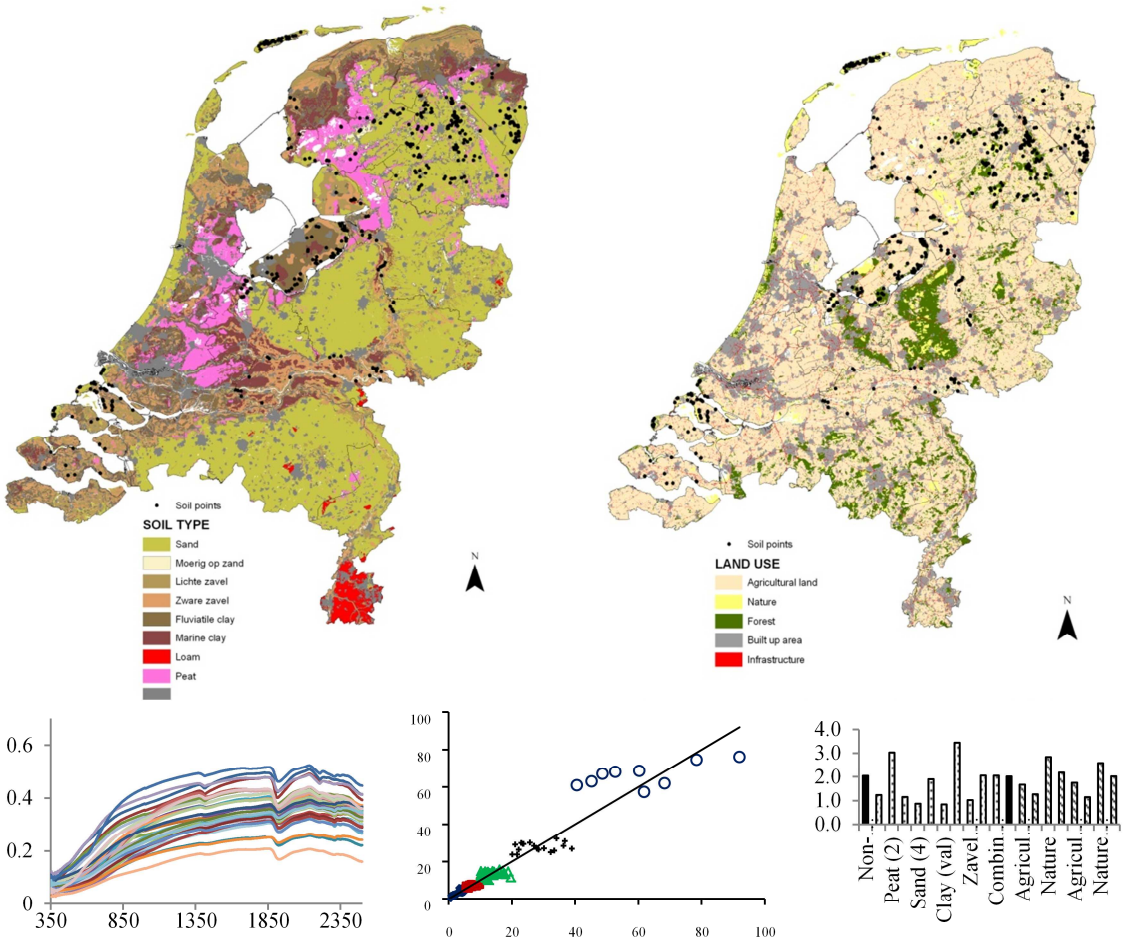
Centre for Geo-Information
Thesis GIRS – 2011–13

Improving soil properties estimations from spectral measurements based on spectral stratification techniques



Clifton R. Sabajo
June 2011

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Improving soil properties estimations from spectral measurements based on spectral stratification techniques

Clifton R. Sabajo

Registration number: 790605724090

Supervisor:
Dr. H. Bartholomeus

June 2011
Wageningen, The Netherlands

WAGENINGEN UNIVERSITY
WAGENINGEN UR

PREFACE

Studies end with research, or do they start with research? Anyway, this topic was chosen after long time of deliberation and consultation with supervisors. I wanted to do something which was not part of my previous background and research should have your attention from start till the end. Remote sensing is interesting as experienced from the Remote Sensing course, so why not a topic on Remote Sensing or something that is related to Remote Sensing. So I made this choice on soil spectroscopy, involving a very small part on lab work followed by lots of analysis.
Furthermore, soil science is interesting and the combination spectroscopy offered me two things: this research topic and my interest in one of my favourite courses from long time ago was awakened again: Soil Science. So, I also refreshed my soil science knowledge for this research. In the end, this research remained interesting throughout the time of conducting it.

To my supervisor: thanks for the opportunity, patience and supervision. This has been a very interesting part of my study. Sitting and thinking, at the start not knowing where to start, in the end not knowing where to end.
To the readers: enjoy this masterpiece.

Clif,

Wageningen, June 2011.

SUMMARY

In this study PLS regression is used to establish a relationship between soil reflectance spectra measured under laboratory conditions and three soil properties (Soil Organic Matter, N-total and pH). The objective of this study was to develop, implement and validate stratification methods to improve soil properties estimations. Three stratification methods were developed and tested. Two stratification methods, i.e. Lithological and Land Use stratification were developed with a Soil and a Land Use map of the Netherlands, and one stratification method was based on Wave-length Based Discriminant Analysis (WBDA) which only uses the spectral data of soil samples. Lithological stratification produced 4 soil clusters: clay, peat, zavel and sand, while Land Use stratification produced 3 clusters: agricultural land, forest and nature areas. WBDA stratification produced three clusters of a specific SOM range (A: <5%, B: 5-10% and C: >10%).

Prediction models were developed with the clustered data sets and the predictive performance of developed models was evaluated with the predictive statistics: R^2, RPD and RMSE.

I expected that the soil properties predictions would improve when they are based on stratified data sets. Stratification indeed resulted in improved predictions. However, the improvement was not achieved with all developed models. Improvement of SOM prediction was achieved with a clay, peat and zavel model but also when SOM content was predicted in agricultural, forest and nature samples with a non-clustered Land Use model. Improvement of Nt prediction was achieved with a peat, agricultural and nature model, and when Nt content in clay, zavel, and forest samples was predicted with a non-clustered Soil Type or Land Use model. Better pH prediction was achieved with a peat model and when pH was predicted in clay, sand, agricultural, forest and nature samples with a non-clustered Soil Type or Land Use model.

Stratification based on WBDA produced partly better predictions. This method has the potential to produce very accurate predictions but due to the risk of allocating spectra in wrong classes this potential is reduced drastically. This method needs to be improved further.

A result of the stratification methods is a decision tree which allows to select a model to predict a soil property of interest (SOM, Nt or pH) with a beforehand indicated accuracy.

Key words: PLSR, ParLeS, stratification, VNIRS, chemometrics, soils, The Netherlands

INDEX

List of figures

List of tables

1. Introduction

1.1 Context and Background

The upper soil horizon (A) contains useful information for farmers and decision makers. Chemical, physical, and mineralogical properties are essential for identifying the soil characteristics and are usually described in detail by soil scientists and then validated and elaborated by laboratory analyses (Ben-Dor *et al.*, 2008a). But when assessing soil quality many soil properties are involved which vary in space and time (Doran *et al.*, 1994; Doelman and Eijsackers, 2004; Cécillon *et al.*, 2009). Researchers acknowledge that in order to meet the amount of data required for statistical deduction in soil monitoring the current soil survey methods have some drawbacks, i.e. complicated, expensive, time and money consuming, destructive and often require use of many chemical reagents (Ben-Dor *et al.*, 2008a). Therefore, other time- and cost efficient methods for soil analysis could be an alternative to replace or complement the current analytical methods (Zornoza *et al.*, 2008).
Other analytical techniques would make rapid sampling and determination of, for example Soil Organic Carbon (SOC) values, at the field and regional level possible (Stevens *et al.*, 2008). One of those techniques is Visible and Near Infrared Spectroscopy (VNIRS) (Zornoza *et al.*, 2008). VNIRS provides an alternative to chemical analysis (Stevens *et al.*, 2008). The advantages of using NIR reflectance spectroscopy include the simple sample pre-treatment (sieving of soils), no use of chemical reagents, its non-destructive nature, and the fact that it is rapid, inexpensive and accurate for analysis (Zornoza *et al.*, 2008).

Using VNIRS has resulted in models which predict soil properties. The problem is that these models do not always predict the soil property correct and consequent.
Van Groenestijn (2009) suggested that a solution for improving model robustness(*) might be to create training sets based on a certain range of the soil property in question. One conclusion of her results (for Soil Organic Matter, SOM) was that by creating stratified training sets for a specified range of soil properties and using them for calibrating soil property prediction models, model robustness can be improved.
Spectral stratification methods (i.e. clustering) which can be used to improve soil property estimations are thus needed, which will be the focus of this research.

1.2 Problem definition

A wide range of soil properties such as the total iron, water content, mechanical compositions, carbonate, and organic carbon can be derived through reflectance spectroscopy under laboratory conditions if advanced analytical techniques such as Artificial Networks(*) and Partial Least Square Regression analysis are used (Udelhoven *et al.*, 2003; Ben-Dor *et al.*, 2009).

One of the main gaps in effective monitoring of soil quality with NIRS is the building of NIRS based regression models capable of assessing soil conditions at the regional scale across various soil types. Shepherd and Walsh (2002) proposed the use of soil spectral libraries as a tool for building risk-based approaches to soil evaluation. In the spectral library approach, soil properties are measured conventionally for a selection of soils representative of the diversity of the studied region, and then calibrated to soil reflectance spectra (Cécillon *et al.*, 2009).
But building NIR spectral libraries for soils raises several problems, one of which is that the quantitative analysis of soil spectra requires complicated statistical techniques to discern the response of soil attributes from spectral characteristics (Gomez *et al.*, 2008b).

(*) marked words are explained in appendix I.

Various authors have shown the effectiveness of NIR reflectance spectroscopy in estimating macro- and micronutrients in soils, physical characteristics and biochemical properties (Zornoza *et al.*, 2008). Most of all studies have been performed with SOM because of many properties and applications of SOM.
Soil organic matter is mainly concentrated in the top AO horizon. It is exposed to the sun's radiation which makes it a perfect property to be assessed by the Imaging Spectroscopy technology. Furthermore, soil organic matter is responsible for many soil chemical and physical properties and processes such as compaction, fertility, soil structure stability, and it constitutes one of the major resources in the global carbon cycle (Stevens *et al.*, 2008; Ben-Dor *et al.*, 2009).

Studies have shown that under laboratory conditions, VNIR spectroscopy coupled with multivariate calibration can accurately determine organic matter (Palacios-Orueta and Ustin, 1998; Salgó *et al.*, 1998; Reeves III *et al.*, 1999; Chang and Laird, 2002; Fidêncio *et al.*, 2002; Reeves *et al.*, 2002; Kooistra *et al.*, 2003; Gomez *et al.*, 2008b) and experiments have shown that it is feasible to use spectral indices derived from laboratory measurements to predict SOC in various soil types (Bartholomeus *et al.*, 2008). However, a large range in SOC is required for the calibration of the model, since extrapolation beyond the SOC range in the training datasets results in large errors (Bartholomeus *et al.*, 2008).
SOC can be spectrally measured with a reasonable accuracy level, depending on the type of instrument and environmental conditions (Stevens *et al.*, 2008). Stevens *et al.* (2008) studied the accuracy between portable and laboratory spectroscopy and concluded that portable spectroscopy is equivalent to laboratory spectroscopy when measuring SOC under specific surface conditions (low variation in moisture content of the soil surface, low roughness, absence of vegetation) and appropriate pre-treatments able to extract information from noisy spectra.

Other soil properties such as texture, metals, pH and EC have been studied to some extent by many researchers, e.g. different forms of C or N such as total-, organic- and inorganic- (carbonate) C and as well as organic- active- and biomass-C or N, but not mineralizable-N (Reeves III *et al.*, 1999; Reeves III and McCarty, 2001; Chang and Laird, 2002). Results with texture (sand, silt and clay) on a variety of data sets have similarly produced excellent results (Reeves III, 2010).

Given the fact that chemical and physical analyses of soils are expensive and time consuming and that soil spectral information can easily be gathered at a non-destructive way makes it attractive to use the soil spectral information for soil property predictions. But soil prediction models based on spectral measurements yield unsatisfactory results when they are applied for soil types which are not included in the calibration phase. Also highly variable data sets on which these models are based limit the implementation of these models and spectroscopy for estimations at local scale. Stratification of the spectral measurements before the development of the model can improve the prediction result.

This thesis will therefore focus on creating training sets based on a certain range of a soil property and use these for calibrating soil property prediction models to improve model robustness.

1.3 Using VNIRS for soil property estimation

Spectral reflectance characteristics of soils are the result of their physical and chemical properties and are influenced largely by the compositional nature of soils in which main components are inorganic solids, organic matter, air and water (van der Meer and de Jong, 2002). Soils may be identified by their reflectance characteristics (Condit, 1970) because specific wavelengths can describe the entire spectral curve by specific correlation with ''soil energies'' that represent the soil chromophore (Ben-Dor *et al.*, 2008b). A chromophore is a parameter or substance (chemical or physical) that significantly affects the shape and nature of a soil spectrum (Ben-Dor *et al.*, 2008b). A given soil sample consists of a variety of chromophores, which vary with the environmental conditions and the status of the five soil formation factors (climate, topography, parent material, organic matter, and time). Often the spectral signals related to a given chromophore overlap with the signals of other chromophores and thereby render the assessment of a signal's chromophore. Whereas the spectral reflectance of a given sample is the result of the entire chromophore interaction with the incident electromagnetic energy, the resulting spectral curve can serve as a footprint to the chromophore's overall existence in the examined matter. (Ben-Dor *et al.*, 2008b).

Soil chromophores can be divided into two categories: chemical and physical (Ben-Dor et al., 1999). Chemical chromophores are those materials that absorb incident radiation in discrete energy levels. Usually the absorption process appears on a reflectance spectrum as a feature whose position is attributed to specific chemical groups in various structural configurations (overtone, combination modes, and electronic processes). All features in the VNIR-SWIR spectral regions have a clearly identifiable physical basis. In soils, three major chemical chromophores can be roughly categorized as follows (Ben-Dor *et al.*, 2008b): (1) minerals (mostly clay, iron oxide, primary minerals-feldspar, salt, and hard to dissolve substances such as carbonates, phosphates), (2) organic matter (fresh and decomposing), and (3) water (solid, liquid, and gas phases).

Physical chromophores are properties that affect the overall spectral region and a particular waveband position, or in other words, do not relate to the chemical functional group. Examples of these are particle size variation and refraction indexes of a material that changes from one illumination condition to another (Ben-Dor *et al.*, 2008b).

Spectral signatures in relation to soil information

Spectral signatures of materials are defined by their reflectance, or absorbance, as a function of wavelength. Under controlled conditions, the signatures are due to electronic transitions of atoms and vibrational stretching and bending of structural groups of atoms that form molecules and crystals. The fundamental vibrations of most soil materials can be found in the mid-infrared (MIR) region, with weaker and broader overtones and combinations found in the near-infrared (NIR) region. For example, the C–H stretch fundamental absorption feature can be found at ~3.4 μm in the MIR, with overtones at ~1.7, 1.15, and 0.85 μm in NIR (Workman and Springsteen, 1998). Similarly, clay minerals have diagnostic overtone and combination absorption features in the NIR region: the O–H stretch 1st overtone at ~1.4 μm; the O–H stretch, H_2O bend combination at ~1.9 μm; the O–H stretch, metal–OH bend combinations at ~2.2–2.3 μm; and many minor absorption features (Hunt, 1977; Rencz, 1999). The secondary Fe-oxyhydroxides hematite and goethite are also easily identified in the VNIR region, with broad electronic absorptions at higher energy NIR wavelengths (0.7–1.0 μm) as well as in the VIS region (0.4–0.7 μm) giving rise to the distinctive red and yellow colours (Scheinost *et al.*, 1998; Scheinost and Schwertmann, 1999). Brown (2007) states that clays and to a lesser extent organic matter have well-recognized diffuse reflectance absorption features in the VNIR region related to their basic chemistry and mineralogy, which gives a reason to believe that combining local samples with a global soil-spectral library could improve on local calibration samples alone for the local prediction of SOC, clay and clay mineralogy.

In the NIR region, the radiation is absorbed by the different chemical bonds, such as C–H, N–H, S–H, C–O and O–H of any chemical compounds present in the sample. Moreover, the radiation is absorbed in accordance with the concentration of these compounds. As a consequence, NIR reflectance spectra basically contain information about the organic composition of a soil sample. Organic matter has distinct fingerprints that relate to several functional groups (e.g., carboxyl C-H, hydroxyl O-H, and amine N-H) in the organic compounds (Ben-Dor and Banin, 1995). The NIR spectrum results from the overtones and combinations of fundamental vibration bands for each of the chemical bonds, which are more strongly absorbed in the mid-infrared (MIR) region (Zornoza *et al.*, 2008).

Organic matter has a very important influence on the spectral reflectance properties of soils because amounts exceeding 2% are known to have a masking effect on spectral reflectance thus reducing the overall reflectivity of the soil and reducing (and sometimes completely obscuring) the diagnostic absorption features. Thus soils with a high (>20%) amount of organics appear dark throughout the 0.4 to 2.5 um range. In contrast, less decomposed soils have higher reflectance in the near-infrared region and enhanced absorption features (van der Meer and de Jong, 2002).

1.4 Difficulties using VNIRS for soil property estimations

Comprehensive studies over the past decade showed that the VIS (400–700 nm), NIR (700–1100 nm), and SWIR (1100–2500 nm) spectral regions serve as a powerful tools for recognizing soils qualitatively and quantitatively (Ben-Dor *et al.*, 2009). But still, as a consequence of overlapping bands, NIR information cannot be directly interpreted from the obtained spectra. NIR reflectance spectroscopy is based on the use of calibrations, coupled with chemometric techniques, which utilize absorbance at many wavelengths to predict particular properties of a sample (Batten (1998) cited in Zornoza *et al*.(2008)).

Partial Least Square (PLS) regression has been commonly used to build prediction models for SOC (Bartholomeus *et al.*, 2008). In the PLS regression approach, the full spectrum is used to establish a linear regression model where the significant information contained in the VNIR/SWIR spectra is concentrated in a few latent variables that are optimized to produce the best correlation with the desired property of interest (Gomez *et al.*, 2008a).
PLS reduces the NIR matrix to a few components, such as in a principal component analysis (PCA), but during the components extraction step in PLS, the data of the target parameter to be estimated is taken into account. The number of PLS components (so-called PLS-vectors) used is the ‘‘factor n’’ of the PLS regression (e.g. n-factor model). The first PLS-vectors are those which provide more information about the target parameter. In general terms, models with few factors (or latent variables) are preferred, because the higher the rank used, the higher the noise included (Zornoza *et al.*, 2008).

An important drawback of multivariate calibration models is the comparatively large number of training set samples required. Some complex materials may require hundreds or even thousands of samples to be tested and spectra measured before a suitable set of training samples can be identified (Roberts and Workman, 2004). In order to give the model the best chance to learn to recognize the information for the constituents of interest, it is important to train it, using samples that match the unknown as closely as possible (Roberts and Workman, 2004). Another reason to use a large number of samples for calibration is to allow more factors in the model. There must be enough samples/factors to account for the variability in the real samples that will be predicted as unknowns (Roberts and Workman, 2004).

Another major drawback is the complexity of the transfer of prediction models from one sensor to another. Sensor characteristics like wavelength position, bandwidth or number of bands, which requires new model calibrations for each sensor (Bartholomeus *et al.*, 2008).
The variance of the sample set used to produce NIR calibration equations also determines both robustness (i.e. applicability to a wide range of samples) and accuracy[(*)] for a particular application. A

training set that includes a wide variety of sample types and a large constituent range will allow a calibration model where a wider range of materials may be analysed, but with a resultant loss in accuracy. If the training set has a small variance in sample type and a narrow constituent range, the accuracy for analysed samples within the range is increased, but fewer unusual samples can be analysed with confidence using this approach. There is generally a trade-off between calibration performance and calibration robustness. The robust (i.e., applicable to a wide range of samples) calibration is used to detect outliers, while the dedicated calibration is used to accurately measure the constituent values of the normal samples (Roberts and Workman, 2004).

Enough samples should be used to model the data variability: the more data, the higher the confidence in the analysis and in the statistics (Duckworth in: Roberts and Workman, 2004). Another reason to use a large number of samples for calibration is to allow more factors in the model. If the training set has a small variance in sample type and a narrow constituent range, the accuracy for analysed samples within the range is increased , but fewer unusual samples can be analysed with the confidence using this approach. Thus, for quality control procedures, one may wish to have both calibrations available. The robust (i.e. applicable to a wide range of samples) calibration is used to detect outliers, while the dedicated one is used to accurately measure the constituent values of normal samples (Westerhaus in Roberts and Workman, 2004).

1.5 Soils of The Netherlands

The soil profile in The Netherlands is arbitrarily defined as the 0 – 120 cm layer of the sediment (excluding the litter layer). The different layers observed in the soil profile can have a geogenetic origin (different formations) or a pedogenetic origin, e.g. as a result of organic matter accumulation in the topsoil or transport of secondary Fe/Al-(hydr)oxides. In the latter case, these are termed horizons which are the basic properties used in further soil classification (van der Veer, 2006) These horizons result from various soil forming processes. These soil forming processes are in turn determined by a variety of soil forming factors, which are: parent material, climate and vegetation, topography and hydrology, time (soil age) and human impact. In the Netherlands, the human impact on soil forming factors as hydrology, topography and soil age, as well as on the geogenic layering of the soil profile has been extensive (van der Veer, 2006). Because of the tremendous impact mankind has had on the soil profile and soil properties, it should be considered one of the most important soil forming factors in the Netherlands.

In the Netherlands there are hardly any soils formed on consolidated rock. Roughly half of the mineral soils are formed in alluvial sediments, mostly marine clay and to a lesser extent river clay. The other half of the mineral soils are derived from aeolean(*) sediments. The latter are mostly loam-poor and slightly loamy cover sands; a small part consists of the transitional sediments between cover sands and loess, viz. the loamy sands and sandy loams, whereas most of the loess comes under the silty loam class (de Bakker *et al.*, 1989).

At the highest level of the classification, five orders are discerned which form 5 major soil types in The Netherlands (van der Veer, 2006): vague soils (soils that show very little soil formation), earth soils (thick A1-horizon), podzolic soils (podsolization), peat soils (strong accumulation of organic matter) and brick soils (illuviation).

Peat soils[1]: have been formed during the Holocene on both marine clays and older sandy deposits. With the current classification, peat soils are defined by having peaty material sensu lato (including ‘moerig’) over a depth of at least 40 cm within the first 80 cm profile. This means that these soils can have a substantial mineral topsoil (either sandy or clayey) and/or have sand or clay deeper in the profile. The sandy deeper soils can have a podzol B-horizon derived from earlier soil formation.

[1] : information on soil classification has been retrieved mainly from the PhD-thesis from Van der Veer (2006)

Peat soils are further subdivided on the basis of the organic matter content of the topsoil, which can be mineral (earthy peat soil) or organic (raw peat soils). The so-called raw peat soils are rare and found locally throughout the Netherlands, often in nature reserves. Earthy peat soils are common and often used as agricultural land. Here the mineral top layer can be of depositional origin (marine/fluviatile) or the result of levelling.

Podzol soils[1]: formed exclusively in the sandy Pleistocene sediments. They are defined by the presence of a clear podzol-B-horizon below a depth of 20 cm and lack of a thick anthropogenic A1-horizon (< 50cm).

Brick soils[1]: were all formed in loamy or clayey material, mainly loess and to lesser extent some old fluviatile(*) deposits. Brick soils are defined by the pre-eminence of a textural-B-horizon (brick layer) that starts within the first 80 cm of the profile. The textural-B-horizon (brick layer) that starts within the first 80 cm of the profile.

Earth soil[1]: have been formed mostly in the sandy Pleistocene deposits, but are also found on Holocene clay and sand deposits. Earth soils are characterized by a substantial mineral (humus rich to moderately humus poor) A1-horizon, which was formed by a biological degradation of organic material, and/or raising with organic material, heath sods and dredged mud.

Vague soils[1]: make up a considerable part of the Netherlands, and give its pedology a rather unique character. Vague soils are characterized by the lack of substantial soil formation and occur commonly in the younger Holocene deposits (both sand and clay) in the NL. Especially in recently reclaimed coastal areas and inland lakes, the time of soil formation is very restricted (700-30 years) and these soils have only developed a shallow A-horizon. Also, the inland and coastal dunes are of very restricted age and show very little horizon formation. As such, the composition of these soils will be largely comparable to that of the unaltered parent material.

Furthermore , the soil classification can be based on the type of parent material or on the texture and mineralogy. Based on the different types of parent material in the Netherlands the soils are classically grouped into five districts: sand, loess, peat, fluviatile and marine clay (de Bakker (1987) cited in van der Veer (2006)). This classification of parent materials is more closely related to the texture properties of the sediment.

A. Soil classification based on type of parent material[1]

Sand:
the parent material of the sand district consists mainly of aeolean deposits of the Late Pleistocene age, the so-called cover sand deposits. Much younger aeolean deposits include the inland and coastal dunes. The inland dunes are medium sized non-calcareous sands. The coastal dunes are restricted to the outermost coastal areas of the Netherlands. In contrast to the inland dunes, the coastal dunes can be calcareous, especially in the deeper profile (van der Veer, 2006).

Loess:
the parent material of the loess district (or loamy soils after Stiboka (1965)) consists of silty aeolean sediments and can texturally be classified as silty loam or sandy loam. The occurrence of the loess close to or at the surface is confined to the southern and south eastern part of The Netherlands and cover roughly 2% of the land surface (van der Veer, 2006).

Peat:
the profiles in the peat district are defined as having a high organic matter contents over at least 40 cm of the first 80 cm of the profile. (van der Veer, 2006). Since Roman times, much of the peat has been excavated. As a result, there are few profiles left that have peat over the length of the profile. The majority of peat lands have a non-organic sandy or clayey top layer, which is often of anthropogenic origin (van der Veer, 2006).

Fluviatile district:
the parent material in the fluviatile districts consists mainly of fluviatile clay and sand deposited by the Rhine, Meuse and their tributaries. These deposits can be calcareous or non-calcareous and often show a wide variety of grain size distributions ranging from coarse sands to heavy clay; also sandy deposits as well as sandy to heavy clays (van der Veer, 2006).

Marine district:
The parent material in the marine clay district consists of tidal, intertidal and perimarine deposits of the North Sea. The deposits are often calcareous (shell fragments) and their texture ranges mainly from fine sand to (heavy) clay. Especially in the South western parts of the Netherlands the layer consists of fine sand and sandy to silty clay. It is also found at the surface of the large polders around the central lake of the Netherlands. Large areas in the marine district consist of either lakes or coastal areas that were reclaimed (van der Veer, 2006).

B. Soil classification based on texture

Soil texture classes are based on the grain size composition of the mineral soil parts. Non-aeolean and aeolean deposits (both sand and other heavier material) are classified according to the clay or loam percentage (Kiestra, 2002).
Based on texture classes soil can be divided into the following categories:
A. based on clay percentage (Table 1)
B. aeolean deposits based on clay content (Table 2)

Table 1 Soil classes based on clay percentage

Clay (%)	**Name**	**Summarizing names**		
0 – 5	Clay-poor sand		Sand	Clay-poor material
5 – 8	Clayey sand			
8 – 12	Very light ‘zavel’	Light zavel	‘Zavel’	Clay-rich material (in relation to sand also referred to as clay)
12 – 17.5	Moderately light ‘zavel’			
17.5 – 25	Heavy ‘zavel’			
25 – 35	Light clay		Clay	
35 – 50	Moderately heavy clay	Heavy clay		
50 – 100	Very heavy clay			

From: Kiestra (2002).

Table 2 Categories of aeolean deposits based on clay content

Loam (%)	**Name**	**Summarizing names**	
0 – 10	Loam-poor sand		Sand
10 – 17.5	Slightly loamy sand	Loamy sand	
17.5 – 32.5	Very loamy sand		
32.5 – 50	Extremely loamy sand		
50 – 80	Sandy loam		Loam
80 – 100	Silty loam		

From: Kiestra (2002).

1.6 Stratification

Stratification is a procedure for subdividing the heterogeneous population into subpopulations which are internally homogeneous. Spectral stratification, i.e. stratification based on spectral characteristics can be used for classification of multispectral remotely sensed data (Padmanabhan *et al.*, 1980).

Stratification is for example applied in large scale geo-chemical surveys where it is rather natural choice to use a stratified sampling approach. Stratification of the target area aims at improving the estimations of the overall statistical parameters and leads at the same time to a better coverage of sample locations. Its success depends on effectively defining more or less homogeneous groups within the target area, and it therefore relies on a priori information about the sources of variation therein. But the choice for stratification is not only driven by statistical motivations, but also by the need to divide the target area into geologically meaningful groups (van der Veer, 2006).

Reasons for stratification to improve soil properties prediction have been proposed by different researchers. Malley and Williams (1997) (cited in Kooistra *et al*., 2003) suggested that the predictive capability of PLS might be improved by deriving calibration models for more homogeneous soil units, resulting in samples that are of a similar type rather than showing a wide range of values. Cobo *et al.* (2010) concluded that the soil properties which were well predicted (i.e. sand, clay, pH, C, N, Ca, Mg and CEC) depended on the success of regional calibration.
Bartholomeus *et al*. (2008) concluded that by including a priori knowledge on expected soil associations, mineral composition reclassification may increase the robustness of the prediction model and its applicability for extended geographical areas.
Also Cécillon *et al*. (2009) imply that some stratification must be applied in the spectral library approach. In this approach soil properties are measured conventionally for a selection of soils representative of the diversity of the studied region, and then calibrated to soil reflectance spectra (Cécillon *et al.*, 2009). When applying this approach Duckworth (in Roberts and Workman, 2004) also implies stratification when mentioning some criteria for training set design: training samples should be as similar as possible to unknowns and the constitute values in the training samples should be both larger and smaller than the expected values in unknown samples. By bracketing the range of concentration, the model will give the most accurate answer possible.

Different methods of stratification can be applied, such as lithological, regional or land use stratification. Another method of stratification is known as ‘Wavelength-based Discriminant analysis’ (WBDA) (Roberts and Workman, 2004). In Discriminant Analysis the assumption is made that samples closest together in wavelength space are very similar to one another. Conversely, samples far apart in wavelength space are thought to be part of separate spectral groups.
The best allocation of samples is to have one-half of the samples at low X value and the other half at high X value. In NIR, samples with spectra associated with low and high values would be selected. Spectra with the most extreme spectra relative to the average spectrum can be selected. This technique will pick the lows and the highs in a very simple sample population, but runs the risk of including spectral ‘mistakes’ and underrepresenting the middle of the population. However, this problem could be solved by augmenting the sample set with some randomly selected samples. The ‘alikeness’ of one test spectrum, or series of spectra, to a reference spectrum can be determined by calculating a point-by-point correlation between absorbance data for each test and reference spectrum (Correlating matching (Roberts and Workman, 2004). The same principles apply to searching a database of spectra to determine if a spectrum of an unknown is present in the database (i.e. identify the unknown by matching it’s spectrum to a known spectrum). This technique is known as spectral searching and is routinely used in other areas of optical spectroscopy (e.g. mid-infra-red) to aid in the identification of unknown compounds (Roberts and Workman, 2004).

1.7 Research Objective and Research questions

Visible Near Infrared Spectroscopy is very attractive for soil property predictions but soil prediction models based on spectral measurements yield unsatisfactory results when they are applied for highly variable data sets. Different researchers have proposed to apply stratification to obtain calibration models for homogenous soil units. Stratification may increase the robustness of the prediction model and make these models applicable for larger geographical areas (Bartholomeus *et al.*, 2008). By creating training samples that are similar to the samples that are to be predicted, prediction models may give accurate answers.

The main objective of this study is to determine if stratification methods will result in improved estimations of soil properties using VNIR spectra.

The specific objectives are to:
1. develop (a) stratification method(s)
2. implement the stratification method(s)
3. validate the developed method(s)

Research questions:

1. Can stratification, based on external data sources, improve the estimations of soil properties from VNIR spectra?

2. Can stratification, based on spectral data only, improve estimations of soil properties from VNIR spectra?

The expectation is that prediction models will perform better if suitable clusters and ranges for the soil property of interest are selected.

2. Methodology

2.1 Study area and available data

A soil spectral library of 263 samples was already available which was compiled during previous research by Van Groenestijn (2009). This library contained the spectral signatures of the soil samples, the *x*- and *y*-coordinates, 15 chemical and 1 physical soil property. This soil spectral library was expanded with an additional 312 spectra of soil samples with the same chemical and physical properties, which were measured in this research bringing the total soil samples in the library to 575 soil samples which were collected from different locations in The Netherlands. Not all *x*- and *y*-coordinates were available: from 426 soil samples the locations where these samples were collected were known (figure 1 and table 3).

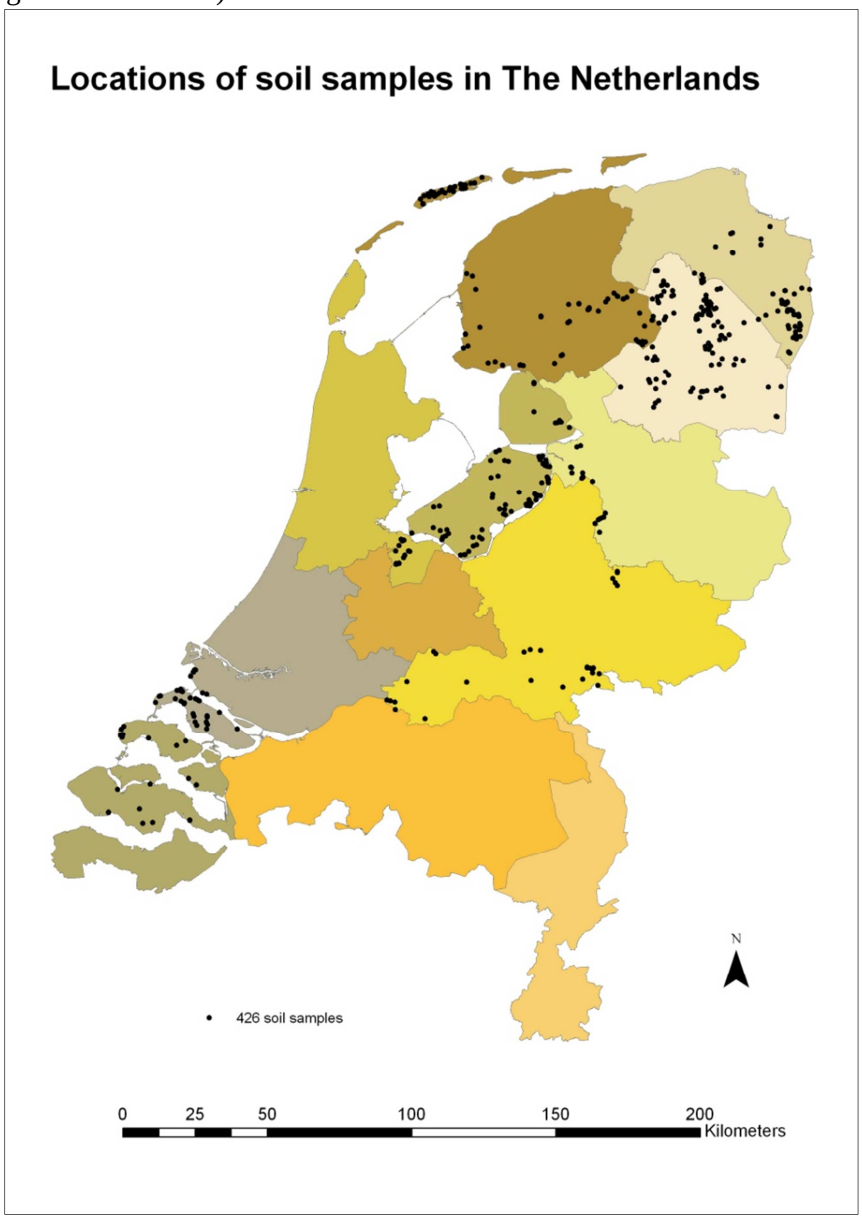


Figure 1. Locations of collected soil samples
From 426 samples the coordinates were known so that their locations could be plotted in the map. Soil samples and *x/y*-coordinates taken by P.A. Slim (Alterra) in the period 2009-2010.

Table 3 Overview of the available data

	Previous research	Additional (this research)	Total for this research
Samples with *x/y*-coordinates	220	206	426
Samples without *x/y*-coordinates	43	106	149
Total	263	312	575

2.2 Sample preparation and spectral measurements

The soil samples used in this research were acquired as follows: 10 samples were taken using a chisel at a depth of max. 10 cm, put in a bag, shaken, and a subsample with the volume and size of a volumetric ring was taken. These samples were all analysed for 16 different soil properties in the laboratory by conventional (extraction) methods.

For the spectrometric measurements basic sample preparation consisted of grinding and sieving 20 grams of homogeneous soil using a 2-mm mesh. To remove the effect of moisture, the soil samples were dried for (for 12 h at 40–45 °C) (Vasques *et al.*, 2008).
The soil samples were scanned with an ASD Fieldspec Pro FR in combination with an ASD contact probe, measuring reflectance in the wavelength range of 350– 2500 nm at 1-nm intervals.
The soil samples were scanned four times at an angle of 90° using the contact probe. Each measurement was recorded as the average of 4 readings to come to an estimation of a homogeneous sample. This average measurement was used for modelling (Vasques *et al.*, 2008).

Spectralon reference measurements, collected prior to the first scan and every 10 samples, were needed to make final conversion to spectral reflectance by dividing the radiance spectra of the soil samples by that of the white Spectralon plate.

2.3 Data analysis

Descriptive statistics

Table 4 shows the available data for this research. Three soil properties, SOM –N-total (Nt) – pH, were used in this research.

Table 4 Descriptive statistics of the soil properties of the soil samples.

	Unit	N	Mean	SD	Min	Max
Nt**	g/kg	575	3.32	3.91	0.05	26.9
Pt	mg/kg	575	575.82	591.90	19	5930
K	mg/kg	575	61.32	77.15	3.8	1066
Na	mg/kg	575	118.59	686.68	1.2	10454
Mg	mg/kg	575	105.25	138.70	3.81	1309
N_NH4	mg/kg	575	5.58	6.98	0.5	110
N(NO3+NO2)	mg/kg	575	5.34	8.12	-0.1	68.2
Ns	mg/kg	575	27.26	20.79	2	166
PPO4	mg/kg	575	1.62	3.34	0	53.1
Al	mg/kg	575	878.76	732.65	2.86	4820
Ca	mg/kg	575	8020.73	11426.92	1.37	85842
S	mg/kg	575	69.99	428.51	1.3	9756
Cl	mg/kg	575	158.47	1046.23	3	16109
SOM**	%	575	11.37	14.13	0.24	95.8
pH**	-	575	6.27	1.61	3.62	9.53
Moisture	%	575	25.62	18.15	0.86	90.2

The soil properties marked with ** (Nt, SOM and pH) were used in this thesis for analysis. Descriptive statistics are based on 575 samples.

Multivariate analysis

The whole dataset was split into a calibration set (2/3) and an independent validation set (1/3). The dataset was sorted from lowest to highest value (of soil property) and every third sample and its data was moved to a separate file for use as a validation set (after Dunn *et al*. (2002)). An extra rule was added i.e. that the range of the soil properties of the validation set should be within the range of the calibration set to avoid extrapolation of the model beyond the range for which it was calibrated. Calibrations were developed on the remaining samples.
The Levene's test for equality of variances and the Student's *t*-test of equality of means were performed between the calibration and validation sets to make sure there was a representative validation set (Vasques *et al.*, 2008).

Analysis and statistics were performed with SPSSv.15/17 and ParLeS version 3.1. ParLeS is software for chemometric analysis of spectroscopic measurements (Viscarra Rossel, 2008). ParLeS was used to calculate multivariate calibration models by PLSR with leave-1-out cross validation to establish relationships between the NIR spectra and the reference data from physical and chemical analyses (Viscarra Rossel, 2008). To reduce non-linearities the reflectance spectra were first transformed to log 1/R. No pre-processing of the data was applied.

The coefficient of determination (R^2), the root mean square error (RMSE), and the residual prediction deviation (RPD) were used to evaluate the models (Vasques *et al.*, 2008; Zornoza *et al.*, 2008). Refer to equations I to III for the model evaluation parameters:

A. Coefficient of determination (R^2):

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad \text{(eq. I)}$$

B. Root Means Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{n}} \quad \text{(eq. II)}$$

Where $\hat{y}$ = predicted values; $\bar{y}$ = mean of observed values; y = observed values; n = number of predicted/observed values with i = 1, 2, …, n.

C. Residual Prediction Deviation (RPD):

$$\text{RPD} = \frac{\text{standard deviation of analysed data}}{\text{RMSE}} \quad \text{(eq. III)}$$

Eq. I, II: Vasques *et al.* (2008)
Eq. III: Zornoza *et al.* (2008)

Based on the values of the coefficient of determination (R^2) and Residual Prediction Deviation (RPD) (RPD) prediction results were ranked in 3 categories (A, B or C) following Chang *et al.* (2001) (table 5), with category A as the best of all three:

Table 5 Model evaluation parameters.

Category	R^2	RPD
A	0.8 – 1.00	> 2.0
B	0.50 – 0.80	1.4 – 2.0
C	< 0.50	< 1.4

After Chang *et al.* (2001).

2.4 Stratification methods

Research question 1:

Can stratification, based on external data sources, improve the estimations of soil properties from VNIR spectra?

Two external data sources, the Land Use map of The Netherlands (LGN-5) and the Soil map 2006 of The Netherlands, were used as a basis for the stratification of the soil samples. With these external data sources the soil samples were clustered based on Land Use or Soil Type.

Lithological stratification (based on Soil Type)
The samples were divided according to type of parent material in the following 4 clusters (table7):
1. Clay (fluviatile and marine clay)
2. Peat
3. Zavel
4. Sand

Land Use stratification
Based on the Land Use map the samples were divided into the following 3 clusters (table 7):
1. Agricultural land
2. Forest
3. Nature area

Per cluster the soil samples were divided in calibration and validation sets. Soil prediction models were based on these clusters and the performance of these models was compared to each other and to 4 reference models, which were models based on non-clustered data.
R^2, RMSE and RPD (see table 5) were used to evaluate performance of the prediction models based on the defined clusters.

Table 6 Stratification with external data themes

	Theme	
	Soil type	**Land Use**
Cluster	1. Clay	1. Agricultural land
	2. Peat	2. Forest
	3. Zavel	3. Nature area
	4. Sand	

In figure 3 an overview is given on how the stratification was applied to the available data, using the two external data sets.

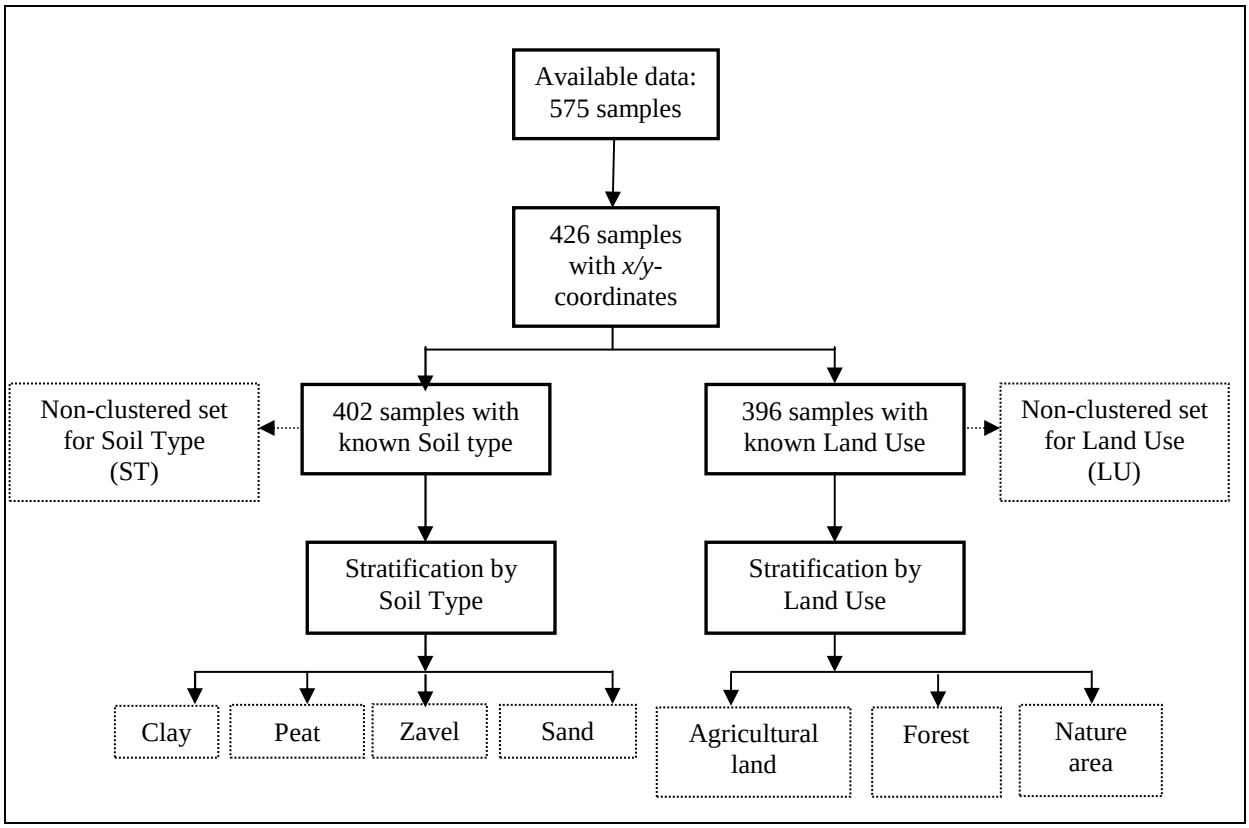


Figure 2. Creating stratified and non-stratified sets.
Stratified data sets were created based on soil type and land use. Dashed boxes indicate the created 2 non-stratified and 7 stratified sets. Stratified and non-stratified sets were split in calibration and validation sets.

Research question 2:

Can stratification, based on spectral data only, improve estimations of soil properties from VNIR spectra?

The soil samples were divided in groups that have spectral signatures which look similar. This method is known as Wavelength-based discriminant analysis (WBDA) (Roberts and Workman, 2004). The assumption is that the concentration of a certain soil property results in the a certain spectral signature. Soil samples which have the same content should have, based on this assumption, the same spectral signature.
This method was applied for SOM. The available information are the spectral signatures of reference samples and the SOM content of these samples. If the spectral signature of an unknown sample looks like the spectral signature of a reference signature this unknown sample might have the same SOM content.

Four reference spectra were selected representing soils with 4 different SOM content: one with a SOM content of 5%, one with a SOM content of 10.1%, one with a SOM content of 20% and one sample with a SOM content of 62.4% (figure 4). The ratio of the reflectance between these 4 different soils was determined. Soils with the largest difference in SOM content should have the largest ratio e.g. ratio spectral signature $SOM_{5\%}/SOM_{62.4\%}$ > ratio spectral signature $SOM_{10.1\%}/SOM_{20\%}$. Figure 5 shows the ratio between these reference spectra which will be the basis for classifying all spectra. Although the relation between these ratio's is exponential (figure 5) linear interpolations were applied to determine the ranges of the ratio to classify the remaining soil spectra (figure 6). Based on the linear interpolations the samples were classified according to the ranges given in table 7.

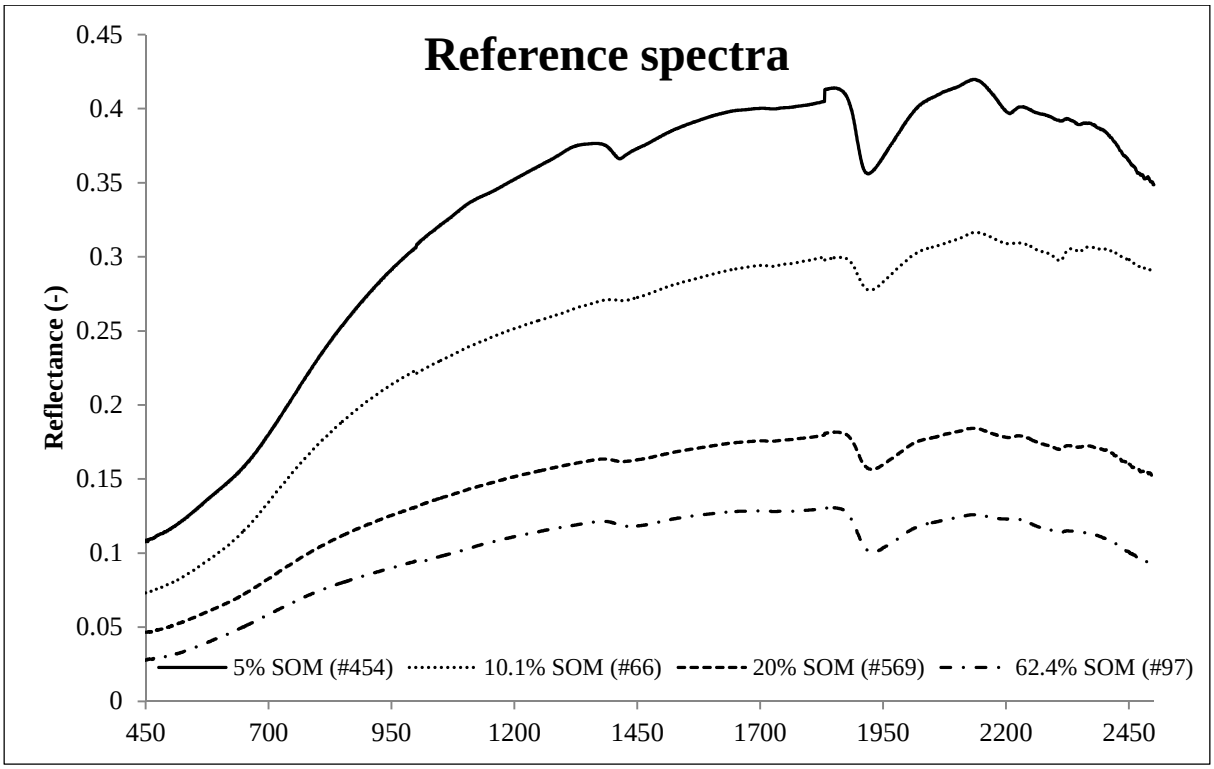
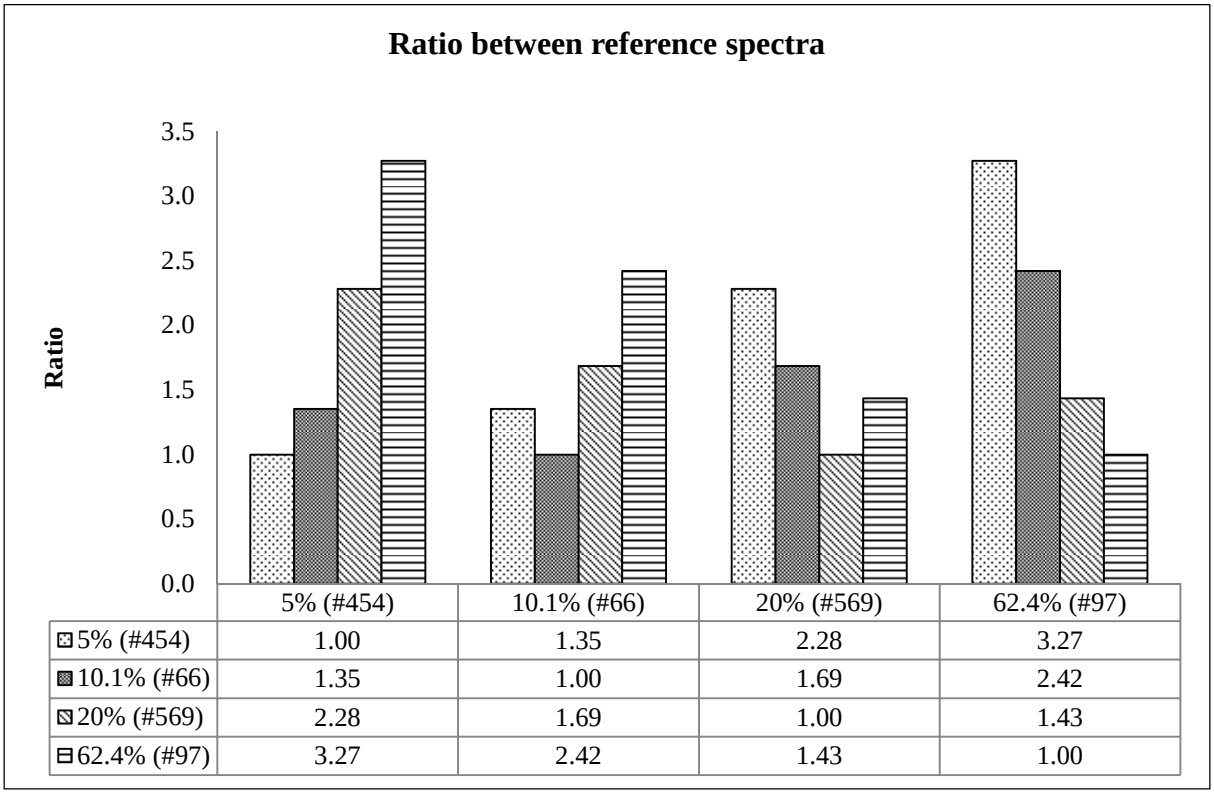


Figure 3. Four reference spectra representing four different SOM content.
The numbers #454, #66, #569 and #97 are the samples that were selected as reference spectra.



	5% (#454)	10.1% (#66)	20% (#569)	62.4% (#97)
5% (#454)	1.00	1.35	2.28	3.27
10.1% (#66)	1.35	1.00	1.69	2.42
20% (#569)	2.28	1.69	1.00	1.43
62.4% (#97)	3.27	2.42	1.43	1.00

Figure 4. Ratio between reference spectra.

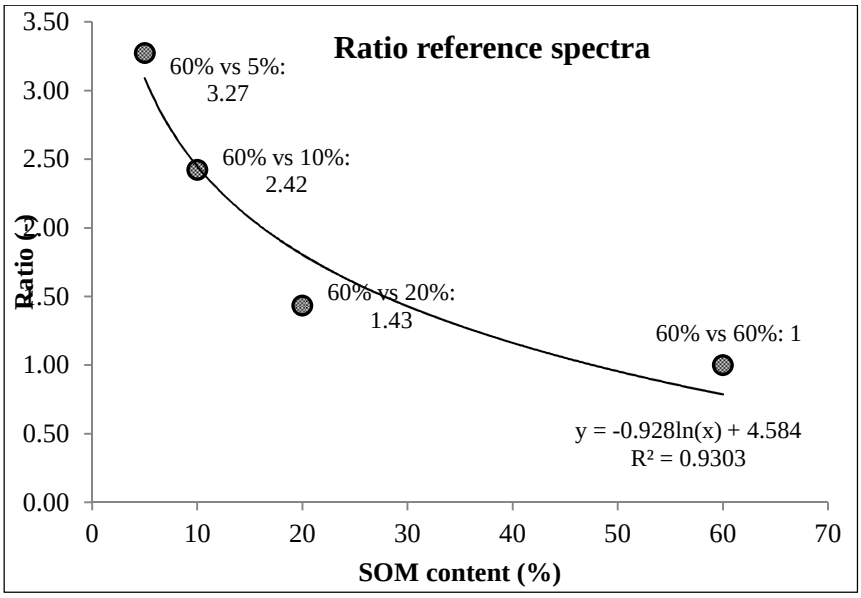


Figure 5. Relation between the ratio of the reference spectra.
All ratios in the picture are compared to the spectrum of the sample with 60% SOM content.

For all the samples the ratio between their spectral signatures and $sample_{SOM=62.4\%}$ was calculated and based on the ratio the samples were split in four classes following the criteria set in table 7, assuming that they would be within the range of SOM specified by the ratio.

Next, the samples per class were divided into a calibration set and a validation set. Every third sample was taken separate into a validation set. The remaining samples formed the calibration set.

Table 7 Classification based on ratios of spectral signatures and reference spectral signature.

Ratio	Cluster	Expected SOM content
> 2.85	A	< 5%
2.85 - 1.93	B	5 - 10%
1.21 - 1.93	C	10- 20%
< 1.21	D	> 20%

Ratio as compared to spectral signature with SOM content of 62.4%.

Figure 6 shows the methodology of this stratification method. Soil prediction models were based on these clusters and the performance of these models was compared to each other and to a reference model, i.e. the non-clustered model. R^2, RMSE(CV) and RPD (see table 5) were used to evaluate the performance of the prediction model based on these defined clusters.

Based on the results of R^2, RMSE and RPD, the effect of this clustering method was evaluated.
Comparing the results based on stratification or based on external sources or based on this method should give the answer of the best method of clustering.

Figure 7 shows a flow chart of the research methodology showing the procedures which were followed to answer the research questions and to arrive to a synthesis.

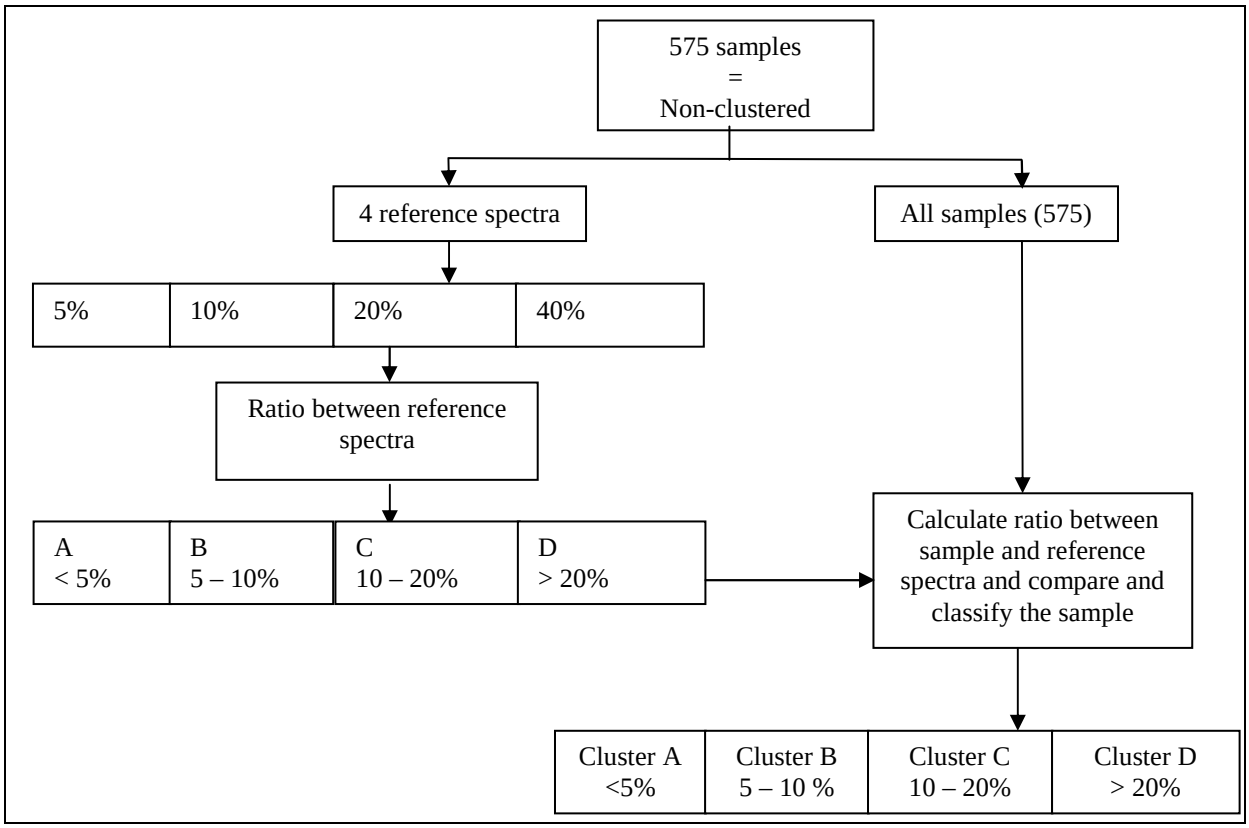


Figure 6. Stratification method based on ‘Wavelength-based Discriminant Analysis’.

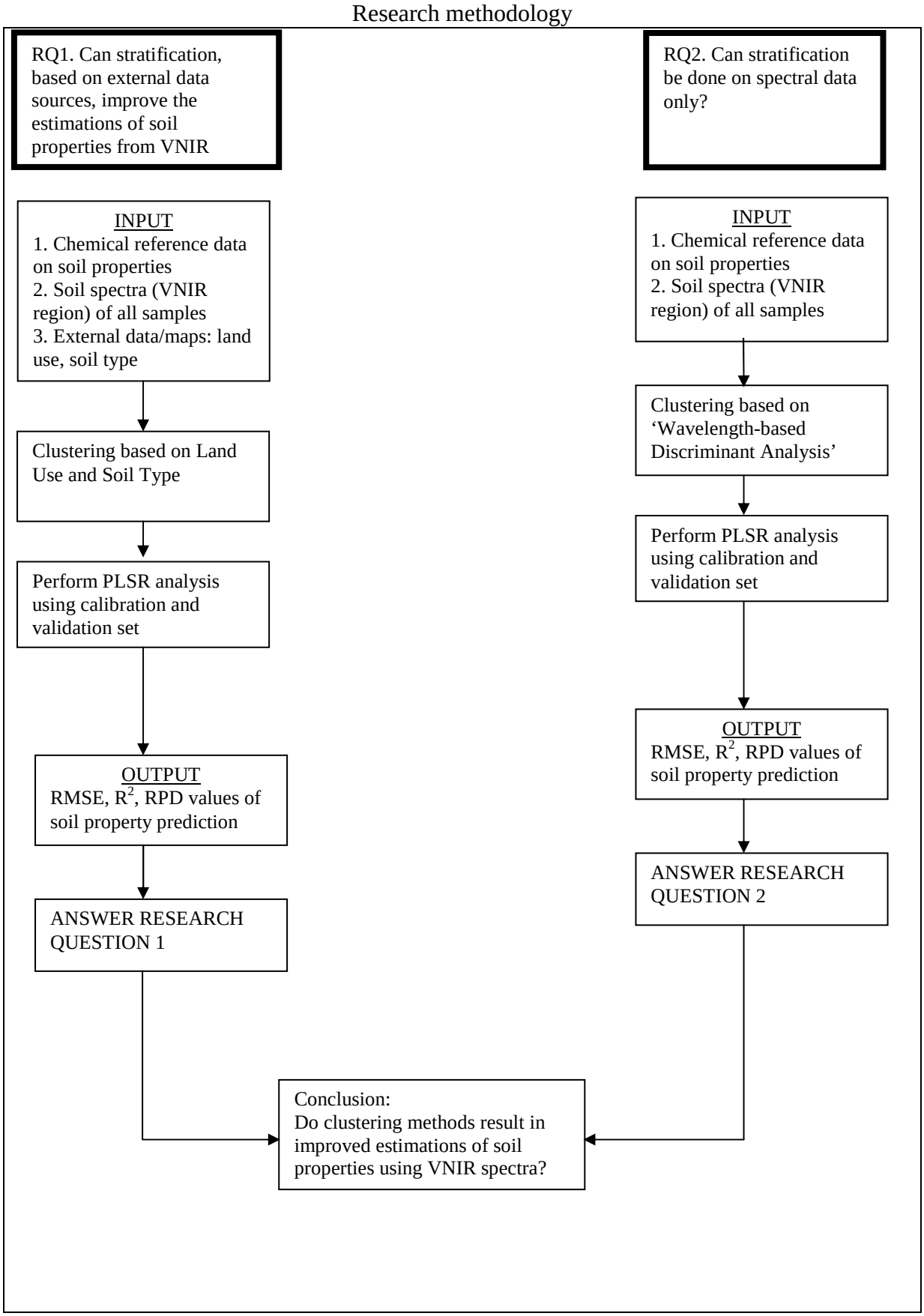


Figure 7. Research methodology flow chart

3. Results and discussion

3.1 Data preparation for stratification

Data for Soil Type and Land use stratification

The available data set was divided into two clusters for use in prediction models: one for Soil type and one for Land Use. For each soil property (SOM, Nt and pH) calibration and validation sets were created out of the total data set. These non-clustered data sets were the reference data sets to which the results of the prediction based on the stratified data sets were compared.

The tables 8 – 11 give an overview of the different clusters and the size of these clusters.
In table 8 the two non-clustered sets are shown. Based on a Soil map the soil type from which the soil samples originated was deducted and based on the Land Use map the land use type from which the soil samples originated was also deducted. The soil samples from which the soil types were derived, were grouped in one set forming the non-clustered set for Soil Type (ST – total of 402 samples) and the different samples from which the Land Use types were known, were grouped in one non-clustered set for Land Use (LU – a total of 396 samples).

Table 8 Non-clustered data sets for Soil type and for Land Use

Cluster	Soil property	N calibration	N validation	Total	Composition
Non-clustered set for Soil Type (ST)	**SOM**	296	133	402	Calibration and validation set contain only samples from which the soil type was known.
	Nt	269	133	402	
	pH	269	133	402	
Non-clustered set for Land Use (LU)	**SOM**	265	131	396	Calibration and validation set contain only samples from which the land use was known.
	Nt	265	131	396	
	pH	265	131	396	

Table 9 contains the calibration and validation sets based on Soil Type stratification. In this stratification, soil samples belonging to ‘water’ or ‘built up’ (n=3), loam, and ‘moerig op zand’ were excluded from clustering because they were too few (loam, n = 3), or difficult to categorize in one of the 4 soil type clusters (e.g. ‘moerig op zand’, n = 18), or not relevant to cluster (e.g. water or built up areas, n = 3). Map accuracy or the accuracy of the position recordings could also have placed or excluded some points in or from a cluster. Based on the soil type the remaining 402 samples were divided in 4 soil clusters: a clay, a peat, a ‘zavel’ and a sand cluster. The sand cluster was the largest cluster containing more than half of all the soil samples (N_{sand} = 219). Even though the clay, peat and zavel clusters were small, calibration and validation sets were created according to the ratio 2/3 calibration and 1/3 validation set. Figure 8 shows the Soil map of The Netherlands and the distribution of the points over the different soil clusters.

Table 9 Clusters based on lithological stratification (Soil Type)

Soil property	Cluster	N calibration	N validation	Total
SOM Nt pH	Clay	34	17	51
	Peat	38	18	56
	Zavel	51	25	76
	Sand	146	73	219
	Total	**269**	**133**	**402**

Table 10 contains the data sets based on Land Use stratification. Excluded from this set were soil samples collected on water, infrastructure and building (n = 30). This could also be due to map inaccuracy or GPS inaccuracy. The remaining soil samples were grouped in 3 different clusters: Agricultural land (n= 60), Forest (n = 152) and Nature area (n = 184). The Nature cluster was the largest with 184 samples followed by the forest cluster with 152 samples. Calibration and validation

sets for the different soil properties were formed according to the 2/3 calibration and 1/3 validation criterion.

Figure 9 shows the Land Use map which formed the basis of this stratification procedure.

Table 10 Clusters based on Land Use stratification

Soil property	Cluster	N calibration	N validation	Total
SOM **Nt** **pH**	Agricultural land	40	20	60
	Forest	102	50	152
	Nature	123	61	184
	Total	**265**	**131**	**396**

Stratification based on spectral data

Table 11 contains the calibration and validation data based on stratification by 'Wavelength-based Discriminant Analysis' (or 'Spectral Similarity'). This was only applied for SOM.

Table 11 Clusters based on stratification by Wavelength-based Discriminant analysis

Soil property	Cluster	SOM range	N calibration	N validation	Total
SOM	A	< 5%	133	83	216
	B	5 - 10%	137	96	233
	C	> 10 %	114	12	126
	Total		**384**	**191**	**575**

The calibration samples are grouped based on a priori knowledge of the SOM content. Two third of the whole data set was assigned to one of the three clusters. The remaining samples were the validation samples which were assigned to a cluster based on only the spectral signature and by arranging them according to a ratio between the spectral signature of the soil samples to the spectral reflectance of the reference spectrum. Prediction models were fitted for each cluster with its calibration and corresponding validation set.

The validation and calibration sets in table 12 are based on a priori knowledge of the SOM content. Based on a priori knowledge of the SOM content the validation points were divided into 5 clusters which have the same range as the calibration strata. The a priori knowledge resulted in an 'ideal' validation set in which each soil samples was placed at once correctly in a the right stratum. Prediction models were fitted for each cluster with its calibration and corresponding validation set.

Table 12 *Ideal* clusters

Soil property	Cluster	SOM range	N calibration	N validation	Total
SOM	I*	< 5%	132	66	198
	II*	5 - 10%	137	68	205
	III*	10 -20 %	59	30	89
	IV*	20 - 40%	38	18	56
	V*	> 40%	18	9	27
	Total		**384**	**191**	**575**

In this data set the validation set are formed based on prior knowledge of the SOM content of the validation sample.

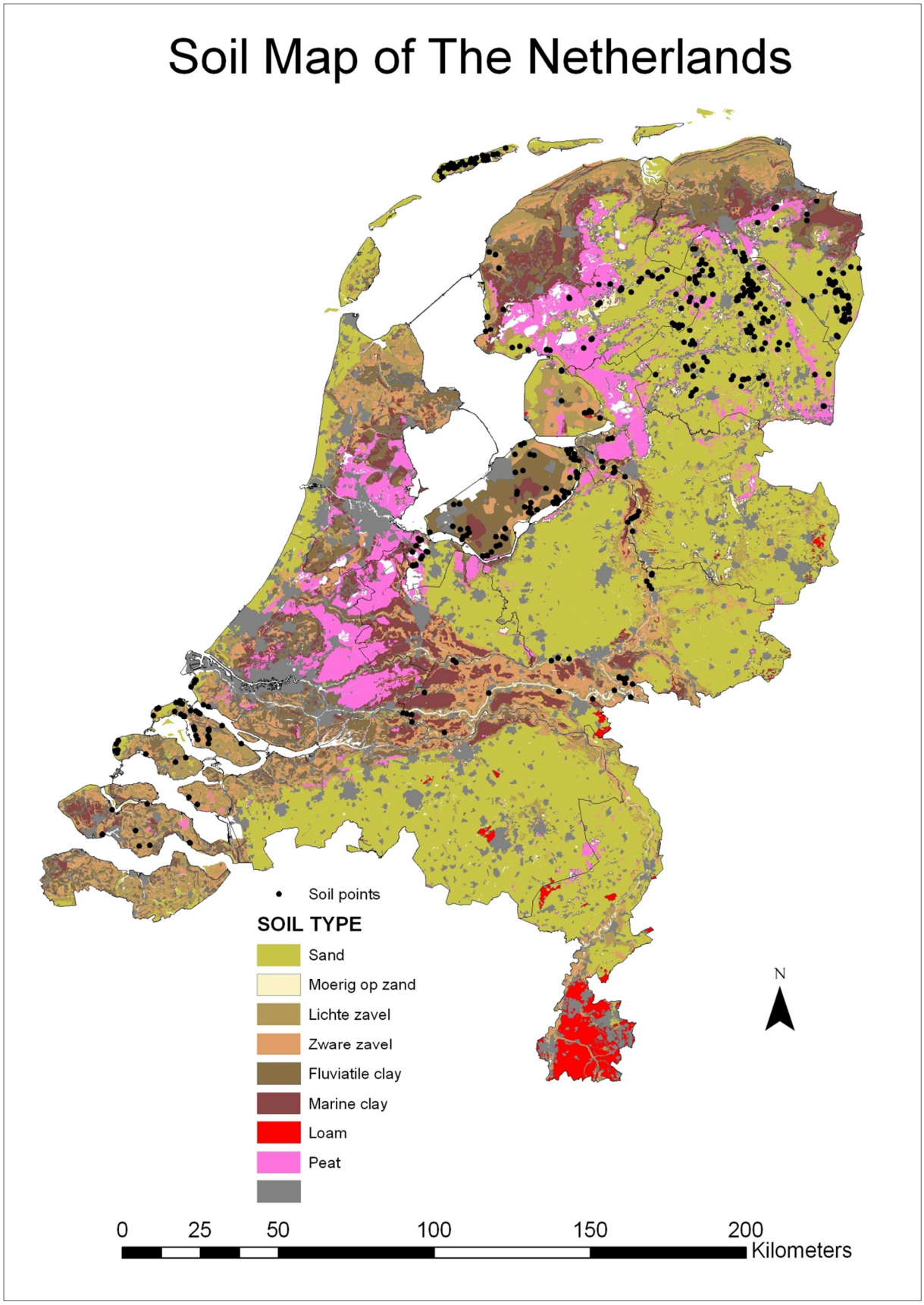


Figure 8. Soil map of The Netherlands with 402 soil points
© WUR-Alterra CGI (2006)

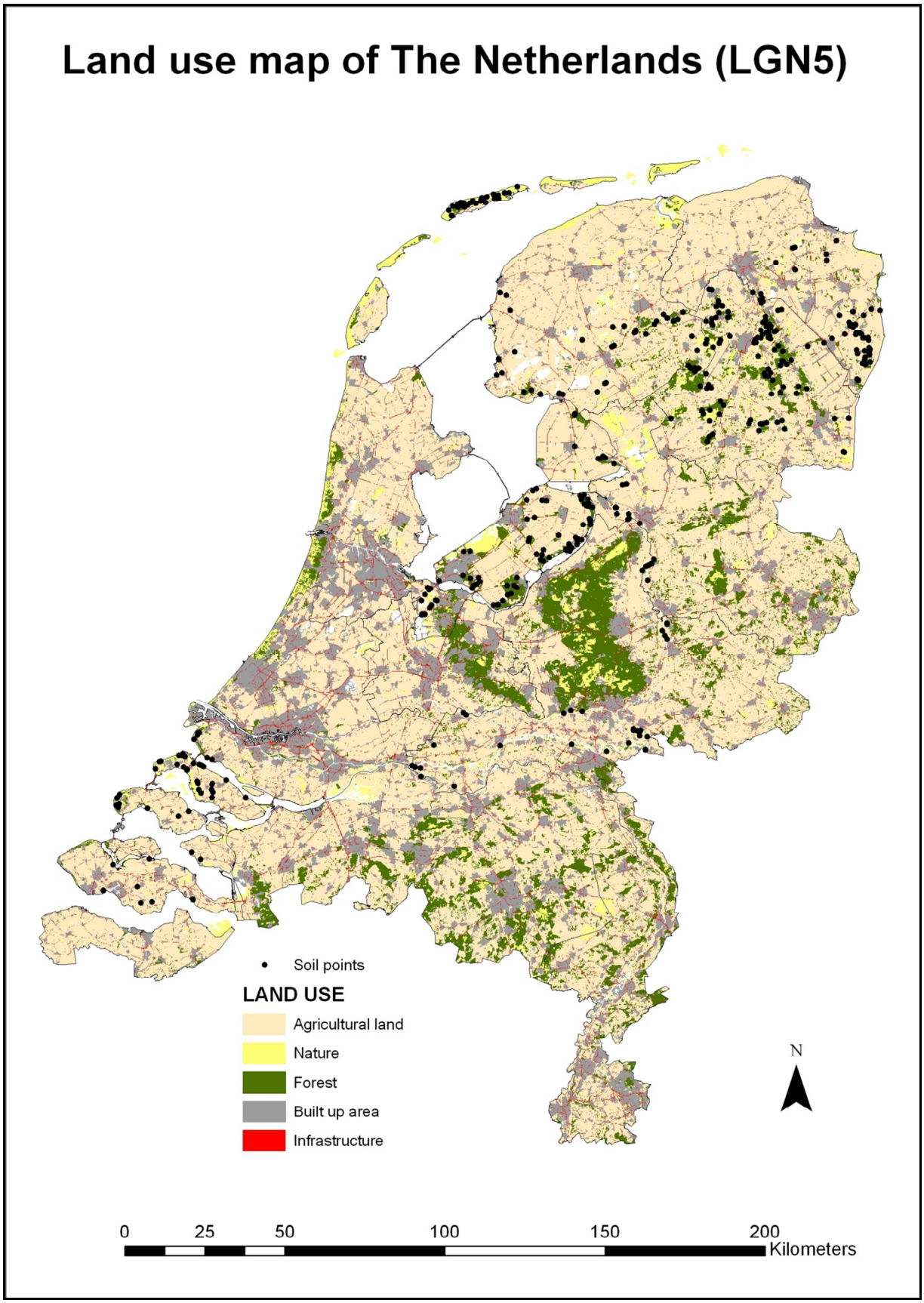


Figure 9. Land Use map of The Netherlands (LGN5) with 396 soil points
© Geodesk (CGI)

Normality tests

All non-clustered and clustered sets and their corresponding calibration and validation sets were tested for normality. Most of the data sets were not normally distributed as can be seen in table 13, still no adjustments or transformations were made to manipulate or force the data into a normal distribution. Further analysis were still performed using these data sets. Appendix III contains the statistics for this test for normality (Kolmogorov-Smirnov & Shapiro-Wilk).

Table 13 Summary of the tests of normality.

SOM (%)	Set	Normal distribution	Nt (g/kg)	Set	Normal distribution	pH	Set	Normal distribution	SOM (%)		Normal distribution
CLAY	Cal	no	CLAY	Cal	no	CLAY	Cal	no	Non-clustered	All	no
	Val	no		Val	no		Val	no		Cal	no
	All	no		All	no		All	no		Val	no
PEAT	Cal	no	PEAT	Cal	no	PEAT	Cal	no	Cat A (<5%)	All	no
	Val	no		Val	yes		Val	yes		Cal	no
	All	no		All	no		All	yes		Val	no
ZAVEL	Cal	no	ZAVEL	Cal	no	ZAVEL	Cal	no	Cat B (5 - 10%)	All	no
	Val	yes		Val	no		Val	no		Cal	no
	All	no		All	no		All	no		Val	no
SAND	Cal	no	SAND	Cal	no	SAND	Cal	no	Cat C (> 10%)	All	no
	Val	no		Val	no		Val	no		Cal	no
	All	no		All	no		All	no		Val	no
AGRIC	Cal	no	AGRIC	Cal	no	AGRIC	Cal	no			
	Val	no		Val	no		Val	no			
	All	no		All	no		All	no			
FOREST	Cal	no	FOREST	Cal	no	FOREST	Cal	no			
	Val	no		Val	no		Val	no			
	All	no		All	no		All	no			
NATURE	Cal	no	NATURE	Cal	no	NATURE	Cal	no			
	Val	no		Val	no		Val	no			
	All	no		All	no		All	no			
NON-CLUSTERED SOIL TYPE	Cal	no	NON-CLUSTERED SOIL TYPE	Cal	no	NON-CLUSTERED SOIL TYPE	Cal	no			
	Val	no		Val	no		Val	no			
	All	no		All	no		All	no			
NON-CLUSTERED LAND USE	Cal	no	NON-CLUSTERED LAND USE	Cal	no	NON-CLUSTERED LAND USE	Cal	no			
	Val	no		Val	no		Val	no			
	All	no		All	no		All	no			

Clay, peat, zavel, sand, agric., forest, nature, non-clustered Soil Type and non-clustered Land Use are the clusters formed based on land use or soil type. Cal= calibration set; val = validation set; All = whole set (i.e. calibration and validation set).

Levene's test and student *t*-test

The Levene's test indicated homogeneity of variance for all calibration and validation sets of the Soil and the Land Use clusters. For clusters formed with the Wavelength based discriminant analysis method the Levene's test indicated unequal variances for the calibration and the validation set of the A and B cluster. Comparison between the mean values of calibration and validation sets also showed significant differences between these them indicating that the calibration and validation sets did not appropriately represent the population under study. All the calibration and validation sets of the Soil and Land Use clusters were representative. The results of the Levene's test and the student's *t*-test are summarized in table 14 and 15. The statistics of these tests can be found in appendix IV.

Table 14 Summary of the Levene's test for calibration and validation clusters (Soil Type and Land use stratification).

Representative calibration and validation sets?			
	SOM	**Nt**	**pH**
Non-clustered Soil Type (ST)	yes	yes	yes
Clay	yes	yes	yes
Peat	yes	yes	yes
Zavel	yes	yes	yes
Sand	yes	yes	yes
Non-clustered Land Use (LU)	yes	yes	yes
Agricultural Land	yes	yes	yes
Forest	yes	yes	yes
Nature	yes	yes	yes

Table 15 Summary of the Levene's test for calibration and validation sets (Wave-length Based Discriminant Analysis).

Representative calibration and validation sets?	
Non-clustered	yes
Cat A (<5%)	no
Cat B (5-10%)	no
Cat C (>10%)	yes

From the results in table 13 – 15 it is clear that the different clusters (whole set, calibration and validation) were not all normally distributed, or were not representative for the population, criteria which are requirements for PLSR analysis. Ideally the data should be transformed (e.g. log-transformation or another transformation) to fit the data into a normal distribution. PLSR analysis would then be performed on the transformed data. In this research this is deliberately avoided because a log-transformation of the data also involves a log-transformation of the units of the soil property and the predictions seem to have high accuracies while in fact they could be large errors which become clear when a back transformation of the error is applied. In order to avoid back transformation of the predictions and prediction errors no data transformation was applied in this research.

SOM

As expected the peat cluster has the highest mean SOM content (table 25, appendix II) but the statistics also show that there is a minimum SOM content of 2.60% which is too small for a sample originating from a peat soil. The clay, zavel and sand cluster have average mean SOM contents of 10.91%, 8.21% and 8.89% respectively. The peat cluster has a large range for SOM content as the maximum SOM content in this cluster is 95.80%. The sand cluster also has a large range for the SOM content from 0.34% - 72.40% SOM.

Of the Land Use clusters the nature cluster has an average higher SOM content than the agriculture and forest cluster (table 26). The nature and forest clusters contain the samples with the highest SOM content (94.70% and 95.80%), probably due to the litter layer which was collected with the sample. The range of the SOM content in the forest and nature cluster is also larger than agricultural land. Agricultural land is probably more homogeneous due to human activity causing less variance in the data.

Cluster based formed by means of the Wave-length based Discriminant analysis show contradictory results (table 27). First of all this stratification method was supposed to result in 4 clusters, but the criteria set for this method resulted in three clusters. The fourth cluster (D >20% SOM) returned only 1 sample. This sample was combined with cluster C resulting in a cluster representing samples with SOM content more than 10%. The descriptive statistics show that the validation set of class A cluster (<5%) was supposed to contain only samples with SOM content smaller than 5% but in the maximum SOM content in this cluster is 52.70%. This sample is definitely misplaced with this method. Still, the mean SOM content of this cluster less than 5%. This also occurred in class B cluster (5% - 10 % SOM) which has a mean SOM content of 10.15% and a maximum SOM value of 92%. In this cluster the minimum SOM content was supposed to be 5%, still the statistics show that there is a minimum SOM content of 2.30%. Samples were correctly placed in cluster C (>10% SOM). In cluster C minimum SOM content of 10.20% was observed.
This result shows that samples can be misplaced when this method is applied. Misplacements of spectra can be attributed to the reference spectra which were arbitrarily chosen from the whole data set. There was no check or other external spectra available to compare the reference spectra with. The ratios of the reference spectra and the ratio of the reference spectra with other spectra are thus influenced by the quality of the chosen reference spectra. Another reason for misplacements of spectra can be influence of other soil particles such as soil mineralogy.

N-total (Nt)

The N-total content of the soil samples varied from 0.16 – 25.40 g/kg (table 28). The different soil clusters have different mean Nt content indicating indeed differences between the different soil types. Peat has the highest mean Nt content of 8.30 g/kg. This is probably related to the high SOM content of this cluster (table 24). Sand has the lowest mean Nt content (2.20 g/kg) which is probably correct for this soil type, although a maximum Nt content of 20.20 g/kg was observed in this cluster. The clay and zavle cluster have a smaller range (12.49 and 8.33 g/kg respectively) compared to the peat and sand cluster which have a range of 24.93 and 20.04 g/kg respectively.
The Land Use clusters (table 29) show that nature areas have the highest mean Nt content (4.04 g/kg), followed by agricultural land (3.79 g/kg) and forest (2.73 g/kg). The difference in mean Nt content between these clusters is not large. The range of the Nt content is the smallest in the agricultural land cluster (0.26 – 16.00 g/kg), indicating a more homogeneous distribution of Nitrogen in agricultural soils. Forest and nature areas have a higher range in the Nt content indicating that these soil samples are more diverse in Nt content. These clusters also have maximum Nt contents of 25.40 g/kg and 21.20 g/kg for forest and nature areas respectively, probably caused by litter.

pH

The peat and sand cluster have the lowest mean pH values (table 30) of 5.31 and 5.43 respectively. This is related to the properties of peat, soils having high SOM and N content are more acidic. Clay and zavel have about the same mean pH value, but the zavel cluster has a smaller range in pH values, indicating that these samples are more homogeneous, contrary to the sand cluster which has a higher range and probably is more heterogeneous. The clay, peat and sand samples are also very diverse as can be inferred from the range of these clusters: for clay the range is from 4.16 – 8.66, for peat the range is form 3.62 – 7.63 and pH values in the sand cluster range from 3.70 – 9.11.
When the Land Use clusters are examined (table 31) we see that the mean pH values do not differ much between the Land Use clusters: from 5.97 (nature) to 6.16 (agricultural land). The ranges are

also not much different from each other: pH values of the agricultural land range from 3.89 – 8.32, the pH of forest range from 3.70 – 8.51 and the pH values in nature areas range from 3.62 – 9.11.
Agricultural land has the highest average pH (6.16) which can be attributed to human intervention in maintaining the pH at a certain level required for agricultural production. But still, minimum pH values of 3.89 are also observed for agricultural land which is very low for this type of land use.

3.2 Stratification based on external data sources

Stratification was based on two external data sources: a Soil Type and a Land Use map.

A. Lithological stratification (Soil Type)

Soil Organic Matter

In table 16 the different PLSR models that were developed for the prediction of Soil Organic Matter (SOM) are shown.

Table 16 Validation statistics for SOM for soil clusters.

SOM				**Validation statistics**		
	N cal	**N val**	**# factors**	**R^2**	**RMSE (%)**	**RPD**
Clay (1)	34	17	4	0.81	**3.59**	1.84
Peat (2)	38	18	9	0.82	**9.99**	2.27
Zavel (3)	51	25	3	0.31	**2.89**	1.18
Sand (4)	146	73	7	0.48	8.24	1.36
Combined (1+2+3+4)	269	133	-	0.72	**7.34**	1.88
Non-clustered Soil Type (ST)	296	133	12	0.63	8.41	1.64
Clay (validation)		17	12	0.27	6.86	0.96
Peat (validation)		18	12	0.80	11.43	1.99
Zavel (validation)		25	12	0.39	4.78	0.71
Sand (validation)		73	12	0.46	8.83	1.27
Combined validation sets		133	-	0.63	8.41	1.64

Bold **RMSE's** indicate that the accuracy has improved compared to the prediction of SOM with the non-clustered model or with the non-clustered model tested with clustered validation sets.

PLSR models were developed for the 4 soil clusters after leave-one-out cross validation using the stratified calibration samples. A 4-factor model was developed for the clay cluster; for the peat, zavel and sand cluster a 9-, 3- and 7-factor model were developed respectively.
A 12-factor PLSR model was developed using 296 non-clustered samples (ST). This non-clustered model was tested with an independent set of 133 non-clustered samples. This model was also tested using independent clustered samples of the clay, peat, zavel and sand cluster.

Of all the 4 soil models, the prediction of SOM with the 9-factor peat model gave the highest R^2 (=0.82) which is higher than the R^2 of the predictions of the non-clustered samples with the non-clustered model (R^2 = 0.63) or with the clustered peat validation samples (R^2 = 0.80). The other models resulted in lower R^2 than the non-clustered model.

The highest R^2 was reached with the non-clustered PLSR model when it was tested with the stratified peat samples (R^2 = 0.80). This validation also yielded a RPD of 1.99, but the accuracy was also the lowest of all predictions as the RMSE = 11.43%.

The highest accuracy (RMSE) of predictions was achieved with the 3-factor zavel PLSR model (RMSE = 2.89%) and the 4-factor clay PLSR model (RMSE = 3.59%). The lowest accuracy was reached with the peat model (RMSE = 9.99%).

The results show that when e.g. SOM content in clay samples is predicted with the clay model the accuracy is higher than when it is predicted with the non-clustered model: SOM predictions in clay sample with the 4-factor clay model are predicted with an accuracy (RMSE) of 3.86%, but when SOM content in clay samples is predicted with the 12-factor non-clustered model the accuracy of the predictions is 6.86%. Prediction of SOM content in peat, zavel and sand samples with their corresponding models results in better accuracies than when SOM content in these samples is predicted with the non-clustered model.

In figure 10 the calibration and validation results of the soil type clusters are shown. The calibration and validations of the clustered models (bottom) are compared to the calibrations and validations of the non-clustered models (top). Remarkable are the negative predictions of SOM which occur during calibration and validation. Calibration of the non-clustered model resulted in 30 negative predictions; in the validation of this model there were 19 negative predictions. Negative predictions were observed for the sand model, which had most negative predictions (12 negative predictions during calibration and 3 negative predictions after validation of the model), the peat (2 negative predictions during calibration and 1 negative prediction after validating the validation) and clay which had only 1 negative prediction in the calibration of the model. These negative predictions reduce the predictive ability of the models.
Negative predictions do not occur for zavel (during both calibration and validation of the model) and clay (only for validation of the model). Negative predictions occur for small values of SOM content (SOM < 3.04% in the calibration set or SOM < 1.11% in the validation model, but there was one sample in the peat model which has a SOM content of 12.4% but was predicted as -7.42%). The sand and peat model also have the highest RMSE (table 16) indicating less accurate predictions, especially for sand this could have been the cause of the negative predictions. Another cause can also be attributed to the range of the different soils clusters. The peat and sand cluster have the largest range in SOM content, while the zavel and clay clusters have smaller ranges. The developed models probably cannot deal with a too much large range of the data, as was the case with the peat and sand samples.

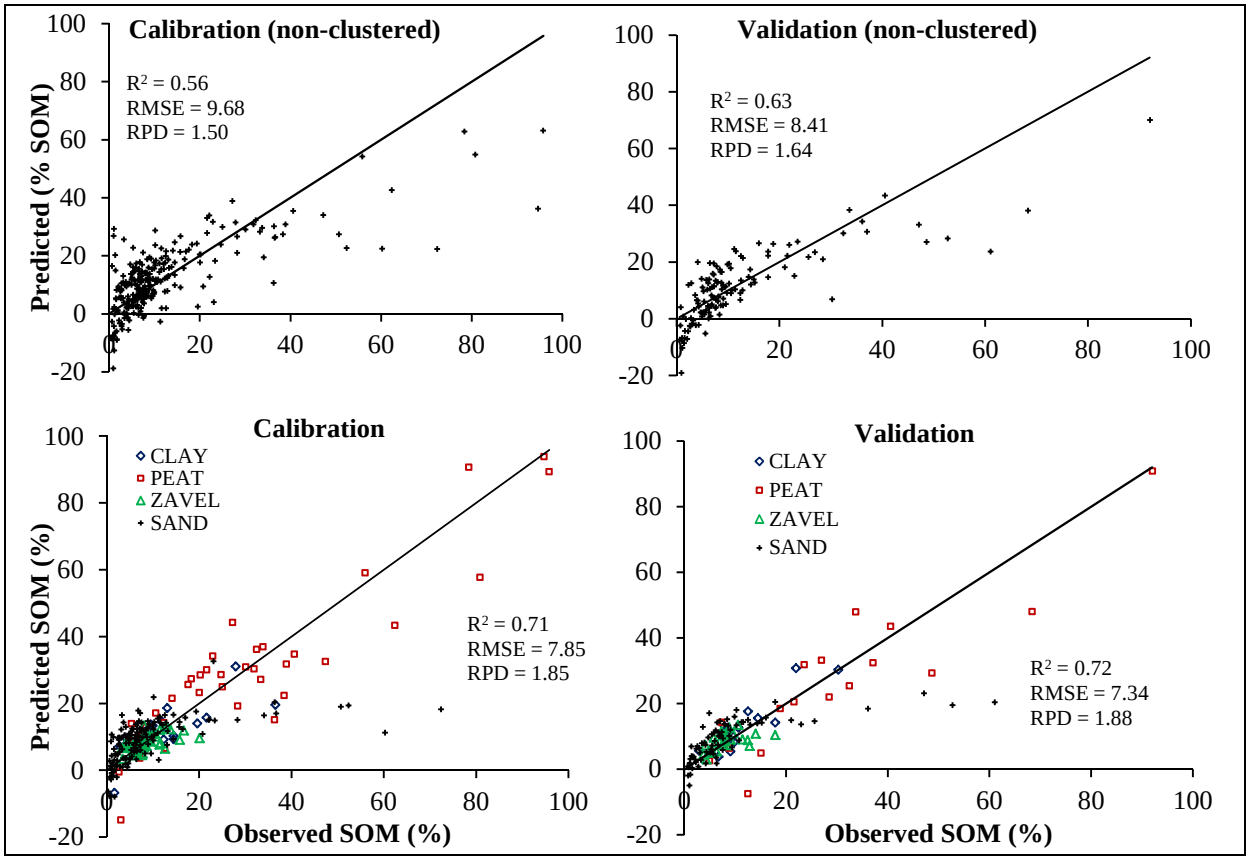


Figure 10. Calibration and validation results for combined soil clusters for SOM (Lithological stratification). The lines represent the 1:1 line.

Of all the models, the zavel model performed the best because it required the fewest factors, 3 PLSR factors which is very parsimonious, did not result in negative predictions and the predictions were achieved with the highest accuracy (RMSE = 2.89%).

N-total

In table 17 the different PLSR models that were developed for the prediction of N-total (Nt) are shown.

Table 17 Validation statistics for Nt for the soil clusters.

N-total				**Validation statistics**		
	N cal	**N val**	**# factors**	**R^2**	**RMSE (g/kg)**	**RPD**
Clay (1)	34	17	4	0.76	1.23	1.99
Peat (2)	38	18	8	0.63	**3.03**	1.69
Zavel (3)	51	25	3	0.29	1.15	1.21
Sand (4)	146	73	7	0.67	**0.87**	1.68
Combined (1+2+3+4)	269	133	-	0.74	**1.90**	1.95
Non-clustered Soil Type (ST)	269	133	15	0.70	2.05	1.81
Clay (validation)		17	15	0.92	0.84	2.93
Peat (validation)		18	15	0.52	3.40	1.48
Zavel (validation)		25	15	0.66	1.02	1.36
Sand (validation)		73	15	0.60	2.05	1.53
Combined validation sets		133	-	0.70	2.05	1.81

Bold **RMSE's** indicate that the accuracy has improved compared to the prediction of Nt with the non-clustered model or with the non-clustered model tested with clustered validation sets.

Four models with 4 – 8 factors were developed for the soil clusters. All these soil cluster models produced R^2 values ranging from 0.29 - 0.76. Predictions of Nt in zavel samples with the zavel model had the lowest R^2 (0.29) of all models.
A 9-factor PLSR model was developed with the 269 non-clustered samples, and tested with 133 independent non-clustered samples. The R^2 (0.70) and RPD (1.81) of the Nt predictions of these samples are higher than the predictions of Nt with the clustered models, except for the predictions of the clay model.

The accuracy of the prediction with the clay and zavel model was not better than prediction with the non-clustered. The accuracy of predictions with the non-clustered model was 0.84 g/kg for clay and 1.02 g/kg for peat (cf. $RMSE_{clay}$ = 1.23 and $RMSE_{zavel}$ = 1.15). Better prediction accuracies were reached when predicting Nt in peat samples with the peat model (RMSE of 3.03 g/kg) or in sand samples with the sand model (RMSE = 0.87 g/kg).

The clustered models that were developed were quite parsimonious: they were developed with 3 – 8 factors, compared to the non-clustered model which was developed with 15 factors. The sand model could be a good model because it produces more accurate predictions (lowest RMSE of 0.85 g/kg).

Figure 11 shows the results of the lithological stratification. The calibration and validations of the clustered models (bottom) are compared to the calibrations and validations of the non-clustered models (top). Also here negative predictions can also be observed. In the calibration of the non-clustered model there were 33 negative predictions and in the validation of the non-clustered model there were 12 negative predictions. These negative prediction occurred for Nt contents smaller than 2.35 g/kg (calibration) or smaller than 2.30 g/kg (validation).
Negative predictions were observed for the sand model, which had most negative predictions (18 negative predictions during calibration and 9 negative predictions after validation of the model). Negative predictions were also observed for the clay and peat model but only for the calibrations of these models: 1 negative prediction for each model. The negative predictions of the model occur for Nt contents smaller than 0.75 g/kg (calibration) or smaller than 1.03 g/kg (validation).

The sand model probably cannot deal with the large range in this cluster.

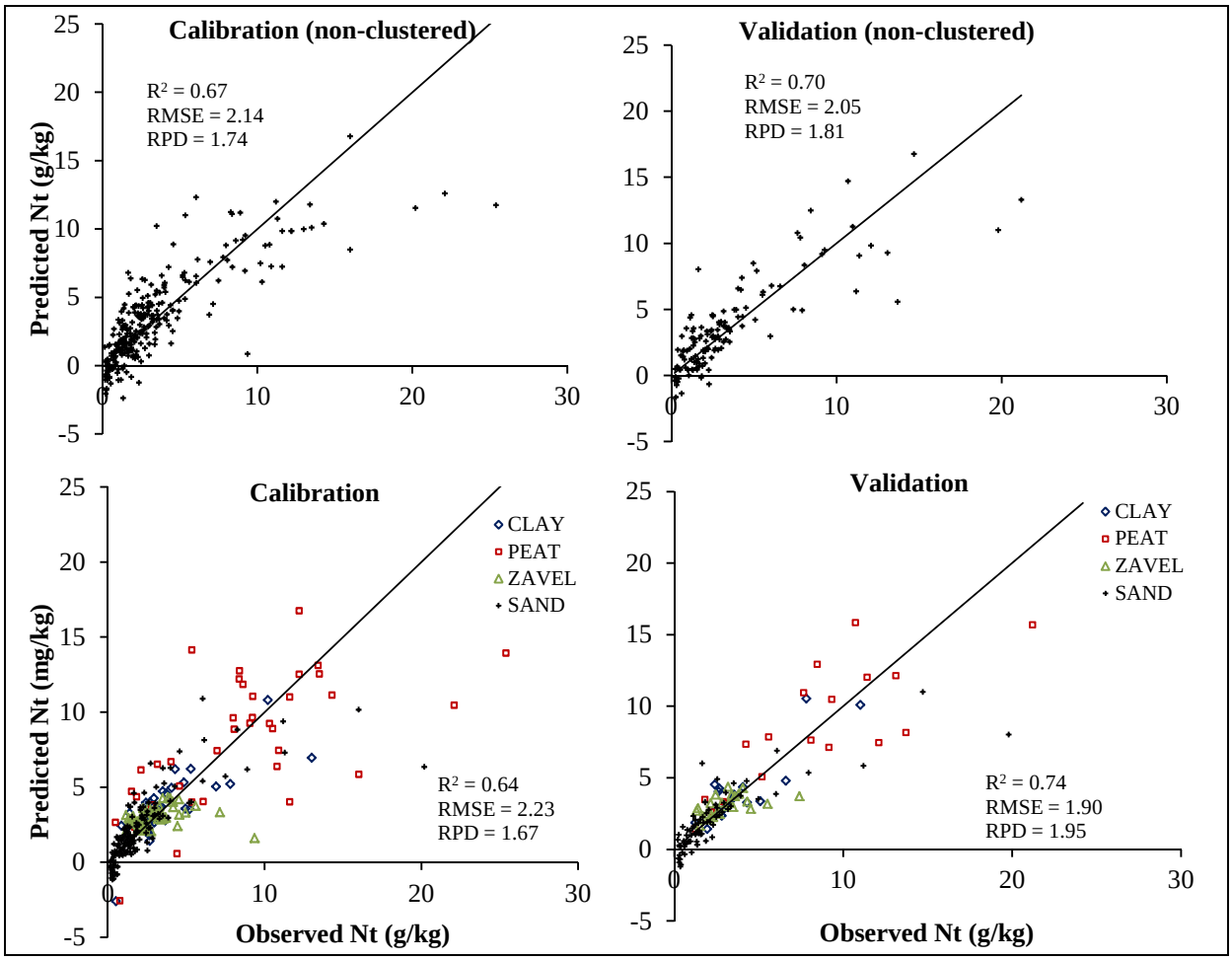


Figure 11. Calibration and validation results for combined soil clusters for Nt (Lithological stratification). The lines represent the 1:1 line.

pH

In table 18 the different PLSR models that were developed for the prediction of pH are shown.

Table 18 Validation statistics for pH for the soil clusters.

pH				**Validation statistics**		
	N cal	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**
Clay (1)	34	17	5	0.67	0.59	1.62
Peat (2)	38	18	6	0.44	**0.54**	1.23
Zavel (3)	51	25	17	0.23	**0.46**	1.00
Sand (4)	146	73	12	0.78	0.78	2.01
Combined (1+2+3+4)	269	133	-	0.85	0.68	2.50
Non-clustered Soil Type	269	133	16	0.86	0.68	2.48
Clay (validation)		17	16	0.76	0.53	1.81
Peat (validation)		18	16	0.54	0.68	0.98
Zavel (validation)		25	16	0.19	0.67	0.69
Sand (validation)		73	16	0.81	0.72	2.18
Combined validation sets		133	-	0.86	0.68	2.48

Bold **RMSE's** indicate that the accuracy has improved compared to the prediction of SOM with the non-clustered model or with the non-clustered model tested with clustered validation sets.

PLSR models with 5 – 17 factors were developed for the soil clusters. The highest accuracy of predictions was achieved with the zavel model (RMSE = 0.46 pH units) but this model was developed with 17 factors. The clay model was developed with the fewest factors (5) but the accuracy of the pH predictions not have not improved compared to the non-clustered model (cf. $RMSE_{clay\ model}$ = 0.59, $RMSE_{non\text{-}clustered}$ = 0.53). Good prediction accuracy was also achieved with the peat model (RMSE = 0.54) which was developed with 6 factors.

Even though the accuracy of the predictions did not improve for all models, or the improvement of prediction accuracy was small, it should be noted that the clustered models were developed with less factors (except for the zavel model) compared to the non-clustered model, which was developed with 16 factors.

Regarding the R^2, clustered soil models performed have lower R^2's than the non-clustered model. Even the 17-factor zavel model predictions returned a very low R^2 of 0.23 and also the lowest RPD of 1.00. The sand model predictions returned the highest R^2 (0.78) of all models and also the highest RPD of 2.01.

In figure 12 the calibration and validation results for pH based on clustering on soil type are given.

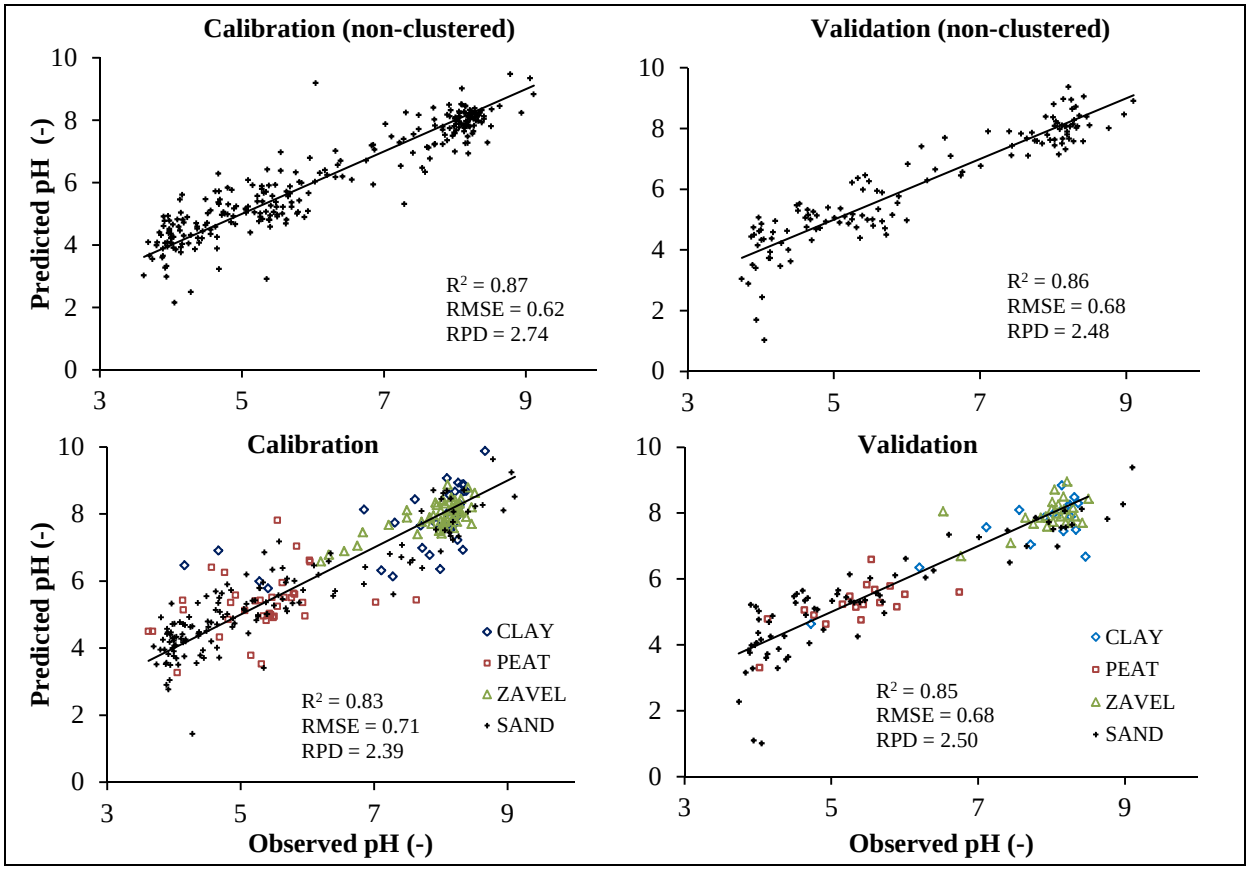


Figure 12. Calibration and validation results for combined soil clusters for pH (Lithological stratification). The lines represent the 1:1 line. The calibration and validations of the clustered models (bottom) are compared to the calibrations and validations of the non-clustered models (top).

Contrary to the other soil properties (SOM and Nt) there are no negative predictions during the calibration and validation of the models. This can be related to the smaller range of the pH (between 3 and 9) and the small RMSE's of the predictions. The variance of the pH of the different soil clusters is small and the models are calibrated and tested with samples which have the same small range and variances. Furthermore, it should be noted that pH values are log-values and that the RMSE's of the predictions may appear to be small, but when transformed back to original concentration values may be large. So care should be taken when interpreting and using the RMSE's of the predictions.

B. Land Use Stratification

Calibration and validation models were also applied for the land use clusters for three soil properties. Prediction statistics for predicting SOM, Nt and pH with stratified data sets can be found in tables 19 – 21.

SOM

In table 19 the different PLSR models that were developed for the prediction of SOM are shown.

Table 19 Validation results for SOM for the Land use clusters.

SOM				**Validation statistics**		
	N cal	**N val**	**# factors**	**R^2**	**RMSE (%)**	**RPD**
Agricultural	40	20	7	0.59	8.50	1.07
Forest	102	50	7	0.82	4.34	2.07
Nature	123	61	16	0.82	7.55	2.35
Combined	265	131	-	0.78	6.69	2.08
Non-clustered Land Use	265	131	17	0.81	6.21	2.24
Agricultural (validation)		20	17	0.61	7.39	1.23
Forest (validation)		50	17	0.89	3.76	2.39
Nature (validation)		61	17	0.84	7.31	2.42
Combined validation sets		131	-	0.81	6.21	2.24

Models based on stratified Land Use data were developed with 7 – 16 factors. The forest and nature models gave SOM predictions with the high R^2 (0.82), but the accuracy of these predictions has not improved compared to the non-clustered predictions, e.g. RMSE of SOM prediction in forest samples with the forest model = 4.34 %, while the accuracy of SOM prediction in these samples is with the non-clustered model is 3.76%. In table 19 can be seen that the RMSE of SOM prediction with clustered models does not result in better prediction accuracy. The RMSE of validation of the non-clustered model with clustered validation samples gives better accuracies and also better RPD's than models built on clustered data. But it should be noted that the non-clustered model was developed with 17 factors, while the clustered models were developed with less factors. The non-clustered model may have better predictions but the number of factors of this model may limit the transferability of this model to other situations.

Figure 13 shows the calibration and validation results of the 3 Land Use clusters combined for predicting SOM and the calibration and validation results of the non-clustered model.
As with the clustering based on Soil Type, there are also negative predictions in the calibration and validation model. There are 33 negative predictions observed for the calibration of the non-clustered model. These negative predictions occur for SOM content less than 8.35%. In the validation of the non-clustered model there are 15 negative predictions which occur for SOM contents smaller than 7.05%.
Negative predictions are also observed for the clustered models during calibration and validation. In the calibration there are 2 negative predictions observed for the agricultural model, 10 for the forest and 13 for the nature model. In the validation of the models negative predictions are only observed for the forest (5) and the nature model (8).
For these models negative predictions occur when the SOM content in the calibration model is smaller than 6.6% (agriculture and forest) or smaller than 6.08% (nature). In the validation model negative predictions occur when SOM content is lower than 8.36% (forest) or lower than 7.05% (nature). The cause of these negative predictions for such high SOM content is unknown. Forest and nature have a very large range in SOM content (0.97 – 94.70% and 0.34 – 95.80%, respectively). Negative predictions occurred only for the forest and nature models in the validation. The very large range of these clusters can be part of the explanation for the negative predictions.

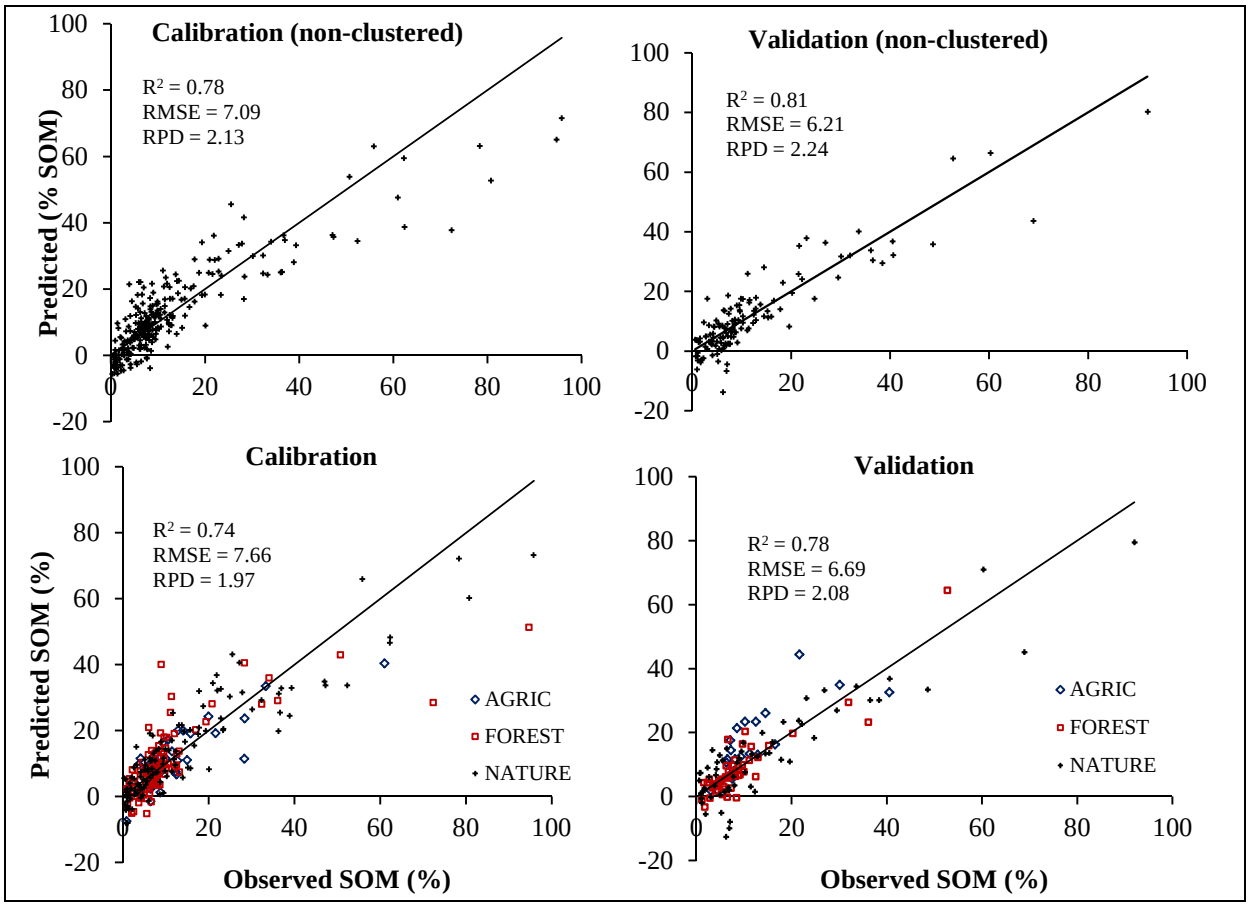


Figure 13. Calibration and validation results for combined Land Use clusters for SOM (Land Use stratification). The lines represent the 1:1 line. The calibration and validations of the clustered models (bottom) are compared to the calibrations and validations of the non-clustered models (top).

N-total

In table 20 the different PLSR models that were developed for the prediction of N-total are shown.

Table 20 Prediction results for Nt for the Land Use clusters.

N-total				**Validation statistics**		
	N cal	**N val**	**# factors**	**R^2**	**RMSE (g/kg)**	**RPD**
Agricultural	40	20	10	0.69	**1.67**	1.70
Forest	102	50	7	0.90	1.26	2.05
Nature	123	61	15	0.65	2.83	1.56
Combined	265	131	-	0.71	2.18	1.66
Non-clustered Land Use	265	131	9	0.70	2.02	1.80
Agricultural (validation)		20	9	0.63	1.75	1.62
Forest (validation)		50	9	0.84	1.14	2.27
Nature (validation)		61	9	0.67	2.59	1.71
Combined validation sets		131	-	0.70	2.02	1.80

Bold **RMSE's** indicate that the accuracy has improved compared to the prediction of SOM with the non-clustered model or with the non-clustered model tested with clustered validation sets.

The highest R^2 was achieved for predictions of Nt content with the forest model (R^2 = 0.90). predictions of Nt content in forest samples with the forest models were also the more accurate than the predictions of the other models ($RMSE_{forest}$ = 1.26 g/kg).
Compared to the non-clustered model, only the agricultural model resulted in an improved accuracy of the predictions ($RMSE_{agriculture}$ = 1.67 g/kg, $RMSE_{non\text{-}clustered}$ = 1.75), but the agricultural model was developed with 10 factors, while the non-clustered model was developed with 9 factors.

In fact, the differences between the accuracies of the prediction between clustered and non-clustered models is not large: for the nature model the difference with the non-clustered model is 0.09 g/kg, for the forest model the difference is 0.12 g/kg. The largest difference in accuracy is between the nature model: 0.24 g/kg.

In figure 14 calibration and validation results of land use clusters for predicting Nt are shown. Also note here the negative predictions which occur for predictions with the agricultural, the nature, the forest and the non-clustered model.
Calibration of the non-clustered model gives 28 negative predictions for Nt contents smaller than 4.40 g/kg. the validation of the non-clustered models gives 10 negative predictions for samples which have a Nt content smaller than 0.92 g/kg.
Calibration of the clustered models produced 28 negative predictions (agriculture: 2, forest: 6 and nature: 20) for samples which have a NT content smaller than 2.11 g/kg. Validation of the clustered models produced 9 negative predictions (forest : 2 and nature: 7) for samples with Nt content smaller than 1.03 g/kg.

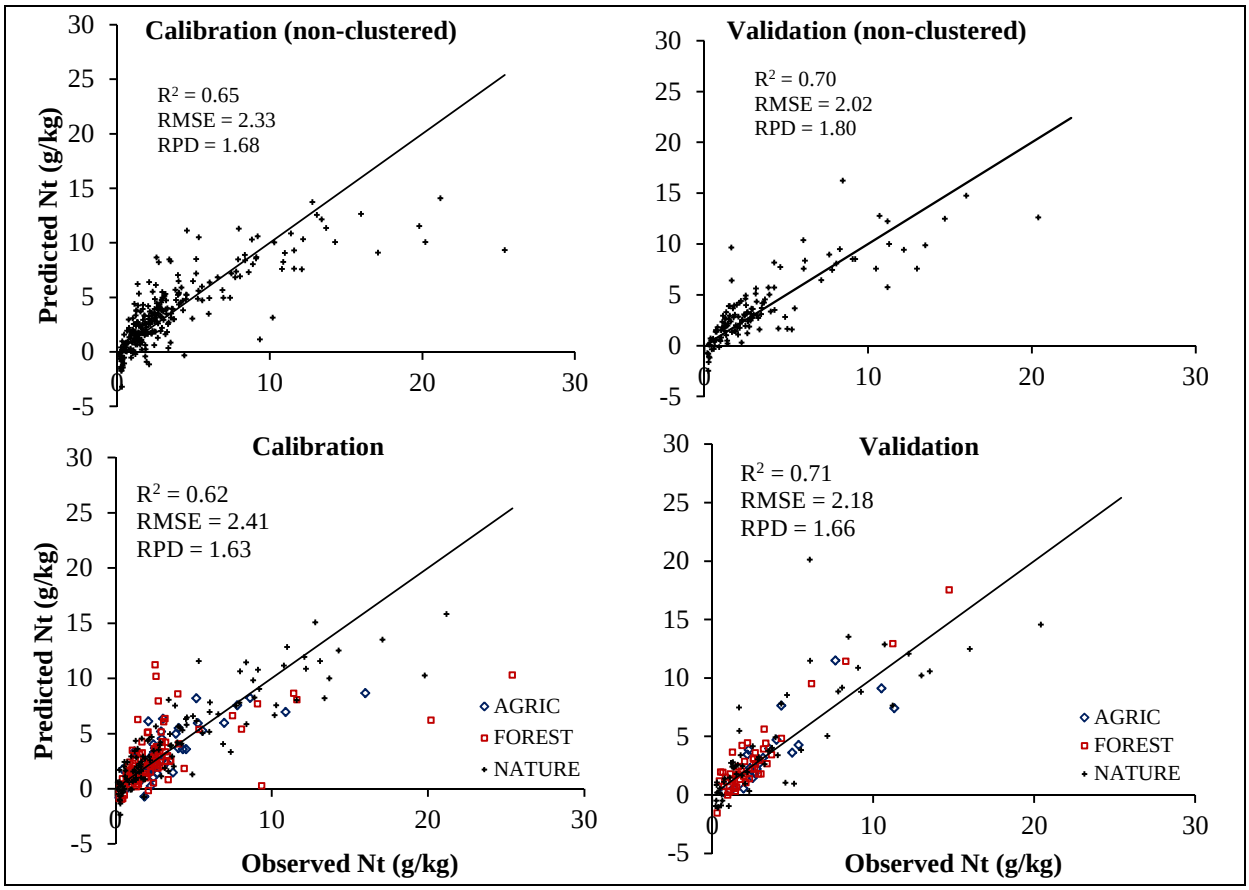


Figure 14. Calibration and validation results for combined land use clusters for Nt (Land Use stratification). The lines represent the 1:1 line. The calibration and validations of the clustered models (bottom) are compared to the calibrations and validations of the non-clustered models (top).

pH

In table 21 the different PLSR models that were developed for the prediction of pH are shown.
A 14-factor non-clustered model was developed and non-clustered models were developed with 9 – 12 factors. Table 21 shows that prediction accuracy improved only for the nature model (RMSE = 0.78), but this is just a very slight improvement compared to the non-clustered model prediction (RMSE = 0.79). Inspection of the prediction accuracy of the other clustered model shows that accuracy did not improve, but the changes are also very small: 0.1 pH units (e.g. between $RMSE_{agriculture}$ (0.53) and $RMSE_{non\text{-}clustered}$ (0.43)).

R2 and RPD values of both the clustered and non-clustered models are good and the differences between the clustered and non-clustered models are very small.

Table 21 Prediction results for pH for Land Use clusters.

pH				Validation statistics		
	N cal	N val	# factors	R^2	RMSE	RPD
Agricultural	40	20	9	0.84	0.53	2.37
Forest	102	50	12	0.91	0.61	3.15
Nature	123	61	14	0.78	**0.78**	1.92
Combined	265	131	-	0.85	0.68	2.38
Non-clustered Land Use	265	131	14	0.85	0.65	2.52
Agricultural (validation)		20	14	0.91	0.43	2.97
Forest (validation)		50	14	0.93	0.50	3.79
Nature (validation)		61	14	0.79	0.79	1.90
Combined validation sets		131	-	0.85	0.65	2.52

Bold **RMSE's** indicate that the accuracy has improved compared to the prediction of SOM with the non-clustered model or with the non-clustered model tested with clustered validation sets.

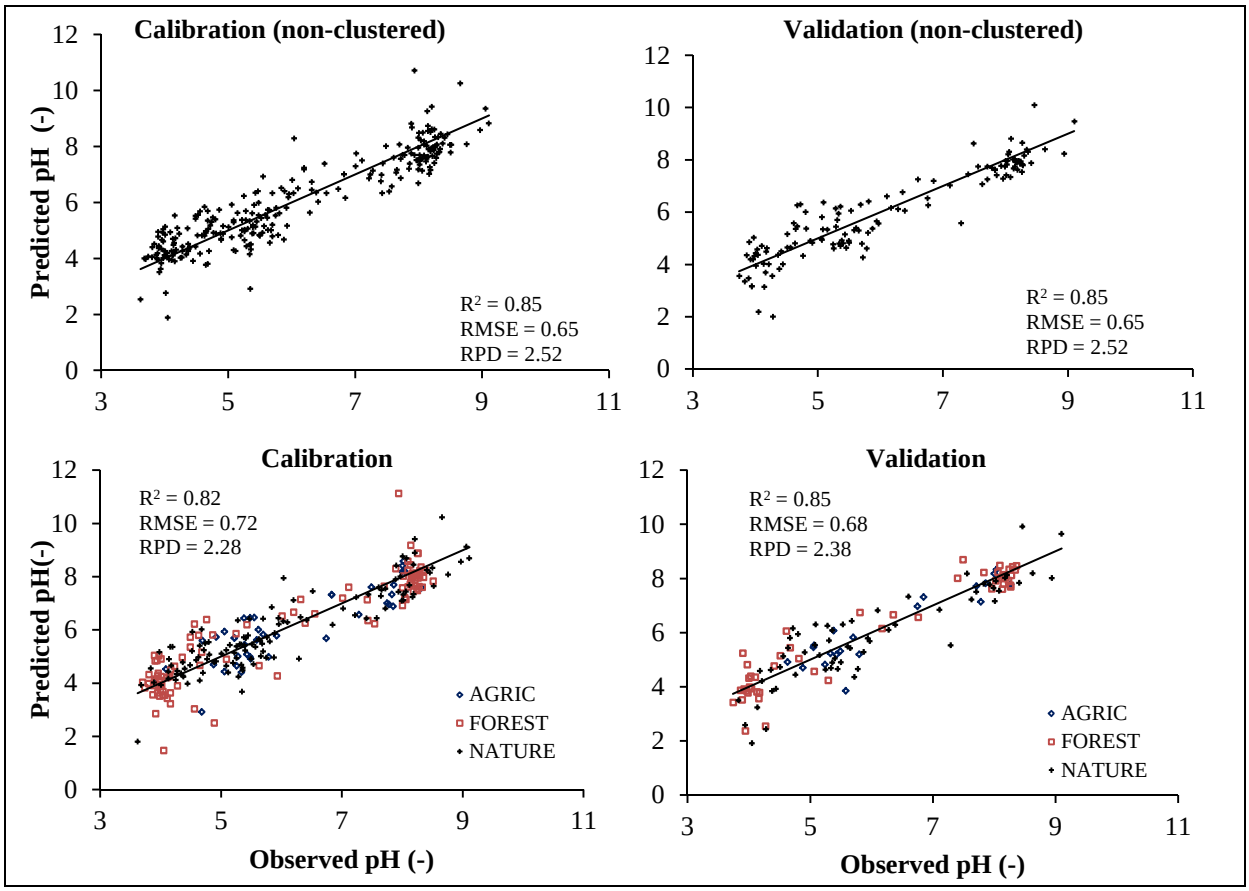


Figure 15. Calibration and validation results for combined Land Use clusters for pH (Land Use stratification). The lines represent the 1:1 line. The calibration and validations of the clustered models (bottom) are compared to the calibrations and validations of the non-clustered models (top).

There are no negative predictions in the calibration and validation models, as can be seen in figure 15. Also here absence negative predictions can be related to the specific range within pH values can vary. In this situation pH values varied from 3.5 – 9.0.

Good predictions can be attributed to the smaller range of the pH values in the calibration and validation samples and the fact that pH values are log-values.

3.3 Stratification based on spectral data

Calibration and validation models were also applied for clusters which were formed based on spectral alikeness. This was only tested for SOM. Prediction statistics for predicting SOM based on this method, Wavelength-based Discriminant Analysis, are shown in tables 22.

Table 22 Prediction statistics based on 'Wavelength-based Discriminant Analysis'.

SOM				**Validation statistics**		
	N cal	**N val**	**# factors**	**R^2**	**RMSE (%)**	**RPD**
A: <5%	**133**	**83**	9	0.57	6.09	1.15
B: 5-10%	**137**	**96**	8	0.13	16.70	0.94
C: >10%	**114**	**12**	8	0.79	**6.89**	1.99
Combined: A + B + C	**384**	**191**	-	0.25	12.62	1.10
Non-clustered Set	**384**	**191**	13	0.79	6.58	2.10
A: <5% (validation)		**83**	13	0.72	6.00	1.16
B: 5% - 10% (validation)		**96**	13	0.82	6.73	2.34
C: > 10% (validation)		**12**	13	0.53	8.86	1.54
Combined validation sets		**191**	-	0.79	6.58	2.10

Bold **RMSE's** indicate that the accuracy has improved compared to the prediction of SOM with the non-clustered model or with the non-clustered model tested with clustered validation sets.

Table 22 shows that there was no improvement of the R^2's for all clusters after WBDA clustering was applied. R^2 was only improved for the C cluster (>10%), which was 0.53 for the predictions with the non-clustered model and was improved to 0.79 after clustering was applied.
After clustering the accuracy of only the C cluster (>10%) was improved from 8.86% to 6.89%.
The accuracy of the prediction with the A cluster model dropped slightly from 6.00% to 6.09% but the accuracy of the prediction with the B-cluster model dropped drastically from 6.73% to 16.70%.

Predictions with the A-cluster model (A: <5%, $RMSE_{CAT\ A}$ = 6.09%) were more accurate than predictions with the other clustered models (B and C) but given the range of SOM in this cluster (0-5%), this accuracy of 6.09% is larger than the range in this cluster. Also for cluster B the RMSE is large compared to the ranges of the clusters: for the B-cluster (5-10%) the RMSE = 16.70%. The accuracy of predictions with the C cluster (>10%) is 6.89%, which is reasonable given the large range of this cluster.

Over all , the R^2, RMSE and RPD of the non-clustered model are better than the clustered models prediction statistics, except for cluster C which has better prediction statistics than the non-clustered model. With these results, clustering based on WBDA did not lead to improvement of the predictions of SOM.

The poor improvements of the accuracy of the predictions can be attributed to the misplacements of several spectra in the clusters (see descriptive statistics in appendix II, table 27). Using the criteria to arrange the spectral signatures in 3 categories resulted in some spectra being paced in the wrong cluster. Spectra corresponding to samples with very high SOM content were placed in the category which contained the lowest SOM content. These outliers have an influence on the accuracy of the predictions. Still the R^2's and RPD's of the predictions are remarkably high, despite these misplacements. Probably the majority of the spectra was placed in a correct class masking the effect of the outliers.

Figure 16 shows the calibration and validation results of the spectral stratification method. A total of 43 negative predictions are observed after calibration of the non-clustered model. These negative predictions occur for SOM contents smaller than 7.66%. After validation of the non-clustered model 22 negative predictions are observed for samples with SOM contents smaller than 5.44%.
After clustering with the WBDA method, 4 negative predictions are observed after calibration of the models: 3 negative in the A-cluster (<5%) for three samples having a SOM content of 0.24, 0.27 and

0.99%, and 1 negative prediction in the C-cluster (>10%) for a sample with a remarkably high SOM content of 12.4%. After validation of these clustered models, only 3 negative predictions were observed for the A-cluster for samples with SOM contents of 0.29, 0.35 and 0.57% SOM. The combination of very low SOM content and a low accuracy of the predictions could have caused the negative predictions. May be the range 0-5% is still too large for this cluster. As there are samples with a SOM content of smaller than 0.5% a further stratification of this cluster could may be improve the predictions of this cluster. For the clusters with a higher SOM content negative predictions do not occur, with the exception of the negative prediction after calibration of cluster C-model.
It should be noticed that the number of negative predictions has deceased after clustering was applied: 42 negative prediction of the non-clustered model were reduced to 3 negative predictions after clustering.

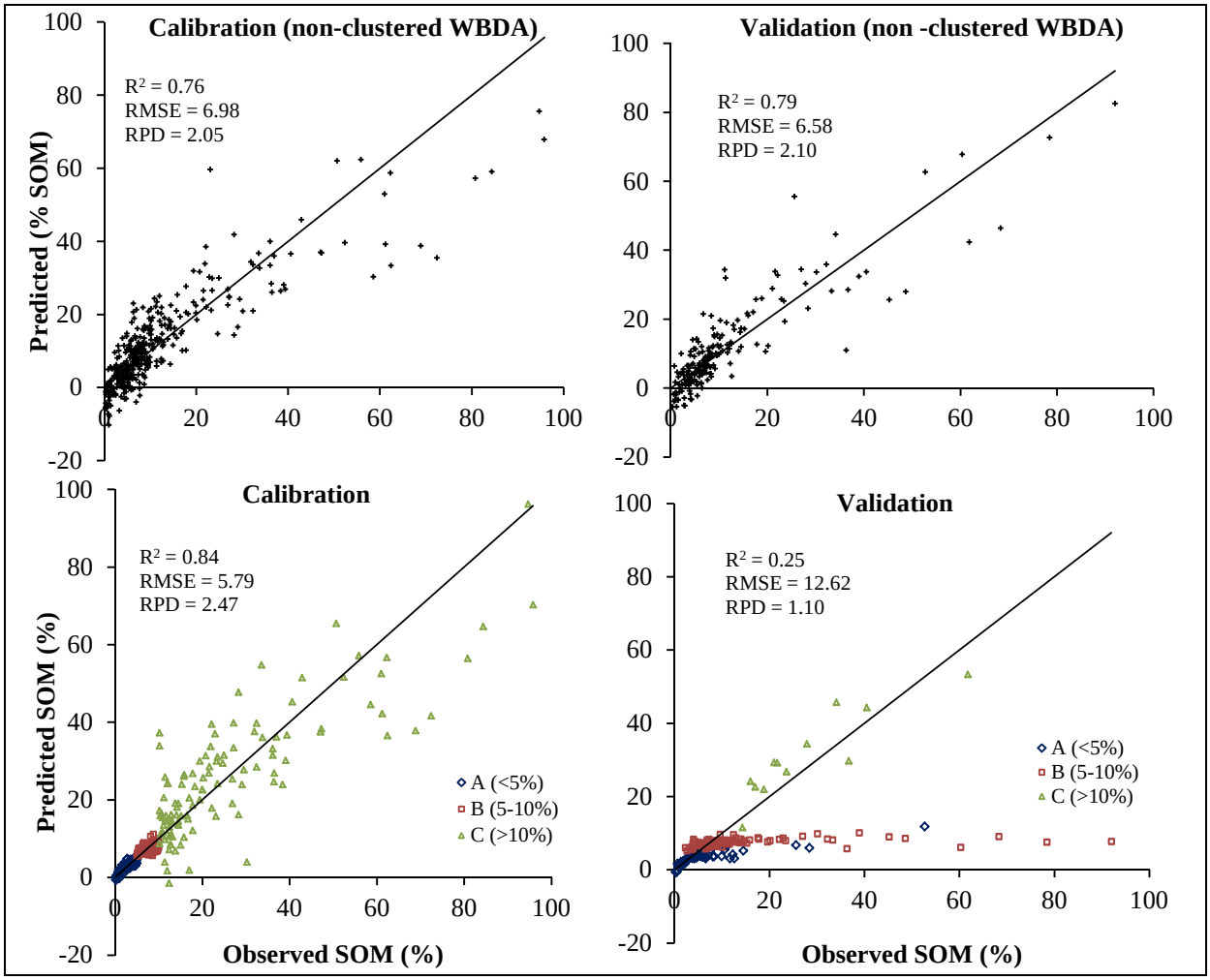


Figure 16. Calibration and validation results for combined spectral clusters for SOM (based on WBDA). The lines represent the 1:1 line. WBDA = Wave-length based Discriminant Analysis. The calibration and validations of the clustered models (bottom) are compared to the calibrations and validations of the non-clustered models (top).

A test was performed with creating ideal calibration and validation clusters with more intervals. The results of this test can be found in table 23. With ideal calibration and validation sets created both on a priori knowledge we see that for the 'ideal I set' (< 5%) set the accuracy is the best largest (RMSE = 0.70) (table 23). The accuracy of the predictions decreases for the next sets depending on their ranges: predictions with the ' ideal II set' (5-10%) have an accuracy of 1.18%, predictions with the 'ideal III set' (10-20%) have an accuracy of 2.65%, predictions with 'ideal IV set' (20-40%) have an accuracy of 5.84% and predictions with 'ideal V set' ((>40%) have an accuracy of 13.74%. These models were also developed with less factors and still resulted in improvement of the prediction accuracies.

Unfortunately, for 2 clusters (group II and IV) the R^2 is very low (<0.10), but the RPD's of the clustered models are better than the non-clustered RPD's. Predictions for high SOM (>10%) cannot be predicted well with PLS regression. the developed models cannot deal with the large SOM content of the samples, even if they were calibrated with samples containing the same range of high SOM content.

Table 23 Prediction statistics based on ideal clusters.

SOM				**Validation statistics**		
	N cal	**N val**	**# factors**	**R^2**	**RMSE (%)**	**RPD**
I: < 5%	**132**	**66**	**9**	0.78	**0.70**	1.99
II: 5-10%	**137**	**68**	**8**	0.25	**1.18**	1.13
III: 10-20%	**59**	**30**	**3**	0.06	**2.65**	1.01
IV: 20-40%	**38**	**18**	**2**	0.04	**5.84**	1.03
V: > 40%	**18**	**9**	**2**	0.43	**13.74**	1.21
Combined	**384**	**191**	-	0.93	3.72	3.71
Non-clustered	**384**	**191**	13	0.79	6.58	2.10
I: < 5% (validation)		**66**	13	0.82	3.96	0.35
II: 5-10% (validation)		**68**	13	0.17	4.25	0.31
III: 10-20% (validation)		**30**	13	0.20	7.54	0.35
IV: 20-40% (validation)		**18**	13	0.01	11.36	0.53
V: > 40% (validation)		**9**	13	0.66	14.84	1.12
Combined validation sets		**191**	-	0.79	6.58	2.10

Both calibration and validation sets were formed based on a priori knowledge of the SOM content.
Bold **RMSE's** indicate that the accuracy has improved compared to the prediction of SOM with the non-clustered model or with the non-clustered model tested with clustered validation sets.

Figure 17 shows the combined calibration and validation results for the ideally formed clusters.

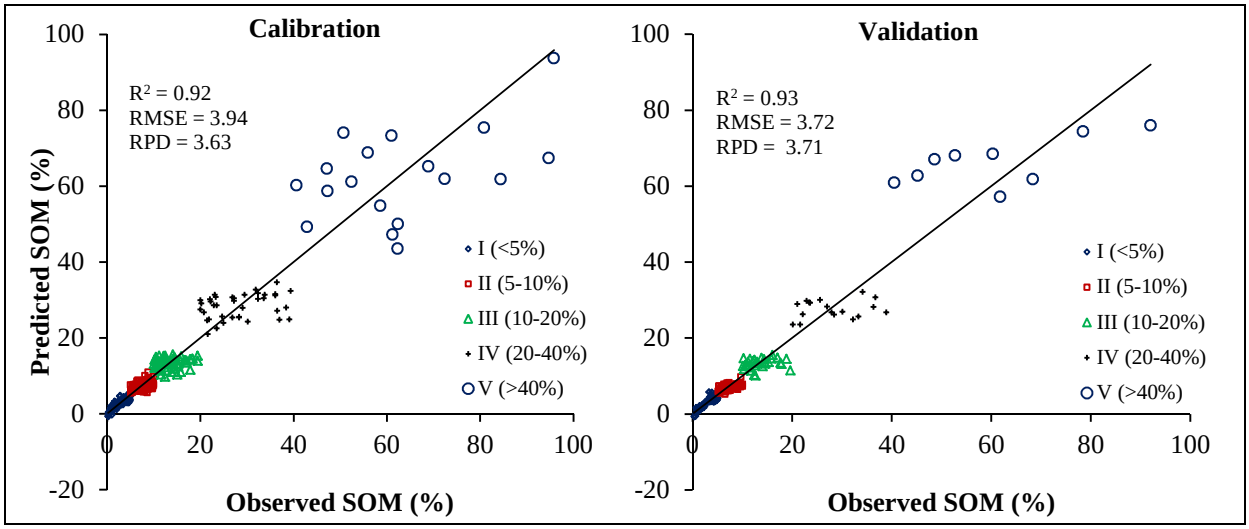


Figure 17. Calibration and validation results for combined ideal spectral clusters for SOM.
I: <5% SOM; II: 5 – 10% SOM; III: 10-20% SOM; IV: 20-40% SOM; V: >40% SOM. The lines represent the 1:1 line.

These ideal sets are free from outliers, so the effect of outliers on the prediction results is cancelled out, thus producing good prediction statistics. Prediction for clusters with high SOM content becomes difficult as these clusters contain very few calibration samples compared to the clusters which have very low SOM content as there are more samples available with low SOM content. Observing the predictions in figure 17 predictions for the first three clusters (I – III) are closer to the 1-1 line than predictions for the clusters with the high SOM content (IV and V).
In previous results low SOM content resulted in negative predictions, but when they are more clusters with smaller intervals, negative predictions occur less, as can be seen in figure 17. Only three negative predictions occurred in the calibration (for samples with SOM content of 0.24, 0.27 and 0.99%) and in

the validation (for samples with SOM content of 0.29, 0.35 and 0.57). This is an indication that proper ranges for each cluster should be set or chosen to avoid negative predictions. Negative predictions for the cluster with the smallest range (<5%) can probably be cancelled out if this cluster is stratified further.

3.4 Effect of clustering on prediction results

In this research 3 soil properties were predicted using clustered and non-clustered calibration and validation sets. In some cases clustering improved the predictions, but in most cases this was not the case. In figure 18 and 19 the prediction results per soil property for the different clusters and non-clustered sets are presented for comparing the effects of the clustering methods on the prediction parameters: R^2, RMSE and RPD.

The legend in figure 18 and 19 is as follows:
- the dark solid bars are the non-clustered models: the first dark bar is the non-clustered Soil Type result, the second bar is the non-clustered Land Use result.
- the speckled bars are models of the Soil clusters: clay, peat, zavel and sand
    - Clay (val), peat (val), zavel (val) and sand (val) are the validation sets used to test the non-clustered Soil type model
- the dashed bars are the Land Use clusters: Agricultural land, Forest and Nature areas.
    - Agricultural (val), Forest (val) and Nature (val) are the validation sets used to test the non-clustered Land Use model.
- striped bar (horizontal) are the cluster formed with the WBDA-method: A (<5%), B (5 – 10%) and C (> 10%)
    - A: <5% (val), B: 5 – 10% (val) and C: >10% (val) are the validation sets used to test the non-clustered WBDA model.

SOM:

R^2:
The highest R^2 is observed when SOM content is predicted for the forest samples (forest val) with the the non-clustered Land Use model: R^2 = 0.89. The lowest R^2 is observed for the zavel model (R^2 = 0.31) and when the SOM in clay samples is predicted with the non-clustered Soil Type model (R^2 = 0.27). R^2 values for models developed with the WBDA method (figure 19) not much better than the R^2's of the models developed with Land Use or Soil type stratification. The non-clustered WBDA-model and model C (>10%) have a R^2 of 0.79. Predictions with the B-model (>5-10%) produce a very low R^2 of 0.13.

RMSE:
The best prediction accuracy is achieved with the clay and zavel model which have RMSE of 3.59% and 2.89% respectively. The lowest accuracy is achieved when SOM in peat samples is predicted with the non-clustered Soil type model (RMSE of 11.43%).
The WBDA-models have about the same prediction accuracy as the RMSE of the non-clustered WBDA-model does not differ much from the RMSE of the A (<5%) and C (>10%) model. (RMSE non-clustered = 6.58%, RMSE (A:<5%) = 6.09%, RMSE (C:>10%) = 6.73%). The C-model (5-10%) has the lowest accuracy of all developed models: RMSE = 10.70%.

RPD:
The highest RPD values are achieved with the peat and nature model (RPD_{peat} = 2.27, RPD_{nature} = 2.35). Prediction of SOM in forest and nature samples with the non-clustered Land use model also results in high RPD's of 2.39 and 2.42.

Lowest RPD values were achieved with the Agricultural model (RPD = 1.07), and when the non-clustered Soil type model is used to predict SOM in clay samples (RPD = 0.96) and in zavel samples (= 0.71).
RPD's of the WBDA models are low compared to the RPD's of the models developed with the other stratification methods, except for the C-model (>10%) whose predictions produce a RPD of 1.99.

N-total

R^2
The highest R^2 was achieved with the forest model (R^2 = 0.90). The lowest R^2 was achieved with the zavel model (R^2 = 0.29).

RMSE
The best accuracies were achieved with the clay, zavel and sand model (RMSE = 1.23 g/kg, 1.15 g/kg and 0.87 g/kg, respectively). The lowest accuracy was achieved with the peat model, with an RMSE of 3.03 g/kg and when Nt content in peat samples was predicted with the non-clustered Soil Type model (RMSE = 3.44 g/kg).

RPD
The highest RPD was achieved with the clay and the forest model (RPD_{clay} = 1.99 and RPD_{forest} = 2.05). Prediction of Nt in these samples with the non-clustered Soil type and Land use model also produced high RPD's of 2.93 (for clay) and 2.27 (forest).

pH

R^2
The highest R^2 were achieved with the forest model (R^2 = 0.91). The agricultural model (R^2 = 0.84) and the non-clustered Soil Type (R^2 = 0.86) and Land Use models (R^2 = 0.85) also have very high R^2 values. Very low R^2 values were achieved with the zavel model (R^2 = 0.23) and when pH in zavel samples was predicted with the non-clustered soil type model (R^2 = 0.19).

RMSE
The best prediction accuracy was achieved with the zavel model (RMSE = 0.46 pH units).The lowest accuracy was achieved with the sand and nature model (RMSE = 0.78 for both models) and when pH was predicted in nature samples with the non-clustered Land use model (RMSE = 0.79 pH units).

RPD
The highest RPD values were achieved with forest model (RPD = 3.15). other models which produced RPD's higher than 2.00 are: the sand model (RPD = 2.01), the agricultural model (RPD = 2.37), the non-clustered Soil type model (RPD = 2.48) and the non-clustered Land use model (RPD = 2.52).
The lowest RPD values was produced with the zavel model (RPD = 1.00).

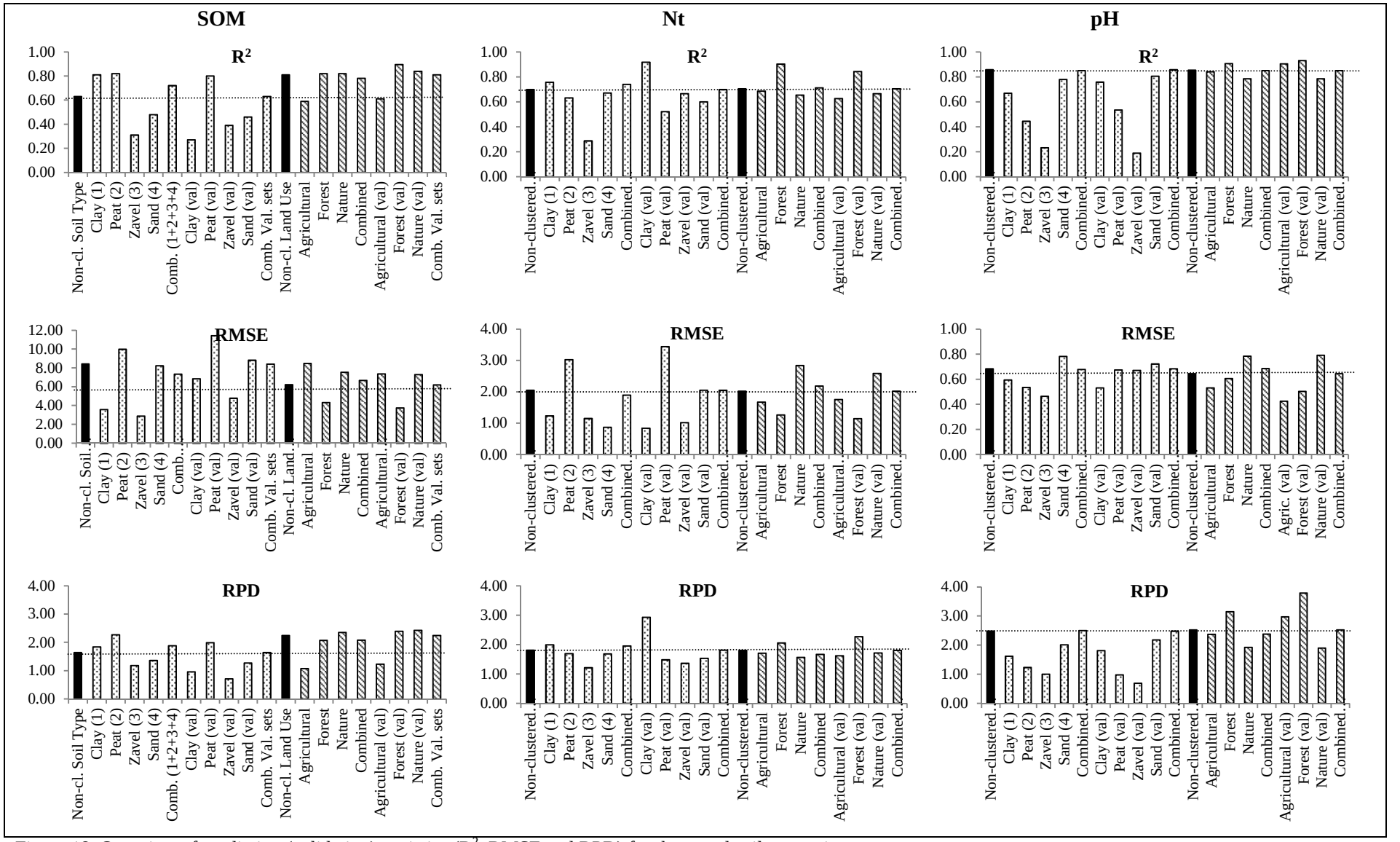


Figure 18. Overview of prediction (validation) statistics (R^2, RMSE and RPD) for the tested soil properties.

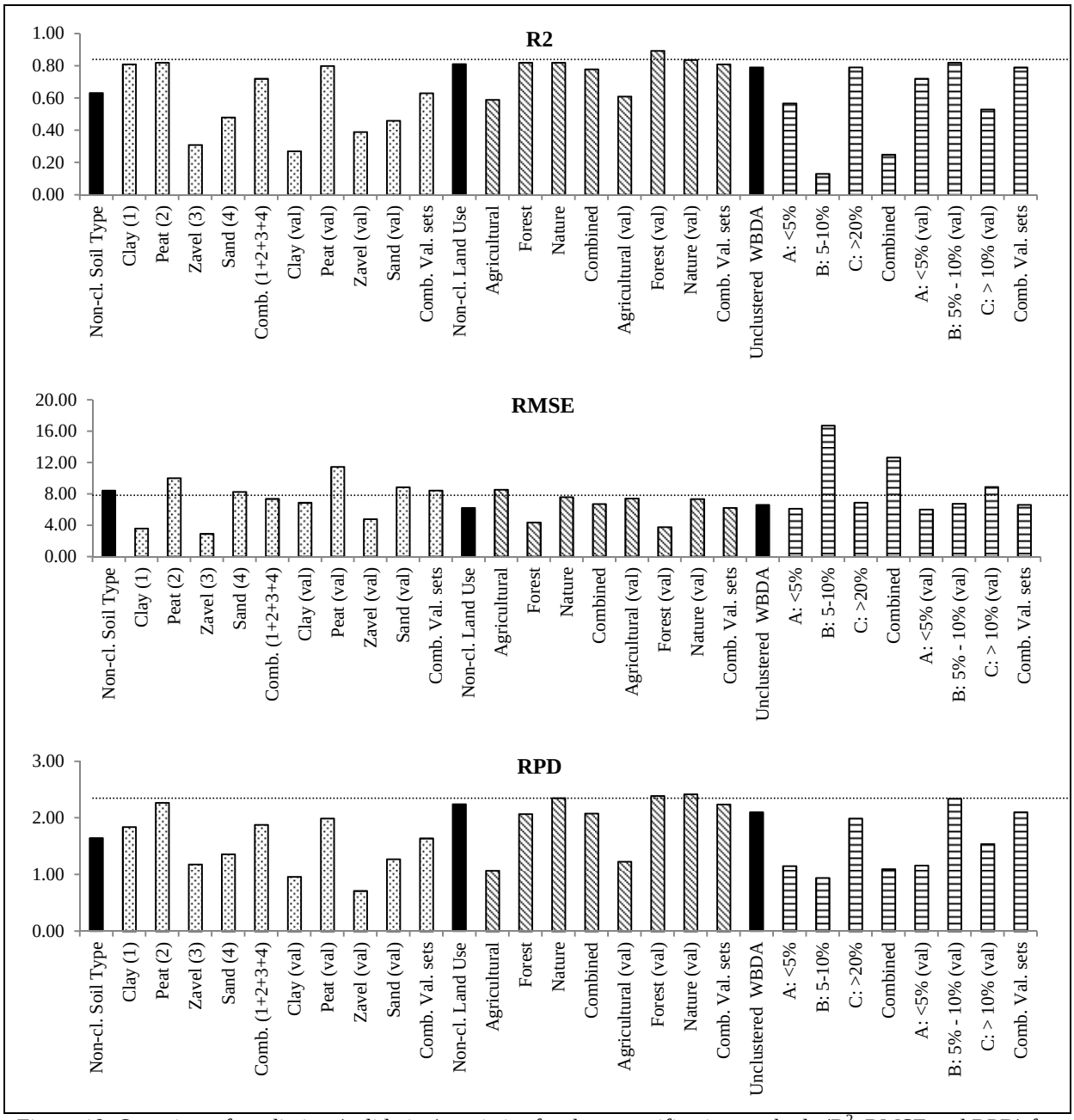


Figure 19. Overview of prediction (validation) statistics for three stratification methods (R^2, RMSE and RPD) for SOM prediction.

Model evaluation

Based on the R^2 and RPD the developed models were categorized in 3 classes following the criteria of Chang *et al.* (2001). The result of this evaluation of the developed models are presented in table 24.
Four models to predict SOM are evaluated as category A models: the peat model, non-clustered Land Use model, the Forest model and the Nature model. Two models developed with the WBDA method were evaluated as class C models (A: <5% and B: 5-10%). Model C (>10%) was evaluated as a class model. As previously mentioned the WBDA method did not improve prediction predictions results.
Only one model to predict Nt content was evaluated as category A model: the Forest model. All other models were evaluated as category B models. The zavel model is the only model that is evaluated as category C model.
For the prediction of pH four models were evaluated as category A models: the non-clustered soil type model, the non-clustered land use model, the forest model and the agriculture model. The clay, sand

and nature model were evaluated as category B models. The peat and zavel models were evaluated as category C models.
Category A models have accurate prediction capabilities, category B model predictions can be possibly improved by using different calibration techniques (Chang *et al.*, 2001), and category C models have no reliable prediction capabilities (Chang *et al*., 2001 & Bartholomeus *et al.*, 2008).

Table 24 Model evaluation.

Validation (SOM)								
Stratification method	**Model**	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**Model evaluation**
Lithological	**Non-clustered**	**296**	**133**	**12**	0.63	8.41	1.64	B
	Clay	**34**	**17**	**4**	0.81	3.59	1.84	B
	Peat	**38**	**18**	**9**	0.82	9.99	2.27	A
	Zavel	**51**	**25**	**3**	0.31	2.89	1.18	C
	Sand	**146**	**73**	**7**	0.48	8.24	1.36	C
	Combined soil models	**296**	**133**	-	0.72	7.34	1.88	B
Land Use	**Non-clustered**	**265**	**131**	**17**	0.81	6.21	2.24	A
	Agricultural	**40**	**20**	**7**	0.59	8.50	1.07	C
	Forest	**102**	**50**	**7**	0.82	4.34	2.07	A
	Nature	**123**	**61**	**16**	0.82	7.55	2.35	A
	Combined	**265**	**131**	-	0.78	6.69	2.08	B
WBDA	**Non-clustered**	**384**	**191**	**13**	0.79	6.58	2.10	B
	A: <5%	**133**	**83**	**9**	0.57	6.09	1.15	C
	B: 5-10%	**137**	**96**	**8**	0.13	16.70	0.94	C
	C: >20%	**114**	**12**	**8**	0.79	6.89	1.99	B
	Combined	**384**	**191**	-	0.25	12.62	1.10	C
Validation (Nt)								
Stratification method	**Model**	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**Model evaluation**
Lithological	**Non-clustered**	**296**	**133**	**15**	0.70	2.05	1.81	B
	Clay	**34**	**17**	**4**	0.76	1.23	1.99	B
	Peat	**38**	**18**	**8**	0.63	3.03	1.69	B
	Zavel	**51**	**25**	**3**	0.29	1.15	1.21	C
	Sand	**146**	**73**	**7**	0.67	0.87	1.68	B
	Combined soil models	**296**	**133**	-	0.74	1.90	1.95	B
Land Use	**Non-clustered**	**265**	**131**	**9**	0.70	2.02	1.80	B
	Agricultural	**40**	**20**	**10**	0.69	1.67	1.70	B
	Forest	**102**	**50**	**7**	0.90	1.26	2.05	A
	Nature	**123**	**61**	**15**	0.65	2.83	1.56	B
	Combined	**265**	**131**	-	0.71	2.18	1.66	B
Validation pH								
Stratification method	**Model**	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**Model evaluation**
Lithological	**Non-clustered**	**296**	**133**	**16**	0.86	0.68	2.48	A
	Clay	**34**	**17**	**5**	0.67	0.59	1.62	B
	Peat	**38**	**18**	**6**	0.44	0.54	1.23	C
	Zavel	**51**	**25**	**17**	0.23	0.46	1.00	C
	Sand	**146**	**73**	**12**	0.78	0.78	2.01	B
	Combined soil models	**296**	**133**	-	0.85	0.68	2.50	A
Land Use	**Non-clustered**	**265**	**131**	**14**	0.85	0.65	2.52	A
	Agricultural	**40**	**20**	**9**	0.84	0.53	2.37	A
	Forest	**102**	**50**	**12**	0.91	0.61	3.15	A
	Nature	**123**	**61**	**14**	0.78	0.78	1.92	B
	Combined	**265**	**131**	-	0.85	0.68	2.38	A

Lithological stratification resulted good prediction (category A) models for the prediction of:
1. SOM with the peat model
2. Nt: no soil type model is suitable
3. pH with the non-clustered soil type model.

Land use stratification resulted in good prediction (category A) models for the prediction of:
1. SOM with the non-clustered soil type model, the forest and agricultural model.
2. Nt with the forest model
3. pH with the non-clustered, agricultural and forest model

Spectral stratification (WBDA) did not produce A models suitable for the prediction of the SOM content in stratified samples.

Chang *et al.* (2001) evaluated models only on the R^2 and RPD. But the accuracy (RMSE) of the prediction is also important. Based on the prediction accuracy of the models the effect of the three stratification methods is evaluated but with only the RMSE as main criterion.

Lithological stratification resulted in good accuracy of predictions of:
1. SOM: with the clay and zavel model, and in zavel samples using the non-clustered soil type model.
2. Nt: with the clay, zavel and sand model, and in clay and zavel samples using the non-clustered soil type model.
3. pH: with the zavel and clay model.

Land Use stratification resulted in good accuracy of predictions of:
1. SOM: with the forest model, and in forest samples using the non-clustered land use model.
2. Nt: with the forest and agricultural model, and in forest samples using the non-clustered land use model.
3. pH: with the agricultural model, and in agricultural and forest samples using the non-clustered land use model.

Spectral stratification (WBDA) did not produce models that reached reasonable prediction accuracy: RMSE for the cluster with the smallest range (0 – 5%) is 6.09%.

Based on the accuracy of the predictions a flow chart (figure 20) is designed to show which model one can use to predict SOM, Nt or pH. Which model one can use depends on the soil property and on the location (or source) the soil sample. The source can either be the soil type or the land use on which the soil sample was taken. If one does not know what the soil type or land use is, x-y coordinates also suffice. With the x-y coordinates one can also determine the soil type or the land use. If one does not have any information about the source of the sample, the non-clustered models can be used. The accuracy of the predictions that are achieved are mentioned in the flow chart.

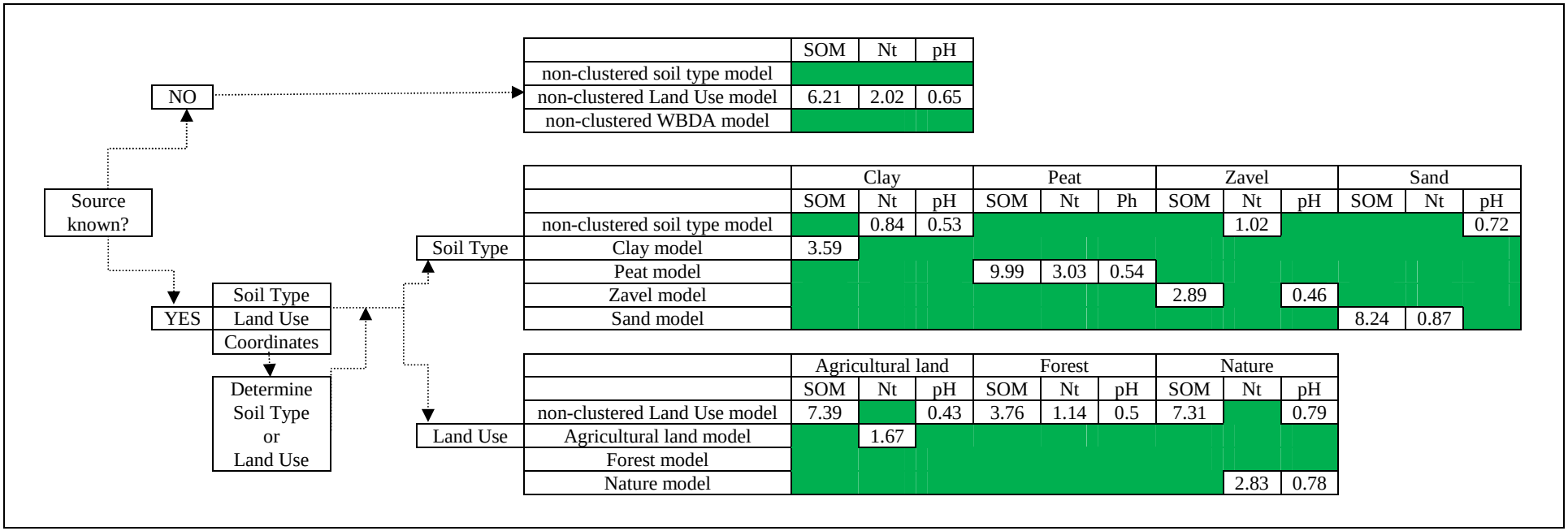


Figure 20: Decision tree for predicting soil properties.
The coloured cells indicate that the model cannot be used for the prediction of the chosen soil property. The open cells indicate that the model can be used. The number in the cells indicate the accuracy that is reached when the model is used. Units of accuracy: SOM (%), Nt (g/kg), pH (-). WBDA= Wave-length Based Discriminant Analysis.

3.3 General discussion

On developing and implementing stratification methods

In this research two external data sources (Soil Type map and Land Use map) were the basis for two stratification methods. A third stratification method was based on Wave-length based Discriminant Analysis (WBDA) or spectral alikeness (similarity).

Lithological stratification divided the data in 4 soil clusters. These 4 soil clusters were aggregated soil classes from a soil classification that distinguishes more soil type classes. By aggregating soil classes that are related to each other or which are sub-classes of each other, the number of soil classes was reduced to four. The advantage of reducing the number of soil classes to four is that more points are allocated to the aggregated soil classes, instead of very few soil points allocated to many soil classes. Especially when creating calibration samples there should be enough samples of each soil type to train the calibration model.

This same approach was applied for Land Use stratification. The many land use types were aggregated and reduced to 3 land use types, ensuring that enough samples would be allocated to each land use class.
The drawback of these aggregating approaches is that the range of the soil property values becomes large as each cluster contains soil samples from different sub classes, which probably have their own specific soil property range. The descriptive statistics indeed showed that the ranges of the soil property per cluster were indeed quite broad, while the purpose of stratification was to create homogeneous sets.

It is worth noting that the Land Use map used for this stratification was LGN5. This is currently an out-dated version as the latest version is LGN6. At the time of this research LGN6 was just introduced and not available for this research. Nothing can be said about the effect of using LGN5 instead of LGN6 because the difference between these two maps was not analysed. If soil samples are placed in a wrong class, this can be attributed to the use of the out-dated land use map.

The accuracy of the GPS coordinates could be another source of possible wrong allocations of the soil points. But this is beyond the control of this research and the fact that the point locations are much smaller than the land use or soil type area, enlarges the chance that the soil points are placed in a correct class. The chance of misplacement is large when the soil samples and the GPS measurements are taken at or near the border of two or more land use or soil types.

The third stratification method did not involve external data sources as basis for stratification. Instead, samples were clustered according to their spectral alikeness. Based on calculated ratios between reference spectra, the different spectra were arranged in clusters representing different content of the soil property (SOM). The pitfall of this method is that between spectra there may be wide differences in chemical composition due to small, yet important absorption bands hidden by a larger band or bands (Roberts and Workman, 2004) and still these two spectra may be nearly identical. The descriptive statistics of the clusters indeed show that misplacements of samples have occurred in all 3 clusters resulting in outliers present in each cluster which influenced the prediction results of the developed models for this stratification method.

On validating the developed stratification methods

The clusters defined by the different stratification methods contained enough samples to develop different calibration models. Some calibration models were developed with more samples than others (e.g. the non-clustered models were developed with twice the number of samples than the sand model, or the sand samples was developed with almost three times twice the number of samples than the peat

model), which could have an effect on the prediction ability of the model as prediction model parameters are estimated more accurately when more samples are used (Roberts and Workman, 2004). Gomez *et al*. (2008) also stated that SOM prediction models seem to be sensitive to the number of soil samples. In my research differences between the predictions of the created models can also be attributed to the number of calibration samples that were used to develop the developed prediction models.

For good predictions, the clusters should have a wide range in composition (Murray and Cowe in (Roberts and Workman, 2004). Bartholomeus *et al.* (2008) also agree with this as they state that a large variance is required for the calibration of the prediction model otherwise extrapolation beyond the range in the training data set results in large errors. But they also state that a greater variability in the training phase of a statistical model may lead to an improved robustness of the model. Gomez *et al.* (2008) also mentioned excellent cross validations when the soil sample set is more comprehensive. Besides a wide range in composition the samples should be typical to those to be analysed (Roberts and Workman, 2004) which means that validation samples should be similar to calibration samples (Brown *et al.*, 2005). The large data set that was used in my research indeed covered a large range of the soil properties. The created clusters also had a wide range in composition, but in fact this is contradictory to the purpose of stratification, as stratification is meant to create homogeneous populations. In fact, the non-clustered sets covered the widest range in composition and given the fact that the non-clustered were developed with the largest training sets it is not surprising that the clustered models not always performed better than the non-clustered models. The purpose of stratification was to overcome non-linearities, but this was partly achieved because they were still very variable but not as variable as the non-clustered data sets. Due to the non-linear behaviour under- or over estimations of soil properties, especially with higher contents were observed.

About the variability of data sets there are different conceptions. Bartholomeus *et al*. (2008) worked with data with a large variability (0.06 - 45.1% SOM); the SOM content of the data in my research ranges from 0.24 to 95.80%, which I consider definitely as highly variable, while Gomez *et al.* (2008) already speak of a wide range or high variability when the samples have an SOM content between 0.002 and 5.1%. These conceptions can be related to differences in site locations and different geographic regions where this type of research is conducted.
Gomez *et al.* (2008) assumed that a high variability of the soil data set and a high number of soil data could be a factor of improvement of the prediction accuracy, while in my research I assumed that by bracketing the data set into cluster prediction accuracy could be improved. The assumption of Gomez *et al.* (2008) also contradicts the assumption of Cécillion *et al.* (2009) who attribute poor prediction results to the heterogeneity of the samples.

The clusters created based on wave-length based discriminant analysis were the only clusters which did not cover a wide range in composition. The range of these cluster was set by predefined ranges. The clusters that were created were more homogenous than all clusters formed by the other stratification methods. Especially the two clusters with the lower SOM content (A, <5% and B, 5-10%) were the most homogenous given the small range that was defined for this cluster. Still, the prediction results based on these homogenous clusters was not satisfying as the RMSE's of the predictions was larger than the range of the clusters.

Negative predictions have been observed in this research. Negative predictions have also been encountered in other research but the occurrence of these predictions was not discussed and also not how was dealt with this. A possible explanation for this is given by Gomez *et al.* (2008) who observed that when the SOM dropped below 1% reflectance spectra are not able to predict the SOM content. The results in my research did not only show negative predictions for SOM, but also for N-total. These negative predictions occurred for N-total contents smaller than 2.11 g/kg or SOM contents smaller than 7.05%. Apparently small contents of the soil properties were indeed difficult to predict in this research.

Outliers could also have had an effect on the predictions. Outliers have been left in the data as was the purpose to also test the model how to deal with outliers. Outliers influence the linear regression because the regression line is forced through the outliers.

Causes of poor prediction results could be the heterogeneity of sample sets (optimal calibration requires limited but sufficient set heterogeneity) (Cécillon *et al.*, 2009), even though each cluster was assumed to be more or less homogeneous given the fact that each cluster was created by aggregating different soil types or land use classes, the clusters could indeed still be quite heterogeneous. Another cause could be the number of calibration samples that were used for some clusters, e.g. sand models were developed with many calibration samples while the zavel, peat and agricultural models were developed with much smaller training sets. But as there was no consistent relation observed between the number of samples that were used to calibrate models and the predictions results of the models, it cannot be confirmed that the number of calibration samples was a true cause for the poor prediction results.

SOM

Zornoza *et al.* (2008) obtained very good results for prediction SOM in 393 samples: R^2 = 0.98, RMSE = 6.25% and RPD = 5.75. This was achieved with 1st derivative and multiplicative scattering applied as pre-processing methods. The soil samples had a wide range of soil characteristics, land use and vegetation and specific climatic conditions in Spain. Gomez *et al.* (2008) also used PLSR on soils which had a maximum SOM content of 5.10%. For the prediction of 146 samples they found R^2 = 0.71 – 0.73, RMSE = 0.52 – 0.53, and RPD 1.87 – 1.92, obtained for models with 6 and 7 factors. Although they state that their samples were highly variable as the range of SOM content was from 0.002 – 5.1%, their data is still not as variable as the data used in my research:, e.g. the range of SOM content in my research was from 0.24 – 95.80%.

Bartholomeus *et al.* (2008) also found for SOM prediction R^2 of 0.80 – 0.81 but these predictions were based on linear relations with $SOC^{1/4}$ content for 40 samples. The range of these samples was from 0.06 – 45.1%. Viscarra Rossel *et al.* (2006) found for testing 118 samples with a 6-factor model an R^2 of 0.72 and a RMSE of 0.15 but the range of these soils was very small (0.81 – 1.98%) with a mean of 1.34%. which is compared to my samples very homogenous. Summers *et al.* (2009) predictions for 228 samples with a 10-factor model produced a R^2 of 0.57, RMSE = 0.35% and RPD = 1.80. The mean SOM content was 1.5%, and the range was from 0.31 – 2.90% SOM. The most accurate predictions in my research for SOM were achieved with the zavel and clay model ($RMSE_{zavel}$ = 2.89%, $RMSE_{clay}$ = 3.59%), but R^2 = 0.81 of SOM predictions with the clay model was higher than the R^2 = 0.31 of SOM predictions with the zavel model. Given the smaller number of samples in these clusters, these predictions are acceptable, but calibrating the models with more samples is needed to improve the predictions.

N-total

Zornoza *et al.* (2008) found high predictions for N-total with 383 samples giving a R^2 of 0.95, RMSE = 0.41 and RPD = 4.69 without applying any pre-processing method. Cobo *et al.* (2010) found for soils in Zimbabwe and for 165 soils, using PLSR with 1st derivative and Vector Normalization as pre-processing methods, R^2 = 0.96, RMSE = 0.02 and RPD = 5.2, which are excellent prediction results. Best predictions accuracy in my research was achieved with the sand model, having a RMSE of 0.87 g/kg and R^2= 0.67.

pH

Zornoza *et al.* (2008) found for pH predictions for 393 soils, only applying Multiplicative Scattering correction, R^2 = 0.72, RMSE = 0.14 and RPD = 1.90. Cobo *et al.* (2010) predicted for 165 soils using PLSR with 1st derivative and straight line subtraction (SLS) acceptable results as R^2 = 0.86, RMSE = 0.24 and RPD = 3.1. Viscarra Rossel *et al.* (2006) predicted pH for 118 soils with a 14-model factor and found RMSE of 0.14 and R^2 of 0.73. In my research the best prediction accuracy was achieved

with a 17-factor zavel model, giving a RMSE of 0.46 pH units but with a $R^2 = 0.23$ this model was not a good model. A good model was the 9-factor agricultural model which produced a RMSE of 0.53 and a R^2 of 0.84.

Relevance of findings

The number of samples per cluster was different for the clusters that were created. If more samples were available for the clusters larger calibration sets could be created. Still the results look promising because better prediction accuracies were achieved after stratification was applied. Unfortunately, not all models that were built with stratified data sets resulted in better prediction accuracies. This shows that clustering indeed can improve the predictions. The cause for inconsistencies still have to be find out. The current library should be extended with more samples, especially with samples from soil types and land use types that were underrepresented in this study (e.g. peat, zavel and clay soils, and agricultural land should be expanded with more samples) and possibly with other geographical regions or soil types of The Netherlands which were not represented in this study (e.g. loam soils were excluded in this study because there were too few loam samples). Furthermore, a decision tree has been created which allows to select a model to predict a soil property of interest (SOM, Nt or pH) with a beforehand indicated accuracy. Models with high accuracies need to be improved by using more training samples to improve the model predictions. Calibration of the models in this research were done without any pre-processing. This leaves different possibilities open to improve the developed prediction models.
The WBDA-method needs to be revised again because this method seems also promising especially in cases when only spectra are available. Due to the misplacement of spectra in wrong classes the potential of this method is reduced. Methods on how to reduce or eliminate the misplacements of spectra need to be find out and tested.

4. Conclusions

In this research 3 stratification methods were developed and tested. Two external data sources were used to develop a Lithological and a Land Use stratification method, and one method was based on the similarity of the soil spectra: the Wave-length Based Discriminant Analysis. The expectation was that the prediction models would perform better than non-clustered models if suitable clusters and ranges were selected. Contrary to what I expected the clustering methods did not improve the prediction results for all 3 tested soil properties, at least not for all clusters.

Stratification, based on external data sources can improve the estimations of soil properties from VNIR spectra:
For the prediction of SOM this was achieved with the clay, peat and zavel model but also when SOM content was predicted in agricultural, forest and nature samples with the non-clustered land use model.
For the prediction of Nt this was achieved with the peat, agricultural and nature model, and when Nt content in clay, zavel, and forest samples was predicted with the non-clustered Soil Type or Land Use model.
For the prediction of pH this was achieved with the peat model and when pH was predicted in clay, sand, agricultural, forest and nature samples with the non-clustered Soil Type or Land Use model.

Stratification, based on spectral data only can improve the estimations of soil properties from VNIR spectra. This was partly proven for one of the three models which were developed with this method: model C (>10%) produced more accurate predictions than the non-clustered model. But care should be taken when this model is used because this stratification method must be improved further to reduce or avoid wrong placements of spectra in the validation sets.

References

Bartholomeus, H. M., M. E. Schaepman, L. Kooistra, A. Stevens, W. B. Hoogmoed and O. S. P. Spaargaren (2008). "Spectral reflectance based indices for soil organic carbon quantification." Geoderma **145**(1-2): 28-36.

Batten, G. D. (1998). "Plant analysis using near infrared reflectance spectroscopy: the potential and the limitations." Australian Journal of Experimental Agriculture **38**(7): 697-706.

Ben-Dor, E. and A. Banin (1995). "Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties." Soil Sci Soc Am J **59**(2): 364-372.

Ben-Dor, E., S. Chabrillat, J. A. M. Demattê, G. R. Taylor, J. Hill, M. L. Whiting and S. Sommer (2009). "Using Imaging Spectroscopy to study soil properties." Remote Sensing of Environment **113**(Supplement 1): S38-S55.

Ben-Dor, E., D. Heller and A. Chudnovsky (2008a). "A Novel Method of Classifying Soil Profiles in the Field using Optical Means." Soil Sci Soc Am J **72**(4): 1113-1123.

Ben-Dor, E., R. G. Taylor, J. Hill, J. A. M. Demattê, M. L. Whiting, S. Chabrillat, S. Sommer and L. S. Donald (2008b). Imaging Spectrometry for Soil Applications. Advances in Agronomy, Academic Press. **Volume 97:** 321-392.

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

Cécillon, L., B. G. Barthès, C. Gomez, D. Ertlen, V. Genot, M. Hedde, A. Stevens and J. J. Brun (2009). "Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS)." European Journal of Soil Science **60**(5): 770-784.

Chang, C.-W. and D. A. Laird (2002). "Near-Infrared Reflectance Spectroscopic Analysis of Soil C and N." Soil Science **167**(2): 110-116.

Chang, C.-W., D. A. Laird, M. J. Mausbach and C. R. Hurburgh, Jr. (2001). "Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties." Soil Sci Soc Am J **65**(2): 480-490.

Cobo, J. G., G. Dercon, T. Yekeye, L. Chapungu, C. Kadzere, A. Murwira, R. Delve and G. Cadisch (2010). "Integration of mid-infrared spectroscopy and geostatistics in the assessment of soil spatial variability at landscape level." Geoderma **158**(3-4): 398-411.

Condit, H. R. (1970). "Spectral Reflectance of American Soils." Photogrammetric Engineering **36**(9): 955-&.

de Bakker, H., W. P. Locher and G. G. L. Steur (1987). Bodemkunde van Nederland : leer- en handboek op hoger-onderwijsniveau. Den Bosch, Malmberg.

de Bakker, H., J. Schelling, D. J. Brus and C. van Wallenburg (1989). Systeem van bodemclassificatie voor Nederland : de hogere niveaus. Wageningen, Pudoc.

Doelman, P. and H. Eijsackers (2004). Vital soil : function, value and properties. Amsterdam, Elsevier.

Doran, J. W., D. C. Coleman and D. F. Bezdicek (1994). Defining soil quality for a sustainable environment. Madison, Soil Science Society of America.

Dunn, B. W., H. G. Beecher, G. D. Batten and S. Ciavarella (2002). "The potential of near-infrared reflectance spectroscopy for soil analysis - A case study from the Riverine Plain of south-eastern Australia." Australian Journal of Experimental Agriculture **42**(5): 607-614.

Fidêncio, P. H., R. J. Poppi and J. C. de Andrade (2002). "Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy." Analytica Chimica Acta **453**(1): 125-134.

Gomez, C., P. Lagacherie and G. Coulouma (2008a). "Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements." Geoderma **148**(2): 141-148.

Gomez, C., R. A. Viscarra Rossel and A. B. McBratney (2008b). "Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study." Geoderma **146**(3-4): 403-411.

Hunt, G. R. (1977). "Spectral Signatures of Particulate Minerals in Visible and Near IR." Geophysics **42**(3): 501-513.

Kiestra, E. (2002). Bodemkundig-hydrologisch onderzoek voor de waardebepaling van de gronden in het ruilverkavelingsgebied Kromme Rijn. Wageningen, Alterra, Research Instituut voor de Groene Ruimte.

Kooistra, L., J. Wanders, G. F. Epema, R. S. E. W. Leuven, R. Wehrens and L. M. C. Buydens (2003). "The potential of field spectroscopy for the assessment of sediment properties in river floodplains." Analytica Chimica Acta **484**(2): 189-200.

Malley, D. F. and P. C. Williams (1997). "Use of near-infrared reflectance spectroscopy in prediction of heavy metals in freshwater sediment by their association with organic matter." Environmental Science and Technology **31**(12): 3461-3467.

Padmanabhan, K., K. L. Majumder and D. S. Kamat (1980). "Estimation of Areas under Different Cover Types by Spectral Stratification." LARS Symposia. Paper 356.

Palacios-Orueta, A. and S. L. Ustin (1998). "Remote Sensing of Soil Properties in the Santa Monica Mountains I. Spectral Analysis." Remote Sensing of Environment **65**(2): 170-183.

Reeves III, J. B. (2010). "Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done?" Geoderma **158**(1-2): 3-14.

Reeves III, J. B. and G. W. McCarty (2001). "Quantitative analysis of agricultural soils using near infrared reflectance spectroscopy and a fibre-optic probe." Journal of Near Infrared Spectroscopy **9**(1): 25-34.

Reeves III, J. B., G. W. McCarty and J. J. Meisinger (1999). "Near infrared reflectance spectroscopy for the analysis of agricultural soils." Journal of Near Infrared Spectroscopy **7**(3): 179-193.

Reeves, J., G. McCarty and T. Mimmo (2002). "The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soils." Environmental Pollution **116**(Supplement 1): S277-S284.

Rencz, A. N. (1999). Manual of remote sensing: Remote sensing for the earth sciences. New York, Wiley.

Roberts, C. A. and J. Workman (2004). Near-infrared spectroscopy in agriculture. Madison, WI, American Society of Agronomy.

Salgó, A., J. Nagy, J. Tarnóy, P. Marth, O. Pálmai and G. Szabó-Kele (1998). "Characterisation of soils by the near infrared technique." Journal of Near Infrared Spectroscopy **6**(1-4): 199-203.

Scheinost, A. C., A. Chavernas, V. Barron and J. Torrent (1998). "Use and limitations of second-derivative diffuse reflectance spectroscopy in the visible to near-infrared range to identify and quantify Fe oxide minerals in soils." Clays and Clay Minerals **46**(5): 528-536.

Scheinost, A. C. and U. Schwertmann (1999). "Color identification of iron oxides and hydroxysulfates: Use and limitations." Soil Science Society of America Journal **63**(5): 1463-1471.

Shepherd, K. D. and M. G. Walsh (2002). "Development of Reflectance Spectral Libraries for Characterization of Soil Properties." Soil Sci Soc Am J **66**(3): 988-998.

Stevens, A., B. van Wesemael, H. Bartholomeus, D. Rosillon, B. Tychon and E. Ben-Dor (2008). "Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils." Geoderma **144**(1-2): 395-404.

Stiboka (1965). De bodem van Nederland : toelichting bij de bodemkaart van Nederland schaal 1:200.000. Wageningen, Stichting voor Bodemkartering.

Summers, D., M. Lewis, B. Ostendorf and D. Chittleborough (2009). "Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties." Ecological Indicators **In Press, Corrected Proof**.

Udelhoven, T., C. Emmerling and T. Jarmer (2003). "Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study." Plant and Soil **251**(2): 319-329.

van der Meer, F. and S. M. de Jong (2002). Imaging Spectrometry : Basic Principles and Prospective Applications, Kluwer.

van der Veer, G. (2006). Geochemical soil survey of the Netherlands : atlas of major and trace elements in topsoil and parent material; assessment of natural and anthropogenic enrichment factors Ook verschenen als proefschrift Universiteit Utrecht, 2006, Koninklijk Nederlands Aardrijkskundig Genootschap.

van Groenestijn, A. (2009). Quantitative analysis of VNIR spectroscopy based soil property prediction models : finding causes of poor model robustness. Thesis report GIRS;2009-15, Wageningen University.

Vasques, G. M., S. Grunwald and J. O. Sickman (2008). "Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra." Geoderma **146**(1-2): 14-25.

Viscarra Rossel, R. A. (2008). "ParLeS: Software for chemometric analysis of spectroscopic data." Chemometrics and Intelligent Laboratory Systems **90**(1): 72-83.

Workman, J. and A. W. Springsteen (1998). Applied Spectroscopy: a compact reference for practitioners. San Diego, Academic Press.

WUR-Alterra (2006). Data set grondsoortenkaart van Nederland 2006.

Zornoza, R., C. Guerrero, J. Mataix-Solera, K. M. Scow, V. Arcenegui and J. Mataix-Beneyto (2008). "Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils." Soil Biology and Biochemistry **40**(7): 1923-1930.

Appendices

I. Glossary

Accuracy = the closeness to the actual results of values predicted on future unknown samples by a calibration.

Aeolean = related to the activity of the wind, causing local geogenesis, viz. locally formed deposits

Artificial Neural Networks = include many nonlinear technique inspired by the neural connections of the human brain. The feed forward network technique is especially useful in developing quantitative models. Many inputs, initially spectral measurements, are multiplied by weights and summed in a neural network "node'.

Eolian = see aeolean

Fluviatile = deposited by river systems

Kurtosis = Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case (from: Engineering Statistics handbook: *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 24 May 2011).

Leave-one-out method = one sample is systematically left out from each cycle of the regression until all the samples have been excluded once.

Levene's and student *t*-test:
If comparison between the mean values of the calibration and validation set do not show a significant difference between them (according to the student's *t*-test at a 0.05 significance level), this similarity between the calibration and validation set is indicative that the randomly separated validation samples appropriately represents the population under study.

Lithogenesis = soil formation.

Precision = the degree of reproducibility of the result.

RPD = residual prediction deviation = ratio of performance to deviation = standard error of performance / standard deviation of the reference data = standard deviation/RMSEP or standard deviation/RMSECV

Robustness: here refers to the characteristic of a calibration where it can be transferred to other instrument without loss of performance (Roberts and Workman, 2004).

Skewness = The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail (from: Engineering Statistics handbook: *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 24 May 2011).

II. Descriptive statistics

Results of relevant descriptive statistics of the different non-clustered and clustered sets and their corresponding calibration and validation sets can be found in table 25 - 31. Tables are displayed per soil type (SOM, Nt, pH) and per type of stratification method applied.

SOM

Table 25 Descriptive statistics for SOM clusters based on Soil Type (ST).

	ST	ST_Cal	ST_Val	clay	clay_cal	clay_val	peat	peat_cal	peat_val	zavel	zavel_cal	zavel_val	sand	sand_cal	sand_val
Observations	**402**	**269**	**133**	**51**	**34**	**17**	**56**	**38**	**18**	**76**	**51**	**25**	**219**	**146**	**73**
Mean	11.96	11.96	11.96	10.91	10.71	11.32	29.98	30.34	29.23	8.21	8.20	8.25	8.89	8.78	9.12
Median	7.63	7.59	7.66	8.76	8.70	8.99	24.85	24.85	25.20	7.73	7.73	7.73	6.21	6.18	6.24
Variance	204.20	211.82	190.28	45.43	47.50	43.75	567.28	605.78	515.98	12.17	12.68	11.63	119.84	117.66	125.81
Std. Dev.	14.29	14.55	13.79	6.74	6.89	6.61	23.82	24.61	22.72	3.49	3.56	3.41	10.95	10.85	11.22
Minimum	0.34	0.34	0.59	1.61	1.61	2.81	2.60	2.60	4.88	3.28	3.28	3.72	0.34	0.34	0.59
Maximum	95.80	95.80	92.00	36.50	36.50	30.20	95.80	95.80	92.00	20.10	20.10	17.80	72.40	72.40	61.00
Range	95.46	95.46	91.41	34.89	34.89	27.39	93.20	93.20	87.12	16.82	16.82	14.08	72.06	72.06	60.41
Skewness	3.14	3.20	3.02	1.99	2.18	1.77	1.34	1.33	1.47	1.11	1.16	1.06	3.22	3.38	3.01
Kurtosis	11.82	12.20	11.26	4.57	5.84	3.40	1.36	1.28	2.43	1.36	1.64	1.17	12.02	13.66	9.91

Table 26 Descriptive statistics for SOM clusters based on Land Use (LU).

	LU	LU_cal	LU_val	agric	agric_cal	agric_val	forest	forest_cal	forest_val	nature	nature_cal	nature_val
Observations	**396**	**265**	**131**	**60**	**40**	**20**	**152**	**102**	**50**	**184**	**123**	**61**
Mean	12.39	12.46	12.23	12.19	12.13	12.31	9.72	10.01	9.14	14.65	14.61	14.74
Median	7.75	7.73	7.79	8.67	8.59	8.92	7.24	7.24	7.23	7.92	7.84	8.00
Variance	215.85	227.44	193.94	100.02	110.84	83.06	138.20	167.34	80.45	308.54	308.47	313.81
Std. Dev.	14.69	15.08	13.93	10.00	10.53	9.11	11.76	12.94	8.97	17.57	17.56	17.71
Minimum	0.34	0.34	0.59	0.84	0.84	3.04	0.97	0.97	1.11	0.34	0.34	0.59
Maximum	95.80	95.80	92.00	61.00	61.00	40.50	94.70	94.70	52.70	95.80	95.80	92.00
Range	95.46	95.46	91.41	60.16	60.16	37.46	93.73	93.73	51.59	95.46	95.46	91.41
Skewness	2.98	3.00	2.91	2.75	3.02	2.06	4.45	4.47	3.31	2.30	2.30	2.36
Kurtosis	10.53	10.56	10.64	9.67	11.61	4.38	24.32	23.41	12.61	5.93	5.96	6.49

Cal = calibration set; val = validation set.

Table 27 Descriptive statistics for SOM clusters based on Wave-length Based Discriminant Analysis.

	whole set	val	cal	Cat A whole set	Cat A cal	Cat A Val	Cat B whole set	Cat B Cal	Cat B Val	Cat C whole set	Cat C Cal	Cat C Val
Observations	**575**	**384**	**191**	**216**	**133**	**83**	**233**	**137**	**96**	**126**	**114**	**12**
Mean	11.37	11.43	11.26	3.92	2.90	5.57	10.15	7.36	14.12	26.40	26.26	27.78
Median	6.99	6.97	6.99	3.48	3.21	4.18	7.73	7.23	8.68	20.00	19.40	22.60
Variance	199.74	204.59	191.01	21.49	1.97	48.72	113.42	1.80	247.25	343.32	361.35	187.06
Std. Dev.	14.13	14.30	13.82	4.64	1.40	6.98	10.65	1.34	15.72	18.53	19.01	13.68
Minimum	0.24	0.24	0.29	0.24	0.24	0.29	2.30	5.04	2.30	10.20	10.20	14.30
Maximum	95.80	95.80	92.00	52.70	5.00	52.70	92.00	10.00	92.00	95.80	95.80	61.80
Range	95.56	95.56	91.71	52.46	4.76	52.41	89.70	4.96	89.70	85.60	85.60	47.50
Skewness	3.08	3.10	3.06	6.77	-0.33	4.58	4.92	0.20	3.00	1.78	1.79	1.52
Kurtosis	11.19	11.32	11.19	62.01	-1.17	26.69	27.94	-0.91	9.85	3.04	2.99	2.50

Cat A:SOM content < 5%; Cat B: 5%< SOM content < 10%; Cat C: SOM content > 10%.
Cal = calibration set; val = validation set.

N-total

Table 28 Descriptive statistics for N-total clusters based on Soil Type.

	ST	ST_Cal	ST_Val	clay	clay_cal	clay_val	peat	peat_cal	peat_val	zavel	zavel_cal	zavel_val	sand	sand_cal	sand_val
Observations	**402**	**269**	**133**	**51**	**34**	**17**	**56**	**38**	**18**	**76**	**51**	**25**	**219**	**146**	**73**
Mean	3.42	3.39	3.47	3.84	3.77	4.00	8.30	8.34	8.22	3.04	3.03	3.06	2.20	2.15	2.32
Median	2.27	2.25	2.30	2.99	2.98	3.17	8.38	8.38	8.26	2.79	2.75	2.82	1.43	1.43	1.43
Variance	13.75	13.83	13.70	6.15	6.39	6.00	29.17	31.38	26.08	2.13	2.27	1.92	8.10	7.26	9.89
Std. Dev.	3.71	3.72	3.70	2.48	2.53	2.45	5.40	5.60	5.11	1.46	1.51	1.39	2.85	2.69	3.14
Minimum	0.16	0.16	0.20	0.51	0.51	1.23	0.47	0.47	1.22	1.03	1.03	1.31	0.16	0.16	0.20
Maximum	25.40	25.40	21.20	13.00	13.00	11.00	25.40	25.40	21.20	9.36	9.36	7.40	20.20	20.20	19.80
Range	25.24	25.24	21.00	12.49	12.49	9.77	24.93	24.93	19.98	8.33	8.33	6.09	20.04	20.04	19.60
Skewness	2.53	2.61	2.40	1.93	2.09	1.79	0.88	0.95	0.74	1.74	1.90	1.40	3.81	3.87	3.72
Kurtosis	7.93	8.68	6.71	4.17	5.30	3.34	1.20	1.45	0.93	4.72	5.72	2.81	17.73	19.14	16.17

Cal = calibration set; val = validation set.

Table 29 Descriptive statistics for Nt clusters based on Land Use.

	LU	LU_cal	LU_val	agric	agric_cal	agric_val	forest	forest_cal	forest_val	nature	nature_cal	nature_val
Observations	**396**	**265**	**131**	**60**	**40**	**20**	**152**	**102**	**50**	**184**	**123**	**61**
Mean	3.50	3.51	3.47	3.79	3.72	3.93	2.73	2.80	2.58	4.04	4.04	4.06
Median	2.26	2.25	2.29	2.84	2.82	2.85	1.96	1.96	1.96	2.16	2.15	2.17
Variance	14.59	15.34	13.18	8.30	8.61	8.06	10.55	12.51	6.69	19.29	19.37	19.46
Std. Dev.	3.82	3.92	3.63	2.88	2.93	2.84	3.25	3.54	2.59	4.39	4.40	4.41
Minimum	0.16	0.16	0.20	0.26	0.26	0.71	0.24	0.24	0.28	0.16	0.16	0.20
Maximum	25.40	25.40	20.40	16.00	16.00	11.30	25.40	25.40	14.70	21.20	21.20	20.40
Range	25.24	25.24	20.20	15.74	15.74	10.59	25.16	25.16	14.42	21.04	21.04	20.20
Skewness	2.38	2.49	2.11	2.15	2.44	1.67	4.19	4.26	3.17	1.65	1.66	1.68
Kurtosis	6.72	7.44	4.81	5.50	7.56	2.25	21.83	21.71	11.53	2.54	2.61	2.71

Cal = calibration set; val = validation set.

pH

Table 30 Descriptive statistics for pH clusters based on Soil Type.

	ST	ST_Cal	ST_Val	clay	clay_cal	clay_val	peat	peat_cal	peat_val	zavel	zavel_cal	zavel_val	sand	sand_cal	sand_val
Observations	**402**	**269**	**133**	**51**	**34**	**17**	**56**	**38**	**18**	**76**	**51**	**25**	**219**	**146**	**73**
Mean	6.17	6.16	6.20	7.68	7.64	7.75	5.31	5.31	5.32	7.93	7.92	7.96	5.43	5.42	5.45
Median	5.70	5.70	5.72	8.14	8.14	8.14	5.42	5.42	5.42	8.06	8.06	8.06	4.88	4.87	4.89
Variance	2.85	2.86	2.87	1.11	1.23	0.93	0.56	0.63	0.44	0.26	0.28	0.21	2.41	2.39	2.48
Std. Dev.	1.69	1.69	1.69	1.06	1.11	0.96	0.75	0.80	0.66	0.51	0.53	0.46	1.55	1.55	1.57
Minimum	3.62	3.62	3.74	4.16	4.16	4.72	3.62	3.62	4.02	6.20	6.20	6.52	3.70	3.70	3.74
Maximum	9.11	9.11	9.10	8.66	8.66	8.46	7.63	7.63	6.74	8.51	8.51	8.50	9.11	9.11	9.10
Range	5.49	5.49	5.36	4.50	4.50	3.74	4.01	4.01	2.72	2.31	2.31	1.98	5.41	5.41	5.36
Skewness	0.09	0.10	0.08	-2.13	-2.08	-2.42	0.12	0.21	-0.20	-1.94	-1.93	-2.03	0.93	0.94	0.93
Kurtosis	-1.58	-1.57	-1.59	3.77	3.54	6.05	1.38	1.57	0.66	3.43	3.34	4.33	-0.45	-0.43	-0.45

Cal = calibration set; val = validation set.

Table 31 Descriptive statistics for pH clusters based on Land Use.

	LU	LU_cal	LU_val	agric	agric_cal	agric_val	forest	forest_cal	forest_val	nature	nature_cal	nature_val
Observations	**396**	**265**	**131**	**60**	**40**	**20**	**152**	**102**	**50**	**184**	**123**	**61**
Mean	6.01	6.00	6.03	6.16	6.13	6.24	6.01	6.01	6.01	5.97	5.96	5.99
Median	5.55	5.54	5.58	5.59	5.57	5.64	5.53	5.53	5.55	5.53	5.52	5.53
Variance	2.66	2.68	2.65	1.60	1.65	1.59	3.60	3.63	3.62	2.25	2.26	2.26
Std. Dev.	1.63	1.64	1.63	1.27	1.28	1.26	1.90	1.90	1.90	1.50	1.50	1.50
Minimum	3.62	3.62	3.74	3.89	3.89	4.63	3.70	3.70	3.74	3.62	3.62	3.83
Maximum	9.11	9.11	9.10	8.32	8.32	8.27	8.51	8.51	8.36	9.11	9.11	9.10
Range	5.49	5.49	5.36	4.43	4.43	3.64	4.81	4.81	4.62	5.49	5.49	5.27
Skewness	0.27	0.27	0.27	0.40	0.37	0.50	0.11	0.11	0.11	0.52	0.52	0.54
Kurtosis	-1.43	-1.43	-1.44	-1.28	-1.23	-1.49	-1.83	-1.84	-1.88	-1.00	-1.00	-0.99

Cal = calibration set; val = validation set.

III. Normality tests results

For the normality test the null- hypothesis to be tested was:
H_0: the data follow a normal distribution
H_1: the data do not follow a normal distribution.

The results of the normality test should be interpreted as follows:
Reject the null-hypothesis (i.e. accept H_1) if $p < 0.05$.

Table 32 Normality test for all soil clusters for SOM.

		Kolmogorov-Smirnov(a)			Shapiro-Wilk			
SOM	**Set**	**Statistic**	**df**	**Sig.**	**Statistic**	**df**	**Sig.**	**Normal distribution**
NON-CLUSTERED SOIL TYPE	**all**	0.250	402	0.000	0.640	402	0.000	no
	cal	0.254	269	0.000	0.631	269	0.000	no
	val	0.247	133	0.000	0.660	133	0.000	no
CLAY	**all**	0.241	51	0.000	0.790	51	0.000	no
	cal	0.248	34	0.000	0.777	34	0.000	no
	val	0.249	17	0.006	0.812	17	0.003	no
PEAT	**all**	0.151	56	0.003	0.861	56	0.000	no
	cal	0.155	38	0.022	0.861	38	0.000	no
	val	0.146	18	0.200	0.871	18	0.018	no
ZAVEL	**all**	0.118	76	0.011	0.924	76	0.000	no
	cal	0.119	51	0.068	0.922	51	0.002	no
	val	0.127	25	0.200	0.929	25	0.081	yes
SAND	**all**	0.239	219	0.000	0.628	219	0.000	no
	cal	0.239	146	0.000	0.624	146	0.000	no
	val	0.242	73	0.000	0.635	73	0.000	no
NON-CLUSTERED LAND USE	**all**	0.246	396	0.000	0.655	396	0.000	no
	cal	0.250	265	0.000	0.647	265	0.000	no
	val	0.244	131	0.000	0.676	131	0.000	no
AGRIC	**all**	0.238	60	0.000	0.709	60	0.000	no
	cal	0.246	40	0.000	0.680	40	0.000	no
	val	0.224	20	0.010	0.767	20	0.000	no
FOREST	**all**	0.300	152	0.000	0.516	152	0.000	no
	cal	0.308	102	0.000	0.494	102	0.000	no
	val	0.273	50	0.000	0.618	50	0.000	no
NATURE	**all**	0.208	184	0.000	0.729	184	0.000	no
	cal	0.208	123	0.000	0.730	123	0.000	no
	val	0.212	61	0.000	0.729	61	0.000	no

Table 33 Normality test for all soil clusters for Nt.

		Kolmogorov-Smirnov(a)			Shapiro-Wilk			
Nt	**Set**	**Statistic**	**df**	**Sig.**	**Statistic**	**df**	**Sig.**	**Normal distribution**
NON-CLUSTERED SOIL TYPE	all	0.219	402	0.000	0.719	402	0.000	no
	cal	0.219	269	0.000	0.714	269	0.000	no
	val	0.227	133	0.000	0.729	133	0.000	no
CLAY	all	0.208	51	0.000	0.801	51	0.000	no
	cal	0.215	34	0.000	0.792	34	0.000	no
	val	0.221	17	0.027	0.813	17	0.003	no
PEAT	all	0.074	56	0.200	0.934	56	0.004	no
	cal	0.088	38	0.200	0.930	38	0.020	no
	val	0.087	18	0.200	0.946	18	0.363	yes
ZAVEL	all	0.112	76	0.020	0.872	76	0.000	no
	cal	0.121	51	0.059	0.857	51	0.000	yes
	val	0.119	25	0.200	0.901	25	0.019	no
SAND	all	0.236	219	0.000	0.584	219	0.000	no
	cal	0.230	146	0.000	0.599	146	0.000	no
	val	0.252	73	0.000	0.570	73	0.000	no
NON-CLUSTERED LAND USE	all	0.218	396	0.000	0.728	396	0.000	no
	cal	0.222	265	0.000	0.718	265	0.000	no
	val	0.216	131	0.000	0.752	131	0.000	no
AGRIC	all	0.203	60	0.000	0.777	60	0.000	no
	cal	0.209	40	0.000	0.757	40	0.000	no
	val	0.209	20	0.022	0.798	20	0.001	no
FOREST	all	0.289	152	0.000	0.551	152	0.000	no
	cal	0.299	102	0.000	0.528	102	0.000	no
	val	0.258	50	0.000	0.639	50	0.000	no
NATURE	all	0.188	184	0.000	0.798	184	0.000	no
	cal	0.189	123	0.000	0.798	123	0.000	no
	val	0.191	61	0.000	0.800	61	0.000	no

Table 34 Normality test for all soil clusters for pH.

		Kolmogorov-Smirnov(a)			Shapiro-Wilk			
pH	**Set**	**Statistic**	**df**	**Sig.**	**Statistic**	**df**	**Sig.**	**Normal distribution**
NON-CLUSTERED SOIL TYPE	all	0.158	402	0.000	0.884	402	0.000	no
	cal	0.159	269	0.000	0.884	269	0.000	no
	val	0.163	133	0.000	0.883	133	0.000	no
CLAY	all	0.252	51	0.000	0.676	51	0.000	no
	cal	0.255	34	0.000	0.682	34	0.000	no
	val	0.266	17	0.002	0.674	17	0.000	no
PEAT	all	0.113	56	0.074	0.950	56	0.021	no
	cal	0.128	38	0.121	0.943	38	0.052	yes
	val	0.126	18	0.200	0.964	18	0.684	yes
ZAVEL	all	0.230	76	0.000	0.765	76	0.000	no
	cal	0.240	51	0.000	0.763	51	0.000	no
	val	0.228	25	0.002	0.781	25	0.000	no
SAND	all	0.148	219	0.000	0.852	219	0.000	no
	cal	0.149	146	0.000	0.851	146	0.000	no
	val	0.151	73	0.000	0.853	73	0.000	no
NON-CLUSTERED LAND USE	all	0.140	396	0.000	0.895	396	0.000	no
	cal	0.137	265	0.000	0.895	265	0.000	no
	val	0.146	131	0.000	0.894	131	0.000	no
AGRIC	all	0.213	60	0.000	0.880	60	0.000	no
	cal	0.205	40	0.000	0.889	40	0.001	no
	val	0.237	20	0.005	0.855	20	0.006	no
FOREST	all	0.208	152	0.000	0.788	152	0.000	no
	cal	0.211	102	0.000	0.789	102	0.000	no
	val	0.210	50	0.000	0.786	50	0.000	no
NATURE	all	0.137	184	0.000	0.916	184	0.000	no
	cal	0.138	123	0.000	0.917	123	0.000	no
	val	0.145	61	0.003	0.914	61	0.000	no

Table 35 Normality test for all soil clusters for SOM for Wavelength based clustering.

		Tests of Normality						
		Kolmogorov-Smirnov(a)			Shapiro-Wilk			
		Statistic	df	Sig.	Statistic	df	Sig.	Normal distribution
Non-clustered	All	0.247	575	0.000	0.637	575	0.000	no
	Cal	0.248	384	0.000	0.634	384	0.000	no
	Val	0.245	191	0.000	0.643	191	0.000	no
Cat A (<5%)	All	0.269	216	0.000	0.476	216	0.000	no
	Cal	0.106	133	0.001	0.935	133	0.000	no
	Val	0.260	83	0.000	0.552	83	0.000	no
Cat B (5-10%)	All	0.353	233	0.000	0.430	233	0.000	no
	Cal	0.057	137	0.200	0.969	137	0.003	no
	Val	0.276	96	0.000	0.611	96	0.000	no
Cat C (>10%)	All	0.191	126	0.000	0.788	126	0.000	no
	Cal	0.199	114	0.000	0.778	114	0.000	no
	Val	0.203	12	0.184	0.853	12	0.040	no

IV. Levene's test and student's *t* test results

For the Levene's test of equal variance the null- hypothesis to be tested was:
H_0: both data sets have equal variances
H_1: both data sets do not have equal variances

The results of the Levene's test of equal variances should be interpreted as follows:
Reject the null-hypothesis (i.e. accept H_1) if $p < 0.05$.

For the student *t*-test the null- hypothesis to be tested was:
H_0: the validation and calibration set represent the population under study (i.e. they are both representative).
H_1: the validation and calibration set do not represent the population under study (i.e. they are not representative).

The results of the student *t*-test should be interpreted as follows:
Reject the null-hypothesis (i.e. accept H_1) if $p < 0.05$.

Table 36 Levene's and student *t*-test results for soil clusters for N-total.

		Levene's		t-test			Repres-
NT		**F**	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**entative?**
Non-clustered Soil type	Equal variances assumed	0.014	0.906	-0.191	400	0.849	yes
	Equal variances not assumed			-0.191	264.215	0.849	
Clay	Equal variances assumed	0.009	0.925	-0.310	49	0.758	yes
	Equal variances not assumed			-0.314	33.022	0.756	
Peat	Equal variances assumed	0.125	0.725	0.081	54	0.935	yes
	Equal variances not assumed			0.084	36.457	0.933	
Zavel	Equal variances assumed	0.029	0.865	-0.075	74	0.940	yes
	Equal variances not assumed			-0.077	51.540	0.939	
Sand	Equal variances assumed	0.404	0.526	-0.417	217	0.677	yes
	Equal variances not assumed			-0.396	126.107	0.692	
Non-clustered Land Use	Equal variances assumed	0.111	0.739	0.099	394	0.921	yes
	Equal variances not assumed			0.101	277.412	0.919	
Agricultural land	Equal variances assumed	0.038	0.847	-0.263	58	0.793	yes
	Equal variances not assumed			-0.266	39.271	0.791	
Forest	Equal variances assumed	0.457	0.500	0.398	150	0.691	yes
	Equal variances not assumed			0.442	127.907	0.659	
Nature	Equal variances assumed	0.001	0.980	-0.030	182	0.976	yes
	Equal variances not assumed			-0.030	119.542	0.976	

Table 37 Levene's and student *t*-test results for soil clusters for SOM.

SOM		Levene's		t-test			Repres-entative?
		F	Sig.	t	df	Sig. (2-tailed)	
Non-clustered Soil Type	Equal variances assumed	0.031	0.861	0.001	400	0.999	yes
	Equal variances not assumed			0.001	276.093	0.999	
Clay	Equal variances assumed	0.013	0.909	-0.300	49	0.765	yes
	Equal variances not assumed			-0.305	33.324	0.763	
Peat	Equal variances assumed	0.151	0.699	0.162	54	0.872	yes
	Equal variances not assumed			0.167	36.044	0.868	
Zavel	Equal variances assumed	0.032	0.858	-0.059	74	0.953	yes
	Equal variances not assumed			-0.060	49.699	0.952	
Sand	Equal variances assumed	0.084	0.772	-0.215	217	0.830	yes
	Equal variances not assumed			-0.213	139.890	0.832	
Non-clustered Land Use	Equal variances assumed	0.179	0.673	0.146	394	0.884	yes
	Equal variances not assumed			0.150	278.359	0.881	
Agricultural land	Equal variances assumed	0.017	0.897	-0.065	58	0.948	yes
	Equal variances not assumed			-0.069	43.400	0.946	
Forest	Equal variances assumed	0.623	0.431	0.424	150	0.672	yes
	Equal variances not assumed			0.479	132.855	0.633	
Nature	Equal variances assumed	0.000	0.996	-0.049	182	0.961	yes
	Equal variances not assumed			-0.048	118.866	0.961	
Non-clustered (WBDA)	Equal variances assumed	0.070	0.792	0.130	573	0.897	yes
	Equal variances not assumed			0.131	391.277	0.896	
Cat A (<5%)	Equal variances assumed	21.578	0.000	-4.281	214	0.000	
	Equal variances not assumed			-3.442	86.164	0.001	no
Cat B (5-10%)	Equal variances assumed	67.518	0.000	-5.011	231	0.000	
	Equal variances not assumed			-4.202	95.972	0.000	no
Cat C (>10%)	Equal variances assumed	1.023	0.314	-0.269	124	0.789	yes
	Equal variances not assumed			-0.350	15.865	0.731	

** WBDA: wavelength based discriminant analysis. This non-clustered set contains all samples which were divided into the three clusters Cat A, Cat B and Cat C.

Table 38 Levene's and student *t*-test results for soil clusters for pH.

PH		Levene's		t-test			Repres-entative?
		F	Sig.	t	df	Sig. (2-tailed)	
Non-clustered Soil Type	Equal variances assumed	0.005	0.946	-0.233	400	0.816	yes
	Equal variances not assumed			-0.233	262.571	0.816	
Clay	Equal variances assumed	0.322	0.573	-0.348	49	0.729	yes
	Equal variances not assumed			-0.365	36.528	0.717	
Peat	Equal variances assumed	0.306	0.582	-0.025	54	0.980	yes
	Equal variances not assumed			-0.027	39.804	0.979	
Zavel	Equal variances assumed	0.354	0.554	-0.334	74	0.739	yes
	Equal variances not assumed			-0.350	54.074	0.727	
Sand	Equal variances assumed	0.031	0.861	-0.159	217	0.874	yes
	Equal variances not assumed			-0.158	141.977	0.875	
Non-clustered Land Use	Equal variances assumed	0.007	0.933	-0.179	394	0.858	yes
	Equal variances not assumed			-0.179	260.312	0.858	
Agricultural land	Equal variances assumed	0.001	0.979	-0.331	58	0.742	yes
	Equal variances not assumed			-0.333	38.723	0.741	
Forest	Equal variances assumed	0.007	0.931	-0.010	150	0.992	yes
	Equal variances not assumed			-0.010	97.471	0.992	
Nature	Equal variances assumed	0.000	0.994	-0.110	182	0.912	yes
	Equal variances not assumed			-0.110	119.721	0.912	

V. Statistics of calibration and validation data (Lithological stratification)

Table 39 Calibration and validation statistics for SOM for soil clusters.

				Calibration statistics			Validation statistics		
SOM	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**R^2**	**RMSE**	**RPD**
Non-clustered Soil Type	296	133	12	0.559	9.675	1.50	0.63	8.412	1.64
Clay (1)	34	17	4	0.584	4.418	1.56	0.81	3.587	1.84
Peat (2)	38	18	9	0.842	9.706	2.54	0.82	9.989	2.27
Zavel (3)	51	25	3	0.350	2.859	1.25	0.31	2.889	1.18
Sand (4)	146	73	7	0.302	9.042	1.20	0.48	8.235	1.36

Table 40 Calibration and validation statistics for Nt for soil clusters.

				Calibration statistics			Validation statistics		
N-total	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**R^2**	**RMSE**	**RPD**
Non-clustered Soil Type	269	133	15	0.67	2.143	1.74	0.70	2.047	1.81
Clay (1)	34	17	4	0.593	1.599	1.58	0.76	1.234	1.99
Peat (2)	38	18	8	0.409	4.329	1.29	0.63	3.028	1.69
Zavel (3)	51	25	3	0.040	1.497	1.01	0.29	1.148	1.21
Sand (4)	146	73	7	0.601	1.706	1.58	0.67	0.868	1.68

Table 41 Calibration and validation statistics for pH for soil clusters.

				Calibration statistics			Validation statistics		
pH	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**R^2**	**RMSE**	**RPD**
Non-clustered Soil Type	269	133	16	0.869	0.618	2.74	0.86	0.684	2.48
Clay (1)	34	17	5	0.390	0.911	1.22	0.67	0.594	1.62
Peat (2)	38	18	6	0.119	0.951	0.84	0.44	0.535	1.23
Zavel (3)	51	25	17	0.557	0.357	1.48	0.23	0.464	1.00
Sand (4)	146	73	12	0.818	0.669	2.31	0.78	0.782	2.01

VI. Statistics of calibration and validation data (Land Use stratification)

Table 42 Calibration and validation statistics for SOM for land use clusters.

				Calibration statistics			Validation statistics		
SOM	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**R^2**	**RMSE**	**RPD**
Non-clustered Land Use	265	131	17	0.778	7.090	2.13	0.81	6.214	2.24
Agricultural	40	20	7	0.664	6.048	1.74	0.59	8.500	1.07
Forest	102	50	7	0.566	8.530	1.52	0.82	4.337	2.07
Nature	123	61	16	0.823	7.363	2.39	0.82	7.554	2.35

Table 43 Calibration and validation statistics for Nt for land use clusters.

				Calibration statistics			Validation statistics		
Ntotal	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**R^2**	**RMSE**	**RPD**
Non-clustered Land Use	265	131	9	0.647	2.325	1.68	0.70	2.020	1.80
Agricultural	40	20	10	0.552	1.965	1.49	0.69	1.670	1.70
Forest	102	50	7	0.333	2.912	1.21	0.90	1.261	2.05
Nature	123	61	15	0.783	2.047	2.15	0.65	2.834	1.56

Table 44 Calibration and validation statistics for pH for land use clusters.

				Calibration statistics			Validation statistics		
pH	**N cal**	**N val**	**# factors**	**R^2**	**RMSE**	**RPD**	**R^2**	**RMSE**	**RPD**
Non-clustered Land Use	265	131	14	0.846	0.650	2.52	0.85	0.645	2.52
Agricultural	40	20	9	0.752	0.668	1.92	0.84	0.531	2.37
Forest	102	50	12	0.838	0.792	2.40	0.91	0.605	3.15
Nature	123	61	14	0.812	0.665	2.26	0.78	0.784	1.92

VII. Statistics of calibration and validation data (based on spectral data stratification)

Table 45 Calibration and validation statistics for SOM for WBDA-clusters.

SOM				**Calibration statistics**			**Validation statistics**		
	n.cal	**n.val**	**# factors**	**R^2**	**RMSE**	**RPD**	**R2**	**RMSE**	**RPD**
Non-clustered data	**384**	**191**	13	0.762	6.979	2.05	0.788	6.581	2.10
A: <5%	133	83	9	0.796	0.635	2.21	0.567	6.090	1.15
B: 5-10%	137	96	8	0.259	1.186	1.13	0.13	16.701	0.94
C: >10%	114	12	8	0.692	10.516	1.81	0.791	6.885	1.99
Combined	**384**	**191**	-	0.836	5.7855	2.47	0.249	12.620	1.095
Group I*: < 5%	**132**	**66**	9	00.791	0.640	2.19	0.776	0.700	1.99
Group II*: 5-10%	**137**	**68**	8	0.239	1.197	1.12	0.247	1.176	1.13
Group III*: 10-20%	**59**	**30**	3	0.023	2.687	0.95	0.061	2.645	1.01
Group IV*: 20-40%	**38**	**18**	2	0.046	5.959	1.03	0.042	5.836	1.03
Group V*: > 40%	**18**	**9**	2	0.229	14.768	1.13	0.427	13.736	1.21
Combined	**384**	**191**	-	0.924	3.94	3.63	0.931	3.72	3.71