A selection of sensing techniques for mapping soil hydraulic properties

M. Knotters, F.M. van Egmond, G. Bakker, D.J.J. Walvoort, F. Brouwer

WAGENINGEN
UNIVERSITY & RESEARCH

A selection of sensing techniques for mapping soil hydraulic properties

Dit onderzoek is uitgevoerd binnen het kader van het programma BIS 2014 in opdracht van het ministerie van Economische zaken, Landbouw en Innovatie (EL&I)
Project code BO-11-017-028

BIS Nederland: Dé bron voor bodeminformatie: www.BISNederland.wur.nl

A selection of sensing techniques for mapping soil hydraulic properties

M. Knotters, F.M. van Egmond, G. Bakker, D.J.J. Walvoort, F. Brouwer

WEnR Report 2853

Wageningen Environmental Research, part of Wageningen UR
Wageningen, 2017

Abstract

M. Knotters, F.M. van Egmond, G. Bakker, D.J.J. Walvoort, F. Brouwer, 2017, *A selection of sensing techniques for mapping soil hydraulic properties;* , Wageningen, Alterra, WEnR Report 2853. 69 blz.; 21 fig.; 21 tab.; 16 ref.

Data on soil hydraulic properties are needed as input for many models, such as models to predict unsaturated water movement and crop growth, and models to predict leaching of nutrients and pesticides to groundwater. The soil physics database of the Netherlands shows several lacunae, and a substantial part of the data were collected more than thirty years ago and thus might not represent actual soil hydraulic conditions. There is a need to fill lacunae in the soil physics dataset, to make the dataset up-to-date and to aggregate soil hydraulic properties observed at point scale to larger spatial units. The relationship between the unsaturated hydraulic conductivity K or the volumetric water content θ and the pressure head h can be described for example by equations such as the Mualem-Van Genuchten equations, the parameter values of which can be predicted from soil properties by using pedo-transfer functions. These soil properties include clay content, loam content, organic matter content and the median grain diameter of the sand fraction (50-2000 μm). Application of proximal sensing techniques in observing soil physics properties in the field might reduce the costs of data collection. The accuracy with which proximal sensing techniques can predict clay content, loam content, organic matter content and the median grain size of the sand fraction (50-2000 μm) needs to be assessed as these are the explanatory variables in the pedo-transfer functions. The aim of this study was to select on-the-go proximal sensing techniques that are able to quantify soil variables that are used in the pedo-transfer functions for prediction of the parameters of the Mualem-Van Genuchten equations. We conclude that near infrared spectrometry, γ-ray spectroscopy and electromagnetic induction methods have a potential in spatial prediction of clay, silt and sand content, and that near infrared spectrometry has a potential in spatial prediction of organic matter content and soil moisture content. The use of pedo-transfer functions requires also the parameters bulk density and median of the sand fraction (M50). Bulk density can be measured by a device called RhoC. However, its performance in predicting bulk density of terrestrial soils has not been validated yet. Most instruments are applied to measure fractions clay, silt and sand. The application of these instruments in the determination of M50 needs to be developed. Possibly the combination of specific surface area and fractions of fine and course sand can give an indication. On the other hand new pedo-transfer functions can be developed neglecting the M50 value while increasing the contribution of other explanatory variables.

Keywords: Proximal sensing, remote sensing, soil hydraulic properties, Mualem-van Genuchten equations, pedo-transfer functions

ISSN 1566-7197

The pdf file is free of charge and can be downloaded at https://doi.org/10.18174/429204 or via the website www.wur.nl/environmental-research (scroll down to Publications - Wageningen Environmental Research reports), Wageningen Environmental Research does not deliver printed versions of the Wageningen Environmental Research reports.

P.O. Box 47; 6700 AA Wageningen; The Netherlands, info.alterra@wur.nl

Alterra-report 2853
Wageningen, October 2017

Contents

Preface

The Key Registry Subsurface of the Netherlands (BasisRegistratie Ondergrond, BRO) provides on-line information of soil and subsoil in the Netherlands. The soil physics dataset discussed in this report is part of the BRO. The research reported here aims to select proximal sensor techniques that can be applied to fill lacunae in the soil physics dataset efficiently and accurately, to make the soil physics information up-to-date and to aggregate soil physics information at a point scale to larger spatial units. The research started in 2010 as a part of the programme 'BIS-Nederland' with an inventory of potentially useful techniques. In 2016 this research was finished with the design of a validation experiment and the selection of a study area to test the selected techniques under field conditions in the Netherlands.

We are grateful to the leader of the BRO-programme, dr J.P. Okx, for his support and advice during this research.

Wageningen, July 2017, the authors

Summary

Data on soil hydraulic properties are needed as input for many models, such as models to predict water flow in the unsaturated zone and related crop growth, and models to predict leaching of nutrients and pesticides to groundwater. The unsaturated hydraulic conductivity k and the volumetric water content θ, both as a function of the pressure head h, have been derived in the laboratory from 832 samples to characterize the average soil hydraulic properties of 36 soil horizons in the Netherlands. These soil hydraulic characteristics have been summarised to 21 soil hydraulic units derived from the soil map of the Netherlands, 1 : 250,000 and recently to 72 soil hydraulic units derived from the soil map of the Netherlands, 1 : 50,000. The soil physics database 'Priapus' with A-status is developing since 2012, but still shows several gaps: aggregation units for which no soil physical data are available yet. Furthermore, a substantial part of the data were collected more than thirty years ago and thus might not represent actual soil hydraulic conditions.

The relationship between the unsaturated hydraulic conductivity k or the volumetric water content θ and the pressure head h can be described for example by equations such as the Mualem-Van Genuchten equations. The parameter values of these equations can be predicted from soil properties by using pedo-transfer functions. These soil properties include clay content, loam content, organic matter content and the median grain diameter of the sand fraction (50-2000 μm).

There is a need to fill the lacunae in the soil physics dataset, to make the dataset up-to-date and to aggregate soil hydraulic properties observed at point scale to larger spatial units. Application of proximal sensing techniques in observing soil physics properties in the field might reduce the costs of data collection. The accuracy with which proximal sensing techniques can predict clay content, loam content, organic matter content and the median grain size of the sand fraction (50-2000 μm) needs to be assessed as these are the explanatory variables in the pedo-transfer functions.

The aim of this study is to select on-the-go proximal sensing techniques that are able to quantify soil variables that are used in the pedo-transfer functions for prediction of the parameters of the Mualem-Van Genuchten equations. Testing these techniques in the field is not part of this literature search.

This study focuses on the selection of proximal sensing techniques that can be applied in mapping explanatory variables of pedo-transfer functions to predict Mualem-Van Genuchten parameters from soil variables: sand, silt and clay content, median grain size of sand fraction (M50), bulk density, and organic matter content. For the sake of completeness applications of these devices in spatial prediction of other variables are reported as well.

Because the proximal sensing devices will be used to collect exhaustive numbers of observations within a short time in the field they must be field portable, water and shock resistant and easy to operate. Furthermore, for application in soil survey the device must provide the soil surveyor with an instantaneous estimate of the soil variable of interest. Besides these operational constraints the proximal sensing methods must answer to a minimum accuracy with which soil variables can be predicted and to a maximum of costs.

Validation results were found in literature on applications of near and mid infrared spectrometry in observing soil properties for the prediction of soil variables that are of interest in spatial prediction of soil hydraulic properties, both under laboratory and field conditions. A problem in predicting soil properties using VIS-NIR under field conditions is that spectral libraries are based on air-dried soil samples, whereas in the field soil moisture content varies in both space and time. External parameter orthogonalization (EPO), partial least squares regression (PLSR), Artificial Neural Network (ANN), Random Forest (RF) and Support Vector Machine (SVM) are methods to account for moisture in predicting soil organic matter content using libraries for air-dried samples. In four out of seven studies found in literature spatial predictions using field measurements of γ-ray spectrometry were validated. Reasonably accurate results were found for the prediction of clay content in Oostelijk Flevoland, the Netherlands. The performance of electromagnetic induction methods in predicting clay content in topsoils has been validated in a study area in northwest East-Flanders, Belgium. Several papers report studies on the application of ground penetrating radar in predicting soil variables such as soil moisture content. Although results look promising, results of validation experiments were lacking. MultiContinous Electrical Profiling (MuCEP) in predicting soil moisture content might be interesting. Results of validation experiments are not presented, however.

Some techniques can be selected which have a potential in predicting soil hydraulic properties. Explanatory variables used in pedo-transfer functions to predict Mualem-Van Genuchten parameters can be observed by several techniques. Validation results were found for the following variables and techniques:

- sand, silt and clay content: near infrared spectrometry, γ-ray spectroscopy, electromagnetic induction

methods

- organic matter content: near infrared spectrometry
- water content: near infrared spectrometry

Validation results of predictions of median grain size of sand fraction (M50) and bulk density by proximal sensing techniques were not found.

Sensors produce signals that potentially contain information on the soil properties of interest. However, the signals merely reflect the interaction of electromagnetic radiation or physical forces with soil constituents. A statistical model is usually needed to infer information on the soil properties of interest from these signals. Several methods are relevant for the soil properties mentioned above, varying from simple univariate methods, via multivariate methods, to more complicated hypervariate methods: linear regression, principal component regression, partial least squares regression, and machine learning algorithms such as bagging, random forest, artificial neural networks, and deep learning. Although linear regression is relatively simple and implemented in free software, it potentially suffers from collinearity and does generally not fully utilize all available information. Principal component regression solves the collinearity problem, potentially extracts more information out of the available data and is implemented in free software. After decomposition into scores and loadings (principal components), regression analysis is straightforward. However, principal components are extracted without reference to the variable of interest. Hence, there is not necessarily a relation between the first set of principal components and the variable of interest. Furthermore, interpretation is more complicated. It has to be done in terms of eigenvectors (loadings) and eigenvalues (scores) instead of the original data. Partial least squares regression is similar to principal components regression but has the advantage that the correlation between the (transformed) explanatory variables and the response is maximized. Machine learning algorithms can handle both linear and nonlinear relations, and interactions between explanatory variables are easily taken into account. Since the methods are usually non-parametric, no model assumptions are needed. Furthermore, machine learning algorithms are implemented in free software. However, machine learning algorithms are more black-box than linear regression or ordination based methods like principal component regression or partial least squares regression. Besides this, some models are computationally intensive, and measures of uncertainty (prediction intervals, realizations, etc.) are not always implemented in software.

The following conclusions can be drawn from the inventory of validated proximal sensing techniques:

1. Near infrared spectrometry, γ-ray spectroscopy and electromagnetic induction methods have a potential in spatial prediction of clay, silt and sand content;
2. Near infrared spectrometry has a potential in spatial prediction of organic matter content;
3. Near infrared spectrometry has a potential in spatial prediction of soil moisture content.

The use of current pedo-transfer functions requires also the parameters bulk density and M50. Bulk density can be measured by a device called RhoC, which was tested predicting bulk density of sediments in marine wetlands. The performance of the method in predicting bulk density of terrestrial soils has not been validated yet.

M50 is the median of the sand fraction. As most instruments have been applied in clayey areas and therefore were only calibrated to fractions clay, silt and sand, their application in the determination of M50 needs to be developed. Possibly the combination of specific surface area and fractions of fine and course sand can give an indication. On the other hand new pedo-transfer functions can be developed neglecting the M50 value while increasing the contribution of other explanatory variables.

1 Introduction

1.1 Background

Soil physics conditions, such as moisture content, air content, temperature and structure, highly determine the interactions between biochemical cycles, nutrient cycles, biodegradation of organic toxic substances and emission or adsorption of greenhouse gasses from or into the soil. Therefore, the functions and design of our environment largerly depends on the soil. Because of intensified land use a good and basic understanding of the soil and its interactions with the environment is needed to guarantee healthy food and a safe and pleasant environment.

Data on soil hydraulic properties are needed as input for many models, such as models to predict crop growth, and models to predict leaching of nutrients and pesticides to groundwater. The unsaturated hydraulic conductivity k and the volumetric water content θ, both as a function of the pressure head h, have been derived in the laboratory from 832 samples to characterize the average soil hydraulic properties of 36 soil horizons in the Netherlands (Wösten et al., 1987, 2001). These so-called soil hydraulic characteristics have been summarised to 21 soil hydraulic units derived from the soil map of the Netherlands, 1 : 250,000 (Wösten et al., 1988, PAWN) and recently to 72 soil hydraulic units derived from the soil map of the Netherlands, 1 : 50,000 (Wösten et al., 2013, BOFEK). Knotters et al. (2011) showed that if a classification based on topsoil and subsoil, organic matter content, texture and geology is rigorously applied the theoretical number of aggregation units is 2364. Knotters et al. (2011) indicated several gaps in the soil physics database 'Priapus' with A-status (Stolte et al., 2007), i.e., aggregation units for which no soil physics data are available yet. Furthermore, a substantial part of the data were collected more than thirty years ago and thus might not represent actual soil hydraulic conditions. Recently Bakker et al. (2015); **?** filled the major lacunae in the Dutch Soil Database (BIS Nederland) by adding 100 new soil hydraulic characteristics of high quality. Furthermore they upgraded the descriptive information of 91 existing soil hydraulic characteristics, resulting in 191 soil hydraulic characteristics of high quality.

The relationship between the unsaturated hydraulic conductivity k or the volumetric water content θ and the pressure head h can be described for example by equations derived by Van Genuchten (1980), the so called Mualem-Van Genuchten equations, or generalized expressions as recently described by Heinen and Bakker (2016). The parameter values of these equations can be predicted from soil properties by using pedo-transfer functions (Bouma, 1989; Vereecken et al., 2010). These soil properties include clay content, loam content, organic matter content and the median grain diameter of the sand fraction (50-2000 μm).

1.2 Problem definition

There is a need to fill lacunae in the soil physical dataset, to make the dataset up-to-date and to aggregate soil hydraulic properties observed at point scale to larger spatial units. However, soil physical measurements are time consuming and thus costly. Application of proximal sensing techniques in observing soil physical properties in the field might reduce the costs of data collection. Therefore the applicability of these techniques needs to be explored. In particular the accuracy with which proximal sensing techniques can predict clay content, loam content, organic matter content and the median grain size of the sand fraction (50-2000 μm) needs to be assessed as these are the explanatory variables in the current pedo-transfer functions.

1.3 Aim

The aim of this study is to select on-the-go proximal sensing techniques that are able to quantify soil variables that are used in the pedo-transfer functions for prediction of the parameters of the Mualem-Van Genuchten equations. Testing these techniques in the field is not part of this literature search.

This overview only focuses on scientifically published studies that included a sound validation of results. It therefore excludes good results in calibration studies or in studies that were not published in a scientific journal but in whitepapers or project reports.

1.4 Outline

Chapter 2 describes the selection of soil variables that are relevant in spatial prediction of soil physical properties. These include soil variables used in the pedo-transfer functions for prediction of the parameters of the Mualem-Van Genuchten equations (Wösten et al., 2001). Chapter 3 presents operational constraints, the minimum accuracy and the maximum of costs to which the methods must answer. Chapter 4 summarises the measurement devices that can be used to observe the variables mentioned in Chapter 2 and discusses their potential for use in spatial prediction of soil physical properties. Chapter 5 reviews methods to estimate values of soil variables from sensor data. Chapter 6 summarizes the conclusions and recommendations following from this study. Appendix A describes the set up of a validation experiment to evaluate the accuracy of soil properties predicted by using sensor technologies. This experiment will take place in 2017.

2 Prioritization of soil variables

This study focuses on the selection of proximal sensing techniques that can be applied in mapping soil physics properties. Vereecken et al. (2010), Wösten et al. (1999) and Wösten et al. (2001) use the following explanatory variables in pedo-transfer functions to predict Mualem-Van Genuchten parameters from soil variables:

- sand, silt and clay content
- median grain size of sand fraction (M50)
- bulk density
- organic matter content

Proximal sensing methods that can be used in spatial prediction of the soil variables mentioned above are reported in this study. For sake of completeness applications of these devices in spatial prediction of other variables are reported as well, in particular soil moisture content which is a key variable in describing the agricultural and ecological potential of the soil.

3 Constraints

This inventory focuses on proximal sensing devices to collect exhaustive numbers of observations within short time in the field. Therefore, the devices must be field portable, water and shock resistant and easy to operate. Furthermore, for application in soil survey the device must provide the soil surveyor with an instantaneous estimate of the soil variable of interest. Apart from these operational constraints the proximal sensing methods must answer to a minimum accuracy with which soil variables can be predicted and to a maximum of costs. In summary, the constraints are:

1. **Applicability in the field**
    - Field portable
    - Water and shock resistant
    - Simple to operate if applied during soil survey
    - Instantaneous estimate of target variable (important in soil survey)
2. **Accuracy**
    - Magnitude of measurement error must be known
3. **Costs**
    - Good balance between costs and gain of accuracy

4 Measurement devices

4.1 Overview of literature

4.1.1 General reviews

Adamchuk et al. (2004) give a general review of proximal sensing devices that can be applied to observe soil properties in the field. Table 4.1 summarises the measurement devices discussed in this review. Note that γ-ray spectrometry is not included in that review, and validation results were not presented.

Table 4.1
General review on proximal sensing techniques for observing soil properties in the field, by Adamchuk et al. (2004)

Instrument	Target variable
Electrical conductivity/resistivity contact sensors	texture, OM content, moisture, salinity
Electrical conductivity proximity sensors (electromagnetic induction method))	?
Electrical conductivity and capacitance contact sensor	moisture, salinity
Capacitance contact sensor	volumetric moisture content
Single wavelength subsurface soil reflectance sensor	OM content
Hyperspectral visual and near infrared subsurface soil reflectance sensor	OM content, texture, moisture, CEC, pH, nutrients
Hyperspectral visual and mid-infrared subsurface soil reflectance sensor	mineral nitrogen
Microwave sensor	moisture
Ground penetrating radar	volumetric soil moisture
Microphone equipped soil shank	clay content
Ion-selective field effect transistors (ISFETs) with flow injection analysis	nitrate concentration in soil extracts
Direct measurement of ion activity using ion-selective electrodes	soluble K, residual nitrate content, pH

Hartemink and Minasny (2014) present an overview of digital morphometrics of attributes measured in a soil profile. Table 4.2 summarises this overview. Note that γ-ray spectrometry was also not considered in that overview. In the following subsections we summarise applications of techniques mentioned by Adamchuk et al. (2004) and Hartemink and Minasny (2014) with a focus on validation results.

Table 4.2
Overview of digital soil morphometrics of attributes measured in a soil profile, from Hartemink and Minasny (2014)

Attribute	Digital morphometrics
Horizon depth, and boundaries	Electrical resistivity; radio-MT; GPR; profile cone penetrometer; XRF
Texture	XRF; laser diffraction; vis-NIR
Matrix color	Vis-NIR; GPR; mobile phones
Structure	Ultrasonics; X-ray CT, SEM
Moisture	TDR; GPR; electrical resistivity
Redoximorphic features; mottles	Hyperspectral scanner; XRF; digital cameras
Rupture resistance, consistence	X-ray CT and standardized drop-shatter
Carbonates	Vis-NIR
Rock fragments	Electrical resistivity; radiometers
Pores	X-ray CT; video digitizing; colored dyes; CAT scanning; image analysis
Roots	Image processing thin sections; GPR

Radio-MT = radio magnetotelluric-resistivity; GPR = Ground Penetrating Radar; XRF = X-ray fluorescence; vis-NIR = visible and near infrared; X-ray CT = X-ray computed tomography; SEM = Scanning Electron Microscope; CAT = computed axial tomography.

Several of the techniques mentioned in tables 4.1 and 4.2 use a specific part of the electromagnetic spectrum. To understand their relation the entire electromagnetic spectrum is depicted in Figure 4.1.

4.1.2 Near and mid infrared spectrometry

Near infrared spectrometry is based on the principle of light in the visible and near infrared part of the spectrum (300-2500 nm) being reflected and absorbed by properties of the material it interacts with, see Figure 4.2.

Reflectance depends on the material the light interacts with. In soil, this will be quartz, feldspars, organic components such as lignin and cellulose and other minerals and chemical/organic bonds. Each component has a unique spectral signature with varying reflectance depending on wavelength. When soil samples are measured on soil properties in the laboratory and measured spectrally with a spectrometer, these data are stored in a spectral library which can be used to train or calibrate a spectral model. This model can be

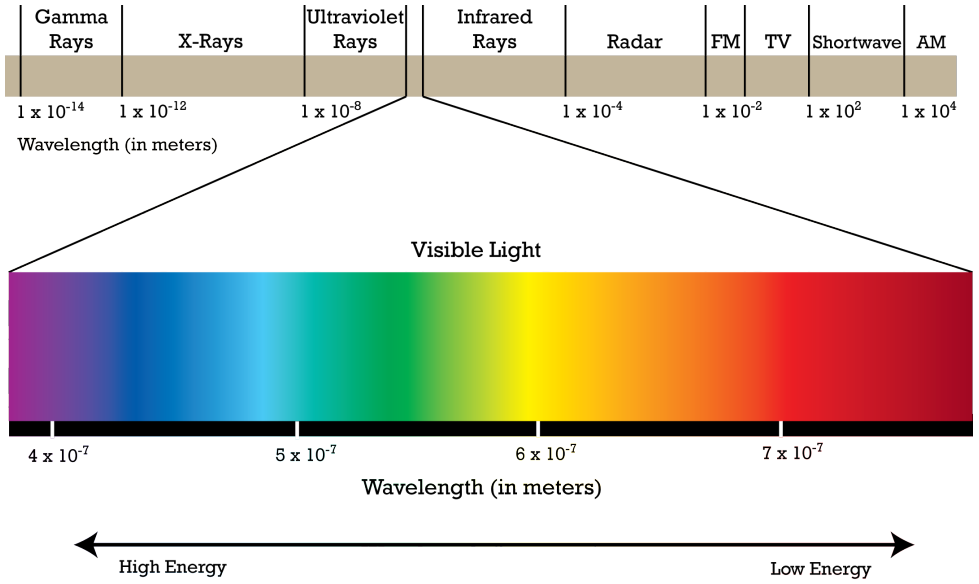


Figure 4.1
The electromagnetic spectrum

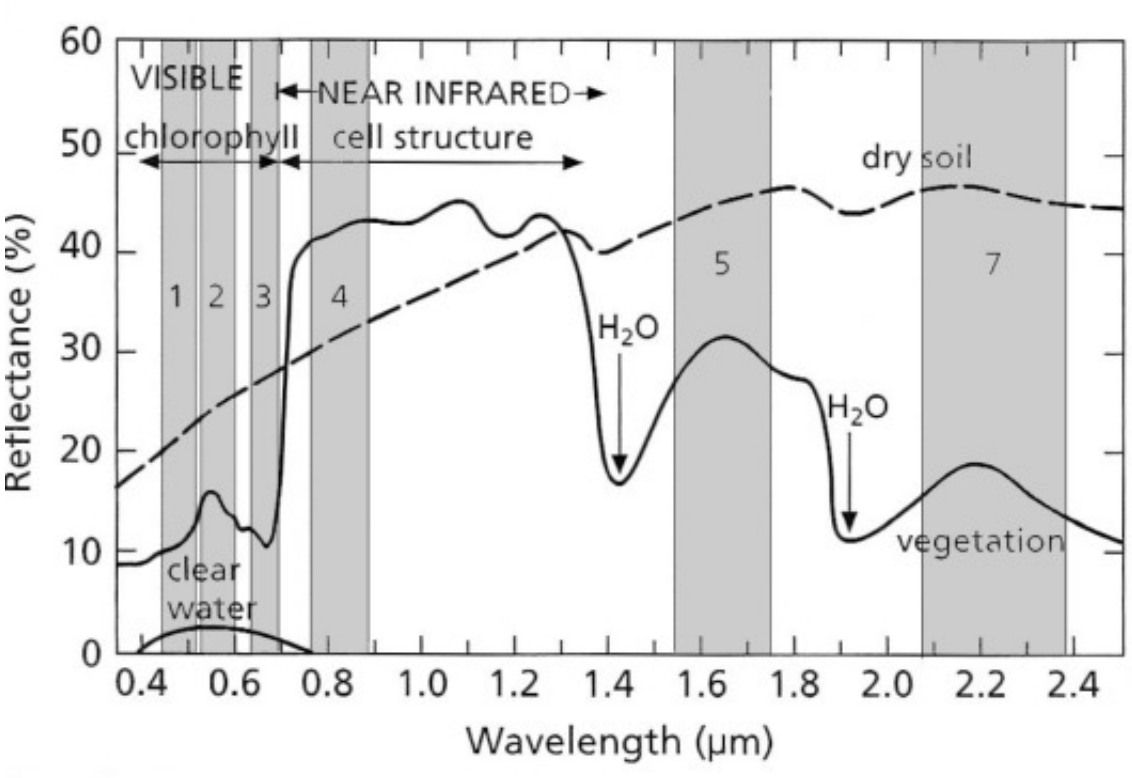


Figure 4.2
The near infrared spectrum

applied to other spectral measurements to estimate soil properties without laboratory measurements. Since water reflections occupy a significant part of the spectrum, rendering it useless for estimation of other properties, this technique works best on air-dried samples. For soil properties with signatures outside the moisture bands, the technique can be used in the field and from satellite imagery to estimate soil properties when bare soil is not obscured by vegetation or clouds. Since wavelengths in this part of the spectrum are relatively short, the penetration of the signal in soil is limited to a few centimeters maximum. Figure 4.3 shows an in situ spectral measurement of soil.

Figure 4.3
Spectral measurement of soil by Hafiz Sultan Mahmood

Table 4.3 summarises literature on applications of near and mid infrared spectrometry in observing soil properties. These applications concern both observations in the field and in the laboratory. In four out of thirteen studies the accuracy of field observations was quantified by validation. Shonk et al. (1991), Ben-Dor et al. (2008), Lagacherie et al. (2008) and Viscarra Rossel et al. (2009) present validation results for the prediction of soil variables that are of interest in spatial prediction of soil hydraulic properties.

Shonk et al. (1991) performed a validation experiment at field scale by comparing predicted OM content with OM content observed in the laboratory from soil samples. The soil samples were taken from agricultural land in Indiana (USA), with OM contents varying from 1 to 6%. The sensors were connected on a tractor tool bar in a housing that excluded ambient light. Reflectances of the soil were measured at some centimeters depth below the ground surface. A fraction of variance accounted for (R^2) of 0.83 was found. Further it was concluded that accuracy decreases with increasing OM content. OM contents larger than 6 % would be hard to differentiate.

Viscarra Rossel et al. (2006) present a review of applications of visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy in predicting various soil properties, as well as cross-validation results. The review includes indications of accuracy by RMSE and R^2. The cross-validation results concern observations under laboratory conditions.

Ben-Dor et al. (2008) developed a device to measure reflectances at the inner soil wall of a borehole. Four soil profiles in Israel were selected and sampled(according to the USDA nomenclature these soils were classified as Rhodoxeralf, Haploxeralf, Haploaquept, and Chromoxerert). For OM content the validation results indicated a $R^2 = 0.92$ and a $RMSE = 7.6\%$, based on 28 samples. For soil moisture content validation results indicated a $R^2 = 0.9$ and a $RMSE = 5.3\%$, based on 35 samples. The validation procedure is not entirely clear, however.

Lagacherie et al. (2008) performed a cross-validation of predictions of clay content (g kg^{-1}) from measurements with a portable spectro-radiometer. The study area was situated in the La Peyne Valley area, southern France. The study area has vineyards as the dominant land cover and a complex pattern of soil types. The parent material is formed by heterogeneous marl, limestone and calcareous sandstones coming from Miocene marine and lacustrine sediments. Soil types observed in this area include Lithic Leptosols, Calcaric Regosols and Calcaric Cambisols (FAO). These sediments can partly covered by successive alluvial deposits ranging from the Pliocene to Holocene and differing in their initial nature and in the duration of weathering conditions, resulting in the occurence of a large range of soil types such as Calcaric, Chromic and Eutric Cambisols, Chromic and eutric Luvisol and eutric Fluvisols. Recent volcanic activity and local transport of colluvial material along slopes added to the complexity of the soil pattern.The cross-validation resulted in a $R^2 = 0.58$.

Viscarra Rossel et al. (2009) predicted clay content from field measurements with a portable spectrophotometer in New South Wales, Australia. The Australian Soil Classification scheme identified the ten observed soil profiles as a Red Ferrosol derived from deeply weathered shale, an Aeric Podosol formed in marine sand, an Oxyaquic Hydrosol formed in fine-grained marine sediment, a Brown Dermosol formed in granitic colluvium, a Red Chromosol formed in colluvium of mixed sources, a Black Vertosol derived from basaltic colluvium, a Grey Vertosol derived from predominantly basaltic colluvium, a Yellow Sodosol formed

in sandstone colluvium, a Yellow Kandosol formed in sandstone alluvium, and a Lithocalcic Calcarosol formed in marl.The accuracy of the predictions was assessed by an independent validation set ($n = 39$). A percentage of variance accounted for of $R^2_{\text{adj}} = 0.78$ was found, and a $RMSE = 7.9\%$.

A problem in predicting soil properties using VIS-NIR under field conditions is that spectral libraries are based on air-dried soil samples, whereas in the field soil moisture content varies in both space and time. Minasny et al. (2011) applied external parameter orthogonalization (EPO) and partial least squares regression (PLSR) to account for moisture in predicting soil organic matter content using libraries for air-dried samples. A validation experiment with soil samples from agricultural areas in southern New South Wales, Australia, indicated increased accuracy by applying EPO: without EPO R^2 varied from 0.558 to 0.768, with EPO from 0.818 to 0.887.

Ge et al. (2014) assessed the accuracy of predictions of organic carbon content and clay content from VIS-NIR measurements using EPO and PLSR, in a validation experiment using moist and intact soil samples from Central Texas. For clay content a RMSE of 9 dag kg^{-1} was found, for organic carbon content a RMSE of 0.73 dag kg^{-1} was found.

Ackerson et al. (2015) predicted clay content of old and highly weathered soils classified as Ferrasols, Nitisols, Lixisols, and Arenosols with mineralogy consisting of a mixture of kaolinite and ironaluminum oxides under field conditions, EPO and PLSR to account for moisture in predicting clay content using libraries for air-dried samples. By using EPO the RMSE in predicted clay content could be reduced from 43 dag kg^{-1} to 12.3 dag kg^{-1}.

Wijewardane et al. (2016) predicted organic matter content and clay and sand content of soil samples from Nebraska using Vis-NIR and EPO combined with PLSR, Artificial Neural Network (ANN), Random Forest (RF) and Support Vector Machine (SVM). The accuracy was assessed by validation using dried soil samples that were rewetted in the laboratory. The combination of EPO with ANN and SVM outperformed the other two slightly. Significant improvements by using EPO were found for organic carbon content. Marginal improvements were observed for clay and sand content.

Dhawale et al. (2015) applied a portable mid-infrared spectrometer and PLSR to predict clay content, sand content and soil organic matter content in a validation experiment on agricultural fields in Quebec. Clay and sand contents of either wet or dry soil samples were predicted with a RMSE of around 10%, soil organic matter content was predicted with RMSE values between 0.76 and 2.24%.

From these results we can conclude that near infrared spectroscopy (interpreted here are visible to SWIR (short wave infrared) 300-2500 nm) is capable of measuring organic matter and clay content, sometimes also sand content in the top 1-3 cm of measured soil.

At present all presented applications above do not allow for instantaneous soil property estimation in the field. Both the estimation of soil properties and the soil moisture corrections require significant data processing after measurement and either the availability of a relevant spectral library or a large number of calibration samples. Fortunately, in the scientific and policy fields there is a trend towards more open data including spectral libraries such as the EU LUCAS database(JRC), the Global Soil Spectral Library (Viscarra Rossel et al., 2016), the ICRAF-ISRIC database etc. At present commercial applications of VISNIR spectroscopy tend to build a private database, thereby assuring consistent measurement quality and methods. This does increase cost however and a further increase in open soil spectral libraries of assured quality and reported metadata should therefore be encouraged to stimulate the uptake of this promising technique for practical applications in the field and using satellite imagery.

4.1.3 γ-ray spectrometry

Gamma-ray spectrometry or radiometry is based on the passive measurement of naturally occurring radioactivity with a scintillation crystal. The low levels of radioactivity are emitted by nuclides present in minerals of the soil and have been formed at the formation of Earth. Potassium (K40), Thorium (Th232) and Uranium (U238) are most commonly used in soil assessment, in relevant regions supplied by Caesium (Cs137) which is a man-made nuclide. This is present in the natural environment in Europe as a result of above-ground nuclear tests in the 1960s and the Chernobyl accident in 1986. The composition of minerals in a soil is dependent on geological provenance (and therefore parent material) and soil texture, with a larger differentiation and abundance of nuclide-rich minerals in the clay fraction compared to sand. Also, the behaviour of nuclides in soil varies per nuclide with some, like uranium, being more soluble than others such as thorium. For these two reasons, measuring the nuclide composition of a soil can be an indicator for parent material and soil (textural) properties.

The nuclide composition of a soil is derived by decomposing the measured total spectrum to nuclide concentrations using either Windows analysis or Full Spectrum Analysis (Hendriks et al., 2001). In Windows

analysis the window with the main energy peak of a nuclide is used as a measure for the amount of radiation of that element at that location while correcting for the contribution of other elements using stripping factors. In full spectrum analysis the spectral signatures or standard spectra (pure response of 1 Bq/kg of that nuclide) of the nuclides are fitted to the measured spectrum using a Chi-squared algorithm. This yields the concentration of the nuclides in the measured spectrum. This is 1.7 times more accurate than Windows analysis (Hendriks, 2001). The nuclide concentrations are then correlated to soil properties mainly by linear regression, although authors have also applied PLSR, PCA etc. Because it is a passive measurement and because radiation is absorbed by matter, the contribution of soil matter to the measured signal decreases exponentially with depth. In general, 90 % of the measured signal comes from the top 30 cm of soil around the sensor, whether used in on-the-go or in borehole configuration. Figure 4.4 shows a field measurement with a γ-ray spectrometer fixed to the bumper of a vehicle.

Figure 4.4
Measurement with gammaspectrometer (front) on agricultural soils by Medusa. GPR is towed at the back.

Table 4.4 summarises literature on applications of γ-ray spectrometry. In four out of seven studies spatial predictions using field measurements of γ-ray spectrometry were validated.

Pracilio et al. (2006) applied γ-ray spectrometry in a study area in western Australia with granite and migmatite terrains with shallow depths to bedrock and rocky fragments in some fields/paddocks, sedimentary terrains (sandstone and shale regional bedrock geology), and plains to long slopes with soils of Red Brown Loam and Yellow Deep Siliceous Sand over ferruginous gravel. Pracilio et al. (2006) used the following multiple linear regression model in predicting clay content:

$$\begin{aligned} \mathrm{Log}_{10}\left(c_{\mathrm{clay,dag/kg}}\right) \quad = \quad & 0.114 \times c_{\mathrm{U}} \quad +0.0634 \times \frac{c_{\mathrm{Th}}}{c_{\mathrm{U}}} \\ & +9.09 \times c_{\mathrm{K}} \quad -4.43 \times 10^{-4} \times \frac{c_{\mathrm{K}}}{c_{\mathrm{Th}}} \\ & +0.181 \end{aligned} \tag{4.1}$$

where $c_{\mathrm{clay,dag/kg}}$ is clay content in dag kg^{-1}, c_{U}, c_{Th} and c_{K} are ground concentrations of U, Th and K, respectively, in mg kg^{-1}, and with $R^2_{\mathrm{adj.}} = 0.75$. The ground concentrations were calculated from counts per second using sensitivity coefficients of 16.5, 10.2 and 0.01929 mg kg^{-1} per count per second for U, Th and K, respectively, using Windows analysis, where the window with the main energy peak is used as a measure for the amount of radiation of that element at that location while correcting for the contribution of other elements using stripping factors. The performance of γ-ray spectrometry in spatial prediction of clay content was validated with an independent validation sample ($n = 91$). A percentage of explained variance of 66% was found, and a RMSE of 2.4 dag kg^{-1}.

Viscarra Rossel et al. (2007) used bagging partial least squares regression (bagging-PLSR) to predict contents of clay, silt, fine sand and coarse sand from γ-ray data. The accuracy of these texture predictions was assessed by Viscarra Rossel et al. (2007) in a cross-validation experiment in two study areas in New South

Wales, Australia. One site has topography of extensive foot slopes (> 2 km) of undulating low hills and hills derived from alluvial fan systems, predominantly consisting of deep deposits of parent materials eroded from quartzose and lithic sandstones, silty sandstones and mudstones. The second site forms part of a formation formed during the late Pleistocene from the alluvial deposition of parent material sourced from quartz rich sandstone. The fields consist mostly of backplain facies formed by the sedimentation of silts and clays during overbank flow, resulting in plains with very low slopes and minimal relief. Table 4.5 summarises the results. The performance in predicting clay content and content of coarse sand (200-2000μm) was relatively good, the performance in predicting content of silt (2-20 μm) and fine sand (20-200 μm) was relatively poor. The RMSE in predicted clay content could be reduced from about 11 dag kg^{-1} when an areal mean was used as predictor to about 6 dag kg^{-1} when linear regression models and γ-ray spectrometry were used in spatial prediction.

Van der Klooster et al. (2011) predicted clay content in the Dutch marine clay area from γ-ray data using linear regression models with the following general structures:

$$\begin{array}{rcl} c_{\mathrm{clay},\%} & = & \beta_0 + \beta_{\mathrm{K}} \times r_{\mathrm{K}} \\ c_{\mathrm{clay},\%} & = & \beta_0 + \beta_{\mathrm{Th}} \times r_{\mathrm{Th}} \\ c_{\mathrm{clay},\%} & = & \beta_0 + \beta_{\mathrm{K}} \times r_{\mathrm{K}} + \beta_{\mathrm{Th}} \times r_{\mathrm{Th}} \\ c_{\mathrm{clay},\%} & = & \beta_0 + \beta_{\mathrm{TC}} \times t_{\mathrm{TC}} \end{array} \tag{4.2}$$

where r_{K} and r_{Th} are the radiatons of ^{40}K and ^{232}Th, respectively, in Bq kg^{-1}, t_{TC} is the total energy count in s^{-1}, and β_0, β_{K}, β_{Th} and β_{TC} are regression parameters. The accuracy of predicted clay content was validated at field scale, regional scale and district scale by random data-splitting. At the field scale, the RMSE in predicted clay content could be reduced from 8-10% when an areal mean was used as predictor to 2-3% when linear regression models and γ-ray spectrometry were used in spatial prediction. These results are in the same range as those found by Viscarra Rossel et al. (2007).

Mahmood et al. (2013) evaluated the accuracy of clay, silt, sand and total organic matter content predicted from γ-ray data by using windows and full-spectrum analysis methods followed by linear regression. The study area of 4 ha was part of the experimental farm "Broekemahoeve" near Lelystad, in the Flevoland province of The Netherlands with marine clay soils. The study area included a field with convential agriculture and a field with organic agriculture. Tables 4.6 and 4.7 summarise the validation results. Only clay content is predicted reasonably accurate.

Coulouma et al. (2016) tested the performance of γ-ray spectrometry in prediction of clay content in a Mediterranean landscape in the Languedoc, France. The study area includes very shallow soils developed over micritic limestone without igneous minerals and with low weathering intensity Hyperskeletic calcaric Leptosol; deep soils on sediments from the slopes in the same environment Colluvic calcaric Regosol; soils on recent marine sediments Sodic Cambisol (colluvic); soils on marine sedimentary sandstones with high mica contents Calcisol (siltic); soils on tertiary sediments of different origins Calcisol (clayic); soils on old colluvial deposits Chromic Cambisol (clayic); and soils on old alluvial deposits with high proportions of igneous minerals and highly weathered igneous rock fragments Skeletic Fluvisol (clayic). The specific activity of ^{232}Th in bq kg^{-1} was used as explanatory variable in a simple linear regression model. Cross-validation results indicated a R^2 of 0.60 for predictions at point scale, a mean error of 0.1 dag kg^{-1} and a RMSE of 4.2 dag kg^{-1} (the mean clay content in the study area is estimated at 24.3 dag kg^{-1}).

Reviewing these results, we can conclude that clay content in 0-30 cm depth can be predicted using γ-ray spectrometry with reasonable accuracies in multiple studies. In one study also the coarse sand fraction could be predicted. Other properties have either not been evaluated or returned lesser results in the papers reviewed here.

Since calibration of spectra to textural properties is provenance dependent, calibrations tend to be regionally stable. Therefore, once determined for a region, the calibration can be applied on-the-go to derive soil property information in real-time.

4.1.4 Electromagnetic induction methods

Electromagnetic induction methods measure the electrical conductivity of the soil by measuring the secondary magnetic field that is induced in the soil as a result of the primary magnetic field created by two coils of the sensor. The strength of this secondary field is a measure of soil electrical conductivity and magnetic susceptibility. This is influenced by the presence of moisture, chargeable soil particles like clay and organic matter, salts and as a result of that is also influenced bulk density and pore size. Measurement depth is dependent on the distance between coils, multiple depths can be measured by an array of coils.

Figure 4.5
Electromagnetic measurement by ORBit, a research group of the University of Ghent, Belgium.

By inversion of the (multiple) signals a soil profile with depth can be measured on electrical conductivity. Figure 4.5 illustrates fieldwork with an electromagnetic measurement device.

Table 4.8 summarises applications of electromagnetic induction methods in observing soil properties. Triantafilis and Lesch (2005) used EM-methods to predict the average clay content between 0 and 7 m, which is too deep for the purpose of this study. Cockx et al. (2009) presented results which might be interesting, because they applied EM-methods and artificial neural networks to predict clay content in the top layer (0-40 cm) in a 10.5-ha arable field in the polder area of northwest East-Flanders, Belgium. The clay content in the topsoil of this field varies from 19 to 24%. A root mean squared error of 1.68% was found, a mean error of 0.459% and $R^2 = 0.655$. It should be noted, however, that the validation set used by Cockx et al. (2009) was not independent which might cause optimistic results. The application of EM-methods in the prediction of soil moisture content (Triantafilis and Monteiro Santos, 2009) might be interesting. Validation results are not presented, however.

Although results look promising, we have not found validation results for EMI techniques in measuring soil textural properties in this review. However, the technique is much used in precision agriculture to map soil spatial heterogeneity on agricultural fields.

4.1.5 Ground penetrating radar

Ground Penetrating Radar (in Dutch: grondradar) consists of a sending and a receiving antenna. The sending antenna emits radio signals between 100 MHz and 2 GHz depending on the GPR used. This signal travels through the soil with a speed that depends on the di-electric properties of the (layer of) soil at that location. This may change with depth and is moisture, texture and density dependent. The signal is reflected at (sudden) shifts in di-electric properties. This can be a sharp textural boundary (sand to clay or peat) as is found in fluvial soils or am object in the soil such as a stone, wall or cable. The reflected signal is measured by the receiving antenna. Depth penetration of the signal is dependent on di-electric properties of the soil and is reversely correlated with depth resolution and emitted frequency. So a low frequency such as 300 MHz has a penetration of 1-3 m in clayey soils, 4-6 m in sandy soils with penetrating being less in moist soils. Depth resolution will be about 5-10 cm. A high frequency of 2 GHz has a penetration of 20-40 cm in dry soils but has a depth resolution of 1-2 cm. This sensor is suitable for measuring changes in soil profiles with depth.

Table 4.9 summarises studies on the application of ground penetrating radar in predicting soil variables. The papers of Huisman et al. (2003) and Lambot et al. (2004) on prediction of soil moisture content might be interesting. Results of validation experiments were lacking, however. We can therefore not conclude

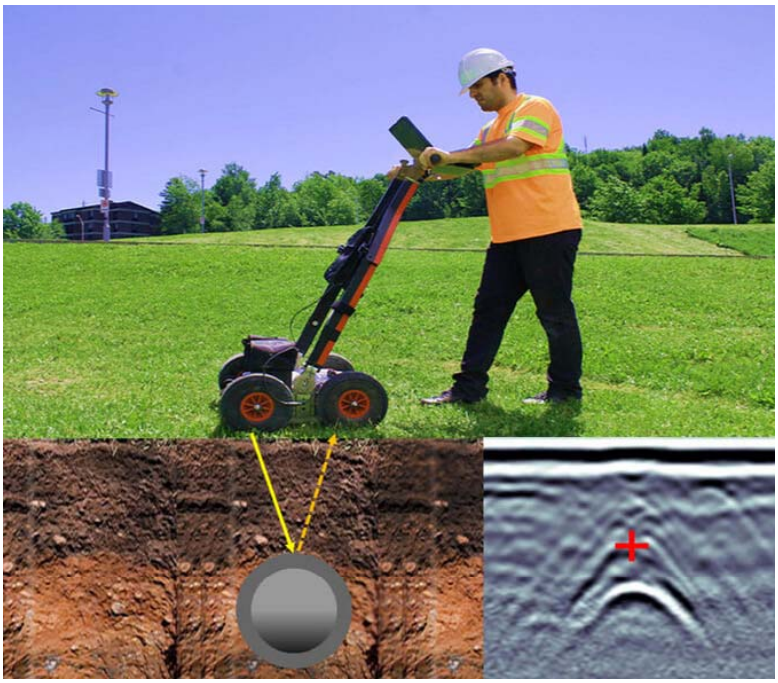


Figure 4.6
Measurement with GPR, in this case of an object.

that prediction of soil textural properties can be performed with GPR based on this review.

4.1.6 Other methods

Table 4.10 lists applications of methods not mentioned in the previous tables. The application of MultiContinous Electrical Profiling (MuCEP) in predicting soil moisture content might be interesting (Cousin et al., 2009; Besson et al., 2010). These studies did not include validation experiments, however.

4.2 Appropriate techniques for predicting soil hydraulic properties

From the overview in the previous section some techniques can be selected which have a potential in predicting soil hydraulic properties. Explanatory variables used in pedo-transfer functions to predict Mualem-Van Genuchten parameters can be observed by several techniques. Reasonable validation results were found for the following variables and techniques:

- sand, silt and clay content: near and mid infrared spectrometry (Lagacherie et al., 2008; Viscarra Rossel et al., 2009), γ-ray spectroscopy (Pracilio et al., 2006; Viscarra Rossel et al., 2007; Van der Klooster et al., 2011), electromagnetic induction methods (Triantafilis and Lesch, 2005; Cockx et al., 2009)
- organic matter content: near and mid infrared spectrometry (Shonk et al., 1991; Ben-Dor et al., 2008)
- water content: near and mid infrared spectrometry (Ben-Dor et al., 2008)

Validation results of predictions of median grain size of sand fraction (M50) and bulk density by proximal sensing techniques were not found.

This overview only focuses on scientifically published studies that included a sound validation of results. It therefore excludes good results in calibration studies or in studies that were not published in a scientific journal but in whitepapers or project reports.

Table 4.3

Literature on sensors for measuring soil properties: **Near Infraread Spectrometry**

Reference	Instrument	Target variable	Validation results (Y/N)
Shonk et al. (1991)	Spectroscopic sensor (luminous flux of 11.43 lumens on 68.52 mm^2)	OM content	Y (field)
Ben-Dor and Banin (1995)	NIR by an Alpha Centauri FTIR spectrophotometer (Mattson Co., Madison, WI), 1-2.5μm	clay content, specific surface area, CEC, hygroscopic moisture, OM content, $CaCO_3$ content	Y (laboratory)
Hummel et al. (2001)	NIR soil sensor (spectral reflectance, bandwidth 45 nm)	OM content, moisture content	Y (laboratory)
Walvoort and McBratney (2001a)	Varian Cary 500 scan spectrophotometer with a Labsphere DRA-CA-50D diffuse reflectance accessory, spectral range 250-2450 nm	clay content, C content, N content	Y (cross-validation, laboratory)
Barnes et al. (2003)	spectrometry	texture, various properties	N (review article)
Islam et al. (2003)	UV-VIS-NIR spectrophotometer (Cary 500) equipped with a diffuse reflectance accessory (Labsphere DRA-CA-50D) at 1.1-nm intervals in the UV-VIS (250–700 nm) range and 3-nm intervals in the NIR (700–2500 nm) range	pH, EC, organic carbon content, air-dry gravimetric water content, free iron, clay content, sand content, silt content, CEC, Ca, Mg, K, Na	Y (laboratory)
Brown et al. (2006)	ASD bFieldspec Pro FRQ VNIR spectroradiometer (Analytical Spectral Devices, Boulder, CO) with a spectral range of 350-2500 nm, 2 nm sampling resolution and spectral resolution of 3 nm at 700 nm and 10 nm at 1400 and 2100 nm	clay content, CEC, organic and inorganic carbon content	Y (cross-validation, laboratory)
Viscarra Rossel et al. (2006)	(i) an ultraviolet visible near infrared (UV-VIS-NIR) spectrometer (Varian Cary 500) equipped with a diffuse reflectance accessory (Labsphere DRA-CA-50D), with a spectral range of 350-2500 nm (28,570-4000 cm^{-1}); (ii) a BioRad FTS 175 rapid scanning Fourier transform (FT) mid infrared (MIR) spectrometer, with an extended range KBr beam splitter and Peltier-cooled DTGS detector with a spectral range of 1200-20,000 nm (8300-470 cm^{-1} and 16 cm^{-1} resolution)	pH_{Ca}, pH_{W}, organic carbon content, lime requirement, sand/silt/clay content, cation exchange capacity (CEC), Ca/Al/NO_3-N/P_{Col}/K content, electrical conductivity (EC)	Y (cross-validation, laboratory)
Ge et al. (2007)	VNIR diffuse reflectance spectroscopy (Cary 500 UV/VIS/NIR spectrophotometer, reflectance spectra from 250 to 2500 nm)	clay content, sand content, Ca, K, Mg, Na, P, and Zn	Y (laboratory)
Ben-Dor et al. (2008)	field spectrometer (analytical spectral device, ASD) (VNIR) (Boulder, CO), 350–2500 nm	moisture, OM content, soil carbonates, free iron oxides, specific surface area	Y (field, procedure not clear)
Lagacherie et al. (2008)	ASD pro FR Portable Spectroradiometer (350-2500 nm)	clay content, $CaCO_3$ content	Y (cross-validation, field)
Melendez-Pastor et al. (2008)	VIS-NIR spectrometry (ASD (Analytical Spectral Devices Inc., Boulder (CO), USA) Field Spec Hand Held VNIR radiometer), 325–1075 nm	silt, sand, E.C., carbonates, OM content	N
Mouazen et al. (2009)	Vis-NIR spectrophotometer (Zeiss Corona 45 visnir fibre), 306.5–1135.5 nm, and 944.5–1710.9 nm	plant available phosphorus	N
Viscarra Rossel et al. (2009)	Vis-NIR spectroscopy (Contact Probe attachment (Analytical Spectral Devices, Boulder Colorado), spectral range 350–2500 nm)	soil colour, mineral composition, clay content	Y (laboratory & field)
Minasny et al. (2011)	AgriSpec instrument with a contact probe (Analytical Spectral Devices, Boulder, Colorado, USA, spectral range 350-2500 nm)	soil organic matter content	Y (laboratory & field)
Ge et al. (2014)	AgriSpec spectroradiometer (Analytical Spectral Devices, Boulder, Colorado, USA)	organic carbon content, clay content	Y (laboratory & field)
Ackerson et al. (2015)	Vis-NIR spectroscopy (ASD FieldspecPro (Analytical Spectral Devices, Boulder CO) spectral range 350–2500 nm)	clay content	Y (laboratory & field)
Dhawale et al. (2015)	Mid-IR spectrometer (variable-filter-array (VFA) diffuse reflectance infrared Fourier transform (DRIFT) spectrometer from Wilks Enterprise, Inc. (East Norwalk, Connecticut, USA))	clay content, sand content, soil organic matter content	Y (laboratory & field)
Wijewardane et al. (2016)	Vis-NIR spectroscopy (ASD LabSpec spectrometer with a mug light (Analytical Spectral Devices, Boulder, Colorado, USA) spectral range 350–2500 nm)	organic, inorganic and total carbon content, clay and sand content	Y (laboratory)

Table 4.4

Literature on sensors for measuring soil properties: **Gamma-ray Spectrometry**

Reference	Instrument	Target variable	Validation results (Y/N)
Cattle et al. (2003)	γ-ray spectrometry (Exploranium GR-820 multi-channel gamma-ray spectrometer coupled to 2 GPX 1024 crystal detectors with a total NaI volume of 33.6 L)	aeolian dust content	N
McBratney et al. (2003)	AGSR (air gamma-ray spectrometer)	'soil properties'	N
Pracilio et al. (2006)	γ radiometry by Universal Tracking Systems Pty Ltd. (UTS) 0.4 to 2.82 MeV range (wavelengths from 8.50×10^{-4} to 4.74×10^{-4} nanometres, and a crystal size detector of 32 L)	clay content, bic-K	Y (field, independent sample, $n = 91$)
Viscarra Rossel et al. (2007)	GR320 portable γ-ray spectrometer (Exploranium$^{\mathrm{TM}}$ Radiation Detection Systems, Toronto, Canada), with a 4.2-litre thallium-activated sodium iodide detector crystal (1 Hz recording frequency)	0-15 cm and 15-50 cm: pH, ED, clay content, silt content, fine sand content, coarse sand content, 0-15 cm: K, Fe	Y (cross-validation, field)
Beckett (2008)	γ-ray spectrometry system	texture, soil unit	N
Buchanan and Triantafilis (2009)	γ-radiometry	water table depth	Y (cross-validation, field)
Van der Klooster et al. (2011)	γ-ray spectrometry	clay content	Y (field)
Coulouma et al. (2016)	γ-ray spectrometry	clay content	Y (cross-validation)

Table 4.5

Results of cross-validation of texture predictions using γ-ray spectroscopy (Viscarra Rossel et al., 2007). RMSE, ME and SDE in dag kg^{-1}

	Nowley					Stanleyville				
texture class	mean	RMSE	ME	SDE	$R^2_{\mathrm{adj.}}$	mean	RMSE	ME	SDE	$R^2_{\mathrm{adj.}}$
Soil 0-15 cm										
Clay	30.65	5.34	0.21	5.48	0.76	10.64	6.56	-1.32	6.59	0.63
Silt	6.22	2.46	-0.16	2.52	0.40	9.52	1.83	0.00	1.88	0.44
Fine sand	21.58	3.96	-0.04	4.06	0.05	16.78	2.28	0.29	2.32	0.15
Coarse sand	41.77	8.28	0.75	8.46	0.73	36.64	6.25	1.66	6.19	0.76
Soil 15-50 cm										
Clay	43.15	8.40	-1.23	8.53	0.54	40.27	6.75	-0.78	6.89	0.61
Silt	5.07	2.29	-0.46	2.30	0.40	9.68	2.90	-0.19	2.97	0.03
Fine sand	18.78	3.23	0.25	3.30	0.31	15.85	2.39	0.26	2.43	0.31
Coarse sand	33.84	10.33	1.99	10.40	0.37	34.78	5.30	0.74	5.39	0.79

Table 4.6

Descriptive statistics of the control data for validation of texture predictions and predictions of total organic carbon content (TOC) using γ-ray spectroscopy (Mahmood et al., 2013)

	Conventional field				Organic field			
Soil property	min	max	mean	SD	min	max	mean	SD
Soil 0-15 cm								
Clay (%)	16.0	22.0	18.9	1.55	17.0	22.4	19.7	1.38
Silt (%)	15.0	25.0	19.5	2.38	9.0	24.0	14.9	3.56
Sand (%)	53.0	68.0	61.5	3.48	56.4	73.2	65.3	4.44
TOC (dag kg^{-1})	1.13	1.27	1.2	0.037	0.94	1.28	1.14	0.092
Soil 15-30 cm								
Clay (%)	15.8	23.0	18.4	1.61	16.0	19.4	18.1	0.96
Silt (%)	16.0	28.0	21.6	2.61	15.0	19.7	17.9	1.06
Sand (%)	50.2	67.0	60.0	3.82	61.3	69.0	64.0	1.59
TOC (dag kg^{-1})	0.86	1.09	0.99	0.057	0.92	1.03	0.97	0.022

Table 4.7

Results of validation of texture predictions and predictions of total organic carbon content (TOC) using γ-ray spectroscopy (Mahmood et al., 2013). EWs: energy windows. FSA: full spectrum analysis.

Soil property	Conventional field				Organic field			
	FSA		EWs		FSA		EWs	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Soil 0-15 cm								
Clay (%)	0.65	0.96	0.59	1.06	0.73	0.81	0.67	0.82
Silt (%)	0.19	2.07	0.27	2.01	0.16	3.42	0.21	3.24
Sand (%)	0.31	2.90	0.38	2.83	0.35	3.67	0.40	3.36
TOC (dag kg^{-1})	0.45	0.027	0.47	0.027	0.17	0.078	0.34	0.069
Soil 15-30 cm								
Clay (%)	0.13	1.52	0.11	1.58	0.55	0.57	0.62	0.55
Silt (%)	0.30	2.33	0.42	2.13	0.22	0.91	0.18	0.93
Sand (%)	0.28	3.34	0.35	3.18	0.52	1.09	0.51	1.12
TOC (dag kg^{-1})	0.03	0.056	0.07	0.54	0.09	0.017	0.08	0.020

Table 4.8

Literature on sensors for measuring soil properties: **Electromagnetic induction methods**

Reference	Instrument	Target variable	Validation results (Y/N)
Sudduth et al. (2001)	EM38 (electromagnetic induction method)	topsoil depth	N
Corwin and Lesch (2005)	EM38	various properties	N (review article)
Triantafilis and Lesch (2005)	EM34, EM38	clay content	Y ($n = 8$, cross-validation $n = 40$
Brevik et al. (2006)	EM38	water content, soil units	N
Vitharana et al. (2006)	EM38DD	topsoil clay content	N
Cockx et al. (2007)	EM38DD	clay lenses	Y
Weller et al. (2007)	EM38	clay content	N
Vitharana et al. (2008)	EM38DD	layer depth	Y
Buchanan and Triantafilis (2009)	EM34, EM38	water table depth	Y (cross-validation)
Cockx et al. (2009)	EM38DD	clay content	Y (however, no independent data)
Kühn et al. (2009)	EM38	geological unit, OM content, $CaCO_3$, clay content	N
Saey et al. (2009b)	EM38DD	clay content	N
Saey et al. (2009a)	EM38DD, DUALEM-21S	depth to clay	Y
Triantafilis and Monteiro Santos (2009)	EM38/EM34	moisture content, soil variation	N
Triantafilis et al. (2009)	EM38, EM31	CEC	Y

Table 4.9

Literature on sensors for measuring soil properties: **Ground penetrating radar**

Reference	Instrument	Target variable	Validation results (Y/N)
Shih et al. (1986)	GPR (SIR-8, 300 MHz)	water table depth	Y
Smith et al. (1992)	GPR (SIR-8 impulse radar with a 120 Mhz antenna)	water table depth	Y
Birkhead et al. (1996)	GPR (SIR-10, 500 MHz)	water table depth	Y
Lapen et al. (1996)	GPR (PulseEKKO IV, two 100 MHz antennas)	water table depth	Y
Freeland et al. (2001)	GPR (SIR System-10A GPR mainframe (GSSI, Inc., North Salem, N.H.) with a 200 MHz antenna)	water table depth	N
Huisman et al. (2003)	Ground penetrating radar	soil moisture content	N
Lambot et al. (2004)	GPR (UWB-SFCW radar system)	water content	N
Doolittle et al. (2006)	ground penetrating radar (GPR) (Subsurface Interface Radar (SIR) System-2)	water table depth, local groundwater flow pattern	N
Freeland and Odhiambo (2007)	GPR (GSSI Subsurface Interface Radar (SIR) System 10-A and a 200 MHz antenna (model 3105, Geophysical Survey Systems, Inc., New Salem, N.H.))	presence/absence of preferential flow paths	N
Gerber et al. (2007)	GPR: GSSI (Geophysical Survey Systems, Inc.) SIR-2000 and a SIR-2 GPR system with five different antennae (100, 200, 400, 500 and 900 MHz)	Pleistocene periglacial slope deposits	N

Table 4.10

Literature on sensors for measuring soil properties: **Other methods**

Reference	Instrument	Target variable	Validation results (Y/N)
Levin et al. (2005)	digital camera	soil colour	N
Besson et al. (2010)	MuCEP (MultiContinous Electrical Profiling)	moisture content	N
Cousin et al. (2009)	MuCEP (MultiContinous Electrical Profiling)	moisture content	N
Chaplot et al. (2001)	radio magnetotelluric-resistivity (Radio-MT)	hydromorphic horizons	N
Kalnicky and Singhvi (2001)	Field portable XRF (various instruments)	heavy metals	N
Bernick et al. (1995)	Field portable XRF (Outokumpu Elctronics Inc. X-MET 880, Spactrace Instruments Spectrace 9000 FPXRF)	Pb	Y
Zhu and Weindorf (2009)	Field portable XRF (Innov-X Systems Alpha series FPXRF with tantalum X-ray tube	Calcium content	Y
Weindorf et al. (2009)	Field portable XRF (Innov-X Systems Alpha series FPXRF with tantalum X-ray tube	$CaSO_4$ content	N

5 From sensor signal to soil physical information: a brief overview of methods

5.1 Introduction

In the previous sections, an overview was given of soil properties that affect water retention and soil hydraulic conductivity. These soil properties are: soil texture (notably clay content, silt content, and the median grain size of sand), soil organic carbon content, and dry bulk density (Wösten, 1997). These soil data are needed for 'the Key Registry of Netherlands Geological Information' (Basisregistratie Ondergrond, BRO).

In addition, an overview was given of proximal sensors that are potentially interesting for measuring these soil properties. In particular visible and near-infrared and γ spectrometers stand out.

Sensors produce signals that potentially contain information on the soil properties of interest. Sensors, however, do not provide this information out-of-the-box. The signals merely reflects the interaction of electromagnetic radiation or physical forces with soil constituents. To infer information on the soil properties of interest from these signals, a statistical model is usually needed. That is illustrated in Figure 5.1.

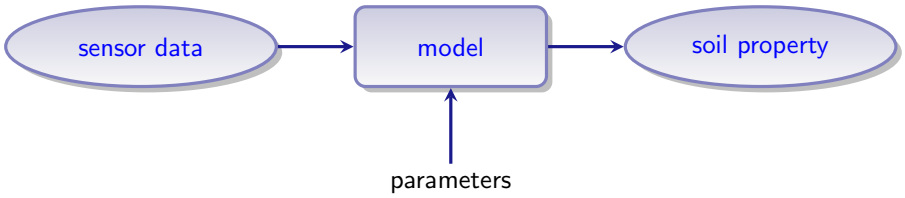


Figure 5.1
Notional example of how sensor signals are converted to soil properties.

As an illustrative example of such a model, consider the well known pH-meter. A pH-meter is basically a voltmeter that converts voltages to pH-units. For this conversion, the pH-meter contains a simple model that needs to be calibrated prior to use. Calibration runs down to measuring the pH of buffer solutions with known pH. After calibration, the pH-meter is ready for use. Figure 5.2 is an adaptation of Figure 5.1, specific for measuring pH in a soil suspension. See Appendix C for details.

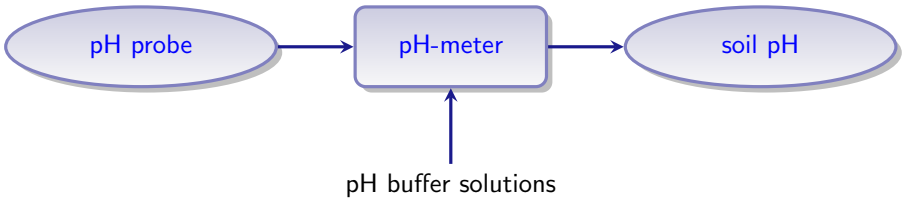


Figure 5.2
As Figure 5.2, but then specific for measuring soil pH.

In this chapter, several techniques are described that are useful for analyzing raw sensor data and converting these data to useful soil information by means of a statistical model. It is not our intention to give a comprehensive overview all available techniques. We will primarily focus on techniques that are relevant for the soil properties mentioned above and therefore for the 'BRO – Key Register of Netherlands Geological Information'. The methods vary from simple univariate methods, via multivariate methods, to more complicated hypervariate methods.

Each section starts with a brief overview of the method. This part should give the reader an intuitive and qualitative understanding of what the method is about. If needed, details are provided in the appendices. Appendix B gives an overview of all symbols used in this chapter and in the appendices.

Most sections also contain illustrative examples. Appendix D gives details on the data that we use to illustrate the methods. Each section is concluded with a summary of the advantages and disadvantages of the method.

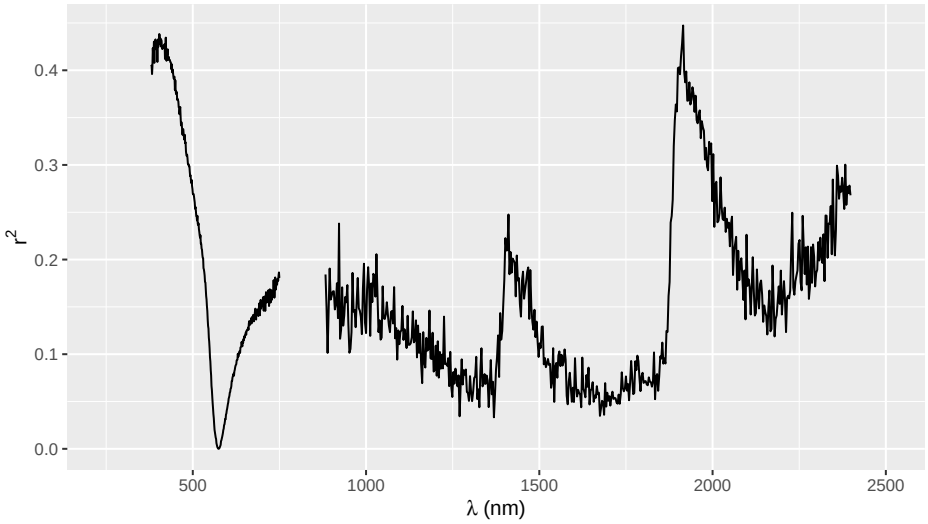


Figure 5.3
The square of Pearson's correlation coefficient for Equation 5.3 as function of wavelength.

5.2 Linear regression

5.2.1 Overview

Linear regression aims at modelling a response y as a linear combination of one (*i.e.*, simple linear regression) or more (*i.e.*, multiple linear regression) explanatory variables x_i:

$$y = b_0 x_0 + b_1 x_1 + \cdots + b_p x_p + \varepsilon \tag{5.1}$$

where b_i are parameters, and ε is the residual error.

Response y can be any soil property that is relevant for the 'BRO – Key Register of Netherlands Geological Information'.

As the name suggests, an explanatory variable can be any variable that explains part of the variation of the response variable. An explanatory variable can be a sensor signal, but also environmental variables like altitude, land use, or the availability of drainage networks, etc.

The residual error ε represents all variation not accounted for by the model (*e.g.*, measurement error, or errors due to invalid model assumptions). If $x_0 = 1$, the model includes an intercept, setting $x_0 = 0$ removes the intercept from the model.

Once the parameters b_i have been estimated, the model can be used to predict y:

$$\hat{y} = \hat{b}_0 x_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p \tag{5.2}$$

where $\hat{y}$ is the prediction, and $\hat{b}_i$ (for $i = 0 \ldots p$) are the estimated parameters. See Appendix E for more details on how to estimate the parameters.

5.2.2 Example on clay content

We will start by predicting soil clay content given the spectra in Appendix D by means of simple linear regression. Our model is:

$$y = b_0 + b_1 x_1 + \varepsilon \tag{5.3}$$

where y is clay content and x_1 is the diffuse reflectance at a single wavelength λ that results in the smallest sum of squared errors ε. This is equivalent to finding the wavelength that maximizes the squared Pearson's correlation coefficient r^2. Figure 5.3 gives r^2 as function of wavelength. Note that there are several peaks. We will select the peak at 404 nm.

Hence, the predictor is

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1$$

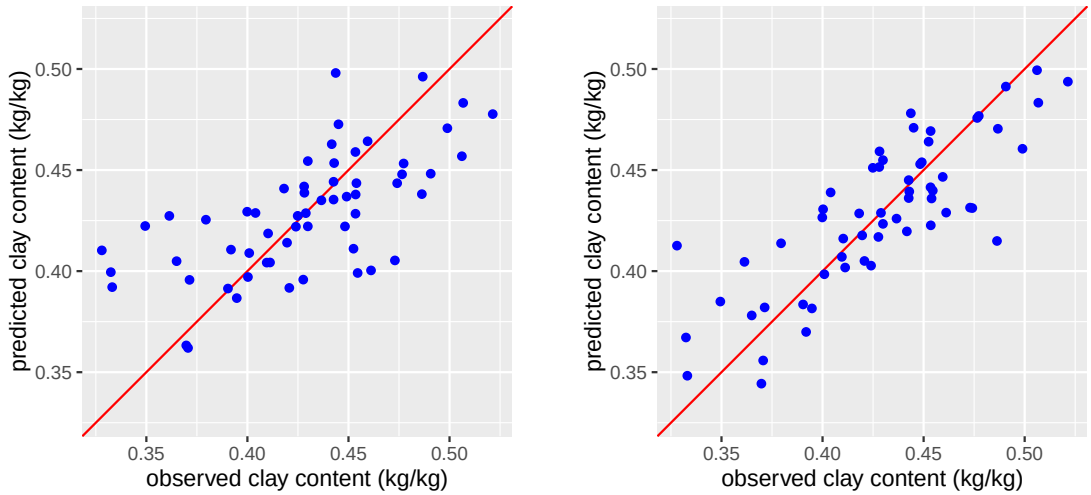


Figure 5.4
Predicted versus observed clay contents. Left based on a single wavelength, right based on two wavelengths. The line is the 1:1 line.

where x_1 is the diffuse reflectance at 404 nm. The square of the correlation coefficient r^2 equals 0.44. This means that the model explains 44% of the variation in y. A scatter plot of predicted versus observed clay content y is given in Fig. 5.4.

In the example above, we only used the reflectance at a single wavelength in our model. The other 842 wavelengths were ignored. Intuitively, this is a bit of a waste. Not all the available data have been used.

By using multiple linear regression, the reflectances at more wavelengths may be used. By extending our model to two wavelengths we get the following predictor:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2$$

If we select the two wavelengths that explain most of the variation in clay content y, we get a model where x_1 is the reflectance at 2185 nm and x_2 is the reflectance at 1912 nm. This model performs better than our previous model. It explains 67% of the variation in the data ($r^2 = 0.67$).

We might as well extend our model to even more explanatory variables x_i. However, this is not always possible due to collinearity problems (https://en.wikipedia.org/wiki/Multicollinearity).

5.2.3 Summary

Advantages

- Relatively simple;
- Implemented in free software (*e.g.*, R, (R Core Team, 2017)).

Disadvantages

- Potentially suffers from collinearity;
- Does generally not fully utilize all available information.

5.3 Principal component regression

5.3.1 Overview

Despite its appealing simplicity, linear regression (Section 5.2) might not be suitable in cases when many explanatory variables are available. This is a common situation in case of sensory data. A spectrometer, for instance, produces reflectances at many wavelengths (see Figure D.1). Multiple linear regression might fail, because the explanatory variables are often highly correlated and therefore nearly collinear (Everitt,

2006, p.265). As a consequence, the matrix inverse in Equations E.3 and E.4 might not exist (Geladi and Kowalski, 1986). In addition, linear regression deals with situations where the number of observations is greater than the number of explanatory variables[1]. In spectrometry on the other hand, the number of explanatory variables often exceeds the number of observations.

One solution is to apply linear regression to a subset of the explanatory variables only, as was done in Section 5.2.2. A disadvantage of this approach is that not all available information will be used, potentially resulting in less accurate predictions. A more appealing solution is to condense the p explanatory variables to a new but smaller set of, say $\ell \ll p$, explanatory variables by filtering out the noise. It is assumed that the smaller data set only contains the relevant information in the original data[2]. Linear regression is then performed on the condensed data set instead of the original data set. This technique is known as principal component regression (PCR). See Jolliffe (2002) for a textbook devoted to principal component analysis or Wold et al. (1987) for an excellent paper on this subject.

Practitioners often find principal component analysis (PCA) hard to understand. The reason might be that PCA uses unfamiliar terminology like 'eigenvectors', 'eigenvalues' or 'latent variables'. In addition, PCA also requires a basic knowledge of linear algebra to fully appreciate the technique. In this section we will try to shed some light on PCA by presenting this technique in a more intuitive way. Details are provided in Appendix F.

5.3.2 Example on organic matter content

Suppose we have a data set consisting of measurements on 100 soil samples, where:

- y is the soil organic matter content;
- and x_1 and x_2, are the diffuse reflectances at two wavelengths measured by means of a spectrometer.

A scatter plot of x_2 versus x_1 is given in Figure 5.5 (left). Each dot represents a soil sample. Clearly, x_1 and x_2 are positively correlated, *i.e.*, x_2 increases as x_1 increases.

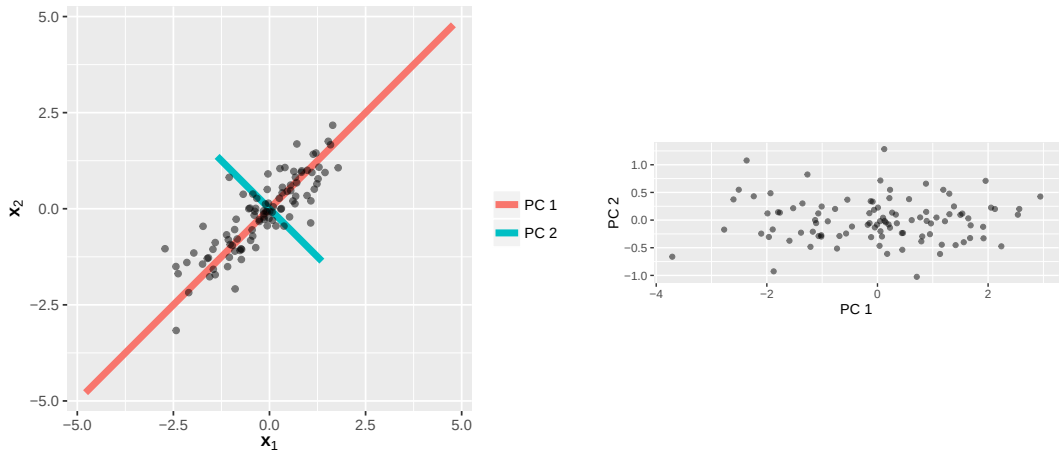


Figure 5.5
Left: Example of 100 soil samples plotted in the space spanned by explanatory variables x_1 and x_2. The corresponding principal components are given in red (PC 1) and bluish-green (PC 2). The right plot gives the corresponding score plot, obtained by rotating the left plot relative to the origin $(x_1, x_2) = (0, 0)$, so that the first principal component axis is now horizontally oriented.

The idea now is to replace the original variables x_1 and x_2 by new orthogonal variables called principal components. The first principal component (PC 1) is drawn in such a way that it explains most of the variation in the data. This is the red line in Fig. 5.5 and is equivalent to an ordinary least squares fit of the

[1] Also known as an overdetermined system of linear equations.
[2] In practice, there is no guarantee that this assumption is always met. See Jolliffe (1982) or Wold et al. (1987).

data (Appendix E). The second principal component (PC 2) will be drawn at a straight angle (orthogonal) to the first principal component and explains the remaining variation in the data.

Rotating the top graph in Figure 5.5 around its origin $(x_1, x_2) = (0, 0)$, and aligning the first principal component horizontally gives the so called score plot (Figure 5.5, right). The first principal component explains about 90% of the variation in the data. The second principal component the remaining 10%. The original data x_1 and x_2 may therefore be represented by the first principal component only, without loosing too much of the variation in the data.

In principal component regression (PCR), the original variables are simply replaced by a smaller set of principal components. In the example above, the multiple linear regression equation:

$$y = b_0 + b_1 x_1 + b_1 x_2 + \varepsilon$$

may be replaced by

$$y = b_0 + b_1 \mathrm{PC}_1 + \varepsilon$$

In this simplified example we only have two explanatory variables x_1 and x_2. Spectrometers often give diffuse reflectances at hundreds of wavelengths. In that case, the number of variables p may even exceed the number of observations n. Ordination regression methods like PCR are then indispensable. Since principal components are orthogonal by construction, collinearity problems are avoided.

5.3.3 Summary

Advantages

- Solves collinearity problem;
- Potentially extracts more information out of the available data;
- Implemented in free software (*e.g.*, R, (R Core Team, 2017));
- After decomposition into scores and loadings (principal components), regression analysis is straightforward.

Disadvantages

- Principal components are extracted without reference to the variable of interest y, but only on the basis of the mutual dependences between the explanatory variables. Hence, there is not necessarily a relation between the first set of principal components and the variable of interest;
- Interpretation is more complicated. It has to be done in terms of eigenvectors (loadings) and eigenvalues (scores) instead of the original data.

5.4 Partial Least Squares Regression

5.4.1 Overview

Principal component regression condenses the explanatory variables to a new but usually smaller set of orthogonal variables called principal components. It is often implicitly assumed that only the first number of principal components contain information on the response variable. The remaining principal components are considered to resemble only noise, and are excluded from further analysis. However, as Jolliffe (1982) and Wold et al. (1987) point out, there is no guarantee that the first set of principal components actually contain useful information on the response variable.

Partial least squares regression (PLSR, Martens and Næs (1989)) is related to principal component regression. It transforms the set of explanatory variables, as in PCR, into a new set of orthogonal variables, but additionally maximizes the relation with the response variable at the same time. In other words, PLSR makes sure that the correlation between a response variable (*e.g.* organic matter content) and the new set of orthogonal variables (*e.g.* derived from a soil spectrum (Figure D.1)) is maximized. It is therefore expected that PLSR results in more accurate predictions than PCR. Details are provided in Appendix G.

5.4.2 Summary

Advantages

- Solves collinearity problems;
- Uses potentially more of the available data;
- Implemented in free software (*e.g.*, R, (R Core Team, 2017));
- Maximizes the correlation between the (transformed) explanatory variables and the response variable.

Disadvantages

- Interpretation is more complicated. It has to be done in terms of eigenvectors (loadings) and eigenvalues (scores) instead of the original data.

5.5 Machine learning

5.5.1 Overview

Machine learning algorithms are a broad class of algorithms that 'learn from data' (*e.g.* soil spectra) to make predictions on for instance soil properties. Examples are bagging (Breiman, 1996), random forest (Breiman, 2001), artificial neural networks, and deep learning.

In this section we will focus on tree like models as these are often used in precision agriculture.

A random forest is an ensemble of Classification And Regression Trees (CART-models) that together predict the soil property of interest. Given a data set of n soil samples, B bootstrap samples of size n are constructed from these data by means of simple random sampling *with* replacement. For each bootstrap sample a CART-model is grown and used to predict the soil property of interest. For classification trees, the ensemble prediction is usually accomplished by a majority vote, for regression trees, the mean of the values in the leaves (terminal nodes of the tree) is usually taken.

Many variations of tree-based models exist. Cubist-models (https://www.rulequest.com/cubist-info.html) for example consist of one or more trees where each terminal leaf contains a linear regression model. Bayesian additive regression trees (BART), is a sum-of-trees ensemble with an estimation approach relying on a fully Bayesian probability model (Kapelner and Bleich, 2016).

5.5.2 Summary

Advantages

- Can handle both linear and nonlinear relations;
- Interactions between explanatory variables are easily taken into account;
- Usually non-parametric, hence no model assumptions are needed;
- Implemented in free software (*e.g.*, R, (R Core Team, 2017))

Disadvantages

- More black-box than linear regression or ordination based methods like PCR and PLSR;
- Some models (like BART) are computationally intensive;
- Measures of uncertainty (prediction intervals, realizations, etc.) are not always implemented in software.

6 Conclusions and recommendations

The following conclusions can be drawn from the inventory of validated proximal sensing techniques:

1. Near and mid infrared spectrometry, γ-ray spectroscopy and electromagnetic induction methods have a potential in spatial prediction of clay, silt and sand content;
2. Near and mid infrared spectrometry have a potential in spatial prediction of organic matter content;
3. Near and mid infrared spectrometry have a potential in spatial prediction of soil moisture content.

The use of pedo-transfer functions requires also the parameters bulk density and M50. Bulk density can be measured by a device called RhoC (Jacobs et al., 2009; Jacobs, 2011). Jacobs et al. (2009) tested this method in predicting bulk density of sediments in marine wetlands and found a good correlation ($R^2 = 0.77$) between bulk density obtained by RhoC and bulk density obtained by traditional methods. The performance of the method in predicting bulk density of terrestrial soils has not been validated yet.

M50 is the median of the sand fraction. As most instruments are at present only calibrated to fractions of clay, silt and sand, their application in the determination of M50 needs to be developed. Possibly the combination of specific surface area (Ben-Dor and Banin, 1995) and fractions of fine and course sand (Viscarra Rossel et al., 2007, e.g.) can give an indication. On the other hand new pedo-transfer functions can be developed neglecting the M50 value while increasing the contribution of other explanatory variables.

Bibliography

Ackerson, J., Demattê, J., and Morgan, C. (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.

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., and Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, 44:71–91.

Bakker, G., Heinen, M., Wesseling, J., de Groot, W., Assinck, F., and Hummelink, E. (2015). Bodemfysische gegevens in BIS. Technical Report 2613, Alterra, Wageningen UR, Wageningen.

Barnes, E. M., Sudduth, K. A., Hummel, J. W., Lesch, S. M., Corwin, D. L., Yang, C., Daughtry, C. S. T., and Bausch, W. C. (2003). Remote- and ground-based sensor techniques to map soil properties. *Photogrammetric Engineering & Remote Sensing*, 69(6):619–630.

Beckett, K. (2008). Multispectral processing of high-resolution radiometric data for soil mapping. *Near Surface Geophysics*, pages 281–287.

Ben-Dor, E. and 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., and 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.

Bernick, M., Getty, D., Prince, G., and Sprenger, M. (1995). Statistical evaluation of field-portable X-ray fluorescence soil preparation methods. *Journal of Hazardous Materials*, 43:111–116.

Besson, A., Cousin, I., Bourennane, H., Nicoullaud, B., Pasquier, C., Richard, G., Dorigny, A., and King, D. (2010). The spatial and temporal organization of soil water at the field scale as described by electrical resistivity measurements. *European Journal of Soil Science*, 61:120–132.

Birkhead, A. L., Heritage, G. L., White, H., and Van Niekerk, A. W. (1996). Ground-penetrating radar as a tool for mapping the phreatic surface, bedrock profile, and alluvial stratigraphy in the Sabie River, Kruger National Park. *Journal of soil and water conservation*, 51(3):234–241.

Bouma, J. (1989). *Using soil survey data for quantitative land evaluation*, volume 9 of *Advances in Soil Science*, pages 177–213. Springer US.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brevik, E. C., Fenton, T. E., and Lazari, A. (2006). Soil electrical conductivity as a function of soil water content and implications for soil mapping. *Precision Agriculture*, 7:393–404.

Brown, D. J., Shepherd, K. D., Walsh, M. G., Mays, M. D., and Reinsch, T. G. (2006). Global soil characterization with vnir diffuse reflectance spectroscopy. *Geoderma*, 132:273–290.

Buchanan, S. and Triantafilis, J. (2009). Mapping water table depth using geophysical and environmental variables. *Ground Water*, 47(1):80–96.

Cattle, S. R., Meakin, S. N., Ruszkowski, P., and Cameron, R. G. (2003). Using radiometric data to identify aeolian dust additions to topsoil of the Hillston district, western NSW. *Australian Journal of Soil Research*, 41:1439–1456.

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

Cockx, L., Van Meirvenne, M., and De Vos, B. (2007). Using the EM38DD soil sensor to delineate clay lenses in a sandy forest soil. *Soil Science Society of America Journal*, 71:1314–1322.

Cockx, L., Vitharana, U. W. A., Verbeke, L. P. C., Simpson, D., Saey, T., and Van Coillie, F. M. B. (2009). Extracting topsoil information from EM38DD sensor data using a neural network approach. *Soil Science Society of America Journal*, 73:1–8.

Corwin, D. L. and Lesch, S. M. (2005). Characterizing soil spatial variability with apparent soil electrical conductivity I. Survey protocols. *Computers and Electronics in Agriculture*, 46:103–133.

Coulouma, G., Caner, L., Loonstra, E. H., and Lagacherie, P. (2016). Analysing the proximal gamma radiometry in contrasting mediterranean landscapes: Towards a regional prediction of clay content. *Geoderma*, 266:127–135.

Cousin, I., Besson, A., Bourennane, H., Pasquier, C., Nicoullaud, B., King, D., and Richard, G. (2009). From spatial-continuous electrical resistivity measurements to the soil hydraulic functioning at the field scale. *C. R. Geoscience*, 341:859–867.

Dhawale, N., Adamchuk, V. I., Prasher, O., Viscarra Rossel, R. A., Ismail, A., and Kaur, J. (2015). Proximal soil sensing of soil texture and organic matter with a prototype portable mid-infrared spectrometer. *European Journal of Soil Science*, 66:661–669.

Doolittle, J. A., Jenkinson, B., Hopkins, D., Ulmer, M., and Tuttle, W. (2006). Hydropedological investigations with ground-penetrating radar (GPR): Estimating water-table depths and local ground-water flow pattern in areas of coarse-textured soils. *Geoderma*, 131:317–329.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 3 edition.

Everitt, B. S. (2006). *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, UK, 3 edition.

Freeland, R. S., Branson, J. L., Ammons, J. T., and Leonard, L. L. (2001). Surveying perched water on anthropogenic soils using non-intrusive imagery. *Transactions of the ASAE*, 44(6):1955–1963.

Freeland, R. S. and Odhiambo, L. O. (2007). Subsurface characterization using textural features extracted from GPR data. *Transactions of the ASABE*, 50(1):287–293.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467.

Ge, Y., Morgan, C., and Ackerson, J. (2014). VisNIR spectra of dried groud soils predict properties of soils scanned moist and intact. *Geoderma*, 221-222:61–69.

Ge, Y., Thomasson, J. A., Morgan, C. L., and Searcy, S. W. (2007). VNIR diffuse reflectance spectroscopy for agricultural soil property determination based on regression-kriging. *Transactions of the ASABE*, 50(3):1081–1092.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17.

Gerber, R., Salat, C., Junge, A., and Felix-Henningsen, P. (2007). GPR-based detection of pleistocene periglacial slope deposits at a shallow-depth test site. *Geoderma*, 139:346–356.

Hartemink, A. E. and Minasny, B. (2014). Towards digital soil morphometrics. *Geoderma*, 230231:305 – 317.

Heinen, M. and Bakker, G. (2016). Implications and application of the Raats superclass of soils equations. *Vadose Zone Journal*, 15(8). doi:10.2136/vzj2016.02.0012.

Hendriks, P. H. G. M., Limburg, J., and de Meijer, R. J. (2001). Full-spectrum analysis of natural -ray spectra. *Journal of Environmental Radioactivity*, 53(3):365 – 380.

Huisman, J. A., Hubbard, S. S., Redman, J. D., and Annan, A. P. (2003). Measuring soil water content with Ground Penetrating Radar: A review. *Vadose Zone Journal*, 2:476–491.

Hummel, J. W., Sudduth, K. A., and Hollinger, S. E. (2001). Soil moisture and organic matter prediction of surface and subsurface soils using an NIR soil sensor. *Computers and Electronics in Agriculture*, 32:149–165.

Islam, K., Singh, B., and McBratney, A. (2003). Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Australian Journal of Soil Research*, 41:1101–1114.

Jacobs, W. (2011). *Sand-mud erosion from a soil mechanical perspective*. PhD thesis, Technische Universiteit Delft.

Jacobs, W., Eelkema, M., Limburg, H., and Winterterp, J. (2009). A new radiometric instrument for *in situ* measurements of physical sediment properties. *Marine & Freshwater Research*, 60:727–736.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):300–303.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2 edition.

Kalnicky, D. J. and Singhvi, R. (2001). Field portable XRF analysis of environmental samples. *Journal of Hazardous Materials*, 83:93–122.

Kapelner, A. and Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40.

Knotters, M., Brus, D. J., Verzandvoort, S., and Heinen, M. (2011). Aanvullende bodemfysische gegevens voor BIS-Nederland. Technical Report 2245, Alterra.

Kühn, J., Brenning, A., Wehran, M., Koszinski, S., and Sommer, M. (2009). Interpretation of electrical conductivity patterns by soil properties and geological maps for precision agriculture. *Precision Agriculture*, 10:490–507.

Lagacherie, P., Baret, F., Feret, J.-B., Madeira Netto, J., and Robbez-Masson, J. M. (2008). Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. *Remote Sensing of Environment*, 112:825–835.

Lambot, S., Antoine, M., van den Bosch, I., Slob, E. C., and Vanclooster, M. (2004). Electromagnetic inversion of GPR signals and subsequent hydrodynamic inversion to estimate effective vadose zone hydraulic properties. *Vadose Zone Journal*, 3:1072–1081.

Lapen, D. R., Moorman, B. J., and Price, J. S. (1996). Using ground-penetrating radar to delineate subsurface features along a wetland catena. *Soil Science Society of America Journal*, 60:923–931.

Lark, R. M. and Cullis, B. R. (2004). Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, 55:799–813.

Levin, N., Ben-Dor, E., and Singer, A. (2005). A digital camera as a tool to measure colour indices and related properties of sandy soils in semi-arid environments. *International Journal of Remote Sensing*, 26(24):5475–5492.

Mahmood, H., Hoogmoed, W., and Van Henten, E. (2013). Proximal gamma-ray spectroscopy to predict soil properties using windows and full-spectrum analysis methods. *Sensors*, 13:16263–16280. doi:10.3390/s131216263.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*. John Wiley & Sons, Chichester.

McBratney, A. B., Mendonça Santos, M. L., and Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117:3–52.

Melendez-Pastor, I., Navarro-Pedreño, J., Gómez, I., and Koch, M. (2008). Identifying optimal spectral bands to assess soil properties with VNIR radiometry in semi-arid soils. *Geoderma*, 147:126–132.

Minasny, B., McBratney, A., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L., and Joalland, S. (2011). Removing the effect of soil moisture from nir diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma*, 167-168:118–124.

Mouazen, A. M., Maleki, M. R., Cockx, L., Van Meirvenne, M., Van Holm, L. H. J., Merckx, R., De Baerdemaeker, J., and Ramon, H. (2009). Optimum three-point linkage set up for improving the quality of soil spectra and the accuracy of soil phosphorus measured using an on-line visible and near infrared sensor. *Soil & Tillage Research*, 103:144–152.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572.

Pracilio, G., Adams, M. L., Smettem, K. R. J., and Harper, R. J. (2006). Determination of spatial distribution patterns of clay and plant available potassium contents in surface soils at the farm scale using high resolution gamma ray spectrometry. *Plant and Soil*, 282:67–82.

Press, H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992). *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saey, T., Simpson, D., Vermeersch, H., Cockx, L., and Van Meirvenne, M. (2009a). Comparing the EM38DD and DUALEM-21S sensors for depth-to-clay mapping. *Soil Science Society of America Journal*, 73(1):7–12.

Saey, T., Van Meirvenne, M., Vermeersch, H., Ameloot, N., and Cockx, L. (2009b). A pedotransfer function to evaluate the soil profile textural heterogeneity using proximally sensed apparent electrical conductivity. *Geoderma*, 150:389–395.

Shih, S. F., Doolittle, J. A., Myhre, D. L., and Schellentrager, G. W. (1986). Using radar for groundwater investigation. *Journal of Irrigation and Drainage Engineering*, 112:110–118.

Shonk, J. L., Gaultney, L. D., Schulze, D. G., and Van Scoyoc, G. E. (1991). Spectroscopic sensing of soil organic matter content. *Transactions of the ASAE*, 34(5):1978–1984.

Smith, M. C., Vellidis, G., Thomas, D. L., and Breve, M. A. (1992). Measurement of water table fluctuations in a sandy soil using ground penetrating radar. *Transactions of the ASAE*, 35(4):1161–1166.

Stolte, J., Wesseling, J. G., and Verzandvoort, S. (2007). Kwaliteitsdocumentatie voor de verkrijging van Status A voor de gegevens van de Staringreeks zoals opgenomen in het gegevensbestand Priapus. Versie 1. Technical Report 1522, Alterra.

Sudduth, K. A., Drummond, S. T., and Kitchen, N. R. (2001). Accuracy issues in electromagnetic induction sensing of soil electrical conductivity for precision agriculture. *Computers and Electronics in Agriculture*, 31:239–264.

Triantafilis, J. and Lesch, S. M. (2005). Mapping clay content variation using electromagnetic induction techniques. *Computers and Electronics in Agriculture*, 46:203–237.

Triantafilis, J., Lesch, S. M., La Lau, K., and Buchanan, S. M. (2009). Field level digital soil mapping of cation exchange capacity using electromagnetic induction and a hierarchical spatial regression model. *Australian Journal of Soil Research*, 47:651–663.

Triantafilis, J. and Monteiro Santos, F. A. (2009). 2-Dimensional soil and vadose-zone representation using an EM38 and EM34 and a laterally constrained inversion model. *Australian Journal of Soil Research*, 47:809–820.

Van der Klooster, E., Van Egmond, F. M., and Sonneveld, M. (2011). Mapping soil clay contents in Dutch marine ddistrict using gamma-ray spectrometry. *European Journal of Soil Science*, (doi: 10.1111/j.1365-2389.2011.01381.x).

Van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44(5):892–898.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M., and Van Genuchten, M. T. (2010). Using pedotransfer functions to estimate the van Genuchten-Mualem soil hydraulic properties: a review. *Vadose Zone Journal*, 9:795–820.

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B., 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., and Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, 155(Supplement C):198 – 230.

Viscarra Rossel, R. A., Cattle, S. R., Ortega, A., and Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by visNIR spectroscopy. *Geoderma*, 150:253–266.

Viscarra Rossel, R. A., Taylor, H. J., and McBratney, A. B. (2007). Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. *European Journal of Soil Science*, 58:343–353.

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

Vitharana, U. W. A., Saey, T., Cockx, L., Simpson, D., Vermeersch, H., and Van Meirvenne, M. (2008). Upgrading a 1/20,000 soil map with an apparent electrical conductivity survey. *Geoderma*, 148:107–112.

Vitharana, U. W. A., Van Meirvenne, M., Cockx, L., and Bourgeois, J. (2006). Identifying potential management zones in a layered soil using several sources of ancillary information. *Soil Use and Management*, 22:405–413.

Walvoort, D. and McBratney, A. (2001a). Diffuse reflectance spectrometry as a proximal sensing tool for precision agriculture. In *Third Eurpean conference on precision agriculture (3rd ECPA)*.

Walvoort, D. J. J. and McBratney, A. B. (2001b). Diffuse reflectance spectrometry as a proximal sensing tool for precision agriculture. In Grenier, G. and Blackmore, S., editors, *Proceedings of the 3rd European Conference on Precision Agriculture (Book and CD-ROM)*, pages 503–508, Montpellier, France. Agro Montpellier, Ecole Nationale Superieure Agronomique.

Weindorf, D. C., Zhu, Y., Ferrell, R., Rolong, N., Barnett, T., Allen, B. L., Herrero, J., and Hudnall, W. (2009). Evaluation of portable X-Ray Fluorescence for Gypsum quantification in soils. *Soil Science*, 174(10):556–562.

Weller, U., Zipprich, M., Sommer, M., Zu Castell, W., and Wehrhan, M. (2007). Mapping clay content across boundaries at the landscape scale with electromagnetic induction. *Soil Science Society of America Journal*, 71(6):1740–1747.

Wijewardane, N., Ge, Y., and Morgan, C. (2016). Moisture intensive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. *Geoderma*, 267:92–101.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:37–52.

Wösten, J., de Vries, F., Hoogland, T., Massop, H., Veldhuizen, A., Vroon, H., Wesseling, J., Heijkers, J., and Bolman, A. (2013). BOFEK2012, de nieuwe, bodemfysische schematisatie van Nederland. Technical Report 2387, Alterra, Wageningen.

Wösten, J. H. M. (1997). Bodemkundige vertaalfuncties bij SC-DLO. State of the art. Technical Report 563, SC-DLO, Wageningen.

Wösten, J. H. M., Bannink, M., and Beuving, J. (1987). Waterretentie- en doorlatendheidskarakteristieken van boven- en ondergronden in Nederland: de Staringreeks. Technical Report 1932, Stiboka, Wageningen.

Wösten, J. H. M., De Vries, F., Denneboom, J., and Van Holst, A. F. (1988). Generalisatie en bodemkundige vertaling van de bodemkaart 1 : 250 000, ten behoeve van de PAWN-studie. Technical Report 2055, Stiboka, Wageningen.

Wösten, J. H. M., Lilly, A., Nemes, A., and Le Bas, C. (1999). Development an use of a database of hydraulic properties of European soils. *Geoderma*, 90:169–185.

Wösten, J. H. M., Veerman, G., de Groot, W., and Stolte, J. (2001). Waterretentie- en doorlatendheidskarakteristieken van boven- en ondergronden in Nederland: de Staringreeks. vernieuwde uitgave 2001. Technical Report 153, Alterra, Wageningen.

Zhu, Y. and Weindorf, D. C. (2009). Determination of soil Calcium using field portable X-Ray Fluorescence. *Soil Science*, 174(3):151–155.

A Appendix 1 Set up of a validation experiment

A.1 Selection target variables and measurement devices

The primary focus in this validation experiment is on the performance of γ-ray spectroscopy in the spatial prediction of variables that are used in pedo-transfer functions for the Mualem-Van Genuchten parameters: clay, silt and sand content and median grain size of sand fraction (M50). The γ-ray device will be applied close to the ground surface using a quad and at various heights above the ground surface using a drone, to evaluate the prediction performance in areas that are not accessible with a quad. The validation experiment can be extended to evaluate the performance of near infrared spectrometry in the spatial prediction of organic matter content, clay, silt and sand content and median grain size of sand fraction (M50).

A.2 Selection of a study area

A.2.1 Method

A study area for testing several types of proximal sensors was selected according to the following general requirements:

1. the study area should be part of Oostelijk Flevoland;
2. the areas should be contiguous;
3. land-use should be a mixture of agriculture and nature (including forest);
4. the study are should have a wide range of soil properties (organic carbon, clay content, loam content, M50, bulk density), varying from low to high. There should be sufficient spatial variation at short distances in the target variables.

Note that the availability of forests has been superseded because these areas are problematic for both quads and drones. As an alternative, we will fly at different (higher) altitudes over agricultural areas. This is a kind of best case for forests. If, at higher altitudes, the results are bad, then the results will be even worse for flying over a forest (due to possible interference of the trees).

The following additional requirements for the use of drones were obeyed in the selection of a study area:

1. no built up area;
2. at least 10 km from an airport;
3. at least 50 m from 80 km roads;
4. at least 150 m from highways;
5. at least 150 m from high voltage pylons and high voltage lines.

The size of the study area depends on the available time to collect data by using a drone. The use of the drone is restricted to three days. A drone can collect data for 15 to 20 hectares in one day, thus the study area should not exceed 45 to 60 hectares.

The following selection procedure has been applied:

1. download the most recent version of the topographic map scale 1:10000;
2. select all regions that should potentially be included (vector map A);
   (a) select agricultural areas;
   (b) select nature areas;
   (c) combine these areas.

3. select all regions that should be excluded (vector map B):

   (a) select all features listed above (e.g., built-up, airports, roads, highways,) that are in the middle of no-go-areas;

   (b) buffer all these features with the specifications given above (e.g., a buffer of 150 meters on both sides of highways);

   (c) combine all buffers to get a map of all areas that should be excluded (no-go areas).

4. remove all areas in vector map A that are part of vector map B. The result is vector map C;

5. rasterize the soil map 1:50000, the map of groundwater table classes 1:50000, and the land-use map 1:50000 for all polygons in map C (rasters are currently more convenient than vector maps for the analysis below);

6. compute a heterogeneity index for each pixel in map C. Areas where this index is high, are potentially interesting study areas.

The soil map 1:50000, the map of groundwater table classes 1:50000 and the land-use map 1:50000 were combined by assigning an integer value to each unique combination of soil type, water table class and land-use category. Next a heterogeneity index was calculated for each pixel by the normalized Shannon entropy H_{n}, wich is calculated as follows:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{A.1}$$

where p_i is the probability of a specific class, in this case a combination of soil type, water table class and land-use category. The total number of classes equals n. The normalized Shannon entropy is given by

$$H_{\mathrm{n}} = \frac{H}{H_{\max}} \tag{A.2}$$

where $H_{\max} = \log_2(n)$ is the maximum entropy. For each pixel, the normalized Shannon entropy is computed based on a search radius of about 357 meters around each pixel, i.e., an area of 40 hectares.

Not only the entropy of the area is important, also its acreage, since we do not prefer a study area that consists of a lot of small patches. Contiguity is therefore also an important selection criterion and is calculated for each pixel as the area of patches in a search radius of 357 m, with a theoretical maximum of 40 hectares. An entropy of 0.35 and an acreage of 35 hectares were used as thresholds in selecting candidate pixels. Since the candidate pixels appeared to be located in clusters the k-mean algorithm was used to calculate the centers of the clusters. Next the study area was selected from these centers using additional checks based on aerial photographs (Google Earth).

A.2.2 Results

Figure A.1 shows the potential study area without no-go areas for drones in Oostelijk Flevoland (referred to as 'map C' in the previous subsection).

In the green area in Figure A.1 areas of spatial variation were designated using the map of combinations of soil types, water table classes and land-use categories (Figure A.2) and the resulting map of normalized Shannon entropies indicating heterogeneity (Figure A.3).

Figure A.4 shows the area of patches in a search radius of 357 m around each pixel, with a theoretical maximum of 40 hectares, indicating contiguity of patches.

Figure A.5 shows all locations with a normalized entropy greater than 0.35 and a contiguous area greater than 35 hat (centers of clustered pixels, calculated by the k-means algorithm).

After additional checks on the basis of aerial photographs (Google Earth) location 3 in Figure A.5 was selected. Figure A.6 shows this location in close-up. After receiving permission from the land owners a study area of ... hectares around the selected location was selected as study area for the validation experiment.

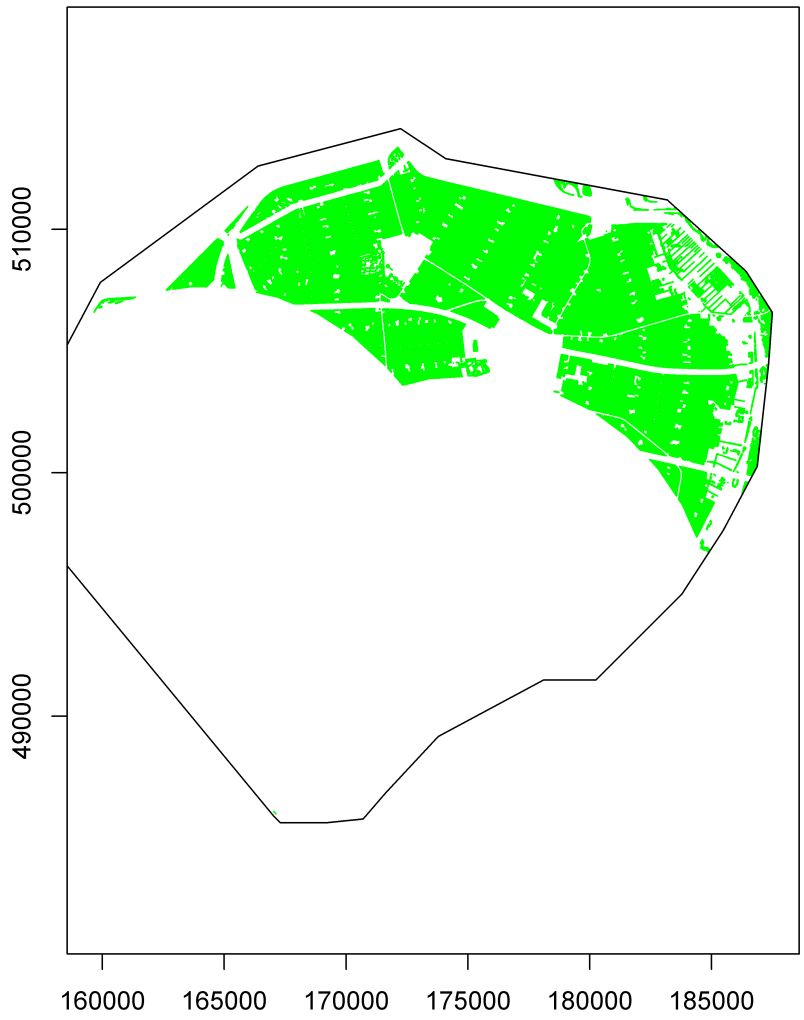


Figure A.1
No-go areas for drones in Oostelijk Flevoland

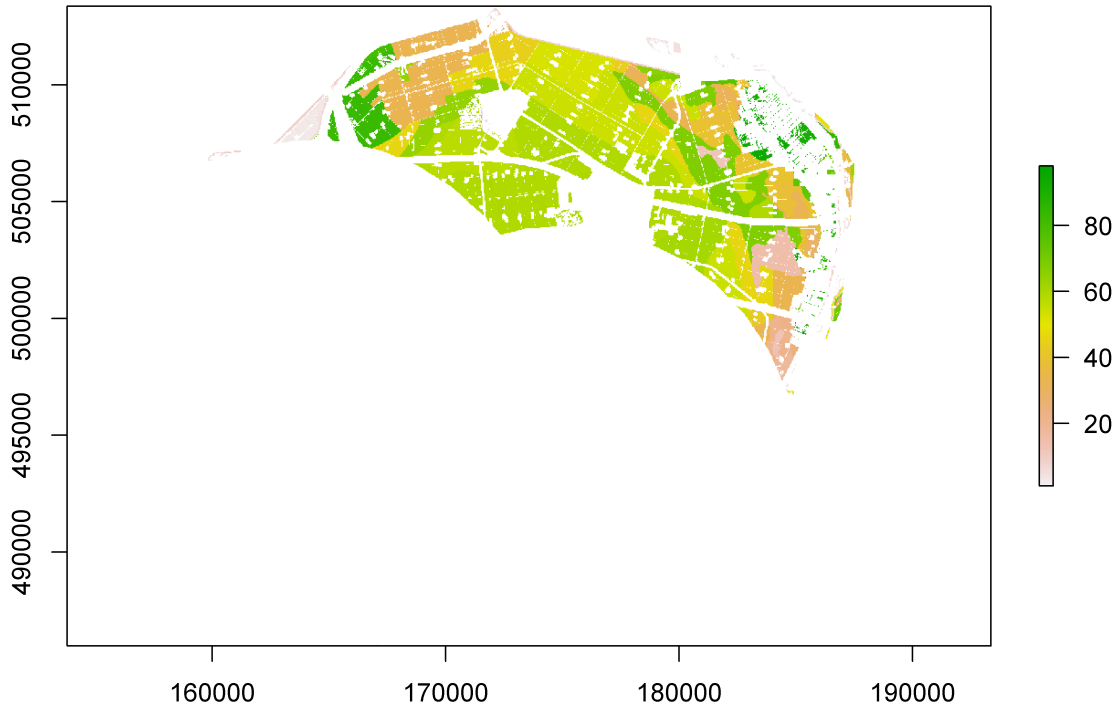


Figure A.2
Map of combinations of soil types, water table classes and land-use categories

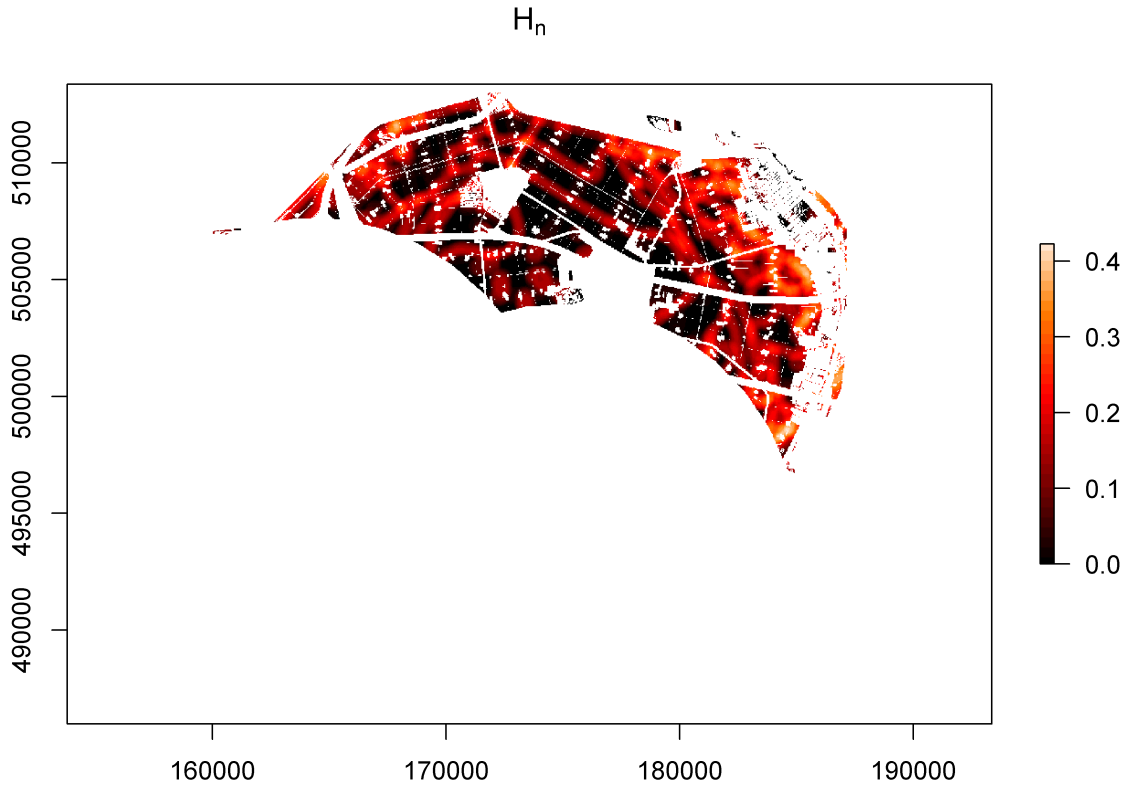


Figure A.3
Map of normalized Shannon entropies, calculated from Figure A.2 usin a search radius of about 357 meter around each pixel

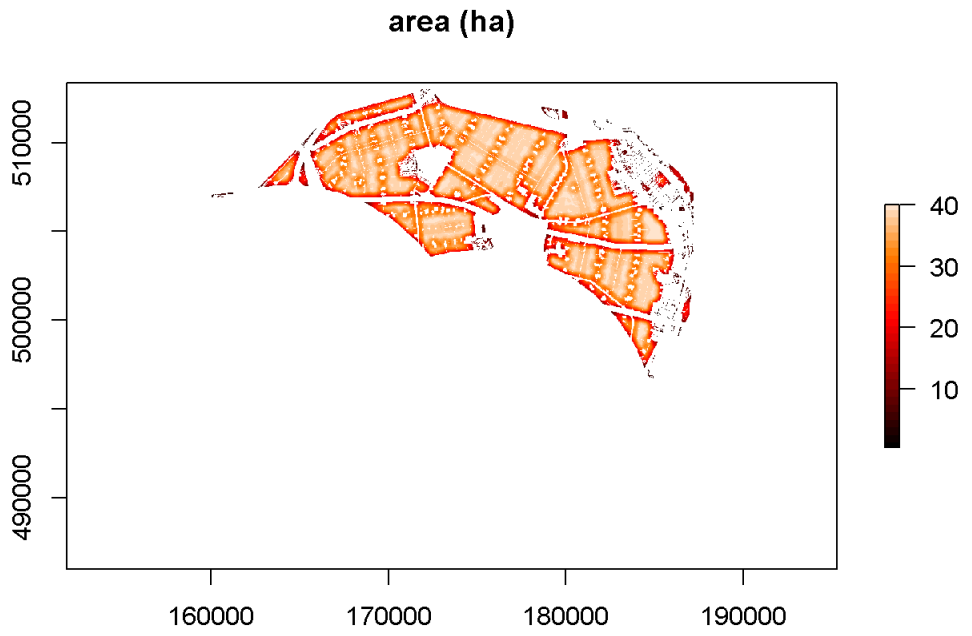


Figure A.4
Area of patches in a search radius of 357 m around each pixel

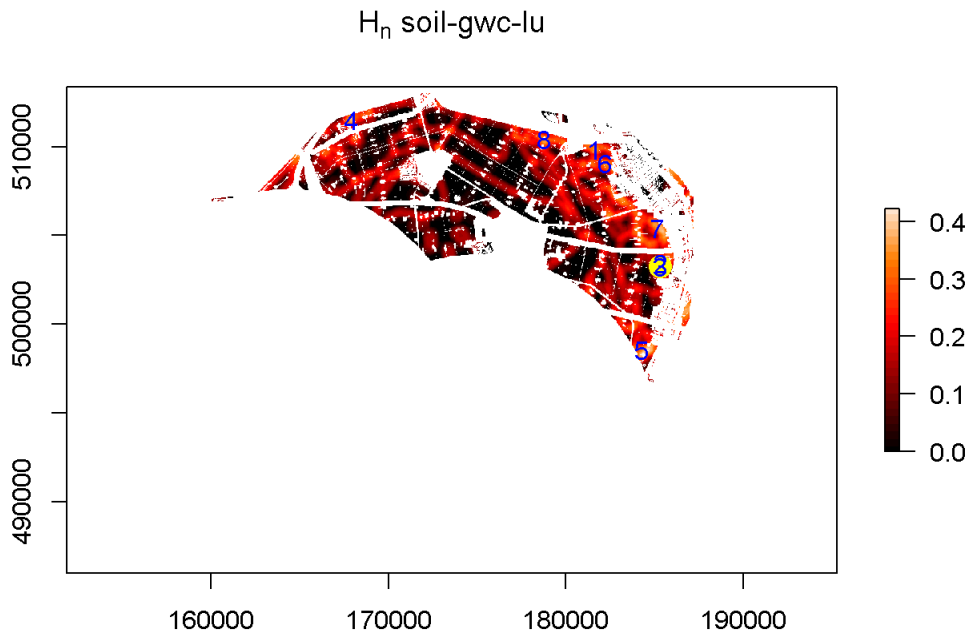


Figure A.5
Locations of areas with a normalized entropy greater than 0.35 and a contiguous area greater than 35 ha. The locations were calculated from clusters of pixels using the k-means algorithm. Location 3 (yellow) has been selected.

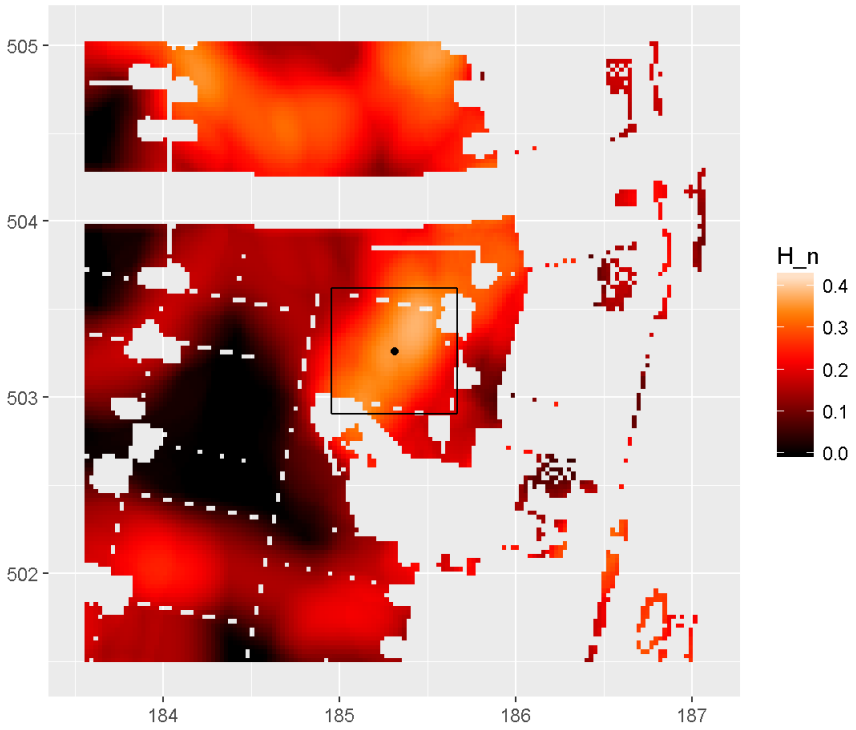


Figure A.6
Close-up of the selected location 3 in Figure A.5

A.3 Sampling strategy

A.3.1 Method

We aimed for a distribution of sampling units in both the geographic and the feature space. To further the distribution of the sampling units in feature space a first stratification of the study area was made on the basis of the soil map. This resulted in seven strata. To further the distribution of the sampling units in the geographic space, these seven strata were next subdivided in compact geographic strata using the R-package Spcosa. Within each of these geographic strata two sampling locations were selected by simple random sampling.

A.3.2 Results

Table A.1 summarizes the stratification of the study area. Figure A.7 shows the distribution of sampling units and strata in the study area.

Table A.1
Division of the study area into strata

Soil stratum	Soil type, water table class	Geographic stratum	Areal size (ha)
1	kVz, VI	1	2.3416
1	kVz, VI	2	2.4470
2	Mn12AF, V	1	3.4207
2	Mn12AF, V	2	2.3275
2	Mn12AF, V	3	3.4309
2	Mn12AF, V	4	2.3720
3	Mn15ApF, V	1	2.2289
3	Mn15ApF, V	2	2.9090
3	Mn15ApF, V	3	2.0997
4	Mn15AF, V	1	2.7192
4	Mn15AF, V	2	2.5233
4	Mn15AF, V	3	2.2654
4	Mn15AF, V	4	3.3591
5	Zn50A, VI	1	0.2181
6	Mv51Ap/Mn25Awp, VI	1	1.8565
6	Mv51Ap/Mn25Awp, VI	2	2.6758
6	Mv51Ap/Mn25Awp, VI	3	3.1003
6	Mv51Ap/Mn25Awp, VI	4	2.9758
6	Mv51Ap/Mn25Awp, VI	5	2.4753
6	Mv51Ap/Mn25Awp, VI	6	2.9227
7	Mn15ApF, VI	1	2.8279
7	Mn15ApF, VI	2	4.0484

kVz: peat soil with a clayey top layer and a sandy subsoil. Mn12AF: Marine, calcareous sandy loam (8-17.5% $< 2\mu$m) with a sandy subsoil starting between 40 and 80 cm, reworked. Mn15ApF: Marine, calcareous sandy loam (8-17.5% $< 2\mu$m) with a Pleistocene sandy subsoil starting between 80 and 120 cm, reworked. Mn15AF: Marine, calcareous sandy loam (8-17.5% $< 2\mu$m), reworked. Zn50A: Medium fine sand (M50 150-210 μm), calcareous. Mv51Ap: Calcareous sandy (clay) loam (8-25% $< 2\mu$m) with a peaty layer of at least 40 cm thickness starting between 40 and 80 cm, and a Pleistocene sandy subsoil starting between 80 and 120 cm. Mn25Awp: Calcareous sandy clay loam (17.5-25%$< 2\mu$m) with a peaty layer of 15-40 cm thickness starting between 40 and 80 cm, and a Pleistocene sandy subsoil starting between 80 and 120 cm. Water table class V: top of seasonal fluctuation 0-40 cm below ground surface, bottom of seasonal fluctuation >120 cm below ground surface. Water table class VI: top of seasonal fluctuation 40-80 cm below ground surface, bottom of seasonal fluctuation >120 cm below ground surface.

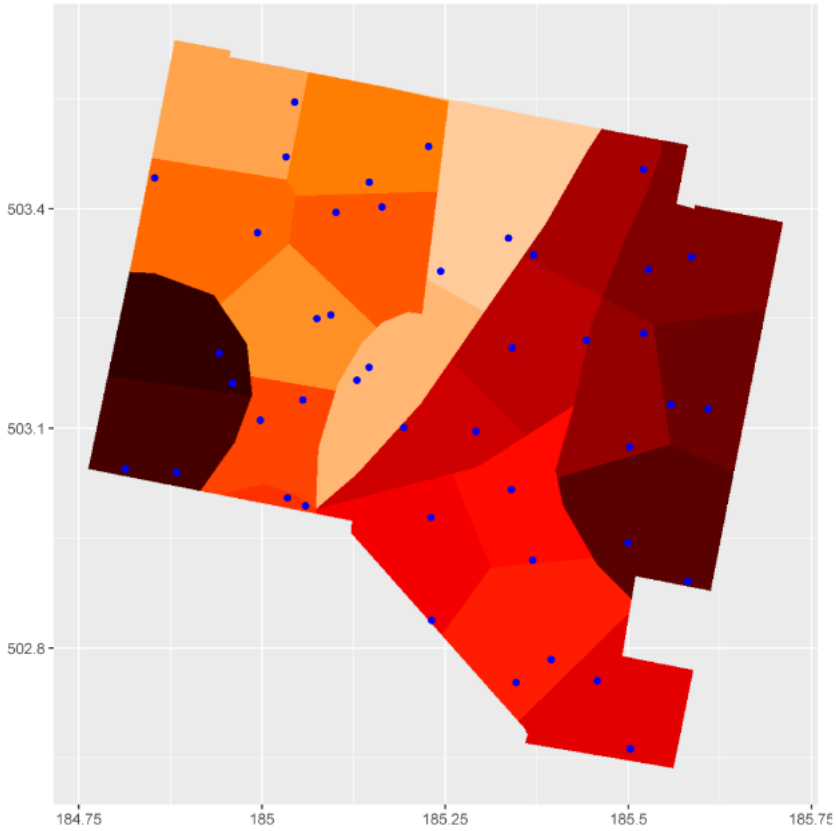


Figure A.7
Locations of the validation points

B Notation

Table B.1
Symbols and description

symbol	size	description
ℓ		number of factors
m		number of responses
n		number of observations
p		number of predictors
$\mathbf{b}$	p	parameter vector
$\boldsymbol{\varepsilon}$	n	error vector
$\mathbf{f}$	n	error vector
$\mathbf{q}$	ℓ	loadings vector
$\mathbf{y}$	n	response vector
$\mathbf{A}$	$p \times p$	correlation matrix
$\mathbf{E}$	$n \times p$	error matrix
$\mathbf{F}$	$n \times m$	error matrix
$\mathbf{I}$	depends on context	identity matrix
$\mathbf{P}$	$p \times \ell$	loadings matrix
$\mathbf{Q}$	$m \times \ell$	loadings matrix
$\mathbf{T}$	$n \times \ell$	score matrix
$\mathbf{X}$	$n \times p$	predictor matrix
$\mathbf{Y}$	$n \times m$	response matrix
$\mathbf{V}$	$p \times p$	covariance matrix

C pH-meter

A pH-meter is basically a voltmeter that converts voltages to pH-units. This conversion is often based on a linear model:

$$\hat{y} = b_0 + b_1 x \tag{C.1}$$

where $\hat{y}$ is the predicted pH, x is the voltage, and b_0 and b_1 are parameters. Prior to pH measurement, these parameters need to be estimated by measuring the voltage in buffer solutions of known pH. These buffer pHs should span the pH range of interest. Suppose that the voltage in buffer 1 is x_1 and the voltage in buffer 2 is x_2, then the parameters b_0 and b_1 can be estimated by solving a set of two linear equations with two unknowns b_0 and b_1:

$$y_1 = b_0 + b_1 x_1$$
$$y_2 = b_0 + b_1 x_2$$

where y_1 and y_2 are the known pHs of buffer solutions 1 and 2 respectively. This procedure is called calibration. After calibration, the parameters b_0 and b_1 are known. By measuring the voltage x in a solution with unknown pH, the pH meter uses Equation C.1 to convert the voltage x to a pH measurement $\hat{y}$.

D Moree Soil Spectra

The methods in this report will be illustrated by means of a case study adopted from Walvoort and McBratney (2001b). A total of 59 soil samples have been collected at a field near Moree, New South Wales, Australia. Each soil sample has been analyzed for total carbon ($\mathrm{g/kg}$) and clay content ($\mathrm{kg/kg}$). In addition, diffuse reflectance spectra have been taken by means of Varian Cary 500 scan spectrophotometer equipped with a Labsphere DRA-CA-50D diffuse reflectance accessory. The spectra and the histograms for clay and total carbon contents are given in Figure D.1.

Each spectrum gives the diffuse reflectance of a light beam on a single soil sample for 1065 wavelenghts. Reflections for wavelengths between 750 and 880 nm have been removed since these are noisy due to changes in filter and grating by the spectrometer.

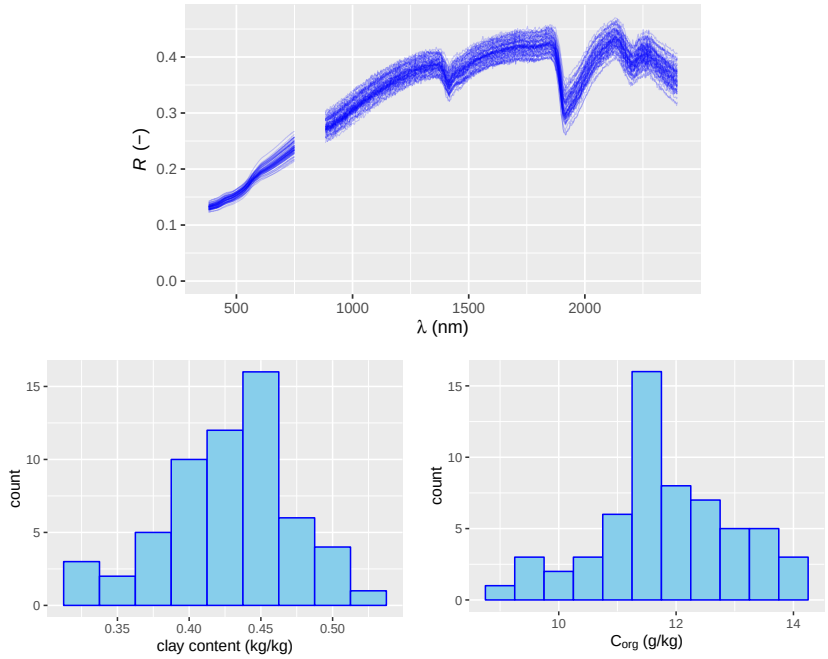


Figure D.1
Top: diffuse reflectance spectra for 59 soil samples for the field near Moree (NSW, Australia); bottom: the corresponding histograms for clay content and total carbon content. See Walvoort and McBratney (2001b) for more information.

E Parameter estimation in linear regression

The parameters in Equation 5.2 can be estimated by substituting n observations y_i and corresponding explanatory variables x_{ij} ($i = 1 \ldots n,\ j = 1 \ldots p$) into the model:

$$y_i = b_0 x_{i0} + b_1 x_{i1} + \cdots + b_p x_{ip} + \varepsilon_i = \hat{y}_i + \varepsilon_i \qquad i = 1 \ldots n \tag{E.1}$$

It is often convenient and more concise to express Equation E.1 in vector-matrix form:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \tag{E.2}$$

where vector $\mathbf{y} = [y_1, \ldots, y_n]'$ contains the response for each observation, $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_p]$ is a $n \times p$ matrix where each column $\mathbf{x}_i$ pertains to a specific variable and each row to a specific observation. If the errors $\boldsymbol{\varepsilon}$ are independent and identically distributed (i.i.d.) then the unknown parameters $\mathbf{b}$ can be estimated by minimizing the sum of squared errors given by:

$$(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

The *ordinary* least squares solution is then given by

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{E.3}$$

However, if the errors are identically distributed with variance σ^2 but correlated with correlation matrix $\mathbf{A}$, then the variance-covariance matrix $\mathbf{V} = \sigma^2\mathbf{A}$ of the errors should also be taken into account. Hence, the expression to minimize is:

$$(\mathbf{y} - \mathbf{X}\mathbf{b})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})$$

The *generalized* least squares solution is given by

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{E.4}$$

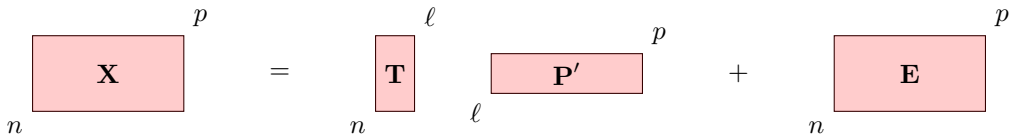
For a new vector of explanatory variables $\mathbf{x}$, predictions $\hat{y}$ are obtained by applying Equation 5.2. See Draper and Smith (1998) for a general text book on regression analysis and Lark and Cullis (2004) for an excellent and concise introduction to least squares, generalized least squares, and maximum likelihood estimation of $\mathbf{b}$.

F Principal component regression

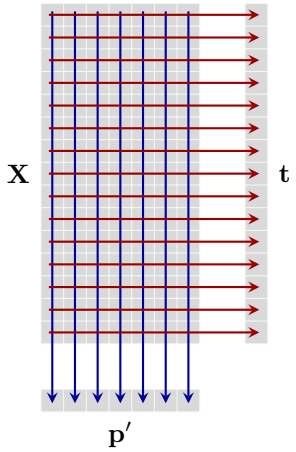
To overcome the problem of collinearity, the explanatory variables in the $n \times p$ matrix $\mathbf{X}$ are decomposed into two terms by means of principal component analysis, an ordination technique first formulated by Pearson (1901). The first term is assumed to contain all the relevant soil information ('signal'), the second term all the noise:

$$\mathbf{X} = \underbrace{\mathbf{T}\mathbf{P}'}_{\text{signal}} + \underbrace{\mathbf{E}}_{\text{noise}} \tag{F.1}$$

where $\mathbf{X}$ is a $n \times p$ matrix of explanatory variables which are column-wise centered and scaled to zero mean and unit variance, $\mathbf{T}$ is a $n \times \ell$ matrix of scores, $\mathbf{P}$ is a $p \times \ell$ projection matrix of loadings, and $\mathbf{E}$ is a $n \times p$ matrix of residual noise. To fully appreciate the potential amount of data compression, one should realize that $\ell \ll p$. Schematically:



The projection matrix $\mathbf{P}$ consists of ℓ (column) vectors $\mathbf{p}_i$ each representing a principal component (a. k. a. loading). As has been illustrated in the figure below, each element of vector $\mathbf{p}_i$ is a projection of the variables (columns) in $\mathbf{X}$ and each element of vector $\mathbf{t}$ is a projection of the observations (rows) in $\mathbf{X}$ (Geladi and Kowalski, 1986; Wold et al., 1987):

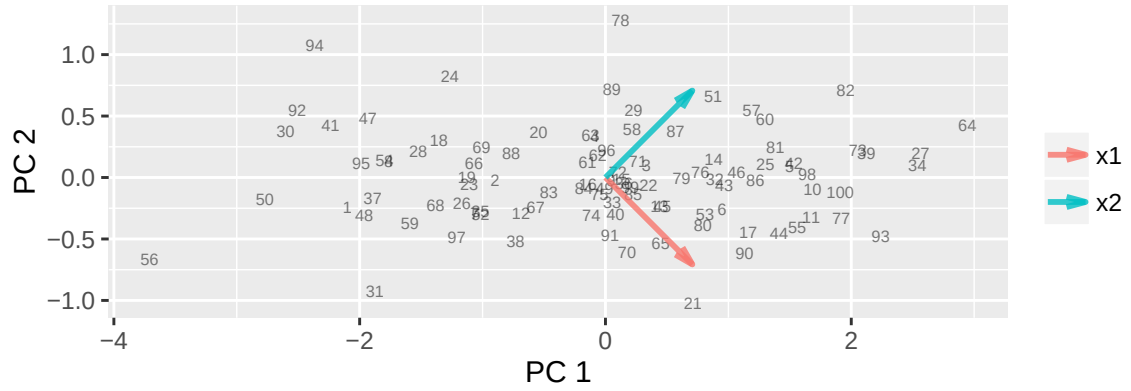


The first vector $\mathbf{p}_1$ is the first principal component and is obtained in such a way that it explains most of the variation in $\mathbf{X}$. Vector $\mathbf{p}_2$ is the second principal component. It is orthogonal to $\mathbf{p}_1$ and explains most of the remaining variation in $\mathbf{X}$ after $\mathbf{p}_1$ has been extracted. The i^{th} vector $\mathbf{p}_i$ is orthogonal to all previously extracted vectors $\mathbf{p}_j$ where $j < i$ and explains most of the remaining variation in $\mathbf{X}$. Since the principal components are at right angles to each other, it follows that

$$\mathbf{P}'\mathbf{P} = \mathbf{I}$$

where $\mathbf{I}$ is the $\ell \times \ell$ identity matrix (with ones on its diagonal, and zeroes on all off-diagonal elements). The elements of each vector can be interpreted as weights with respect to the original variables, *i.e.*, the column vectors of $\mathbf{X}$. The columns of the loadings matrix are also referred to as eigenvectors.

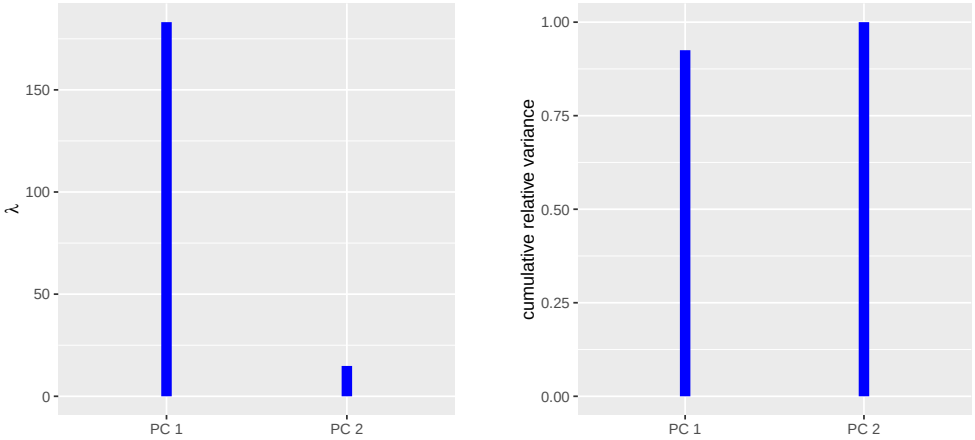
The score matrix $\mathbf{T}$ gives the original data projected onto the column vectors of $\mathbf{P}$. This has been illustrated in Figure 5.5. In other words, the scores are the original data projected onto the new axes formed by the principal components in $\mathbf{P}$. A plot of the data in the space spanned by a pair of principal components is called a score plot. An example of a score plot is given in Figure 5.5 (middle plot). A score plot in which also the original axes are given is called a biplot (Gabriel, 1971). This potentially facilitates the interpretation of the principal components. An example of a biplot is given below. Here the objects are labeled from 1 to 100 to simplify identification.



A scree plot gives for each principal component, the amount of variance in the data that it explains. For component i, the variance is given by

$$\lambda_i = \mathbf{t}_i'\mathbf{t}_i$$

where $\mathbf{t}_i$ is the i^{th} column vector of score matrix $\mathbf{T}$. Variance λ_j is also referred to as the eigenvalue. An example of a scree plot is given on the left of the figure below. The figure on the right gives an alternative representation in which the relative contribution of each eigenvalue is given. This corresponds to the cumulative fraction of variance accounted for by the principal components. The first principal component explains 90% of the variation in $\mathbf{X}$, the first two principal components together explain all variation in $\mathbf{X}$.



One method to estimate $\mathbf{P}$ is by eigen decomposition of $\mathbf{X}'\mathbf{X}$. Note that $(\mathbf{X}'\mathbf{X})/(n-1)$ is the covariance matrix of $\mathbf{X}$, if $\mathbf{X}$ is mean centered, and the correlation matrix of $\mathbf{X}$ if the columns of $\mathbf{X}$ are both mean centered and scaled to unit variance. Let $\mathbf{A}=\mathbf{X}'\mathbf{X}$, then the eigen decomposition of $\mathbf{A}$ is given by

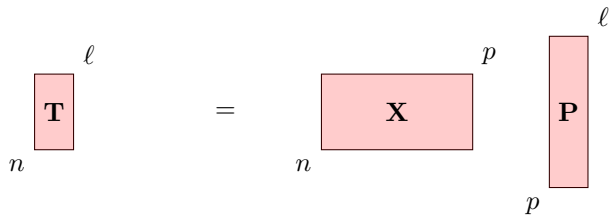
$$\mathbf{AP} = \mathbf{P\Lambda}$$

or equivalently,

$$\mathbf{A} = \mathbf{P\Lambda P}^{-1}$$

where the columns of projection matrix $\mathbf{P}$ contains the eigenvectors (or principal components in case of PCA), and $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix of eigenvalues. The scores $\mathbf{T}$ can then be obtained by projecting the data in $\mathbf{X}$ onto the new axes in $\mathbf{P}$:

$$\mathbf{T} = \mathbf{XP}$$



At first sight, eigenvalue decomposition seems rather cryptic. However, the expression above becomes more clear for the simplified case where we only want to extract a single eigenvector and corresponding eigenvalue:

$$\mathbf{Ap} = \lambda\mathbf{p}$$

where $\mathbf{p}$ is an eigenvector (column of $\mathbf{P}$) and λ is a scalar representing the corresponding eigenvalue.

More numerically accurate decompositions can be obtained by singular value decomposition (Press et al., 1992), or by the NIPALS-algorithm (Nonlinear Iterative Partial Least Squares, Geladi and Kowalski, 1986; Wold et al., 1987; Martens and Næs, 1989). The NIPALS-algorithm is more efficient than singular value decomposition in the general case when only a few eigenvectors need to be extracted.

G Partial least squares regression

Partial least squares regression is explained in Martens and Næs (1989). In this appendix we show how to calibrate a PLS model and how to apply it.

G.1 Multivariate calibration

The bilinear model to calibrate is given by:

$$\begin{aligned} \mathbf{X}_0 &= \mathbf{T}\mathbf{P}' + \mathbf{E} \\ \mathbf{y}_0 &= \mathbf{T}\mathbf{q} + \mathbf{f} \end{aligned}$$

where subscript 0 denotes mean centered with unit variance. Next, specify the number of bilinear factors to extract. The following algorithm (NIPALS) is repeated for each bilinear factor:

1. Use the variability remaining in $\mathbf{y}_0$ to solve for the loading weights $\mathbf{w}$ in

   $$\mathbf{X}_0 = \mathbf{y}_0\mathbf{w}' + \mathbf{E}$$

   The solution is

   $$\hat{\mathbf{w}} = \mathbf{X}_0'\mathbf{y}_0$$

   Scale $\hat{\mathbf{w}}$ to length 1 by diving it by $\sqrt{\hat{\mathbf{w}}'\hat{\mathbf{w}}}$

2. Estimate the bilinear factor scores $\mathbf{t}$ by solving:

   $$\mathbf{X}_0 = \mathbf{t}\hat{\mathbf{w}}' + \mathbf{E}$$

   The least squares solution is given by

   $$\hat{\mathbf{t}} = \mathbf{X}_0\hat{\mathbf{w}}$$

3. Solve

   $$\mathbf{X}_0 = \hat{\mathbf{t}}\mathbf{p}' + \mathbf{E}$$

   for the bilinear factor loadings for $\mathbf{X}$, *i.e.*, the spectral loadings $\mathbf{p}$. The solution is:

   $$\hat{\mathbf{p}} = \frac{\mathbf{X}_0'\hat{\mathbf{t}}}{\hat{\mathbf{t}}'\hat{\mathbf{t}}}$$

4. Solve

   $$\mathbf{y}_0 = \hat{\mathbf{t}}q + \mathbf{f}$$

   for the bilinear factor loading for the $\mathbf{y}$-variable, *i.e.*, chemical loading q. Its solution is given by:

   $$\hat{q} = \frac{\mathbf{y}_0'\hat{\mathbf{t}}}{\hat{\mathbf{t}}'\hat{\mathbf{t}}}$$

5. Compute the remaining variance in $\mathbf{X}_0$ and $\mathbf{y}$ by the subtracting the effect of the current bilinear factor:

   $$\begin{aligned} \hat{\mathbf{E}} &= \mathbf{X}_0 - \hat{\mathbf{t}}\hat{\mathbf{p}}' \\ \hat{\mathbf{f}} &= \mathbf{y}_0 - \hat{q}\hat{\mathbf{t}} \end{aligned}$$

6. Set $\mathbf{X}_0 = \hat{\mathbf{E}}$ and $\mathbf{y}_0 = \hat{\mathbf{f}}$ and return to point 1 until all bilinear factors have been extracted.

G.2 Prediction

The response y can be predicted by means of:

$$\hat{\mathbf{y}} = \mathbf{1}\hat{b}_0 + \mathbf{X}\hat{\mathbf{b}}$$

where

$$\begin{aligned} \hat{\mathbf{b}} &= \hat{\mathbf{W}}(\hat{\mathbf{P}}'\hat{\mathbf{W}})^{-1}\hat{\mathbf{q}} \\ \hat{b}_0 &= \bar{y} - \bar{\mathbf{x}}'\hat{\mathbf{b}} \end{aligned}$$

and $\hat{\mathbf{W}}$ and $\hat{\mathbf{P}}$ are column matrices with vectors $\hat{\mathbf{w}}$ and $\hat{\mathbf{p}}$ respectively.

To explore the potential of nature to improve the quality of life

Wageningen Environmental Research
P.O. Box 47
6700 AB Wageningen
The Netherlands
T +31 (0) 317 48 07 00
www.wur.eu/environmental-research

Report 2853
ISSN 1566-7197

The mission of Wageningen University and Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 5,000 employees and 10,000 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.