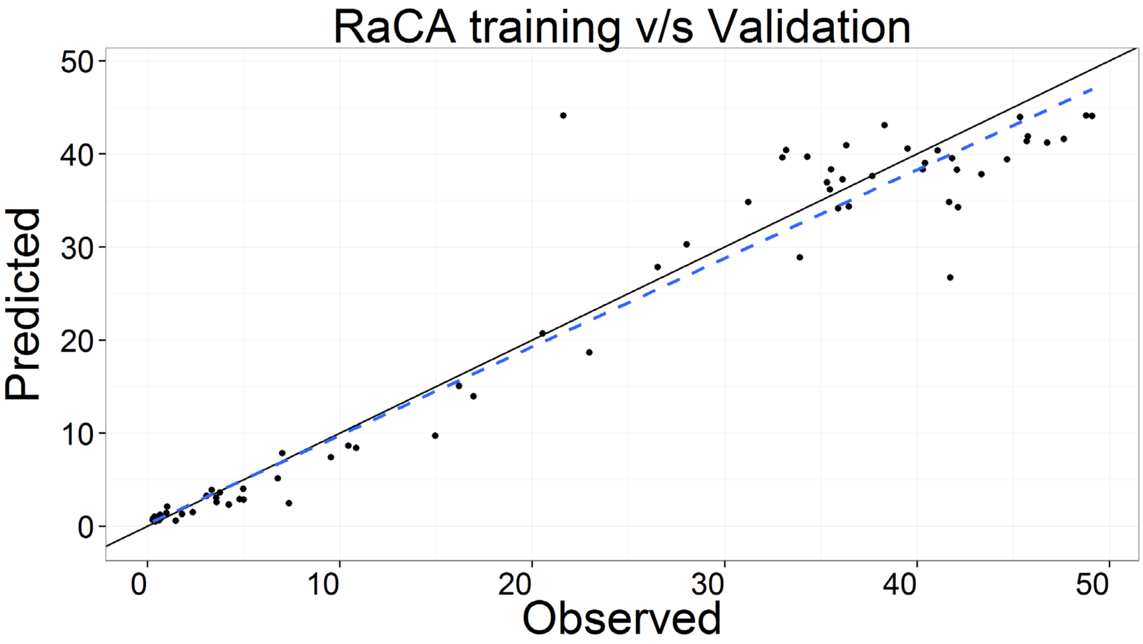


Figure 3-14 Observed vs predicted values of model using RaCA training. Continuous line represents 1:1 observed - predicted ratio and dashed line represents linear regression of predictions.

3.4.6 RaCA local approaches

Among all the models, the ones created with the RaCA dataset from local sites (LUGR 9) were the only ones that produced a significant reduction in bias (Figure 3-15) as expected from their geographic locations (Figure 3-13) and subsequently similar distribution of %TC in training and validation datasets (Figure 3-16).

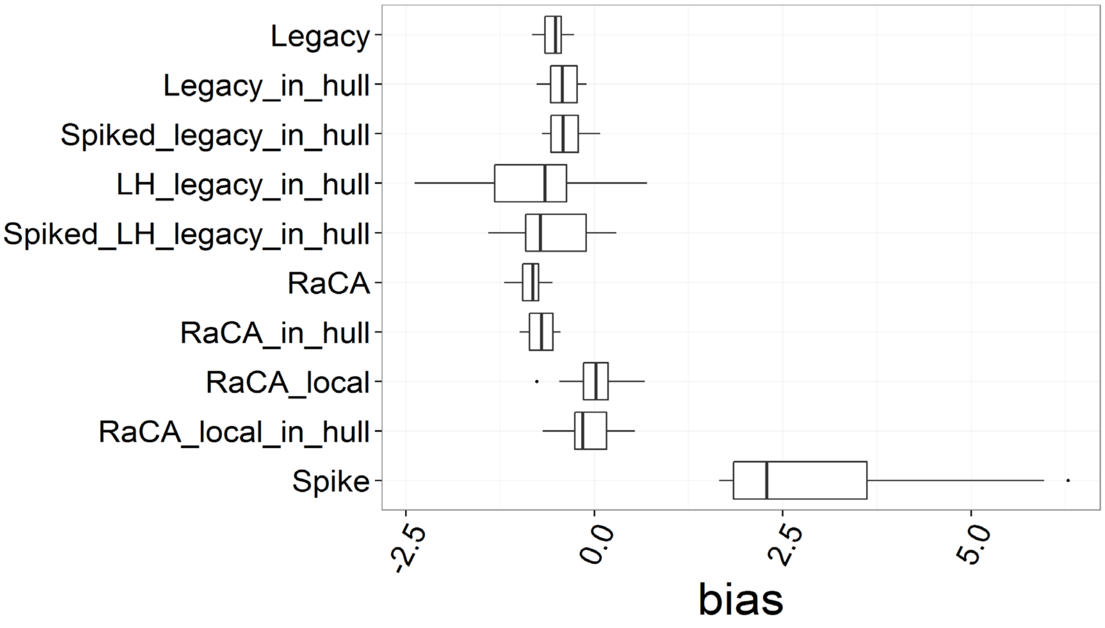


Figure 3-15 Bias values for all the approaches. Boxes represent the values of twenty iterations

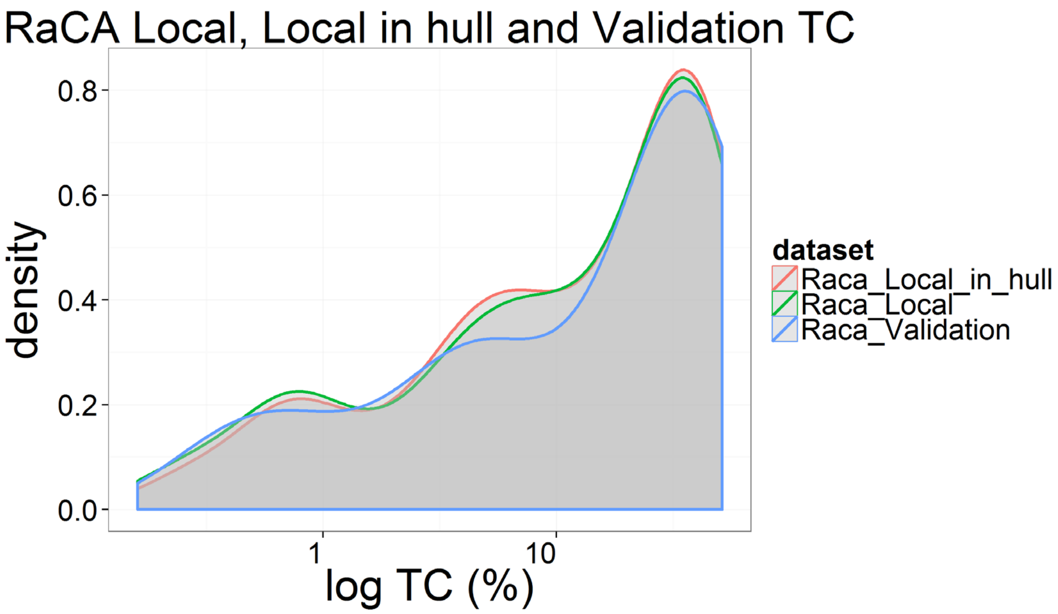


Figure 3-16 TC% for RaCA local, RaCA local in hull and validation dataset.

In a revision of various studies involving the prediction of local samples using large spectral libraries, Stenberg et al. (2010b) noticed the importance of the geographical scale of the calibration, highlighting the influence of the overall variation of the training dataset, implying with this that a large library will produce a less precise outcome.

Despites the benefits of low bias and even though the models were considered very good, their average performance, in terms of R^2 and RMSE, was inferior to the previous approaches. A possible explanation for this slightly lower performance could be due the limited number of samples in relation with the geographical area of the target site (LUGR 9) i.e., 163 and 205 for RaCA local and RaCA local in hull respectively vs more than 15,000 observations in legacy approach for a total area of more than 800,000 km^2.

3.4.7 Spectral variability considerations and discussion

In general, the importance of a spectrally and geographically similar dataset for an accurate prediction of %TC was observed. Also, the fact that a minimal number of samples (spike of 20 samples) could produce a substantial effect over final results, makes the selection of an appropriate spike a critical task which could both benefit or add error to a final result if the target area is not properly characterized e.g., skewed dataset in target area.

This effect can be explained by the fact that when a spectrally different sample is merged into a dataset, the total variance of the spectral space of the training dataset is modified, or in a more intuitive way, the projected convex hull area will change, as observed in Figure 3-17.

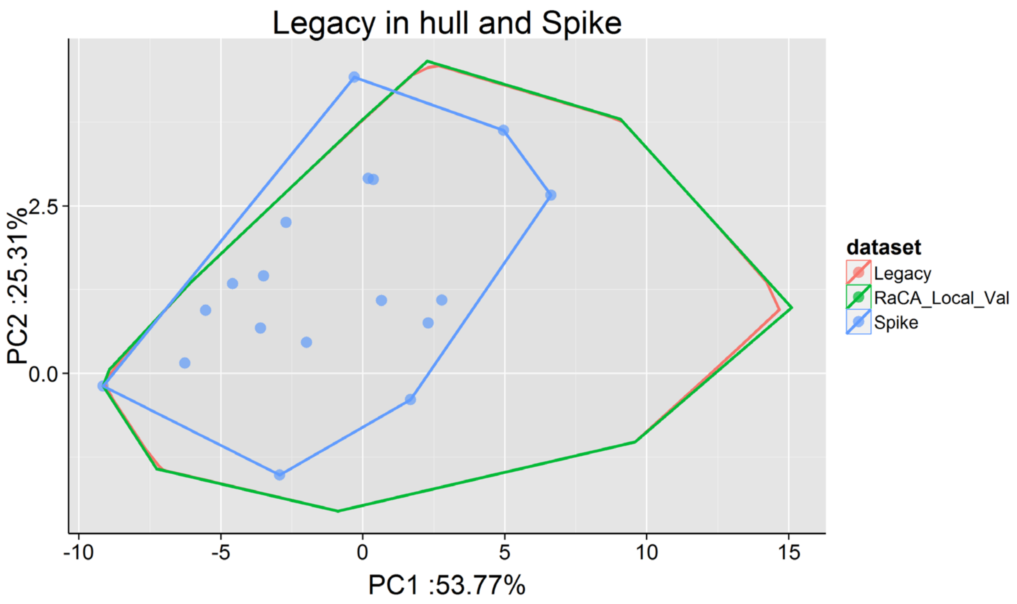


Figure 3-17 Example of change in convex hull in legacy in hull dataset after addition of a 20 samples spike. *Note that legacy in hull dataset (in red) had almost 12,000 samples*.

Further, a possible inconvenience with the addition of large variability into a training dataset just from a few samples resides in the possibility of introducing a large error, if the spike samples contain a big proportion of outliers, because the relative weight of each sample is incremented as they capture variability not considered in the training dataset.

Shepherd and Walsh (2002) observed a similar effect in models predictability when spectrally different samples were introduced to the training dataset (Figure 3-18). They found that when outliers of the local datasets where introduced (equivalent to the boundary samples of local dataset convex hull) in the calibration dataset, the range of the calibration was expanded, and even more when these plus an extra random sample of the local dataset was added, the improvement was more significant. They did not analyse though, the effect in bias that this approach could produce, which was observed in this chapter.

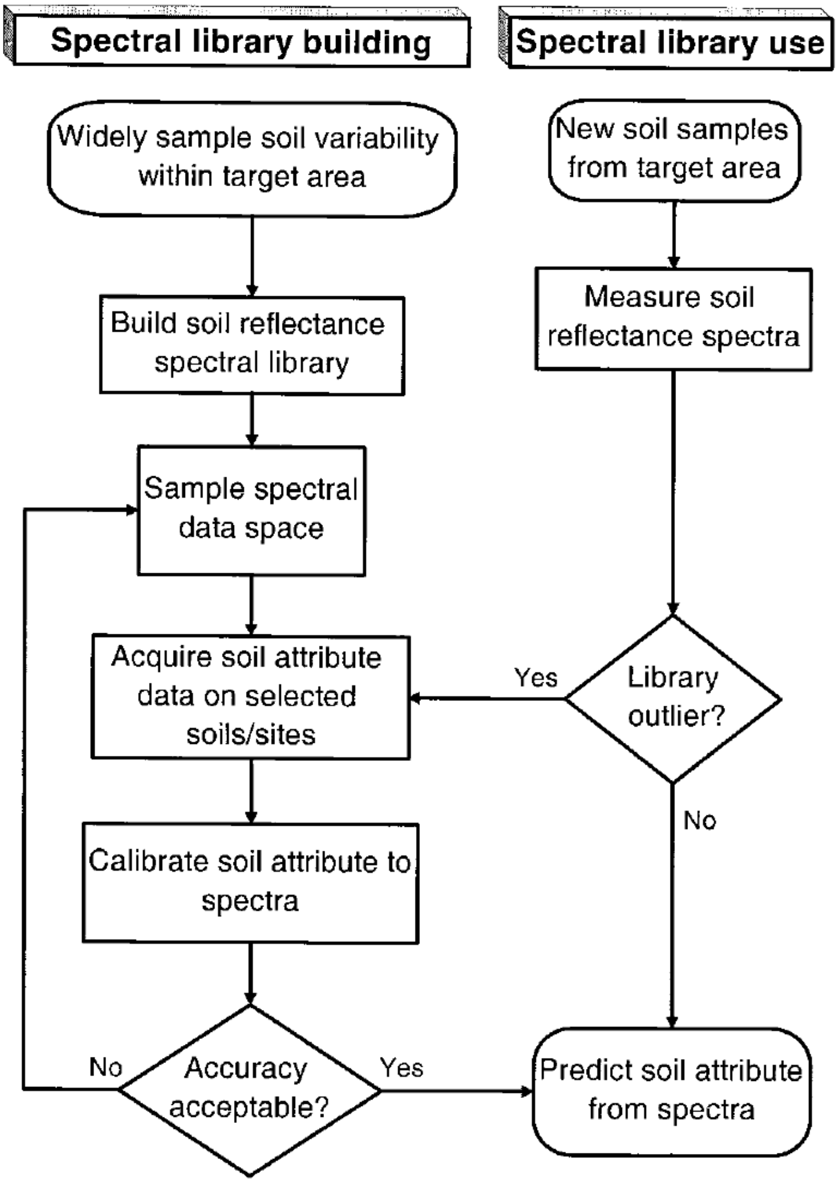

Spectral library building
Spectral library use
Widely sample soil variability within target area
Build soil reflectance spectral library
Sample spectral data space
Acquire soil attribute data on selected soils/sites
Calibrate soil attribute to spectra
Accuracy acceptable?
No
Yes
New soil samples from target area
Measure soil reflectance spectra
Library outlier?
Yes
No
Predict soil attribute from spectra

Figure 3-18 Logical scheme for use of reflectance spectral libraries in a risk-based approach to prediction of soil functional attributes. Extracted from (Shepherd and Walsh, 2002).

On the other hand, the process of adding a complete dataset or "updating" datasets has being studied by Sequeira et al. (2014) and recently by Viscarra Rossel et al. (2016) and it has the advantage of easily identifying samples that could introduce undesired error since the variability of the foreign dataset is well represented.

Finally, the results showed that a good geographic coverage will drastically reduce the bias as observed in the RaCA local approaches. In relation with this, Sudduth and

Hummel (1996) and Wetterlind and Stenberg (2010) had similar conclusions. Wetterlind and Stenberg (2010) reported that models developed from a local dataset, produced better results for six different properties (pH, SOC, clay, sand, silt and ammonium acetate lactate-extractable P) compared with models developed with a spiked national dataset, even if the local dataset had as little as 25 samples.

3.5 Conclusions

Large Vis-NIR spectral libraries in terms of the number of samples, spectral and geographical extent were used for predict TC% on samples from a particular area in the U.S. Even though all the models (except the one created with just 20 samples from the spike) performed good, the results showed that models which considered the larger number of samples and also were spectrally closer i.e., legacy library, legacy within convex hull and spiked legacy within convex hull, produced the best R^2 and RMSE values. The models that included RaCA dataset as a global library were successful, however due to the skewness of the dataset, they produced biased results. Models including RaCA samples from the target site i.e. RaCA local and local in hull were the ones that produced the least bias, due to the geographic and spectral similarity. Finally even though the benefits of spiking have been shown in other similar studies e.g., (Guy et al., 2015; Guerrero et al., 2016), in this work, the improvement on predictions of the spiking procedure was related to an addition of positive bias.

Chapter references

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. European Journal of Soil Science 65(5), 718-729.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Appl. Spectrosc. 43(5), 772-777.

Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140(4), 444-453.

Chang, C.-I., 1999. Spectral information divergence for hyperspectral image analysis, Geoscience and Remote Sensing Symposium, 1999. IGARSS'99 Proceedings. IEEE 1999 International. IEEE, pp. 509-511.

Domingos, P., 1998. Occam's two razors: The sharp and the blunt, KDD, pp. 37-43.

Filzmoser, P., Garrett, R.G., Reimann, C., 2005. Multivariate outlier detection in exploration geochemistry. Computers & Geosciences 31(5), 579-587.

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R.A., Maestre, F.T., Mouazen, A.M., Zornoza, R., Ruiz-Sinoga, J.D., Kuang, B., 2014. Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. European Journal of Soil Science 65(2), 248-263.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-Sinoga, J.D., Zornoza, R., Viscarra Rossel, R.A., 2016. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil and Tillage Research 155, 501-509.

Guerrero, C., Zornoza, R., Gómez, I., Mataix-Beneyto, J., 2010. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. Geoderma 158(1–2), 66-77.

Guy, A.L., Siciliano, S.D., Lamb, E.G., 2015. Spiking regional vis-NIR calibration models with local samples to predict soil organic carbon in two High Arctic polar deserts using a vis-NIR probe. Canadian journal of soil science 95(3), 237-249.

Islam, K., McBratney, A., Singh, B., 2005. Rapid estimation of soil variability from the convex hull biplot area of topsoil ultra-violet, visible and near-infrared diffuse reflectance spectra. Geoderma 128(3-4), 249-257.

Knadel, M., Deng, F., Thomse, A., Greve, M., 2012. Development of a Danish national Vis-NIR soil spectral library for soil organic carbon determination. Digital Soil Assessments and Beyond, 403-408.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2014. Cubist: Rule- and Instance-Based Regression Modeling.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32(9), 1378-1388.

National soil survey center staff, 2013. 1998 Dominant U.S. Soil taxonomy soil orders.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R.A.V., Demattê, J.A.M., Scholten, T., 2013. Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma 199, 43-53.

Roudier, P., 2011. clhs: a R package for conditioned Latin hypercube sampling.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36(8), 1627-1639.

Sequeira, C.H., Wills, S.A., Grunwald, S., Ferguson, R.R., Benham, E.C., West, L.T., 2014. Development and update process of VNIR-based models built to predict soil organic carbon. Soil Science Society of America Journal 78(3), 903-913.

Shepherd, K.D., Walsh, M.G., 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. Soil Science Society of America Journal 66(3), 988-998.

Soil Survey Staff, 2013. Rapid Assessment of U.S. Soil Carbon (RaCA) project, United States Department of Agriculture Natural Resources Conservation Service. Available online. June 1, 2013 (FY2013 official release)

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010b. Visible and Near Infrared Spectroscopy in Soil Science. 107, 163-215.

Stevens, A., Nocita, M., Toth, G., Montanarella, L., van Wesemael, B., 2013a. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. PloS one 8(6), e66409.

Stevens, A., Nocita, M., van Wesemael, B., 2013b. Analysis of large scale soil spectral libraries, International Workshop "Soil Spectroscopy: the present and future of Soil Monitoring". FAO HQ, Rome, Italy.

Stevens, A., van Wesemael, B., Vandenschrick, G., Touré, S., Tychon, B., 2006. Detection of Carbon Stock Change in Agricultural Soils Using Spectroscopic Techniques. Soil Science Society of America Journal 70(3), 844.

Sudduth, K., Hummel, J., 1996. Geographic operating range evaluation of a NIR soil sensor. Transactions of the ASAE 39(5), 1599-1604.

Van der Meer, F., 2006. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. International Journal of Applied Earth Observation and Geoinformation 8(1), 3-17.

Viscarra Rossel, R.A., 2009. The Soil Spectroscopy Group and the development of a global soil spectral library, EGU General Assembly Conference Abstracts, pp. 14021.

Viscarra Rossel, R.A., 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. Journal of Geophysical Research 116(F4).

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world’s soil. Earth-Science Reviews.

Wermund, E., 1996. Physiographic map of Texas. University of Texas at Austin. Bureau of Economic Geology.

Wetterlind, J., Stenberg, B., 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. European Journal of Soil Science 61(6), 823-843.

Chapter 4 *Soil Morphology*

Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles

"I take a pride in probing all your secret moves

my tearless retina takes pictures that can prove"

Judas Priest, 1982

Chapter 4 published as:

Fajardo, M., McBratney, A., Whelan, B., 2016. Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles. Geoderma 263, 244-253.

4.1 Summary

In the past decades the use of Visible and Near Infrared (Vis–NIR) spectra information applied to soil science studies has seen an exponential growth, especially in predicting commonly used soil properties. We used the ability of Vis–NIR for detecting physico-chemical characteristics along with fuzzy clustering techniques to discriminate spectrally homogeneous zones in soil cores and applied a Digital Gradient (*DG*) to define its boundaries i.e., Spectrally derived horizon (*SPD hor)*. We tested this methodology in 59 air dried soil cores varying between 85 and 130 cm depth from the Hunter Wine Country Private Irrigation District (HWCPID), New South Wales (NSW), Australia. We observed that *SPD hor* had great similarity with traditional horizons. The *SPD hor* were more homogeneous in terms of Vis–NIR spectral variability and also offered more information about the relationship between the different spectral classes. Because of the intrinsic characteristics of the methodology it can be easily applicable with or in conjunction with other proximal sensing devices which can add further detail when recognizing soil morphological horizons.

4.2 Introduction

It has been almost 100 years since Professor Curtis Marbut stated that soil studies would not thrive as a science until a generally accepted classification system was developed, suggesting with this, the use of soil horizons as a key element of it (Bockheim et al., 2005; Hartemink and Minasny, 2014). Diagnostic soil horizons have been commonly accepted since then, however it is of common knowledge in the soil science community that the identification of soil horizons and their boundaries could in many situations be inaccurate or biased due to varying description criteria. Furthermore, to classify a diagnostic horizon could require additional laboratory analysis (Weindorf et al., 2012) and given analytical procedures may change in time, this could eventually lead to biased observations (Ciampalini et al., 2013) and in the end misleading interpretations.

For these reasons there is a general challenge in homogenizing soil description criteria and a considerable amount of resources exclusively assigned for this purpose worldwide e.g., Soil Taxonomy, World Reference Base for Soil Resources, Australian Soil Classification, and German Soil Classification, among others (Schoeneberger, 2002; Ad-Hoc-AG, 2005; Jahn et al., 2006; CSIRO, 2009).

As noted by Hartemink and Minasny (2014) soil science is witnessing a historic moment, where a vast amount of new technologies, Vis–NIR stands as one of the most widely used in both remote sensing and proximal sensing.

One of the biggest advantages in using Vis–NIR is that it can easily capture a great part of the physico-chemical variability of the sample which can be used later when comparing between different types of materials.

The objective of the present work is to use Vis–NIR to recognize different materials in soil profiles and to apply a methodology for detecting their relative patterns in depth, to finally establish in a quantitative way, boundaries between homogeneous groups of those materials i.e., soil horizons. Previous studies have used quantitative approaches to distinguish between different soil materials and/or soil horizons (Rooney and Lowery, 2000; Grunwald et al., 2001; Ben-Dor et al., 2008; Weindorf et al., 2012). The main contribution of the present work resides in the creation of a semi-automated soil morphological description procedure where the final SPD *hor* are comparable with others through their membership to global spectral classes which themselves work as a basic example of a classification system.

4.3 Materials and Methods

4.3.1 Study area

The area of study was located approximately 140 km north of Sydney in the HWCPID in the lower Hunter Valley (Figure 4-1). Geologically the area is situated in the Sydney basin, a depositional area formed by both Permian and Triassic materials with thick successions of mainly siliciclastic rocks demonstrating a rhythmic pattern of sedimentation followed by uncommon volcanic units and carbonate rocks in a few areas (Percival, 2012). The dominant soil types according to the Australian Soil Classification

(Isbell, 2002) are Red and Brown Dermosols (Depending in the base saturation value, equivalent to some Udults, Udalfs and Udepts in Soil Taxonomy) and on some hill summits Red Calcarosols (equivalent to some Typic Calciudepts in Soil Taxonomy) (Odgers et al., 2011b).

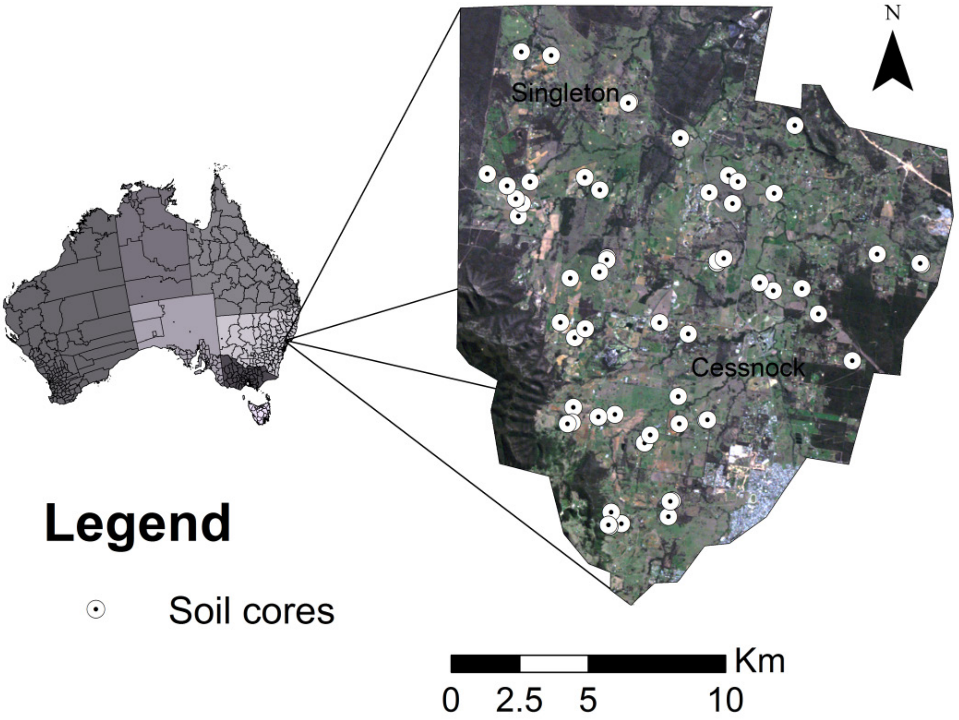


Figure 4-1 Lower Hunter valley study area and sample locations.

4.3.2 Sampling Design

The dataset consisted of 59 soil cores varying between 85 and 130 cm depth taken 50m away from a previous soil survey which followed a Latin hypercube sampling design where compound topographic index, parent material and normalized difference vegetation index were used as environmental variates, in order the maximize the variability of the samples (Minasny and McBratney, 2006). The cores were air dried and scanned with an ASD Agrispec 350-2500 spectrometer using a Spectralon® panel as a reference, every 2 cm resulting in a datasets of 3190 separate soil scans, additionally the soil cores were morphologically described following CSIRO (2009) specifications (Figure 4-1).

4.3.3 Processing of Vis-NIR spectra for SPD hor detection

The following treatments were employed on the dataset in the order below.

1. Step correction between Vis-NIR sensors overlap in bands 1000 nm and 1800 nm.
2. Selection of spectral region between 500 nm and 2450 nm.
3. Conversion to absorbance from raw reflectance data.
4. A second order Savitzky-Golay filter with a smoothing window of 11 bands to each spectrum.
5. Based on the fact that soil spectral features change smoothly with depth we used a running median smoother on each wavelength of the spectrum depth-wise using a smoothing filter described in Hardle and Steiger (1995) and implemented by Martin Maechler in R language (R Core Team, 2013). The smoother basically works as a moving window of variable size throughout the series of numbers i.e., the values of each band through the soil profile. The selected size of the window was 10cm (5 observations every 2cm) after considering the observed spectral variation in the sampled soil cores and the different windows size tested.

6. Selection of every 10th band in order to reduce correlation between variables and high dimensionality.
7. Standard Normal Variate transformation of the spectra.
8. Outlier detection using a Mahalanobis distance criterion (Filzmoser et al., 2005), cores with more than 10 outliers were excluded from the following analyses.

A Principal Component analysis was performed to each processed spectrum and the first 5 components were used (> 95 % of variance explained) for the next stage of fuzzy classification.

4.3.4 Fuzzy clustering

A Fuzzy clustering algorithm (Maechler et al., 2014) was performed to the entire dataset. The algorithm aims to minimize the objective function (Equation 4)

$$\sum_{k=1}^{c} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ik}^{r} u_{jk}^{r} d(i,j)}{2 \sum_{j=1}^{n} u_{jk}^{r}} \qquad (4)$$

where μ_{ik} and μ_{jk} are the memberships of samples i and j to class k, n is the number of observations, C is the total number of classes, r is the membership exponent and $d(i,j)$ is the dissimilarity between observations i and j. Note that if r tends to 1 it gives increasingly crisper clustering whereas *r* tends to infinite it leads to complete fuzziness as specified in Maechler et al. (2014).

4.3.5 Classification performance assessment

In order to measure the quality of the fuzzy clustering we used two different metrics. A Fuzzyness Performance Index (*FPI*) which is a measure of the degree of fuzziness of the final classification for an increasing number of classes (Equation 5).

$$\boldsymbol{FPI = 1 - \frac{c}{c-1}\left[1 - \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^2\right]} \qquad (5)$$

and a Confusion index (*CI)* which is the ratio between the second and the highest membership class μ_{max}, which measures the degree of uncertainty of the final classification (Burrough et al., 1997) , having values closer to 0 as a result of classes totally separated and closer to 1 a membership spread between classes or “high confusion” (Equation 6).

$$\boldsymbol{CI = 1 - (\mu_{max} - \mu_{max-1})} \qquad (6)$$

4.3.6 Digital gradients for horizon boundary detection

After performing the fuzzy clustering classification we used a modification of a *DG* first presented in (Powell et al., 1991). The original approach uses the mean squared differences of the fuzzy memberships of two contiguous points, whereas the modified approach uses the membership's average of points within a window of size λ (Equation 7)

$$\boldsymbol{DG_{k+\lambda}} = \left\{ \frac{\left(\frac{\sum_{j=1}^{C} \left(\mu_{i,k+j}\right) - \left(\mu_{i,k+j+\lambda}\right)}{\lambda} \right)^{2}}{2} \right\}^{0.5} \qquad (7)$$

with *C* equals number of classes, λ equals to the size of the window for comparing μ memberships of class i at depth k with $K = \{k1 = 1\lambda, k2 = 2\lambda, k3 = 3\lambda, \ldots, kn = n\lambda\}$. Small values of *DG* suggest points of stasis while numbers closer to 1 sharp changes, i.e., soil boundaries. In those cases where $D/\lambda \ni \mathbb{N} \Rightarrow kn = D - \lambda \wedge \lambda n = D - kn$ with *D* equal to maximum depth of the soil profile. For consecutive points in the profile with high *DG* values we choose a conservative approach, i.e., the last point with a value over the threshold was chosen, leaving the previous one as part of a transitional zone.

4.3.7 Spectrally derived horizon identification performance

In order to measure the performance in terms of spectral homogeneity of the new *SPD hor* using the *DG*, we used an Interclass Correlation Coefficient (*ICC*) (Wolak et al.,

2012) first presented by Fisher (1925), which measures the proportion of variance of a given outcome variable i.e., the absorbance values of each wavelength, explained by a factor of interest i.e., the observed and the *SPD hor*, measuring the relative homogeneity within groups (Equation 8).

$$\boldsymbol{ICC = \frac{\sigma^2 b}{\sigma^2 b + \sigma^2 w}} \qquad (8)$$

where $\sigma^2 b$ is the variance between groups and $\sigma^2 w$ the variance within groups, from this equation we can deduce that low variance within groups i.e., spectrally homogeneous horizons, will result in higher ICC values and vice-versa.

After calculating the *ICC* on every wavelength of the spectra grouped by traditional horizons and *SPD hor* respectively, we took the mean value of *ICC* by profile i.e., mean of all *ICC* values by wavelength for both types of classification i.e., spectrally derived and traditional, and finally the distribution of *ICC* values was compared.

4.4 Results and discussion

4.4.1 Smoothing by wavelength to capture spectral features

The absorbance values in the different wavelengths changed smoothly down the profile, reflecting the continuous distribution of soil materials. In order to test this assumption we re-scanned the soil cores again, but this time the cores were split at the same interval (2cm), ground, sieved to 2mm and scanned three times each. After this procedure we compared the original scans with the ground sample scans and with the smoothed scans as seen in Figure 4-2.

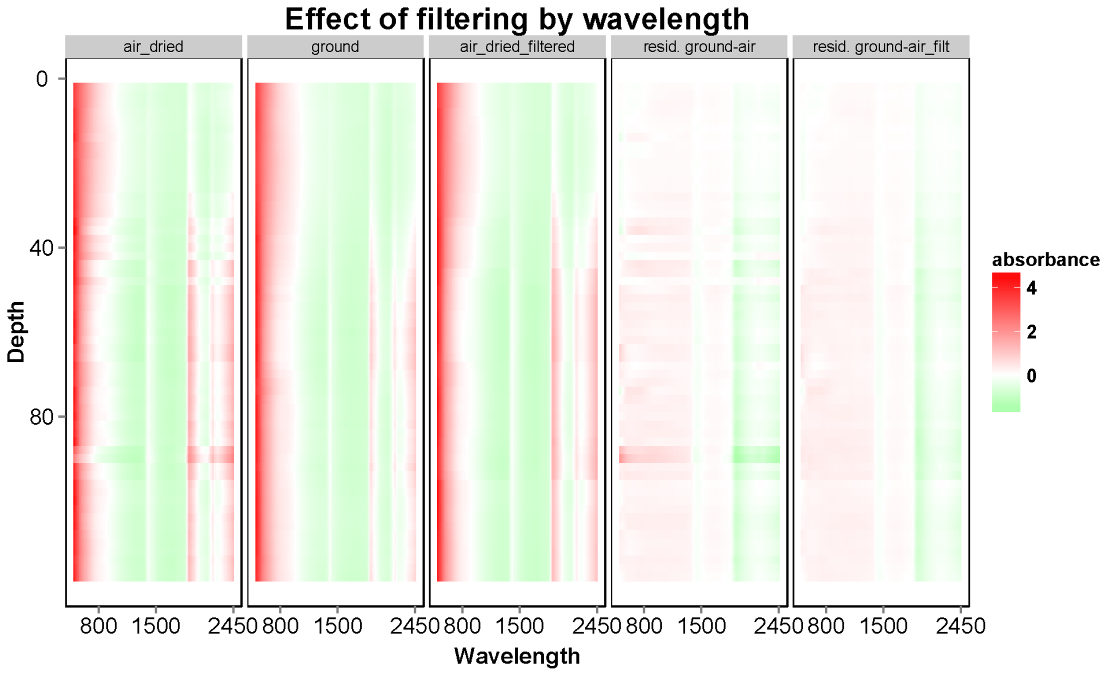


Figure 4-2 Spectral variation of intact, processed cores and smoothing procedure.

It is important to mention that the smoothing process does not replace the grinding process. As we can see in the residuals obtained when subtracting the absorbance values of intact and ground samples, the ground sample absorbs more in the initial part (500 to 1000nm) of the spectrum and less in the final part (2000 to 2450nm). The results were consistent in the 59 soil cores (data not presented). The main effect of the applied filter was to reduce the spectral variation or noise in depth, resulting in clearer spectral patterns along the soil profile, which helps in the following stages of horizon identification.

4.4.2 Relations between horizons and Vis-NIR values

Generally only A horizons formed a defined group as seen in Figure 4-3 illustrating the difference in composition of these horizons compared with lower depths. In the lower part of the profile there is a mixture of soil materials rather than separate classes, reflecting gradational transitions with depth.

As noted Ben-Dor et al. (2008), the larger absorption at 2200 nm related to increasing clay content, helped when distinguishing between A horizons and subsequent illuviation horizons. Also, as expected, the visible part of the spectra (390–700 nm) played an important role when distinguishing types of materials (Figure 4-3).

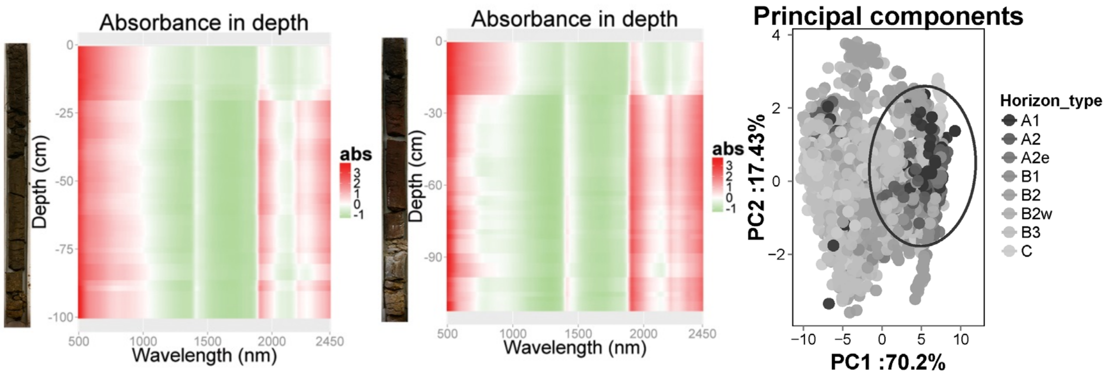


Figure 4-3 *left* and *center*, Relations between horizons origin and associated spectra; *right* showing A1 horizons in the ellipse.

4.4.3 Clustering the dataset: global classes in the soil core

After morphologically describing the soil cores, only 5% of total horizons boundaries were abrupt (5mm to 2cm), 40% were clear (2 - 5cm), 43% gradual (5 - 10cm) and 12% were diffuse. Based on this and in the fact that soil materials change gradually along the soil core, we attempted to imitate this slightly skewed distribution towards a gradual change of materials, using a membership exponent of 1.4 and dividing the entire dataset in 12 fuzzy classes, based on the first local minima of the FPI values for 2 to 20 classes (Figure 4-4 and Figure 4-5)

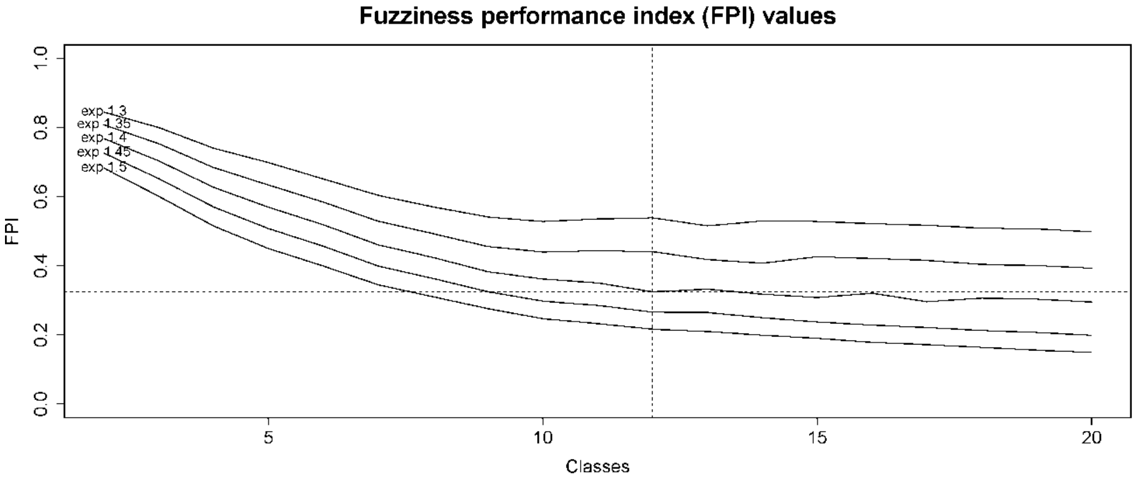


Figure 4-4 Chosen number of fuzzy classes.

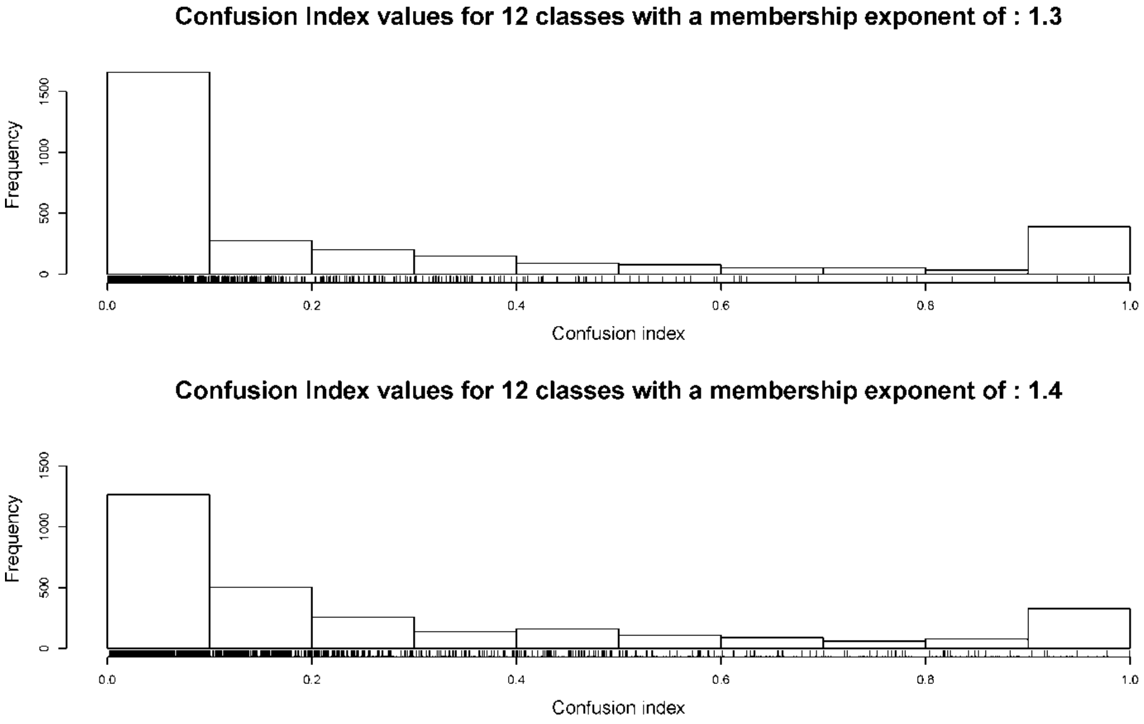


Figure 4-5 Effect of membership exponent in overall confusion index values.

After classifying the dataset into 12 fuzzy classes, 3 out of the 12 classes (classes *a*, *b*, *c*) contained almost all of the near-surface observations i.e. A1-A2-A2$_e$-B1 horizons and only a few in B2$_w$ horizons (Figure 4-6) Classes *d*, *e* and *f* included the rest of the near surface horizons but also some B2 and B3 and B2$_w$ horizons. Classes *g*, *h*, *i* and *j* were coincident with the highest proportion of B2$_w$ horizons.

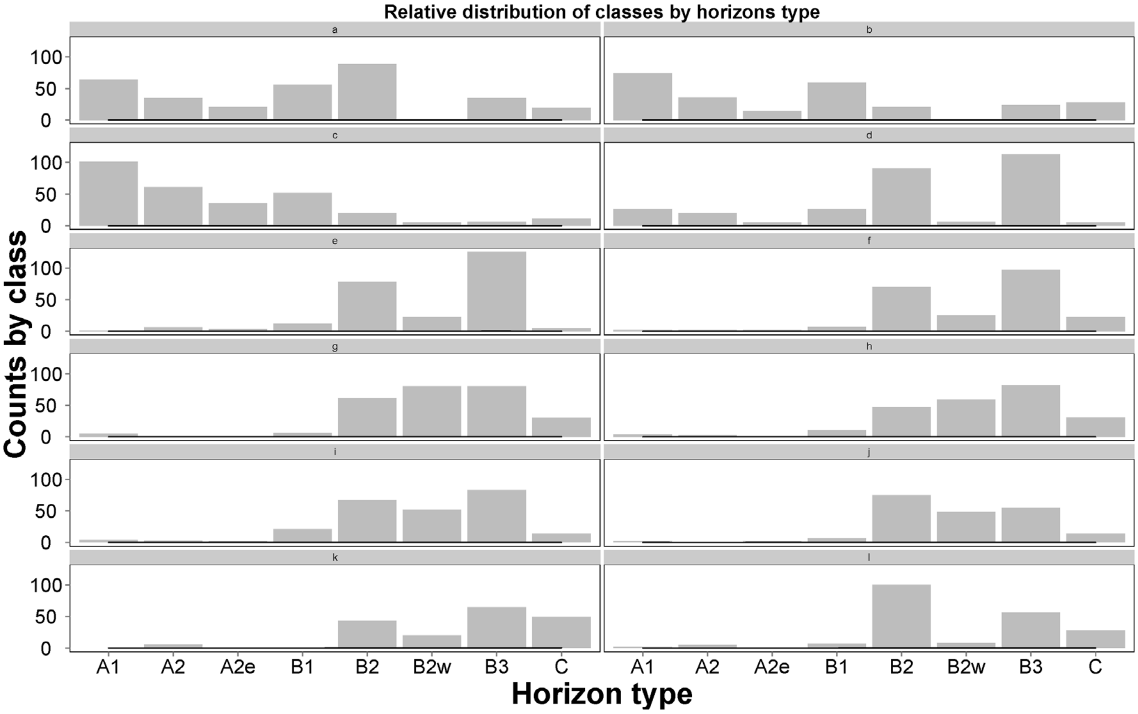


Figure 4-6. Distribution of classes by type of horizon.

Classes belonging to C horizons had a larger percentage in classes *k* and *l*, however they were also present in almost every horizon reflecting rather a high within horizons variability in C horizons compared with A horizons or a misclassification commonly observed in clayey soils with redoximorphic features that can be easily confounded with parent material to the naked eye, or simply, materials that are indeed present in most horizons e.g., parent material in different stages of weathering.

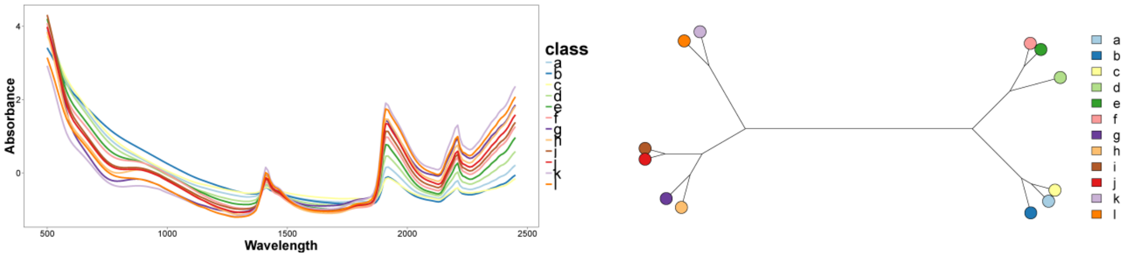


Figure 4-7 Fuzzy classes centroids. *Left*, Spectra of Centroids; *Right*, Euclidean distance between centroids.

The Euclidean distance between the centroids showed that there was a clear cut between near surface horizon materials i.e., classes *a*, *b*, *c* and *d* and the materials found at depth (Figure 4-7) and also an evident organization in the centroids by depth confirming the observed occurrences by horizons.

Despite the clear organization between surface and subsurface materials (fuzzy classes) and horizons, to be able to differentiate between sub-surface horizons we need to analyse the relative changes in composition in depth i.e., relative membership to different classes.

It is important to highlight that in the first step of clustering, we are classifying materials, not horizons. In the present study the methodology explicitly grouped materials (spectra) in a gradual way reflected as an example in the smooth change in the 1400 and 1900 nm peaks' transition related to different O–H bounds in the soil adsorbed water and also correlated to the content of the clay fraction (Ben-Dor et al., 2002; Demattê et al., 2012) (Figure 4-8), in this sense, a single material can be present in two completely different horizons, however two different horizons will always have at least one material exclusive to that horizon.

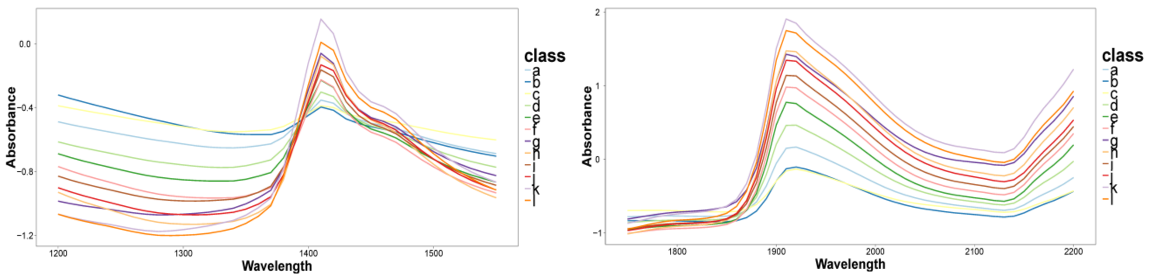


Figure 4-8 Detailed spectra of centroids selected regions.

In order to discriminate between sub-surface horizons e.g., B and C horizons, a different analysis is required, this is to measure the relative composition and homogeneity of that combination of classes in depth, this is reviewed in the following section.

4.4.4 SPD *hor* detection

In the previous sections we discussed the ability of the fuzzy clustering to detect different types of materials and how they occur along the profile, in the present section we discuss the usage of the *DG* to delimit boundaries between groups of materials i.e., discriminate between horizons and their distinctness.

Figure 4-9 shows a soil core with two evident contrasting horizons with a clear to gradual boundary distinctness.

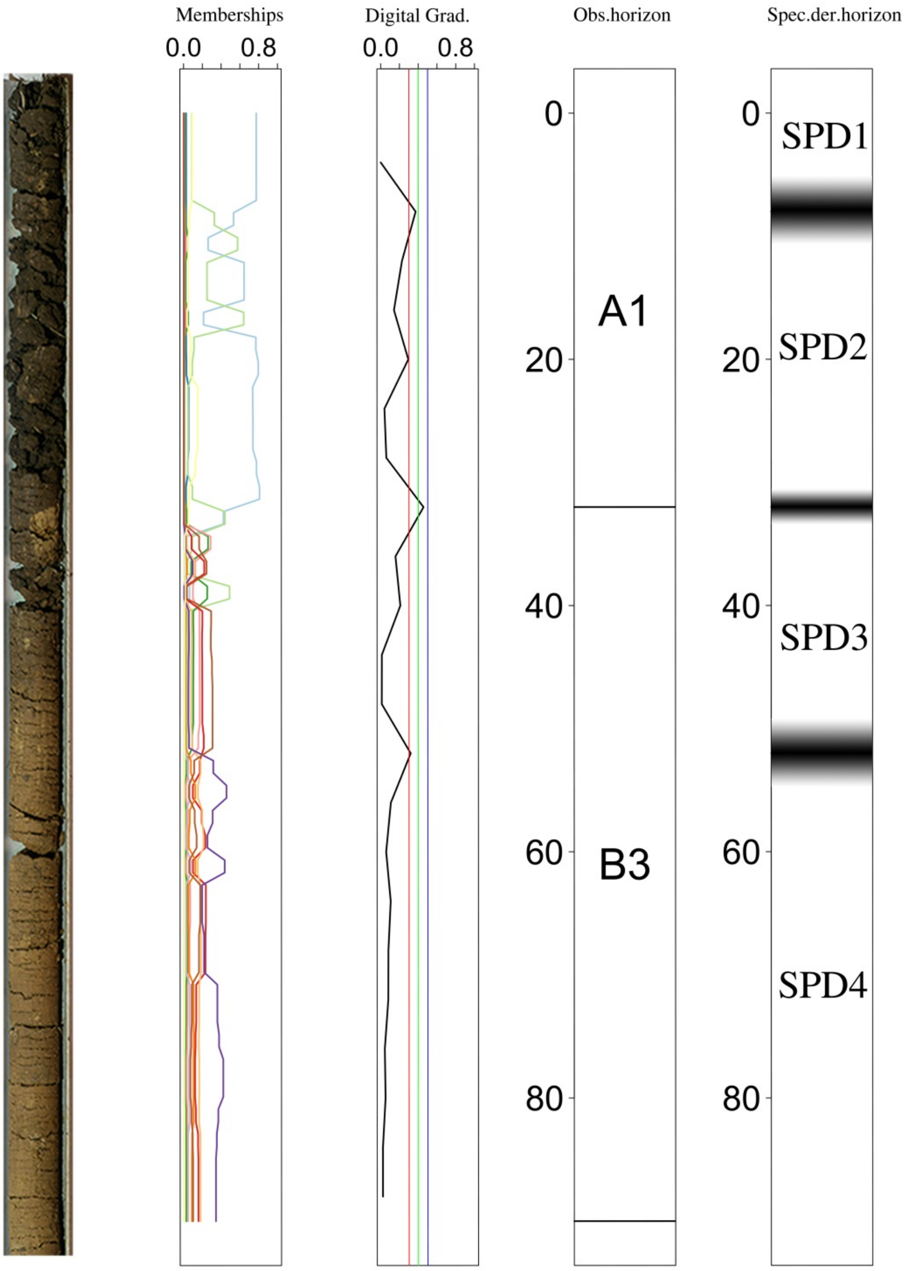


Figure 4-9 Digital gradients in a soil core. The value of the DG comparison window λ was set to 4 cm. Boundaries distinctness is relative to the thickness of the boundary.

As expected the DG was useful when detecting changes of membership gradients along the soil core, with the fuzzy membership values acting as indicators of the proportional composition of each material within the *SPD hor* limits e.g., in the example, the composition in terms of membership to the 12 global fuzzy classes of each *SPD hor* is shown in Table 4-1.

Table 4-1 Example of class percentages in one soil profile.

	SPD hor1	***SPD hor2***	***SPD hor3***	***SPD hor4***
	(0-8 cm)	*(8-32 cm)*	*(32-52 cm)*	*(52-90 cm)*
a	*77.40*	*63.80*	*6.20*	*0.80*
b	*2.70*	*3.30*	*1.00*	*0.50*
c	*9.10*	*8.50*	*1.40*	*0.50*
d	*8.70*	*21.00*	*12.80*	*1.40*
e	*1.00*	*1.80*	*12.20*	*2.70*
f	*0.50*	*0.80*	*15.20*	*4.30*
g	*0.10*	*0.10*	*4.40*	*32.90*
h	*0.10*	*0.10*	*6.60*	*17.20*
i	*0.20*	*0.40*	*22.40*	*10.50*
j	*0.20*	*0.20*	*15.20*	*16.20*
k	*0.00*	*0.00*	*0.90*	*4.60*
l	*0.10*	*0.10*	*1.70*	*8.40*

From Table 4-1 and Figure 4-9 we can easily describe 4 *SPD hor*. The first one with a higher membership to class *a*, mainly present in surface horizons, a second *SPD hor* with a 21% of membership to class *d* present in B3 horizons (Figure 4-6) and the two final *SPD hor* showing a different composition in terms of classes memberships.

We also were able to determine the boundary distinctness. Our methodology was able to detect the more evident change between horizons i.e., between A1 and B3 at 32cm and also to establish an extra boundary in the horizon B3 where the differences in the membership values of classes *d*, *e*, *f* and *g* where particularly high (Table 4-1) but hardly visible to the bare eye.

The comparison window λ used in the DG calculations was of 4cm. The *DG* threshold values were empirically set to 0.5 for clear boundaries, to 0.4 for gradual and 0.3 for diffuse boundaries based on the values of the observed cores. These were used to set a comparison with traditional descriptions and are not intended to be used as a standard, since *DG* values are continuous.

4.4.5 Overall performance

Out of 56 soil cores, 20 (35%) had similar horizon distributions compared with the traditional descriptions, 27 (48%) were different mainly because of a greater number of *SPD hor* (horizon subdivisions) and 9 (16%) did not have a match with traditional descriptions (examples in Appendix 4 and Appendix 5).

The *ICC* distribution for the two types of horizon designation i.e., spectrally derived horizon recognition and traditional description, showed a higher spectral homogeneity in the *SPD hor* compared with the traditional description (Figure 4-10). This result is important, because many modern soil surveys use properties estimated by Vis–NIR spectroscopy calibration models and their training samples often comprise a composite sample taken within individual traditional horizons.

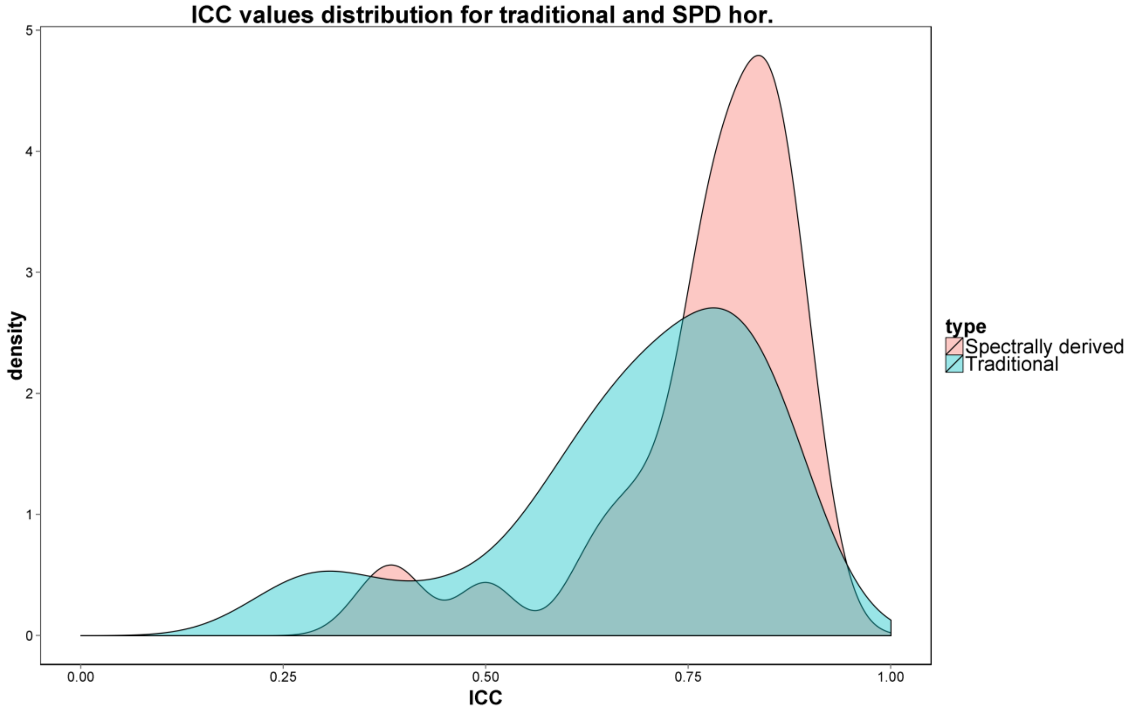


Figure 4-10 Distribution of ICC values for 56 described soil cores (after outlier removal) of traditional horizons designation and *SPD hor.*

4.5 Final considerations

This study has presented a numerical procedure for identifying spectrally homogeneous zones within a soil core and also created a basic classification system i.e., global Vis–NIR fuzzy classes that help to identify each *SPD hor* and also to create a baseline when distinguishing between horizons of different soil cores.

The calculated centroids plus the *DG* proved their usefulness in splitting homogeneous zones in a soil core i.e., soil morphological horizons, however, since they are based on an unsupervised clustering algorithm they don't necessarily constitute an ideal classification.

In order to create more representative (and functional) centroids it may be necessary to build a global spectral database i.e., a higher spectral variability in the database, and also to measure related key soil properties for potential inclusion in the clustering process.

The methodology can eventually be used with, or in addition to, other variables e.g., X-ray fluorescence, penetrometer resistance values or others like measured chemical or structural properties, furthermore, the process may be automated, once the set of fuzzy classes is established, the *DG* values are calculated for each core in few seconds and the boundaries are set automatically; For example, the required time of calculations of all the boundaries for the 56 analysed soil cores was less than a minute.

Further research is needed to address the variability in key soil properties within each SD *hor* (e.g., Clay content, pH, ECEC, Carbon, etc.) as well as lateral variation and in-situ conditions.

4.6 Conclusions

- A new methodology for detection of soil morphological horizons and their boundaries has been devised. This approach offers more information than the traditional description and avoids observer bias when describing a soil core. However it depends on various parameters that influence the final result e.g., number of classes, fuzziness of the groups, size of the window and threshold for the *DG* calculation.
- To achieve the best results, the measurement resolution should be equal or less than 2cm in order to completely capture the soil changes necessary to establish differences.
- Despites the success of the smoothing by wavelength to reduce spectral noise, it is recommended to add replicates to each scan in conjunction with the smoothing procedure to improve the results and reduce the need for spectral pre-treatments as much as possible.
- The procedure represents a promising and reasonably easy replicable methodology, that can be used in conjunction with observations with other proximal sensing devices e.g., penetrometer resistance values, X-ray fluorescence, which would add greater detail i.e., structure, and in the end result in more effective and informative horizon recognition.

Chapter references

Ad-Hoc-AG, B., 2005. Bodenkundliche Kartieranleitung. Ad-hoc-AG der Staatlichen Geologischen Dienstes und der Bundesanstalt für Geowissenschaften und Rohstoffe 5.

Ben-Dor, E., Heller, D., Chudnovsky, A., 2008. A novel method of classifying soil profiles in the field using optical means. Soil Science Society of America Journal 72(4), 1113-1123.

Ben-Dor, E., Patkin, K., Banin, A., Karnieli, A., 2002. Mapping of several soil properties using DAIS-7915 hyperspectral scanner data-a case study over clayey soils in Israel. International Journal of Remote Sensing 23(6), 1043-1062.

Bockheim, J.G., Gennadiyev, A.N., Hammer, R.D., Tandarich, J.P., 2005. Historical development of key concepts in pedology. Geoderma 124(1–2), 23-36.

Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77(2–4), 115-135.

Ciampalini, R., Lagacherie, P., Gomez, C., Grünberger, O., Hamrouni, M.H., Mekki, I., Richard, A., 2013. Detecting, correcting and interpreting the biases of measured soil profile data: A case study in the Cap Bon Region (Tunisia). Geoderma 192(0), 68-76.

CSIRO (Ed.), 2009. Australian Soil and Land Survey Field Handbook (3rd Edition),The National Committee on Soil and Terrain. CSIRO Publishing.

Demattê, J.A.M., Terra, F.d.S., Quartaroli, C.F., 2012. Spectral behavior of some modal soil profiles from São Paulo State, Brazil. Bragantia 71, 413-423.

Filzmoser, P., Garrett, R.G., Reimann, C., 2005. Multivariate outlier detection in exploration geochemistry. Computers & Geosciences 31(5), 579-587.

Fisher, R.A., 1925. Statistical methods for research workers. Genesis Publishing Pvt Ltd.

Grunwald, S., Lowery, B., Rooney, D.J., McSweeney, K., 2001. Profile cone penetrometer data used to distinguish between soil materials. Soil and Tillage Research 62(1–2), 27-40.

Hardle, W., Steiger, W., 1995. Algorithm AS 296: Optimal Median Smoothing. Journal of the Royal Statistical Society. Series C (Applied Statistics) 44(2), 258-264.

Hartemink, A.E., Minasny, B., 2014. Towards digital soil morphometrics. Geoderma 230, 305-317.

Isbell, R.F., 2002. The Australian soil classification / R.F. Isbell. Australian soil and land survey handbook ; vol. 4. CSIRO Publishing, Collingwood, Vic. :.

Jahn, R., Blume, H.P., Asio, V.B., Spaargaren, O., Schad, P., 2006. Guidelines for soil description, 4th edition. FAO, Rome.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2014. cluster: Cluster Analysis Basics and Extensions.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32(9), 1378-1388.

Odgers, N.P., McBratney, A.B., Minasny, B., 2011b. Bottom-up digital soil mapping. II. Soil series classes. Geoderma 163(1–2), 30-37.

Percival, I., 2012. Geological Survey of New South Wales, Permian fossils and palaeoenvironments of the northern Sydney Basin, New South Wales Maitland, N.S.W. : Geological Survey of New South Wales.

Powell, B., McBratney, A.B., MacLeod, D.A., 1991. The application of fuzzy classification to soil pH profiles in the Lockyer Valley, Queensland, Australia. CATENA 18(3–4), 409-420.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rooney, D.J., Lowery, B., 2000. A profile cone penetrometer for mapping soil horizons Soil Sci. Soc. Am. J. 64(6), 2136-2139.

Schoeneberger, P.J., 2002. Field book for describing and sampling soils, Version 3.0. Government Printing Office.

Weindorf, D.C., Zhu, Y., Haggard, B., Lofton, J., Chakraborty, S., Bakr, N., Zhang, W., Weindorf, W.C., Legoria, M., 2012. Enhanced Pedon Horizonation Using Portable X-ray Fluorescence Spectrometry. Soil Sci. Soc. Am. J. 76(2), 522-531.

Wolak, M.E., Fairbairn, D.J., Paulsen, Y.R., 2012. Guidelines for estimating repeatability. Methods in Ecology and Evolution 3(1), 129-137.

Chapter 5 *Pedodiversity*

Use of spectral information and new indices to quantify pedodiversity

"*…El bosque precede al hombre, pero lo sigue el desierto…*"

Patricio Manns y Horacio Salinas

Chapter 5 in revision as:

Fajardo, M., McBratney, Minasny, B., 2016. Measuring functional pedodiversity using spectroscopic information. Catena.

5.1 Summary

Pedodiversity studies greatly depend on the discretization of the soil continuum via various soil classification standards. Soil taxonomic systems are constructed considering soil properties and their organization in the soil profile. Therefore, it seems logical that the pedodiversity of an area could be calculated by considering the simultaneous variation of multiple soil properties within that area. With the aim of creating a bridge between these two lines of thinking, this study presents the development of two new indices of pedodiversity (*HULLdiv* and *EIGENdiv*) which consider continuous variables as input. The soil input information tested was a) Visible and Near Infrared (Vis-NIR) reflectance values and b) values of five soil properties predicted from Vis-NIR spectra. Both indices were employed to measure the pedodiversity at different extents in two perpendicular transects containing 27 (North to South) and 22 (East to West) locations respectively in New South Wales (NSW), Australia. Each location considered natural and intervened land use. The indices successfully represented the pedodiversity of the area of study, however they were different depending on the input soil information i.e., raw Vis-NIR and predicted properties. *HULLdiv* was affected by extreme soil observations and as a result its discrimination power between areas with different soil diversity was inferior. On the other hand, *EIGENdiv* represented well the pedodiversity of an area, despite the extreme observations. The results showed good agreement with conventional methods for assessing pedodiversity i.e., Shannon's and Simpson's indices. However, the new indices were more discriminating by being able to better represent the landuse effect. This study represents the first attempt to measure pedodiversity in a continuous way using Vis-NIR information.

5.2 Introduction

The basis of every type of diversity measurement resides in describing the composition of a community or more specifically the *abundance* of its individual units i.e., how many of them are, the *richness* in terms of the amount of different units contained in the community and finally their relative distribution or their *evenness* (Whittaker, 1972; Orlóci et al., 2002; Pavoine and Bonsall, 2011). For soil sciences this concepts applies in the same way, if we consider the soil resource as the sum of separate soil entities or soil types.

The first works that focused in the concept behind pedodiversity or diversity of soil, described the composition of discrete soil entities in the landscape (Fridland, 1974; Jacuchno, 1976; Linkeš et al., 1983). Since then, studies in pedodiversity have followed a similar outline.

Traditional methodologies for describing pedodiversity usually make use of methods employed in the fields of Ecology and Biology. Simply, the concept of pedodiversity involves the quantification of the different soil types in a determined area. From the previous, it can be seen that the calculation of the pedodiversity on a particular area will require first a previous classification of soils in order to quantify their distribution. Nevertheless, McBratney and Minasny (2007) observed that conventional diversity measures only contemplate the relative abundance of soil classes, producing a limited representation of the pedodiversity of an area thus, recommending the use of a taxonomic distance in the calculations; in other words, to consider the unique properties of each soil when calculating the pedodiversity of an area. Despite of the previous, the idea that the pedodiversity of an area is associated with the inherent variability of soil properties is still a controversial concept (Jie et al., 2001).

Conversely, Toomanian and Esfandiarpoor (2010) observed that the reason of the apparent lack of correlation between pedodiversity and variation of soil properties was that, the geostatistical methods used in the pedodiversity studies trying to establish a link between them, dealt with simple vectors i.e., separate soil properties. Suggesting with this that pedodiversity analyses will require the use of a big matrix of interrelated variables, which takes us to a possible solution. If a link between pedodiversity and soil properties variability is intended, then the use of multivariate information may well be required.

The use of multivariate and continuous information for describing diversity is not completely new, and there are several methods to describe species in multidimensional spaces e.g., (Petchey and Gaston, 2006; Maire et al., 2015); Furthermore, Tilman (2001) described "Functional diversity" as a branch or a subset of biological diversity, that deals with the components that influence how an ecosystem operates and functions, and that can be measured by the values (and ranges in the values), of the organismal "traits" that influence one or more aspects of the functioning of an ecosystem. In simpler words, the analysis of the multiple properties or "traits" that makes a particular species to function in a determined way. The term functional pedodiversity has been previously referred by Ibáñez (1996) as a way to measure "*what the soil does*", also called land versatility or land variability, however its study seems to be underdeveloped compared with its counterpart in Ecology i.e., functional diversity.

With the intention to describe soil properties variability or “functional pedodiversity” as understood by its connotation in functional ecology, the objectives of this work are a) to use Vis-NIR spectroscopy to characterize soil in fine detail by analysing soil profiles scanned with a Vis-NIR spectrometer with a vertical spacing of 2 cm, to measure its functional diversity at different depths and scales. b) to observe the performance of this new way of measuring pedodiversity in discriminating between sites with different land uses, and finally c) to explore the relations between this new way of describing pedodiversity and conventional methods or “taxonomic pedodiversity” (Bockheim and Haus, 2013).

5.3 Materials and Methods

5.3.1 Dataset

The dataset used in the present chapter considers 98 soil cores of 1 m depth extracted at the same locations of the dataset described in Chapter 2. Each soil core was air dried and scanned with an ASD Agrispec 350-2500 spectrometer using a Spectralon® panel as a reference in three different parts of the soil core every 2cm, resulting in 14,040 single observations as shown in Figure 5-1, it is worthy to mention that previous to each observation the surface of the soil core that was in direct contact with the extraction tube (corer) was scraped with a clean knife in order to reduce the presence of material from other depths.

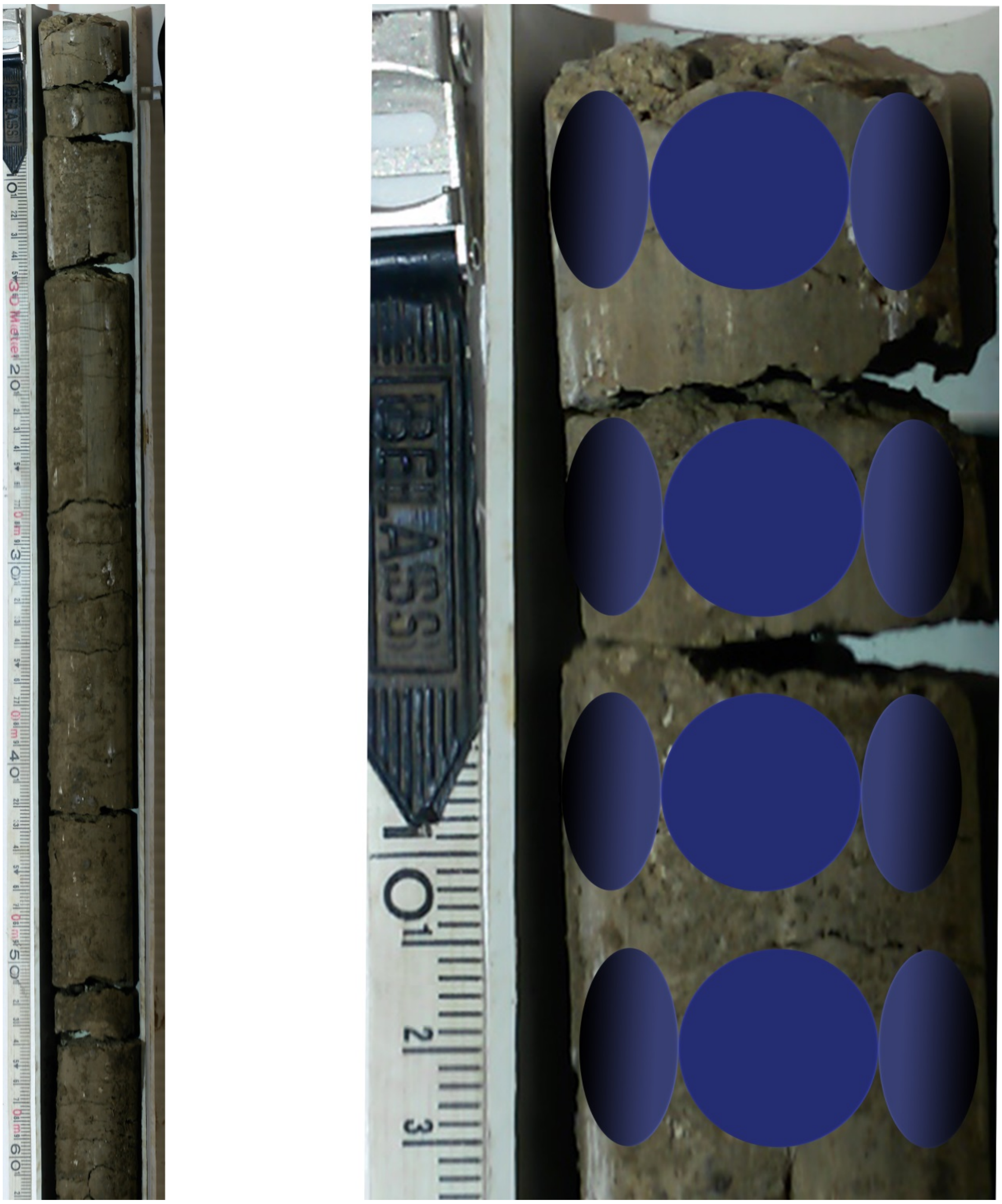


Figure 5-1 Vis-NIR soil scans by core. *Left*: First 60cm of 1m core, *Right*: Schema of Vis-NIR observations.

5.3.2 Pedodiversity measurements

5.3.2.1 In-hull diversity

The use of the area of the convex hull defined in a multidimensional variable space i.e., Principal component space, to estimate soil variability has being previously used by other authors e.g., (Islam et al., 2005; Petersen et al., 2010). Islam et al. (2005) concluded that the area of the convex hull was directly related with the variation of soil materials, hence, the convex hull area formed by the extreme samples projected in the two first principal components (usually explaining more than 85% of the variance when dealing with Vis-NIR spectra), was used as an index of pedodiversity from now will be known as "*HULLdiv*".

Two different principal component scores were used. First, the whole spectra was used for the calculation of the principal components, and second, 5 soil properties i.e., TC (%), CEC (cmol(+) kg $^{-1}$), pH, EC (dS m^{-1}) and Clay (%), predicted from the spectra (calibration models were built using samples from the same sites) were used for creating the principal components. Finally, the convex hull areas were measured using the R package "tripack" (Renka et al., 2015).

5.3.2.2 Eigenvalue diversity

The second measure of pedodiversity from now to be known as *EIGENdiv* (Eigenvalue diversity) also considered multiple soil attributes (raw Vis-NIR spectra and properties predicted from spectra) in a soil observation. Again, a principal component analysis was performed using both spectra and properties (same five properties used in the previous measure) and the score matrices were used as follows.

Let $x_{i,j}$ be the score or the coordinate of observation i in a principal component j within a scores matrix explaining more than 95% of the variance of the original data. The *EIGENdiv* can be calculated as follows in Equation 9:

$$\boldsymbol{EIGENdiv} = \frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{k}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}\left(\boldsymbol{x}_{\boldsymbol{i},\boldsymbol{j}}-\overline{\boldsymbol{x}}_{\boldsymbol{j}}\right)^{\mathbf{2}} \tag{9}$$

with $\bar{x}$ equals to the mean value or “centroid” of principal component j and k equals the number of principal components that explain at least 95% of the variance of the original dataset. Note that if the original data is in the same scale i.e., spectra, the score matrix resulting from the principal component analysis is performed based on the correlation matrix of the original dataset, on the other hand, if the units of the original data are different e.g., a matrix with rows as observations and columns as soil variables, the covariance matrix is used instead.

It can be seen from the previous, that the value of *EIGENdiv* of a sampling area with observations *i,i+1,i+2,…,n* will be the sum of all the score sample variances (or eigenvalue) of those locations, and will represent the variance from the original data, explained by the selected observations in a multidimensional space.

5.3.2.3 Conventional diversity indices

In order to assess the correspondence of the newly proposed pedodiversity indices with known pedodiversity measures, two commonly indices used in Ecology and pedodiversity studies were used, these are a) Shannon index (Shannon (1948) as follows in Equation 10:

$$\boldsymbol{H_{c_l} = -\sum_{i=1}^{k} p_{i_{c_l}} \ln p_{i_{c_l}}} \qquad \textbf{(10)}$$

with p_{icl} as the proportional abundance of soil types at site *i* relative to the total amount of soil categories *c* of taxonomic level *l* found, with *l* equals to soil orders, sub-orders, families and so on.

Intuitively, this formula will produce a higher number as the soil classes in a determined area increase. In other words a large Shannon's index indicates the "pedodiversity" in that area is large.

Another measure is the Simpson index or "Simpson's index of diversity" (Simpson (1949), as shown in Equation 11:

$$\boldsymbol{D_{c_l} = 1 - \sum_{i=1}^{k} {p_{i_{c_l}}}^{2}} \qquad (11)$$

with p_{icl} as the proportional abundance of soil classes *c* of level *l*, at site *i* relative to the total amount of soil types found.

Both conventional diversity indices were calculated using the R package “vegan” (Oksanen et al., 2014). The category used for these indices was the Australian Soil Classification soil order of each site as classified by ABARES et al. (2014). Further studies will include sub-levels i.e., sub-orders, families, etc., as they will add more complexity to the analysis.

5.3.3 Area–pedodiversity relationships

The pedodiversity indices were calculated both at point (by site) and large scale (by increasing areas i.e., more than one site) as specified in the next subsections.

5.3.3.1 Point scale pedodiversity

5.3.3.1.1 Soil surface, 0 to10 cm

Point scale pedodiversity from only the first 10 centimetres was first calculated, this procedure was used for testing the performance of the new pedodiversity indices only at surface (i.e., 9 scans contained at that depth) and with this, to observe a possible land-use effect over pedodiversity.

5.3.3.1.2 Soil profile, 0 to 100 cm

Second, an analysis of the entire profile depth i.e., ~150 scans in a profile was performed, by doing this, each site will be characterized in a more representative way.

5.3.3.2 Large scale pedodiversity

A further analysis involved calculating the indices by increasing areas i.e., number of locations, from 50 to 500 km over the two transects and by land-use, respectively. By following this procedure the result of each index calculated at each separation distance was the mean of all the observations in a transect included in that distance e.g., in a separation equals to 100 km the result of *EIGENdiv* will be equal to*: Mean {EIGENdiv (Site1+ Site2+ Site3), EIGENdiv (Site2+ Site3+ Site4), EIGENdiv (Site3+ Site4+ Site5)…},* where each site is separated by approximately 50 km (Figure 5-2)

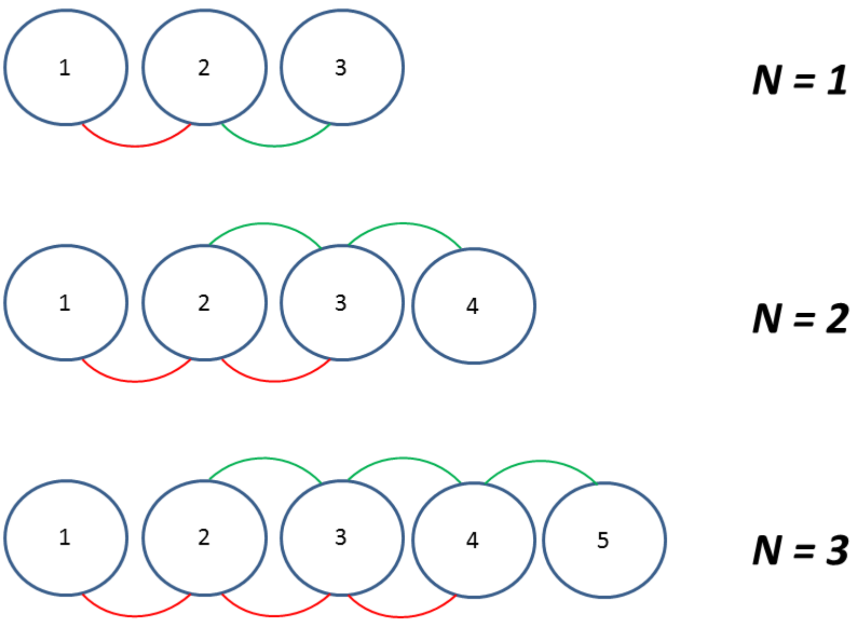


Figure 5-2 Schema of area-pedodiversity calculations, *N represents the number of sites considered.*

As with the point scale pedodiversity, the indices for large scale were calculated for surface and whole profile observations.

5.4 Results and Discussion

5.4.1 Point scale pedodiversity: Soil surface

In order to highlight the efficacy of the new indices in quantifying soil surface's diversity, the first analysis was performed only on those observations contained within the first 10 centimetres of the core i.e., 9 surface soil scans per site-land use combination. Figure 5-3 shows a graphical example of three sites with two land uses respectively, projected in a principal component space using the information of the entire spectra. The polygons from the same site are closer to each other according to their geographical distances i.e., in this example each site is separated by 50 km.

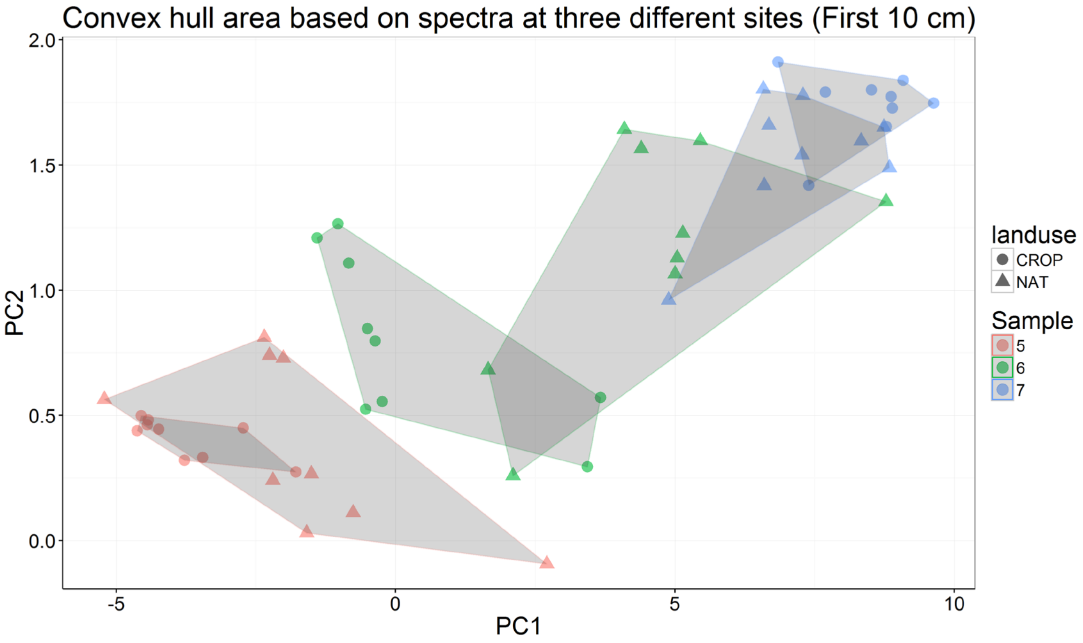


Figure 5-3 Three sites example of convex hulls created using a score matrix with spectra as input (First 10 cm observations). The three selected samples are separated by approximately 50km each.

It was also possible to see in *HULLdiv* (visually equivalent to polygons areas in Figure 5-3 and Figure 5-4) that depending on the principal component (PC) scores used, different area values were obtained and also their distribution in the multidimensional space changed.

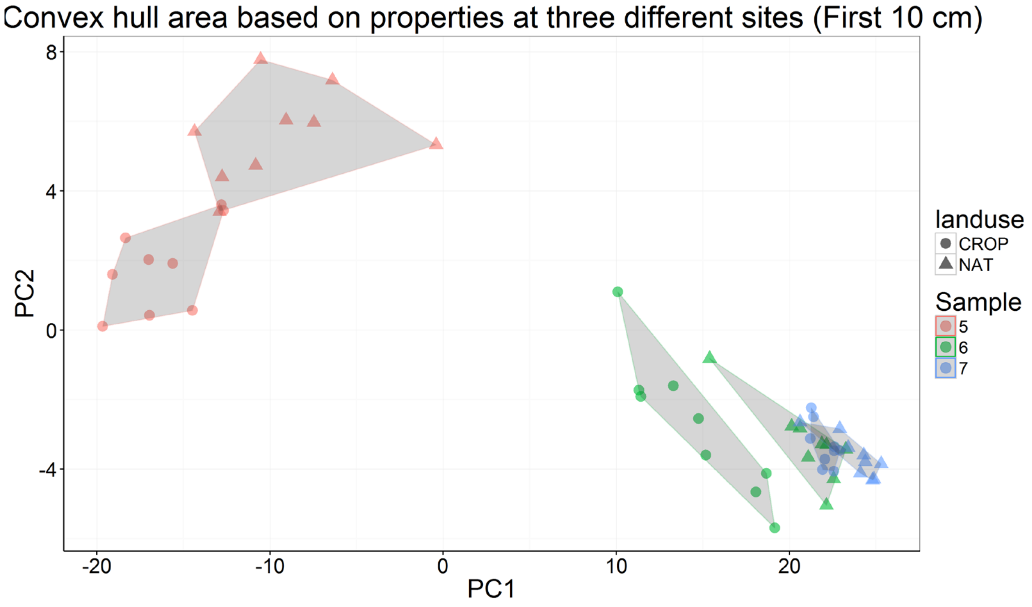


Figure 5-4 Three sites example of convex hulls created using PC scores of five properties (First 10 cm observations). The three selected samples are separated by approximately 50km each.

These results were also observed by Islam et al. (2005), where they observed different values of convex hull areas per site, both in PC scores from the spectra and observed properties (pH, %Clay and %OC). In their work, they concluded that more variability attributed to mineral composition was added by using the PC scores of the Vis-NIR spectra.

The effect in this study was an apparent higher discrimination when using just the selected soil properties (Figure 5-4), suggesting that as more information was introduced in the analysis (Vis-NIR spectrum), also more information common to the different samples could have been added as well e.g., same parent material, same colour, etc. resulting in “closer” samples in a multidimensional space.

In the case of *EIGENdiv* the effect was similar, since this index uses the same two first principal components (which have the highest variance explained). It was also noticeable that in both indices, when the PC scores created with the spectra were used, they had smaller values compared with the indices calculated with PC scores calculated from soil properties, confirming the higher discrimination when using soil properties for the creation of the input scores matrix (Table 5-1).

Table 5-1 Example of pedodiversity indices for three selected samples using the first 2 PC scores of the spectra and soil properties.

	HULLdiv		***EIGENdiv***	
Site	*Spectra*	*Properties*	*Spectra*	*Properties*
5 CROP	0.33	14.49	0.96	8.30
5 NAT	2.85	33.04	4.34	20.13
6 CROP	2.29	14.60	3.81	15.72
6 NAT	4.35	7.76	4.48	7.00
7 CROP	0.72	1.61	0.84	0.84
7 NAT	1.41	2.48	1.67	2.45

5.4.2 Point scale pedodiversity: Profile

The same procedure previously analysed was performed using the whole set of observations i.e., 14,040 separate soil scans, with each soil profile enclosing between 100 to 150 soil observations depending on their depth.

5.4.2.1 Within-sample variation

Figure 5-5 shows the corresponding Vis-NIR scans of a single profile projected in principal component space and coloured by depth. As observed by other authors (Demattê et al., 2012; Hartemink and Minasny, 2014; Vasques et al., 2014; Fajardo et al., 2016a) there is a strong relation between depth and spectral behaviour.

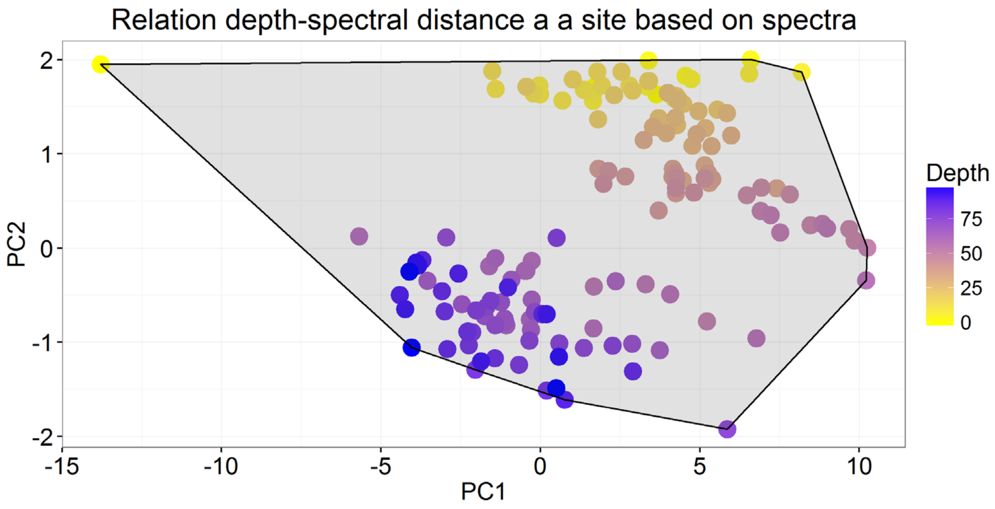


Figure 5-5 Convex hull created using PC scores of spectra as input for one site (Profile).

Also, and as observed in previous section, the distance between the points projected in a principal component space was higher for the case of the scores matrix calculated with properties (*see Y* and *X* axis in Figure 5-6 compared with Figure 5-5). Despite of the previous, the discrimination between samples from different horizons was similar, reflecting the continuous variation of soil with depth, as revised in other studies (Ponce-Hernandez et al., 1986; Bishop et al., 1999; Malone et al., 2009).

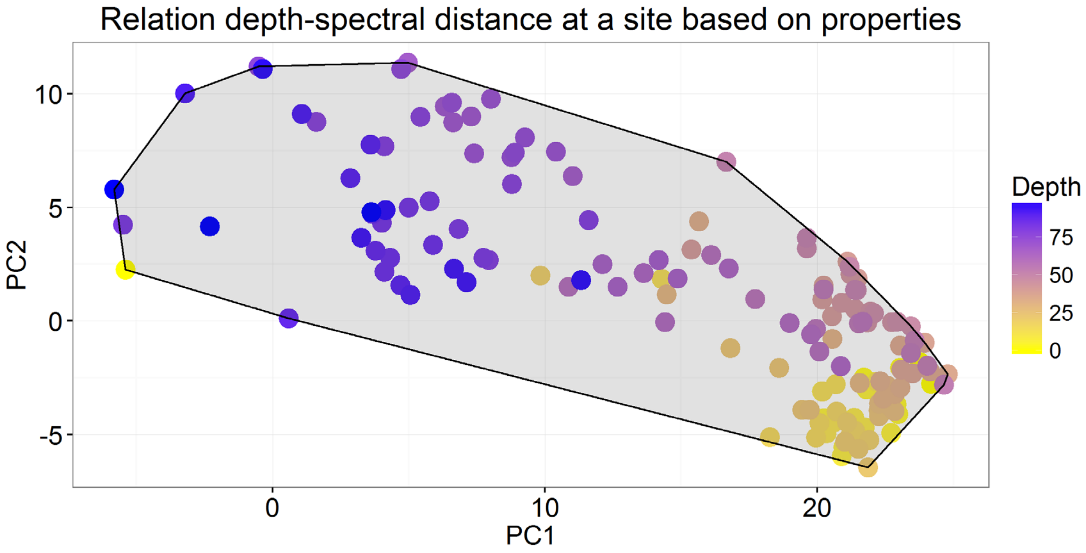


Figure 5-6 Convex hull created using a score matrix with properties as input for one site (Profile).

5.4.2.2 Between-samples variation

Figure 5-7 and Figure 5-8 show the same three sites example shown in section 5.4.1, where the same behaviour can be observed, relative to the different PC scores used. That is, those samples projected using the whole Vis-NIR information had a higher overlap compared to those projected using the properties PC scores.

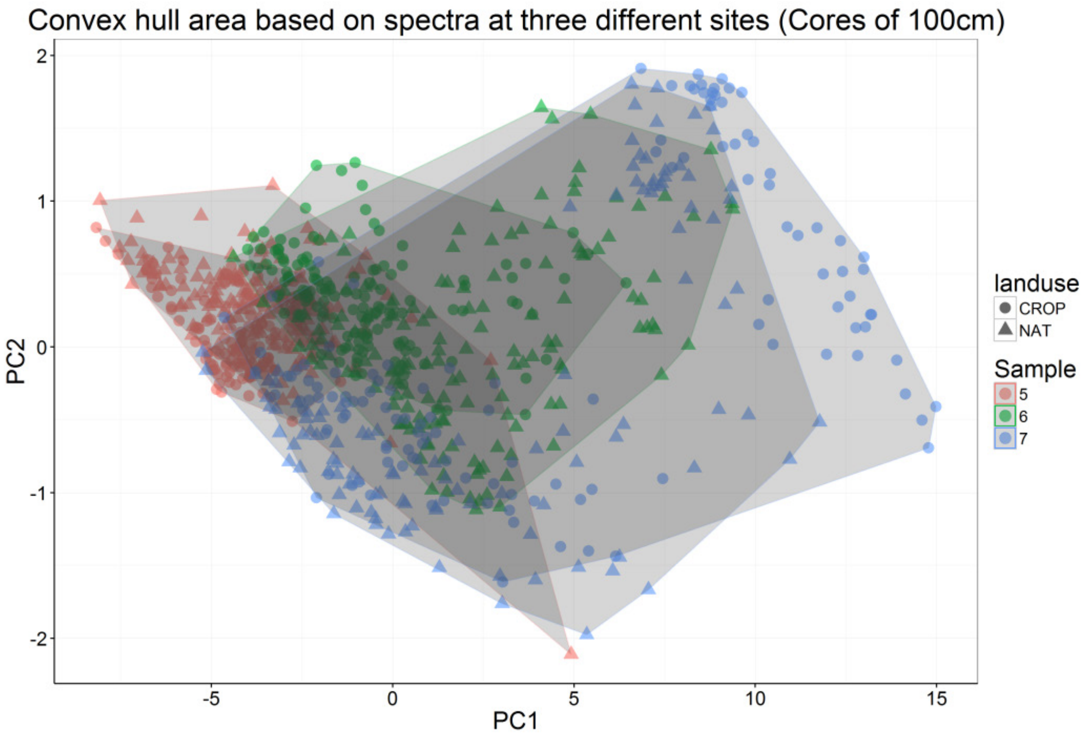


Figure 5-7 Convex hulls created using with the first 2 PC scores of the spectra for three sites (Profile). The three selected samples are separated by approximately 50km each.

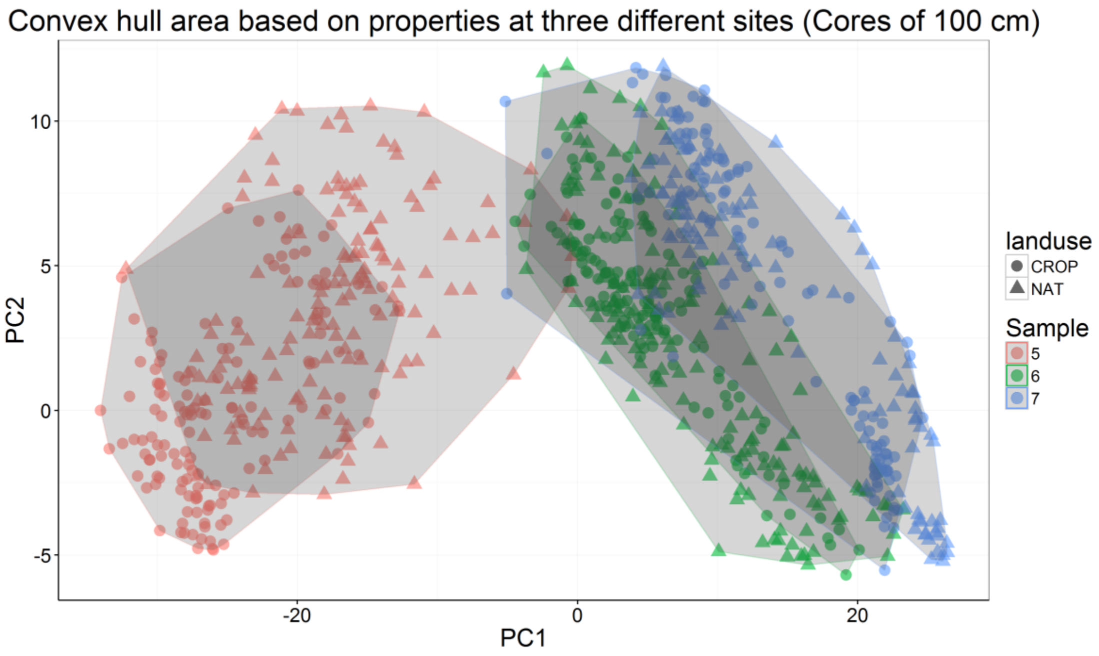


Figure 5-8 An example of convex hulls for 3 sites created using PC scores of five soil properties (Profile). The three selected samples are separated by approximately 50km each.

It is also evident that the area of the convex hulls was now bigger due to the higher internal variability. There is great value in these results, since they clearly show the relations within and between profiles, picturing soil attributes variation and pedodiversity in a detail that has not been observed before.

5.4.3 Large-scale pedodiversity: Topsoil

Even though the results of the indices were useful when describing differences between sites using information of a very small area (first 10 cm of soil surface and about 2 cm in surface area), a multi-scale analysis creates a better picture of the overall performance of the indices; this is the behaviour of the pedodiversity indices over increasing areas, as presented by other authors (Guo et al., 2003; Saldaña and Ibáñez, 2004; Ibáñez and Gómez, 2016).

From studies of diversity, particularly biodiversity, there is a common pattern between diversity and area of sampling or "sampling effort" usually following a power function shape (Preston, 1962; Harte et al., 2009).

By examining the variability of some of the soil formation factors in the area of study i.e., temperature, relief and precipitation (Figure 5-9, Figure 5-10 and the previously shown Figure 2-1 *in chapter 2*), it is possible to anticipate a probable result in terms of the different soil types in the region. The transect East-West presents a higher total variation, especially in the precipitation. In the case of the North-South transect, precipitation was intentionally kept as a constant (Figure 2-1 *in chapter 2*).

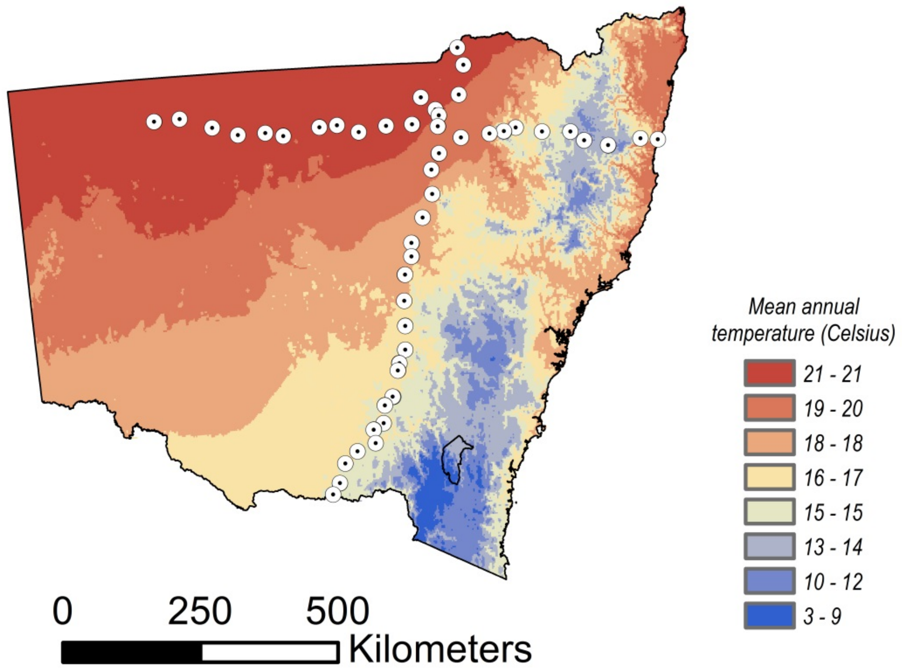


Figure 5-9 Mean annual temperature map of NSW. Source: (ABARES et al., 2014).

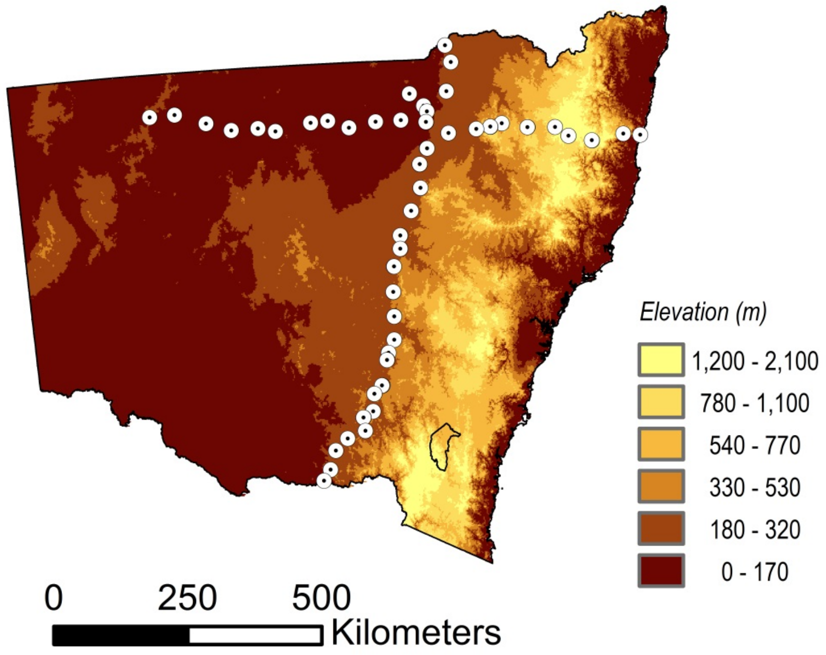


Figure 5-10 Elevation map of NSW. Source: (ABARES et al., 2014).

Accordingly, Figure 5-11 presents the different soil types of NSW, where it is evident the greater richness of different soil types in the East-West transect compared with the North-South transect.

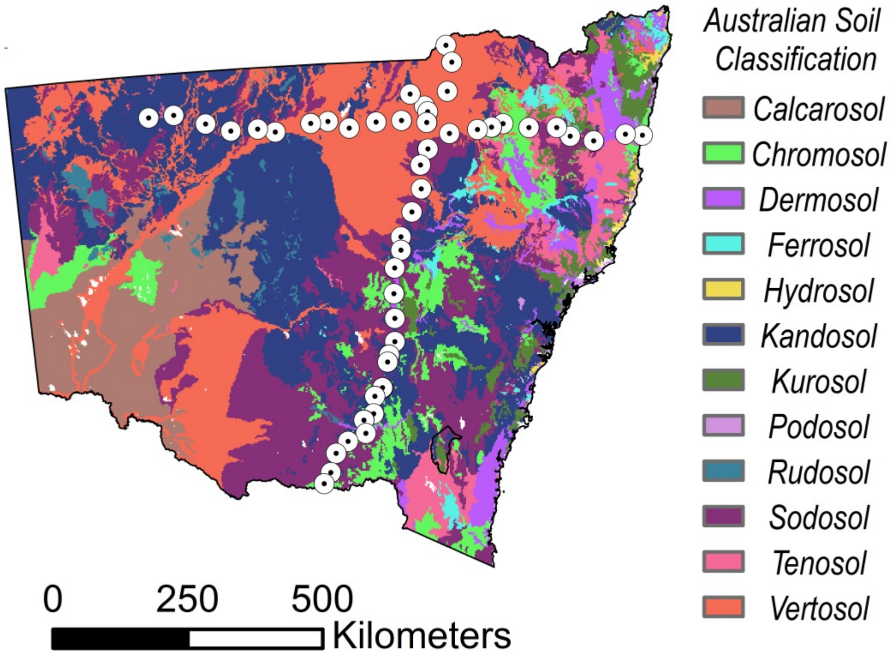


Figure 5-11 Australian soil classification soil types of NSW. Source: (ABARES et al., 2014).

After calculating *HULLdiv* for surface observations at increasing scales, three recognizable patterns were noticeable. First, when using the entire spectra, the calculated indices were significantly smaller than those calculated with the five selected soil properties as input (*see Y* axis *in* Figure 5-12 and Figure 5-13).

Second, and as expected from the pedodiversity at local scale, the discrimination power between the different transects was evidently amplified when using only five properties instead of the entire spectra (*see* differences of E-W transect from N-S transect *in* Figure 5-12 and Figure 5-13), where redundant information is added as other authors have noticed, mainly when using spectra as predictor for soil properties e.g., (Vohland et al., 2014).

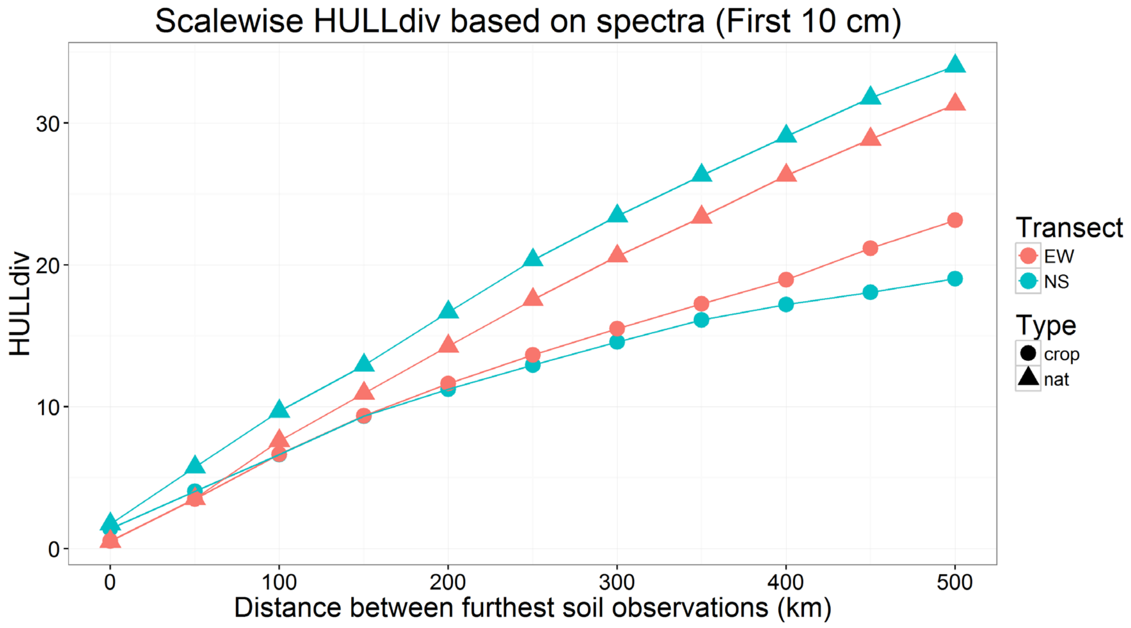


Figure 5-12 HULLdiv index calculated using Vis-NIR spectra at different scales across the two transects.

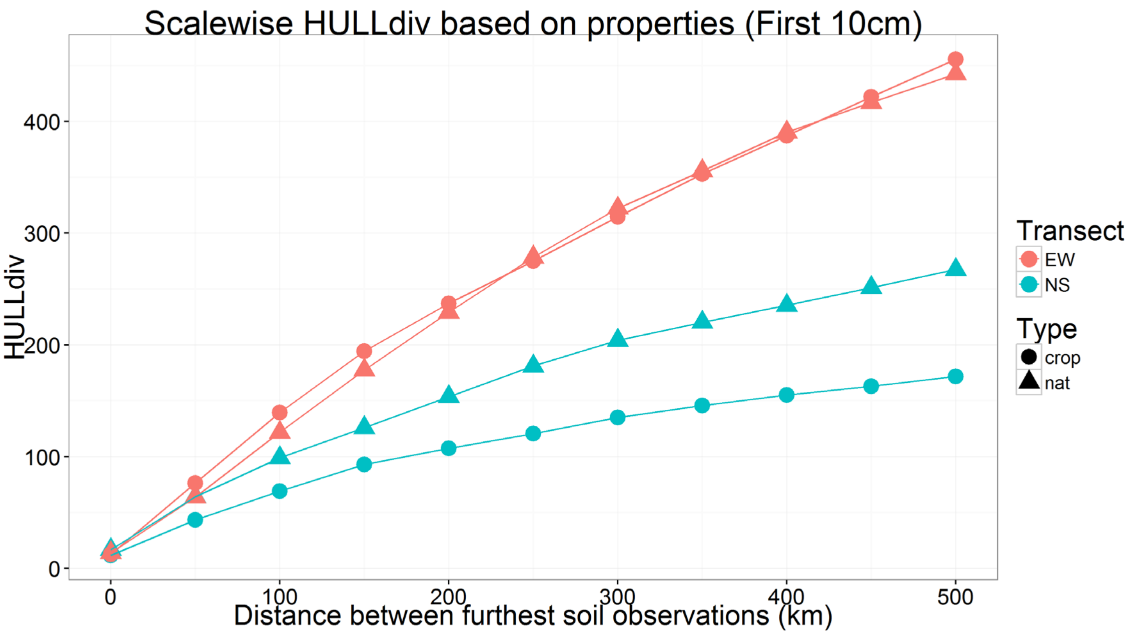


Figure 5-13 HULLdiv index calculated using 5 soil properties at different scales for the two transects.

And third, despite the lower discrimination power between transects when using the whole spectra as a source of soil variation (Figure 5-12), *HULLdiv* with spectra and properties, identified lower soil diversity in the North-South cropping systems when compared against natural conditions. These results are consistent with the land use history of the area, considered as the most intensively cropped area in Australia (Cox et al., 2001) confirming the human triggered detrimental effect on pedodiversity in croplands (Lo Papa et al., 2011).

In the case of *EIGENdiv*, the two sources of information i.e., spectra and properties, resulted in an evident discrimination between transects and also land-use (greater discrimination between land-use *in* Figure 5-14 and between transects *in* Figure 5-15).

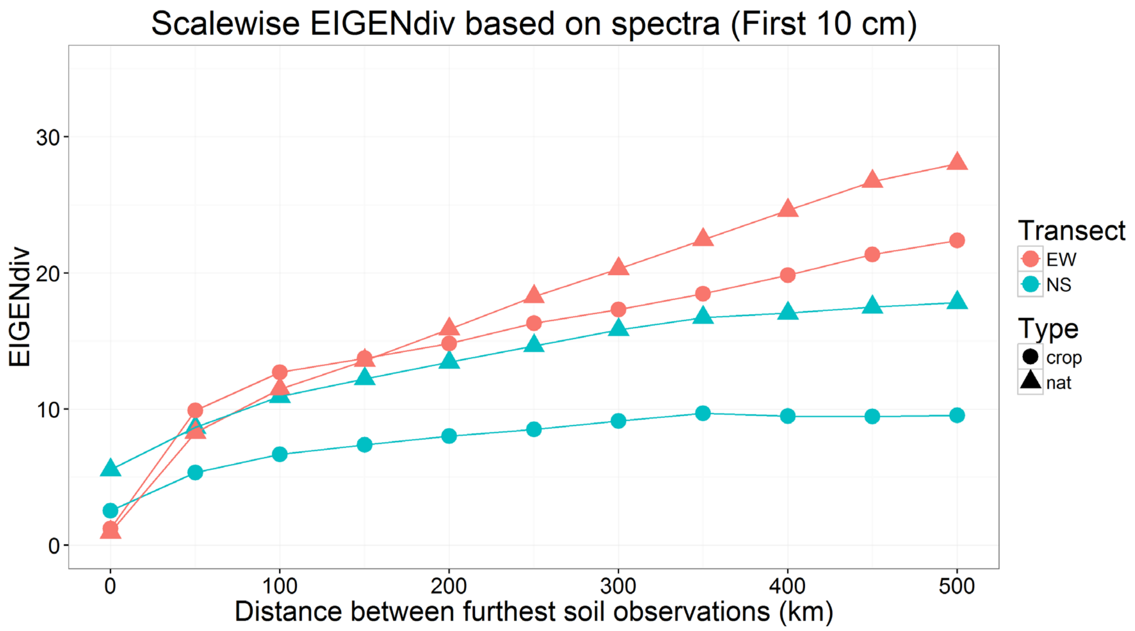


Figure 5-14 EIGENdiv index calculated using Vis-NIR spectra at different scales for the two transects.

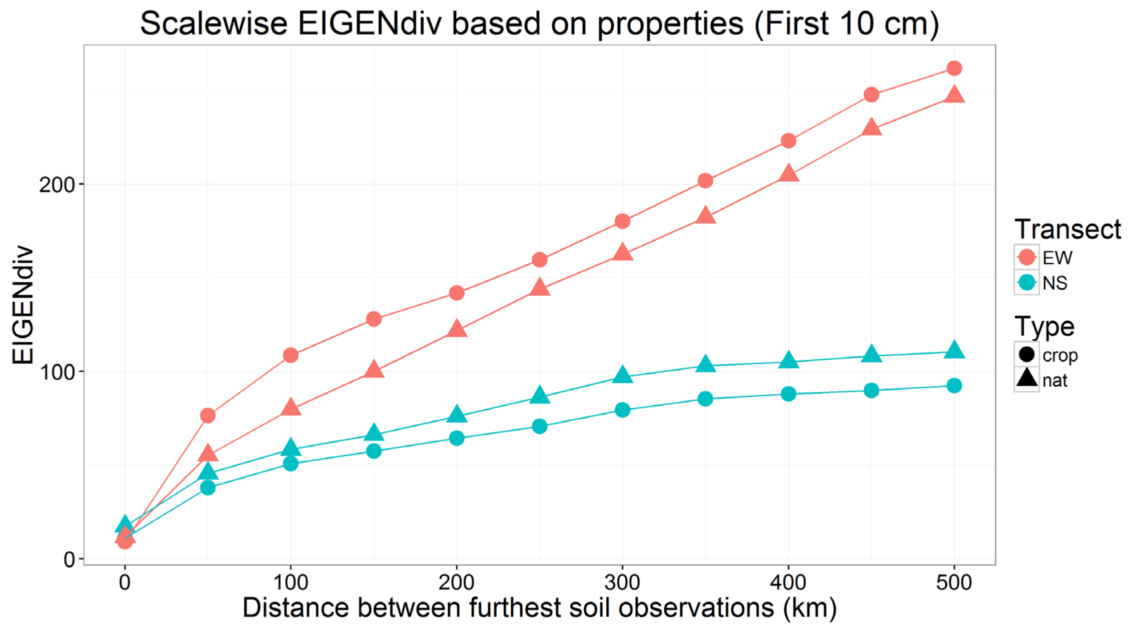


Figure 5-15 EIGENdiv index calculated using predicted properties at different scales for the two transects.

These results compare with those obtained by Maire et al. (2015) in their work on assessing the quality of a multidimensional space on functional ecology, where fish communities (in our cases soil observations) are characterized with a defined number of "traits", (in our case soil properties or selected Vis-NIR wavelengths). They found that a good “multidimensional functional space” in our case a scores matrix built with spectra or selected properties, has to allow a proper discrimination between different functional strategies (in our case landuse or transects). Following their ideas, it is clear that the selected properties represent good discriminators between transects, but on the other hand, when using the whole Vis-NIR spectra, there is a clearer discrimination between land use.

5.4.4 Large scale pedodiversity: Profile

When analysing the different pedodiversity indices at different scales it was observable that in the profile, the effect of land use was masked by the whole profile variability (Figure 5-16), which makes absolute sense, when considering that the variability within a single sample can exceed the variability of soil profiles separated by 50 km as previously shown in Figure 5-8. This masking effect was especially evident in *HULLdiv*, most probably due to the fact that this index considers a projected area and does not consider the distribution of soil observations within that area, depending in the last instance, on the extreme values (or rare specimens in biodiversity terminology) e.g., Surface samples and parent material as soil sciences analogues (Figure 5-16 and Figure 5-17)

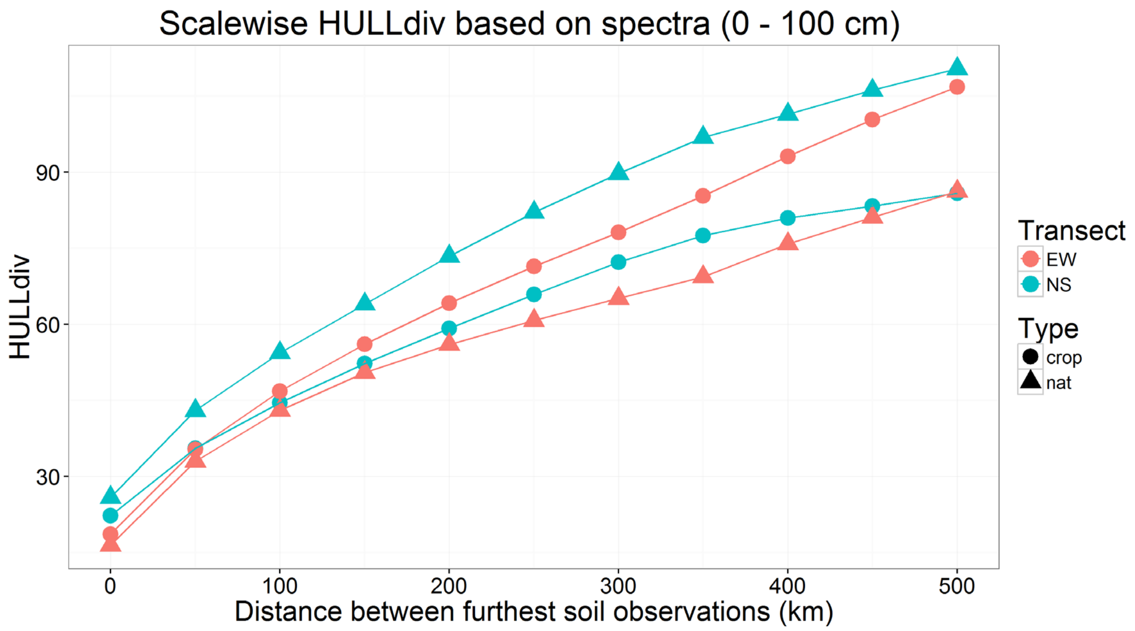


Figure 5-16 HULLdiv index calculated using Vis-NIR spectra at different scales for the two transects (Profile).

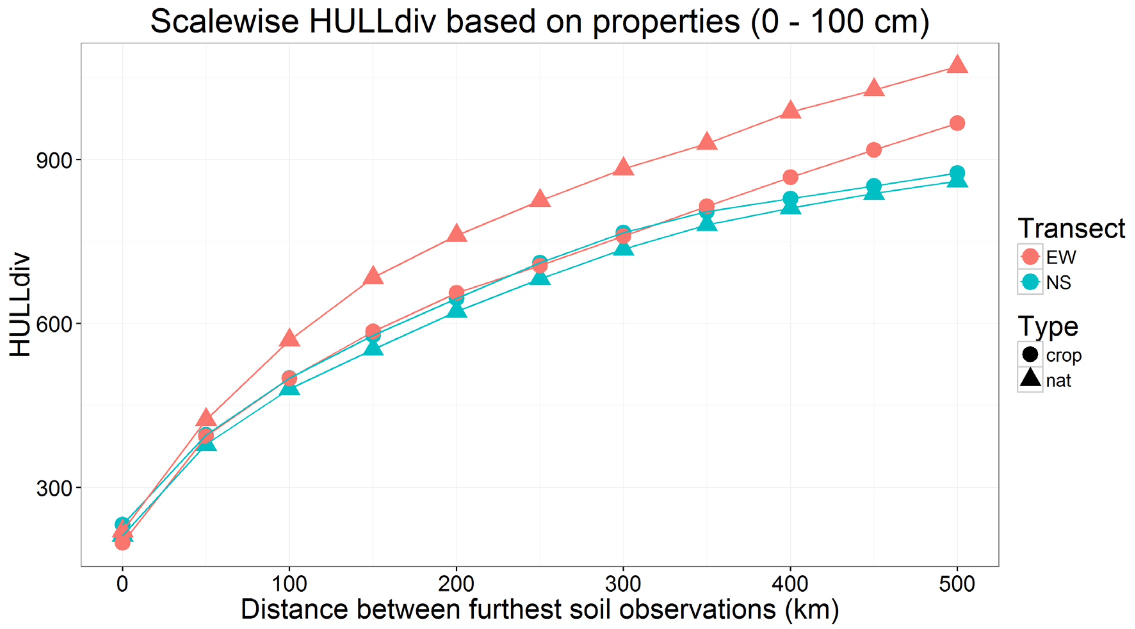


Figure 5-17 HULLdiv index calculated using soil properties at different scales across the two transects (Profile).

For the case of *EIGENdiv* the discrimination between transects and also between different land use was successful, and followed the same pattern presented at the analysis of surface. That is, when using a whole Vis-NIR as input, the discrimination between land uses was exacerbated, and when using properties the discrimination between transects was better assessed suggesting that there is information pertaining to land management within the Vis-NIR range which is not captured by the revised five properties (Figure 5-18 and Figure 5-19).

The differences between both indices i.e., *HULLdiv* and *EIGENdiv,* can be explained by the greater influence of the distribution of soil observations with different characteristics within a profile, since for example *EIGENdiv* measures the overall dispersion of all the soil observations from a hypothetical centroid (*see* Equation 9) and gives less importance to outlier or extreme values compared with *HULLdiv*.

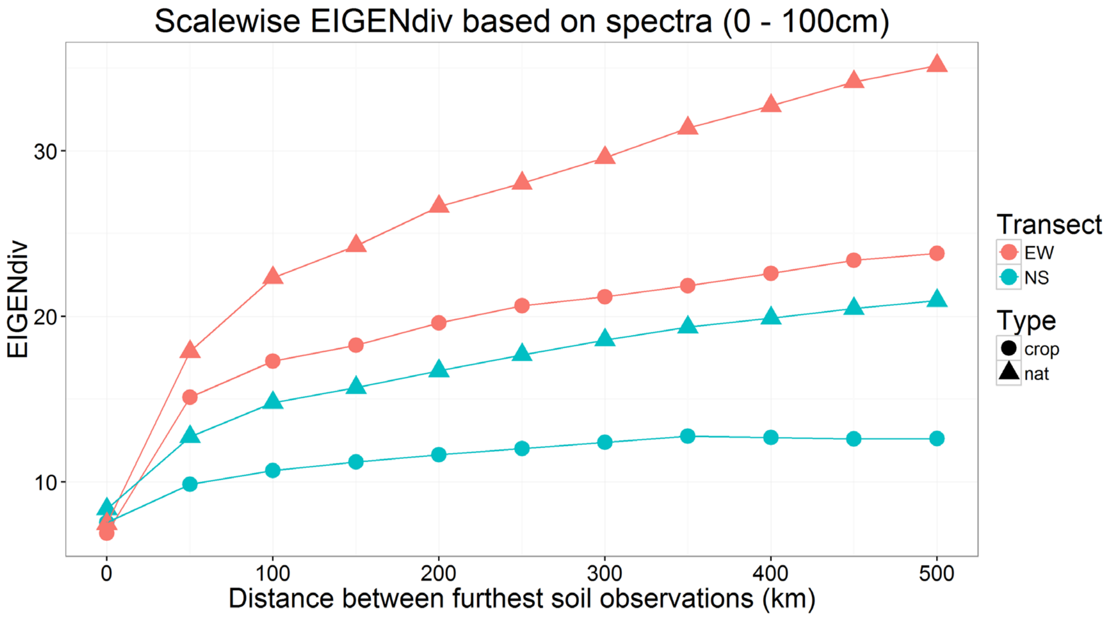


Figure 5-18 EIGENdiv index calculated using Vis-NIR spectra at different scales for the two transects (Profile).

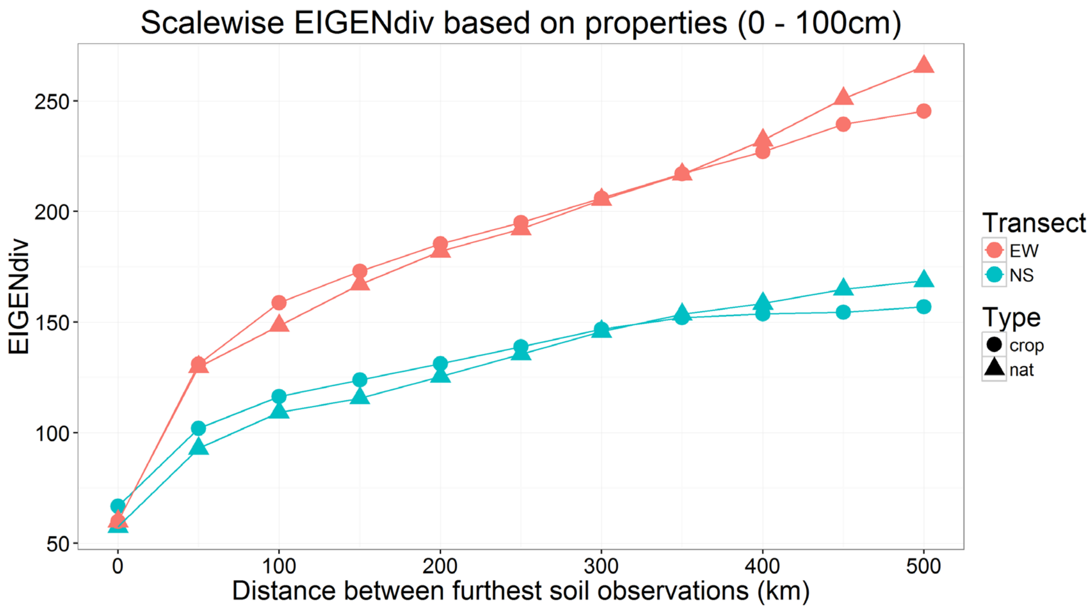


Figure 5-19 EIGENdiv index calculated using predicted properties at different scales for the two transects (Profile).

5.4.5 Large scale pedodiversity: Conventional Indices

Finally, after calculating Shannon’s and Simpson’s indices it was evident that a higher pedodiversity was maintained in the East-West transect even at a 50km scale as may be expected by simple observation of the stronger variations of the selected soil formation factors (temperature, rainfall and relief). The values for different land use were almost identical, basically due to the same soil order denomination on managed and natural sites caused by the pixel resolution of the map used (Figure 5-20 and Figure 5-21).

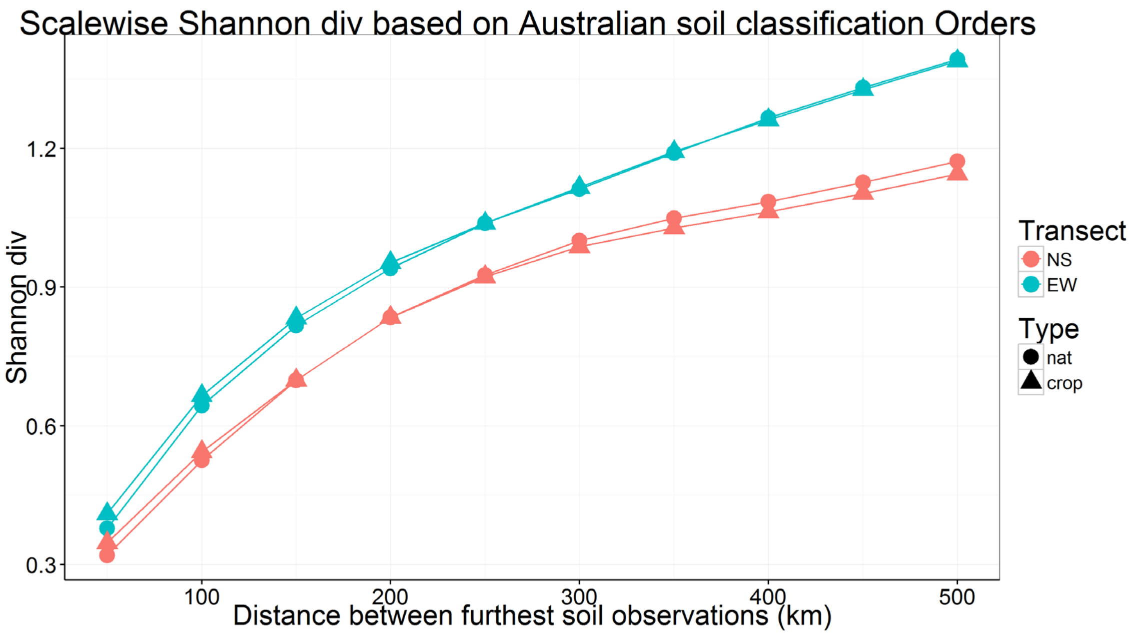


Figure 5-20 Shannon index calculated using Australian Soil Classification orders for two transects and land uses.

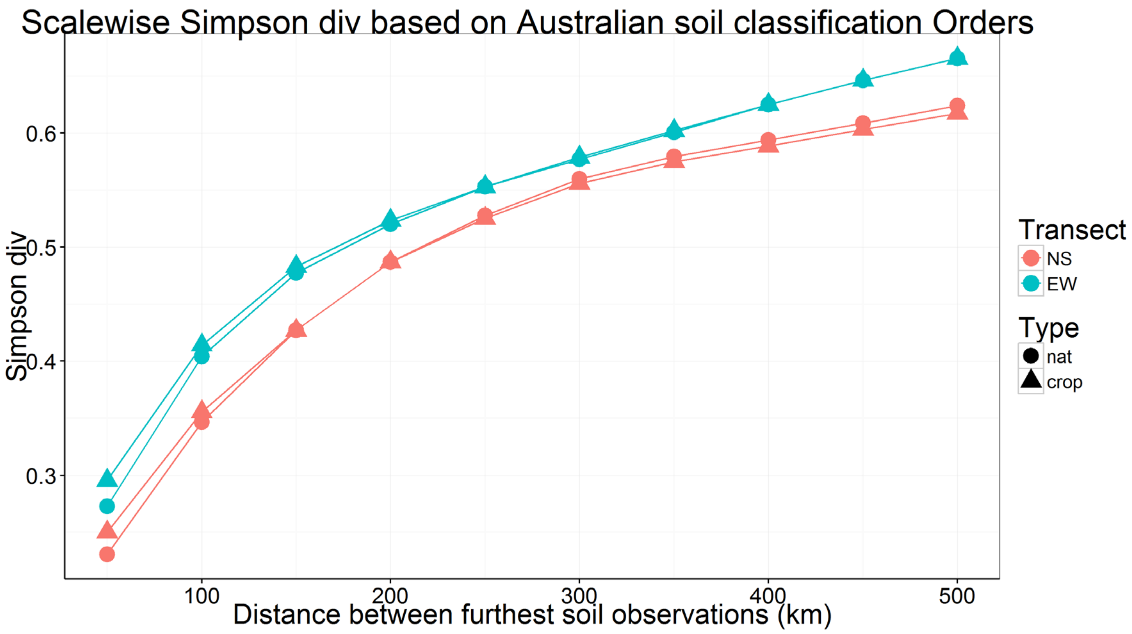


Figure 5-21 Simpson index calculated using Australian Soil Classification orders for two transects and landuse.

5.4.6 Functional and taxonomic pedodiversity

A final comparison between *EIGENdiv* measured with the spectra of whole profile observations (0 to 100cm) and conventional indices (Simpson's index) calculated at different scales (same than previous section), shows evident linear relations for both transects and land-uses (Figure 5-22). These results are expectable for a simple reason, because every soil taxonomy system is based on soils characteristics or functionality.

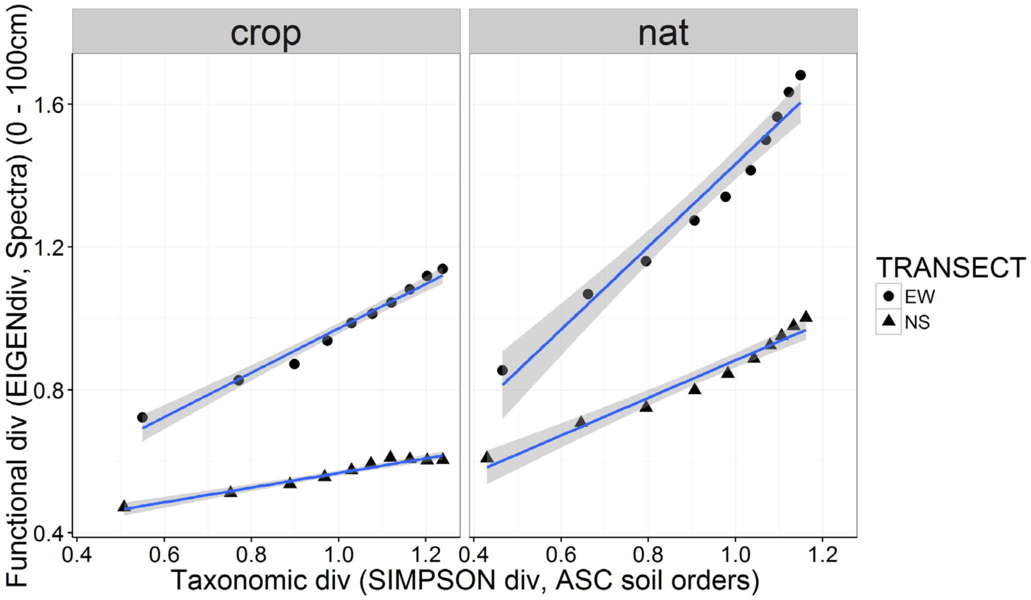


Figure 5-22 Functional vs Taxonomic pedodiversity. Note that values were standardized in order to show the results in the same scale.

Finally, this study has shown that there are linear relations between taxonomical diversity and the variation of soil properties when they are considered as a group by means of a multidimensional index e.g., *EIGENdiv*.

5.5 Conclusions

After reviewing the performance of the newly developed continuous indices of pedodiversity in two transects in NSW with different taxonomic pedodiversity indices, it is possible to highlight some interesting results.

The newly developed indices were effective in identifying pedodiversity; however they performed in a different way depending on the characteristics of the input variables. It was observed that, when using the entire Vis-NIR spectrum as input, the discriminative power between land uses was enhanced, these results suggest that human influence is related with an homogenization of the soil characteristics detected by Vis-NIR reflectance and that this change involves more than just the 5 selected soil properties.

On the other hand, when using five selected soil properties as input, the discrimination between transects was superior, the explanation for this behaviour can be difficult to interpret and remains as an open question since it is possible that this discrimination could be even higher by the addition of other properties.

It was also recognizable that *HULLdiv* was greatly influenced by soil observations with extreme behaviour, basically due the nature of the index, which is based on a polygon that is determined by its extreme values or rare specimens. This phenomenon was exacerbated when using all the information contained in the whole soil profile. *EIGENdiv* was more discriminative both in surface and in depth, mainly because it reflects the overall behaviour of an area giving less importance to outliers. It is also good to remark on the high value of using Vis-NIR spectroscopy. Since, it provided the means for analysing in great detail the soil diversity.

Finally, the results of this study showed linear relations between soil properties variation or functional pedodiversity and taxonomical diversity.

Chapter references

ABARES, Australian Bureau of Agricultural and Resource Economics and Sciences, Geoscience Australia, CSIRO Land and Water, Department of the Environment, Australian Bureau of Statistics,
National Centre for Social Applications of Geographic Information Systems, 2014. 2014 Australian national map layers.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91(1–2), 27-45.

Bockheim, J.G., Haus, N., 2013. Soil endemism and its importance to taxonomic pedodiversity. Pedodiversity. CRC press, Boca Raton, FL, 195-210.

Cox, S.J., Sivertsen, D.P., Bedward, M., 2001. Clearing of native woody vegetation in the New South Wales northern wheatbelt: extent, rate of loss and implications for biodiversity conservation.

Demattê, J.A.M., Terra, F.d.S., Quartaroli, C.F., 2012. Spectral behavior of some modal soil profiles from São Paulo State, Brazil. Bragantia 71, 413-423.

Fajardo, M., McBratney, A., Whelan, B., 2016a. Fuzzy clustering of Vis–NIR spectra for the objective recognition of soil morphological horizons in soil profiles. Geoderma 263, 244-253.

Fridland, V.M., 1974. Special Issue Soil Science in the U.S.S.R Structure of the soil mantle. Geoderma 12(1), 35-41.

Guo, Y., Gong, P., Amundson, R., 2003. Pedodiversity in the United States of America. Geoderma 117(1-2), 99-115.

Harte, J., Smith, A.B., Storch, D., 2009. Biodiversity scales from plots to biomes with a universal species-area curve. Ecology letters 12(8), 789-97.

Hartemink, A.E., Minasny, B., 2014. Towards digital soil morphometrics. Geoderma 230, 305-317.

Ibáñez, J.J., 1996. Pedodiversity, pedometrics and ecological research. Pedometron, Newsletter of the International Society of Soil Science Working Group on Pedometrics (4-5).

Ibáñez, J.J., Gómez, R.P., 2016. Diversity of Soil-Landscape Relationships: State of the Art and Future Challenges. In: J.A. Zinck, G. Metternicht, G. Bocco, H.F. Del Valle (Eds.), Geopedology. Springer International Publishing, pp. 183-191.

Islam, K., McBratney, A., Singh, B., 2005. Rapid estimation of soil variability from the convex hull biplot area of topsoil ultra-violet, visible and near-infrared diffuse reflectance spectra. Geoderma 128(3-4), 249-257.

Jacuchno, V., 1976. K voprosu opredelenia raznoobrazija struktury pocvennogo pokrova. Tezisy dokl. III. VSES Sov. Po structure pocv. Pokrova. Vaschnil, Moskvao.

Jie, C., Xue-lei, Z., Zi-tong, G., Jun, W., 2001. Pedodiversity: a controversial concept. Journal of Geographical Sciences 11(1), 110-116.

Linkeš, V., Simonyová, S., Tobrman, D., 1983. Zložitosť štruktúry pôdneho pokryvu hodnotena pomocou entropie. Ved. Pr. VU podozalec. Vysivy rastl. Bratislave, 133-145.

Lo Papa, G., Palermo, V., Dazzi, C., 2011. Is land-use change a cause of loss of pedodiversity? The case of the Mazzarrone study area, Sicily. Geomorphology 135(3–4), 332-342.

Maire, E., Grenouillet, G., Brosse, S., Villéger, S., 2015. How many dimensions are needed to accurately assess functional diversity? A pragmatic approach for assessing the quality of functional spaces. Global Ecology and Biogeography 24(6), 728-740.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154(1–2), 138-152.

McBratney, A., Minasny, B., 2007. On measuring pedodiversity. Geoderma 141(1–2), 149-154.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2014. vegan: Community Ecology Package.

Orlóci, L., Anand, M., Pillar, V.D., 2002. Biodiversity analysis: issues, concepts, techniques. Community Ecology 3(2), 217-236.

Pavoine, S., Bonsall, M.B., 2011. Measuring biodiversity to explain community assembly: a unified approach. Biological reviews of the Cambridge Philosophical Society 86(4), 792-812.

Petchey, O.L., Gaston, K.J., 2006. Functional diversity: back to basics and looking forward. Ecology letters 9(6), 741-58.

Petersen, A., Gröngröft, A., Miehlich, G., 2010. Methods to quantify the pedodiversity of 1 km2 areas — results from southern African drylands. Geoderma 155(3-4), 140-146.

Ponce-Hernandez, R., Marriott, F.H.C., Beckett, P.H.T., 1986. An improved method for reconstructing a soil profile from analyses of a small number of samples. Journal of Soil Science 37(3), 455-467.

Preston, F.W., 1962. The Canonical Distribution of Commonness and Rarity: Part I. Ecology 43(2), 185-215.

Renka, R.J., Gebhardt, A., Eglen, S., Zuyev, S., White, D., 2015. tripack: Triangulation of Irregularly Spaced Data. R package version 1.3-7.

Saldaña, A., Ibáñez, J.J., 2004. Pedodiversity analysis at large scales: an example of three fluvial terraces of the Henares River (central Spain). Geomorphology 62(1-2), 123-138.

Shannon, C., 1948. BA mathematical theory of communication,[Bell System Tech. J.

Simpson, E.H., 1949. Measurement of diversity. Nature.

Tilman, D., 2001. Functional diversity. Encyclopedia of biodiversity 3(1), 109-120.

Toomanian, N., Esfandiarpoor, I., 2010. Challenges of pedodiversity in soil science. Eurasian Soil Science 43(13), 1486-1502.

Vasques, G.M., Demattê, J.A.M., Ramírez-López, L., Terra, F.S., 2014. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. Geoderma 223-225, 73-78.

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

Whittaker, R.H., 1972. Evolution and Measurement of Species Diversity. Taxon 21(2/3), 213-251.

Chapter 6 *General discussion, conclusions and future work*

"Lo sospeché desde un principio…"

Roberto Gómez Bolaños

6.1 Studying soil complexity

It has been shown that the study of soils is a complex task, from the micro-scale (e.g., colloidal composition of a soil aggregate of 2mm in Chapter 2), to the site scale (e.g., by distinguishing two horizons in a meter of depth in Chapter 4), and finally to state wide and even continental scale in Chapter 3 and Chapter 5 looking for an efficient way to compare observations or looking for geographical patterns in samples separated by hundreds of kilometres.

So far, there is a plethora of methods that can successfully characterize the different aspects of soil, like soil structure or morphology and others. However, it has been shown in the current work, that the study of these sub-disciplines (and probably many others) share something in common, this is that they study the interaction of multiple soil properties that need to be considered simultaneously.

By following this line of thought, it is possible to foresee an optimistic horizon in the use of methodologies that deal with the inherent multi-variability of soils. This thesis has presented some of the possibilities for an increasingly studied area of soil sciences: the use of soil spectroscopy or more specifically the soil variation captured by measuring the reflectance in the visible, near and middle infrared part of the electromagnetic spectrum.

6.2 Electromagnetic spectrum

As reviewed by Kaniu and Angeyo (2015), the use of spectroscopy in soil science holds great potential for the assessment of the soil condition. However, the study of the regions within the Vis-NIR and MIR part of the electromagnetic spectrum is only a fraction of what could be characterized by spectral means.

For example, as was observed in Chapter 4, when spectrally derived soil horizons were identified. Even though they resembled with traditional soil morphological horizons, they were not exactly the same, and this was due the different information that was used for discriminating between soil horizons e.g., human senses like sight and touch vs the information contained in the Vis-NIR regions of the electromagnetic spectrum.

Following this idea, the same chapter emphasized the possibility of adding more information to the algorithm e.g., X-Ray fluorescence (XRF) and with this, adding variability previously not considered.

In the same way, Part 2 of Chapter 2, revealed that, as more information was used by spectral models to predict Slaking Index coefficient *a* (SI*a*) values, other regions of the spectra were considered as important (i.e., MIR, *see* section 2.10.4.2).

The same effect was discussed in Chapter 5, where the success in discriminating between different land-uses pedodiversity was affected by the different source of variation (spectra vs selected soil properties, *see* section 5.4.4).

Current methodologies of spectroscopy not studied in this thesis, and that have proven to have great potential in characterizing soils, are for example the use of the previously mentioned XRF for the identification of heavy metals or soil contaminants (Horta et al., 2015), Laser induced breakdown (LIBS) and Laser induced fluorescence (LIF) (Hilbk-Kortenbruck et al., 2001) or Raman spectroscopy (Luna et al., 2014) among other methodologies which consider other regions of the electromagnetic spectrum.

The value of this thesis is that each of the methodologies presented can be implemented considering other spectral regions, increasing with this, our capability of analysing soils in their inherent multivariate nature.

6.3 Getting the big picture

As the latest studies in soil spectroscopy have shown e.g.,(Viscarra Rossel et al., 2016), the use of spectroscopic tools is mainly used for the prediction of particular soil properties. This phenomenon is comprehensible, since the study of separate soil properties have proven to have great influence over major issues, like soil nitrogen content over wheat yields, or the soil organic carbon by reducing the risk of erosion in large areas, among many others.

After reviewing the different methodologies here presented, it can be seen that for example Chapter 2 and Chapter 3 followed the same schema that other authors have used when dealing with spectroscopic tools.

Chapter 2 identifies those specific features of the spectra and relate those with a target feature, in this thesis case, the hypothetical maximum slaking of a soil sample submerged in water or SI*a*. Following this idea, it can be said that this chapter focuses in detecting those known soil attributes e.g., pH, that influence the soil aggregation and once they are detected, they could be modified and with this it could be possible to "control" the slaking behaviour of soils. Furthermore, Chapter 3 deepens in the use of spectroscopy as a tool for prediction of a certain soil property (%TC) at a local scale and revises the applicability of some of the existing methods for the use of large spectral libraries.

On the other hand if the intention is to discriminate between horizons or between soil types, the single property approach appears to be inadequate. Consequently, Chapter 4 and Chapter 5 follow a different approach, this is to use all the information contained in the spectra as a source of variation or at least a variety of strategic soil properties at the same time.

Based on the results of Chapter 4, it can be said that there is still much to learn about how to effectively characterize the soil profile. Even though our conventional morphological descriptions are useful in describing a soil profile and in communicating that information, it has been shown that our known soil morphological horizons are designated based only on what we can see plus the soil properties that our methods can detect e.g., percentage of organic matter, exchangeable aluminium, etc. However, if those soil attributes that cannot be perceived by traditional methods are included, e.g., the information that is contained in the NIR part of the spectra, somewhat different soil horizons or "spectrally derived soil horizons" could be detected.

Interestingly, having observed that conventional soil horizons and spectrally derived horizons are slightly different, one would expect in Chapter 5 that the variation of soil orders in the landscape or “Taxonomic pedodiversity”, which is based on described soil morphological horizons and the variation of spectra in space or the “Functional pedodiversity”, would not be similar or at least show some disagreement. This was not the case, as linear relations between different soil types in the landscape and their respective spectral attributes were found. These are the first studies that bring these questions, and with them different research opportunities can be anticipated in the short and the long term.

6.4 General conclusions

Regarding the main hypothesis of this thesis it can be concluded that it was possible to measure and represent in a practical, objective and replicable way, the variation of different hierarchical levels of soil organization by using spectroscopic information.

From the colloidal and microaggregation level, Chapter 2 presented a new methodology for describing and measuring the slaking of soil aggregates submerged in water expressed in a Slaking Index (SI). It was shown that is possible to characterize and predict the values of the SI methodology by spectral means. The results were coherent with previous research on soil aggregation, showing the effect of known colloidal fractions involved in the soil aggregation recognized in this study by spectral means. Also, it was shown that there was an improvement in the characterization of the slaking process by adding more information from different parts of the electromagnetic spectrum.

Regarding the use of spectroscopic modelling techniques and large spectral libraries for the prediction of a single property in a determined area, it was observed that the use of this type of spectral libraries holds great potential, however it was revealed that it is greatly influenced by the variability or range of the spectra used for training models. Due to the previous, it is possible to introduce a large amount of variation with a comparatively small amount of samples which can result in a possible addition of error.

Regarding the use of spectroscopy for the characterization of a higher level of organization, e.g., soil morphological horizons, it was shown that is possible to describe the variation contained in the soil horizons by spectral means and also to establish in an objective way, their dividing boundaries. In order to use a spectroscopic approach for describing soil horizons, a specific framework needs to be followed. First, the profile is described at regular intervals by using the multivariate information contained on each observation. Second, those observations can be grouped by their respective spectral similarity and third, by examining their variation in depth, the soil horizon boundaries can be effectively established.

Finally, it was shown that the use of the multivariate information contained in the Vis-NIR region of the spectra, efficiently described the variation of the different soil types from the point-scale to a large-scale. The new indices of pedodiversity represent a contribution towards the characterization of the soil as a continuous, showing that its use can provide results attuned with the conventional methods of describing pedodiversity by discretising soil types but without its inherent bias.

6.5 Opportunities and future work

Each of the research chapters in this thesis has shown that is possible to effectively describe by spectral means the soil variation contained at different levels of soil organization. In the same way, it was shown that there was still room for improvement, which can be summarized as follows.

6.5.1 Practical significance of the Slaking Index

The first part of Chapter 2 presented a new method for assessing soil aggregates stability in fast wetting conditions and also was able to describe the kinetics of slaking of soil aggregates immersed in water. It was shown that the method represented a considerable improvement in terms of cost and time compared to conventional methods like measurement of Mean Weight Diameter (MWD) with a wet sieving apparatus. Despite this, it was shown in the second part of the chapter that the methodology is still time consuming in comparison with even faster methodologies like the use of Vis-NIR or portable MIR spectroscopic.

In respect to the previous, and based on the uncertainty related to each of the methodologies (image recognition and spectroscopy) a key question arises, which is what is the practical threshold e.g., agronomical significance, for the new soil slaking index, consequently, further research in this area will need to address this question.

For example, in a scenario where the agronomical significance of a soil with a Slaking Index (SI) of 1 does not differ substantially from another with a SI of 3, the spectroscopic method should be preferred, considering first, the spectral similarity of the samples where the prediction is intended, since as it was shown, the uncertainty of the spectroscopic approach was not bigger than 2 SI units.

6.5.2 Spectral information related to soil slaking

The second part of the same chapter showed that as other regions of the electromagnetic spectrum were added to the predictive models, their performance improved considerably. As also mentioned before, there are many other methodologies which represent compositional changes by using other parts of the spectra. In this regard, observing which parts of the electromagnetic spectrum are good predictors of SI could give a better understanding of the soil slaking process. Even more, considering that only the asymptotic value of the slaking process was predicted (Coefficient *a*), then the other coefficients i.e., Coefficient *b*, related with macro aggregation and Coefficient *c* with the rate of slaking, can be also analysed by spectral means.

6.5.3 Large spectral databases

Viscarra Rossel et al. (2016), remarked on the need for more research to optimally use large spectroscopic databases for local predictions of soil attributes. With respect to this, Chapter 3 analysed the use of state-of-the-art spectroscopy techniques for the prediction of an important soil property (TC %) of samples from a certain geographic region.

The results showed the importance of the spectral variability of soil samples and the high impact that a few samples with a different spectral behaviour can produce in a predictive model i.e., spike sample. In this regard, it was shown that the inclusion of spectrally different samples can be a “double edged tool”, since this procedure “expands” the predictive space of the models and it was seen that if the distribution of values of included samples is biased, this will be also inherited by the predictive models.

It was also shown the high influence of a geographically closer sample over the final prediction bias on local samples. The results of this study have shown that, as other authors have concluded e.g., (Guerrero et al., 2016), the creation of large spectral libraries is still a challenge.

6.5.4 Soil morphological horizons

Chapter 4 showed that the framework presented in the Rationale (Chapter 1) was adequate for the analysis of soil horizons. This chapter showed that as the complexity of soil organization increased, i.e., soil colloids < soil aggregates < soil horizons, the approach for describing the variation within soil horizons had to involve the analysis of the vertical variation of multiple properties (represented by the spectral information of a single scan).

The fact that the spectrally derived horizons (SPD *hor*) were somehow different to the ones described by a pedologist, showed that the delineation of soil horizon boundaries is indeed subject of the variation perceived, therefore it makes sense that if the soil information used in the discrimination of soil horizons is different e.g., human senses vs Vis-NIR, the resultant soil morphological horizon boundaries will be different as well.

It was shown that it is possible to discriminate between horizons, however there is still an open question, which is the possibility of allocating those already identified soil horizons to a certain class or as conventional taxonomy describes "soil diagnostic horizons". The challenge and opportunity of further studies, involves the analysis of the type of spectra (or a spectral range) that could serve as a proxy for possible "spectrally derived diagnostic horizons" contributing with this to a more objective soil classification. In this regard Triantafilis et al. (2001), Odgers et al. (2011a) and recently Hughes et al. (2014) have shown that is possible to create representative classes or "centroids" considering multiple soil properties and they can be successfully used for classification purposes.

6.5.5 Pedodiversity

It was observed that the multivariate information contained in the soil spectrum was as hypothesized, useful in describing the variation of soil types in the landscape. The results in this chapter are very promising, if we consider that at the moment, many authors and organizations are specially interested in the creation of large Vis-NIR spectral libraries (as commented in Chapter 3). The presented indices can be straightforwardly applied in those global spectral libraries and with this, the soil diversity of local areas, countries, continents and the current global pedodiversity can be calculated if it's desired. This idea was also noted by Viscarra Rossel et al. (2016) for a possible use of the newly created Global spectral library in their statement: "*We showed that the information it contains can be used to characterize soil and its variability and diversity globally* ".

Because of the high influence of the input information i.e., soil properties or Vis-NIR spectra, on the final pedodiversity results, more research on identifying those properties or spectral regions, that contribute to discriminate between different “soil types” is crucial. The research needed in this regard, is intrinsically related with the one mentioned on previous section, considering an efficient discrimination between classes that could represent in a practical way the current soil resource.

As a final conclusion, it was shown that the soil variation can be efficiently described by spectral means, since this type of information captures the intrinsic multi-variability of soils. Furthermore, as the complexity of the soil attribute increases, it was shown that it is also possible to use the same spectral information in addition to other statistical techniques like spectral modelling or fuzzy clustering, to describe the soil variation in depth and in the landscape and with this, creating a complete representation of soil variation from the micro-scale to the global-scale.

Chapter references

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-Sinoga, J.D., Zornoza, R., Viscarra Rossel, R.A., 2016. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil and Tillage Research 155, 501-509.

Hilbk-Kortenbruck, F., Noll, R., Wintjens, P., Falk, H., Becker, C., 2001. Analysis of heavy metals in soils using laser-induced breakdown spectrometry combined with laser-induced fluorescence. Spectrochimica Acta Part B: Atomic Spectroscopy 56(6), 933-945.

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

Hughes, P.A., McBratney, A.B., Minasny, B., Campbell, S., 2014. End members, end points and extragrades in numerical soil classification. Geoderma 226-227, 365-375.

Kaniu, M.I., Angeyo, K.H., 2015. Challenges in rapid soil quality assessment and opportunities presented by multivariate chemometric energy dispersive X-ray fluorescence and scattering spectroscopy. Geoderma 241-242, 32-40.

Luna, A.S., Lima, I.C.A., Rocha, W.F.C., Araujo, J.R., Kuznetsov, A., Ferreira, E.H.M., Boque, R., Ferre, J., 2014. Classification of soil samples based on Raman spectroscopy and X-ray fluorescence spectrometry combined with chemometric methods and variable selection. Analytical Methods 6(22), 8930-8939.

Odgers, N.P., McBratney, A.B., Minasny, B., 2011a. Bottom-up digital soil mapping. I. Soil layer classes. Geoderma 163(1–2), 38-44.

Triantafilis, J., Ward, W.T., Odeh, I.O.A., McBratney, A.B., 2001. Creation and interpolation of continuous soil layer classes in the lower Namoi valley. Soil Science Society of America Journal 65(2), 403-413.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Science Reviews.

Appendices

Transec	***#***	***Land-use***	***Land cover obs.***	***Detailed obs.***
NS	*0*	*nat*	*River shore*	*High density river shore grassland medium density of small to medium trees and*
NS	*0*	*crop*	*Cereal*	*Low density dry wheat cane on clayey soil.*
NS	*1*	*nat*	*Woodland*	*Regular density of small to medium trees. Medium to low shrubs.*
NS	*1*	*crop*	*Fallow*	*Fallow bare cracked soil chickpeas were the last crop.*
NS	*2*	*nat*	*Woodland*	*Regular density small to medium trees, broad areas of medium and long very green*
NS	*2*	*crop*	*Cereal*	*Regular density dry wheat cane with clover*
NS	*3*	*nat*	*Shrub land*	*High to medium density shrubs (mostly young acacias).*
NS	*3*	*crop*	*Cereal*	*Regular density dry wheat cane.*
NS	*4*	*nat*	*Woodland*	*Medium density of tall trees with low grass.*
NS	*4*	*crop*	*Fallow*	*Fallow*
NS	*5*	*nat*	*Woodland*	*Low density medium trees regular to low density medium dry/green grass.*
NS	*5*	*crop*	*Cereal*	*Low density dry wheat cane.*
NS	*6*	*nat*	*Woodland*	*High density Pilliga scrub and Acacia cypress.*
NS	*6*	*crop*	*Ploughed*	*Ploughed.*
NS	*7*	*nat*	*Forest*	*High density tall casuarina trees with high density tall grass*
NS	*7*	*crop*	*Ploughed*	*Ploughed land*
NS	*8*	*nat*	*Woodland*	*Low to medium density tall trees regular low grass.*
NS	*8*	*crop*	*Cereal*	*Regular density dry barley cane harvested rotated with lupine for 15yrs.*
NS	*9*	*nat*	*Woodland*	*Regular density medium to tall trees high density tall grass*
NS	*9*	*crop*	*Pasture*	*High density tall grass pasture premier digit grass*

NS	10	nat	Forest	High density tall trees high density leaf litter medium to short grass
NS	10	crop	Pasture	High density medium grass pasture natural growth
NS	11	nat	Forest	High density tall casuarina trees medium density short green grass
NS	11	crop	Cereal	Regular to high density wheat in anthesis. 2.3 ton calc. carbonate. Changing to
NS	12	nat	Woodland	Regular to high density medium to tall trees regular density short grass
NS	12	crop	Alfalfa	Regular density alfalfa in flowering period rotated with canola lupine
NS	13	nat	Woodland	Low density _ravine vegetation high density medium to tall grass
NS	13	crop	Cereal	High density dry wheat cane
NS	14	nat	Woodland	Regular density tall trees, high density medium to tall grass
NS	14	crop	Cereal	High density fresh wheat cane mixed with clove.
NS	15	nat	Woodland	Regular density tall trees, high density medium to tall grass.
NS	15	crop	Cereal	Regular density dry canola cane harvested.
NS	16	nat	Woodland	Low to Regular density medium trees low density short grass.
NS	16	crop	Cereal	Low density fresh wheat cane harvested remaining stubble.
NS	17	nat	Forest	Regular density tall trees low to medium density short grass .
NS	17	crop	Cereal	Medium density wheat cane harvested in the season .
NS	18	nat	Woodland	Regular density medium to tall trees, high density medium to tall grass.
NS	18	crop	Cereal	High density fresh wheat cane harvested. Crop of only two years.
NS	19	nat	Forest	High density tall trees with leaf litter and low density tall grass.
NS	19	crop	Cereal	Low density dry wheat cane harvested years ago remaining stubble.
NS	20	nat	Grassland	High density tall natural grass.

NS	*20*	*crop*	*Cereal*	*Regular density dry wheat cane harvested years ago mixed with short natural grass.*
NS	*21*	*nat*	*Woodland*	*Regular density tall trees with leaf litter and tall grass.*
NS	*21*	*crop*	*Pasture*	*Regular density short grass pasture.*
NS	*22*	*nat*	*Woodland*	*Low density medium trees and shrubs, high density tall grass*
NS	*22*	*crop*	*Cereal*	*High density fresh wheat non-harvested.*
NS	*23*	*nat*	*Forest*	*High density tall trees ,regular density leaf litter, high density tall grass*
NS	*23*	*crop*	*Pasture*	*High density medium to tall grass pasture natural growth.*
NS	*24*	*nat*	*Woodland*	*Low density medium trees and scattered shrubs high density tall grass.*
NS	*24*	*crop*	*Cereal*	*Regular density fresh wheat cane harvested in the season.*
NS	*25*	*nat*	*Woodland*	*Low to regular density medium tree, high density medium grass .*
NS	*25*	*crop*	*Cereal*	*High density fresh wheat cane harvested years ago (remaining stubble).*
NS	*26*	*nat*	*Forest*	*High density tall trees and shrubs, high density grass.*
NS	*26*	*crop*	*Cereal*	*High density fresh wheat cane harvested remaining stubble and cow faeces.*
EW	*27*	*nat*	*Grassland*	*High density tall natural grass.*
EW	*27*	*crop*	*Bananas*	*High density banana (>2m).*
EW	*28*	*nat*	*Forest*	*High density tall trees high density leaf litter no grass*
EW	*28*	*crop*	*Pasture*	*High density short grass improved pasture.*
EW	*29*	*nat*	*Woodland*	*Regular density tall trees, regular density short grass.*
EW	*29*	*crop*	*Pasture*	*High density short grass improved pasture.*
EW	*30*	*nat*	*Woodland*	*Regular to high density medium to tall trees, regular density short grass.*
EW	*30*	*crop*	*Pasture*	*High density short to medium grass improved pasture.*

EW	*31*	*nat*	*Woodland*	*Regular density medium to tall trees, regular density short to medium grass.*
EW	*31*	*crop*	*Pasture*	*Low density short grass pasture. Rill erosion.*
EW	*32*	*nat*	*Forest*	*High density tall trees, high density tall grass and leaf litter.*
EW	*32*	*crop*	*Pasture*	*High density short to medium pasture.*
EW	*33*	*nat*	*Woodland*	*Low density medium trees high density medium grass.*
EW	*33*	*crop*	*Ploughed*	*Ploughed land with improved pasture (Pilliga 3 years ago)*
EW	*34*	*nat*	*Woodland*	*Regular density tall trees, regular density short to medium grass.*
EW	*34*	*crop*	*Pasture*	*Regular to high density short to medium grass*
EW	*35*	*nat*	*Forest*	*High density tall young trees high density tall grass and leaf litter*
EW	*35*	*crop*	*Ploughed*	*Ploughed land*
EW	*36*	*nat*	*Woodland*	*High density tall trees high density leaf litter.*
EW	*36*	*crop*	*Fallow*	*Fallow bare cracked soil.*
EW	*37*	*nat*	*Woodland*	*Low density small trees regular density young shrub low density tall grass.*
EW	*37*	*crop*	*Cereal*	*Regular to low density wheat in cracked soil.*
EW	*38*	*nat*	*Woodland*	*Low density_small trees_regular density tall grass cracked soil*
EW	*38*	*crop*	*Cereal*	*Regular density dry wheat cane harvested (remaining stubble).*
EW	*39*	*nat*	*Woodland*	*Low density medium trees, regular density tall grass. Cracked soil.*
EW	*39*	*crop*	*Fallow*	*Fallow bare cracked soil wheat stubble. Cracked soil*
EW	*40*	*nat*	*Woodland*	*Low density small trees, regular density tall grass.*
EW	*40*	*crop*	*Cereal*	*Regular to low density wheat. Cracked soil*
EW	*41*	*nat*	*Woodland*	*Low to regular density medium trees regular density tall grass. Cracked soil*

EW	*41*	*crop*	*Fallow*	*Fallow bare in cracked soil, low density remaining wheat stubble*
EW	*42*	*nat*	*Grassland*	*High density tall natural grass.*
EW	*42*	*crop*	*Fallow*	*Fallow bare cracked soil, low density wheat stubble.*
EW	*43*	*nat*	*Woodland*	*Regular to low density medium trees low density leaf litter and grass.*
EW	*43*	*crop*	*Cereal*	*Regular density oats in silty soil*
EW	*44*	*nat*	*Woodland*	*Regular density medium trees regular density small shrubs.*
EW	*44*	*crop*	*Dryland*	*No cover.*
EW	*45*	*nat*	*Woodland*	*Low density medium trees low density leaf litter and grass*
EW	*45*	*crop*	*Pasture*	*High density tall grass*
EW	*46*	*nat*	*Woodland*	*Low density shrub with apparent sodic conditions.*
EW	*46*	*crop*	*Dryland*	*No cover.*
EW	*47*	*nat*	*Woodland*	*low density medium trees low density leaf litter and grass*
EW	*47*	*crop*	*Human intervention*	*Poor drainage.*
EW	*48*	*nat*	*Dryland*	*No cover.*
EW	*48*	*crop*	*Woodland*	*Low density medium trees.*

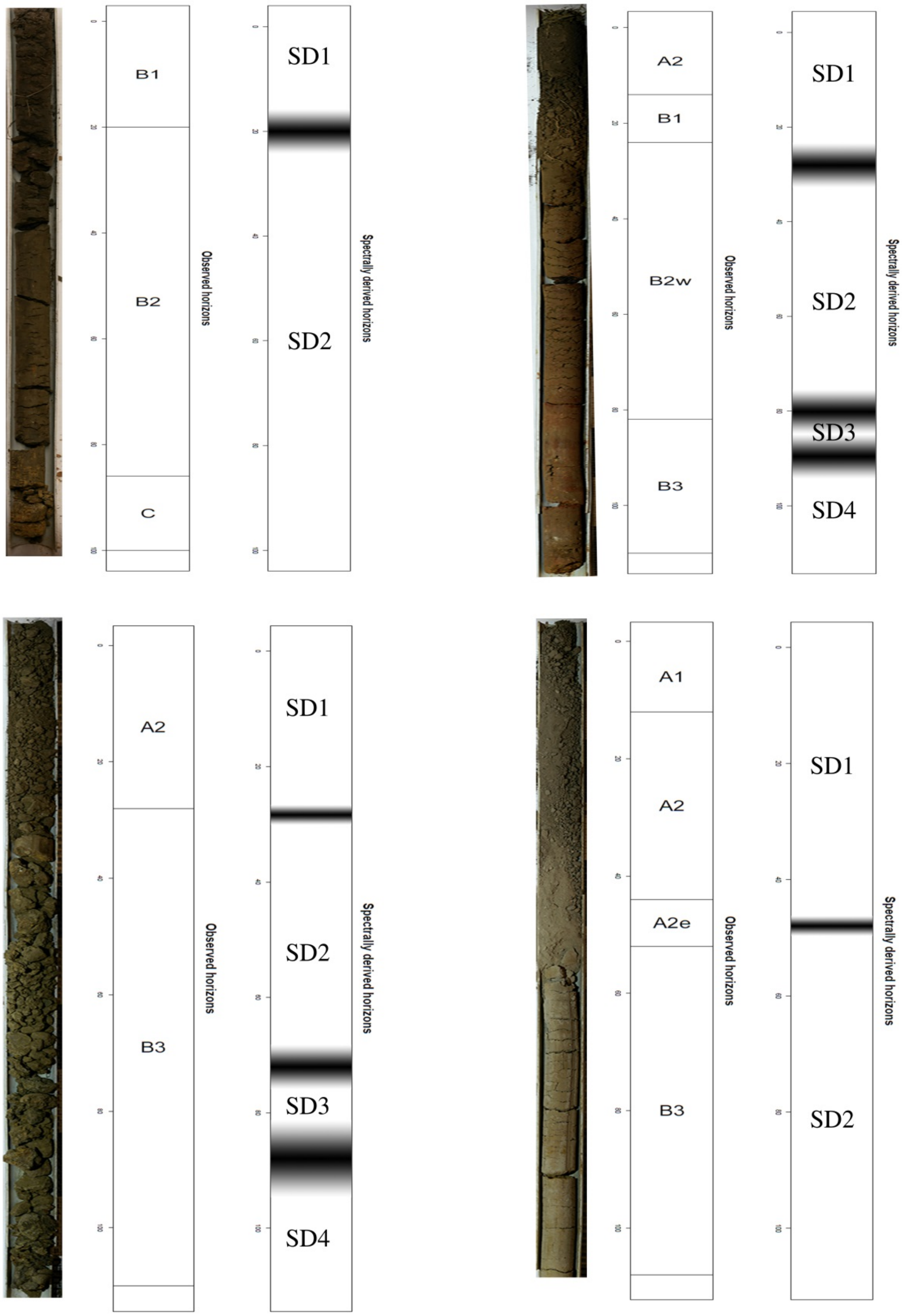
Appendix 1. Land use details of NSW transect sites.

```
#run if packages are not present#
# source("https://bioconductor.org/biocLite.R")
# biocLite("EBImage")
# install.packages('plyr',dependencies = T)

require(EBImage)
require(plyr)

#######INITIAL PARAMETERS#######
Sample_id <- 'Example_Image.JPG'
threshold <- .9
num_agregates <- 5
band <- 3 #blue band
cutting_area<-list(c(1000:2000),c(800:1600))
################################
Initial_image<-readImage(Sample_id)
Cutted_image<- !Initial_image[cutting_area[[1]],cutting_area[[2]],3]>threshold

#Apply filters for taking the '0' values inside the aggregate#
Filtered_image <- closing(Cutted_image, makeBrush(5, shape='disc'))

## Recognize and label each aggregate as a different object##
Labelled_image <- bwlabel(Filtered_image)
colored_image <- colorLabels(Labelled_image)
display(colored_image,method = 'raster')

######Select bigger aggregates######
agregates_id<-names(sort(table(colored_image@.Data[,,3]),decreasing = T))[0:num_agregates+1]
Ag_count <- Image(colored_image@.Data[,,3]*(matrix(colored_image@.Data[,,3]%in%c(agregates_id),ncol=dim(colored_image)[2])),colormode = 2)
result<-list()
result$Ag_count_colored <- bwlabel(Ag_count)
result$original <- Initial_image

#check#
display(result$original,method = 'raster')
display(colorLabels(result$Ag_count_colored),method = 'raster')
computeFeatures.shape(result$Ag_count_colored)[,1] #Final areas(in pixels) of 5 aggregates
#end#
```

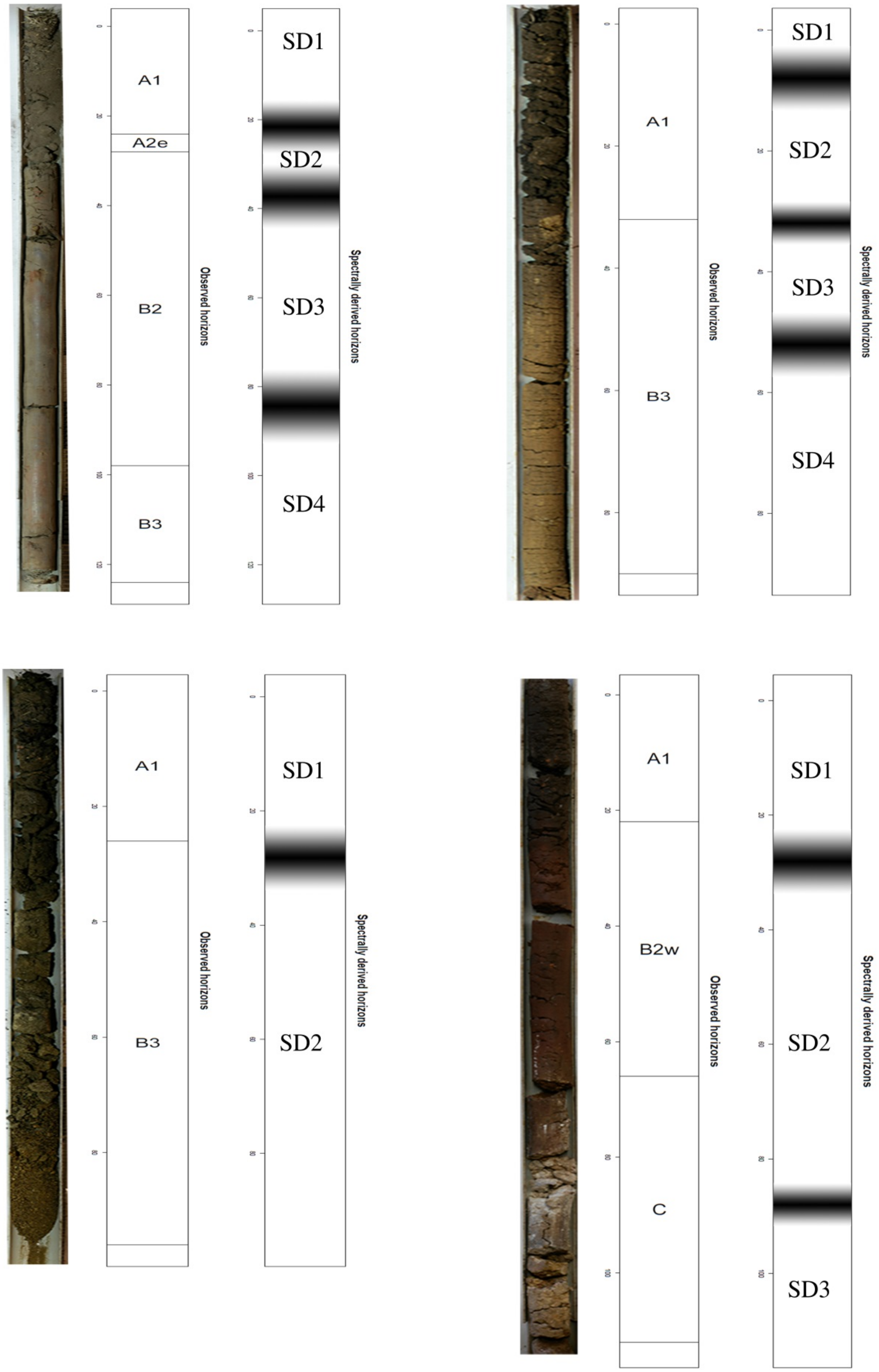
Appendix 2. Example of R code for image recognition of 5 soil aggregates.

Property	*Training* (n = 118)						*Independent Validation* (n = 30)						*External Validation* (n = 23)					
	Min	1Q.	Median	Mean	3Q.	Max	Min	1Q.	Median	Mean	3Q.	Max	Min	1Q.	Median	Mean	3Q	Max
NH4 (mg kg^{-1})	0.1	2.0	4.0	6.0	7.0	53.0	0.1	3.0	4.0	6.3	6.8	40.0	1.0	1.0	1.0	1.6	2.0	3.0
NO$_3$ (mg kg^{-1})	1.0	3.0	9.5	20.5	20.0	137.0	1.0	7.8	21.5	30.6	40.5	202.0	1.0	1.0	1.0	2.1	1.0	10.0
P (mg kg^{-1})	3.0	13.8	29.5	39.7	49.3	273.0	6.0	23.5	38.5	39.1	48.5	94.0	3.0	5.0	10.0	12.2	17.5	32.0
K (mg kg^{-1})	74.0	287.5	442.5	462.6	571.2	1232.0	132.0	321.0	483.0	479.1	563.8	995.0	81.0	148.5	176.0	179.0	213.5	301.0
EC (dS m^{-1})	0.0	0.0	0.1	0.1	0.1	0.3	0.0	0.1	0.1	0.1	0.1	0.5	0.0	0.1	0.1	0.1	0.2	0.5
pH CaCl2	**4.0**	**5.0**	**5.9**	**5.9**	**6.6**	**7.8**	**4.3**	**4.9**	**5.8**	**6.0**	**6.9**	**7.6**	**6.3**	**6.9**	**7.7**	**7.4**	**8.0**	**8.3**
pH H2O	**4.8**	**5.9**	**6.5**	**6.7**	**7.4**	**8.6**	**5.1**	**6.0**	**6.4**	**6.7**	**7.8**	**8.6**	**7.4**	**8.1**	**8.9**	**8.6**	**9.1**	**9.4**
Exc. Al (cmol(+) kg^{-1})	0.0	0.0	0.1	0.1	0.1	1.8	0.0	0.0	0.1	0.1	0.1	1.0	0.1	0.1	0.1	0.1	0.2	0.2
Exc. Ca (cmol(+) kg^{-1})	0.3	4.5	8.3	10.5	16.7	30.6	0.9	5.1	9.5	12.0	20.0	25.5	9.1	11.4	13.6	13.5	15.1	19.8
Exc. Mg (cmol(+) kg^{-1})	**0.3**	**1.3**	**2.9**	**4.1**	**6.7**	**13.1**	**0.5**	**1.5**	**2.6**	**4.0**	**7.1**	**9.7**	**10.8**	**12.6**	**14.2**	**14.7**	**16.4**	**20.0**
Exc. K (cmol(+) kg^{-1})	0.2	0.7	1.1	1.1	1.5	3.0	0.3	0.8	1.2	1.2	1.4	2.2	0.2	0.4	0.5	0.5	0.6	0.8
Exc. Na (cmol(+) kg^{-1})	**0.0**	**0.0**	**0.2**	**0.4**	**0.5**	**3.9**	**0.0**	**0.0**	**0.2**	**0.4**	**0.5**	**2.6**	**0.8**	**1.5**	**2.2**	**3.3**	**4.6**	**8.2**
TC (%)	0.1	0.9	1.6	2.1	2.3	17.8	0.5	1.2	1.6	2.2	2.4	9.3	0.1	0.3	0.4	0.5	0.7	1.0
CEC (cmol(+) kg^{-1})	**1.7**	**7.0**	**12.2**	**16.2**	**26.6**	**38.3**	**3.8**	**8.0**	**12.2**	**17.8**	**30.5**	**35.0**	**21.9**	**27.5**	**32.2**	**32.0**	**36.5**	**42.0**
Clay (%)	1.7	11.9	21.1	26.0	41.9	65.1	6.5	13.7	19.2	27.2	46.5	53.6	23.0	28.5	31.0	30.5	33.0	38.0
Slaking Index	**0.0**	**1.0**	**2.5**	**2.5**	**3.8**	**10.6**	**0.0**	**0.8**	**2.4**	**2.4**	**3.9**	**6.3**	**1.1**	**3.8**	**5.1**	**5.9**	**7.6**	**15.4**

Appendix 3. Descriptive statistics for all the measured properties.



Appendix 4 Example of SD hor. Boundaries distinctness is relative to the thickness of the boundary.



Appendix 5. Examples of SD hor. Boundaries distinctness is relative to the thickness.