Centre for Geo-Information
Thesis Report GIRS-2013-05

Mapping Plant Indicator Values using Airborne LiDAR and Hyperspectral Imagery

Jorg J. van Amerongen

March, 2013

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Mapping Plant Indicator Values using Airborne LiDAR and Hyperspectral Imagery

Jorg J. van Amerongen

Registration number: 891001 014 020

Supervisors:
Hans Roelofsen, MSc
Dr.ir. Lammert Kooistra

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

March, 2013
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2012-05
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Foreword

I searched for a topic using remote sensing for an application in the natural environment. This research combines remote sensing and GIS to predict environmental factors with statistical methods. This method offers a lot of opportunities and possibilities for other predictions with different input data and in different areas.

My thesis process went well, since I achieved my objectives and most of the planned deadlines. However, the writing of this report took me more time than expected. The thesis remained interesting for me through the half year, because many different steps were involved containing different programs and methods. Therefore, the six months were over before I realised, but I am satisfied with the results and I enjoyed the time working on this project.

Thanks to Hans Roelofsen and Lammert Kooistra for the supervision, clear feedback and many suggestions throughout this research. Thanks to the students in the thesis room who helped me when needed. Finally, thanks to Natuurmonumenten for access to the area for the measurements of vegetation plots.

Jorg,

Wageningen, March 2013

Summary

In this research, plant indicator values (IVs) were predicted for herbaceous vegetation in the nature reserve Kampina, Noord-Brabant, the Netherlands. The IVs are an indication of soil properties and are determined based on plant species occurrence in an environmental condition. The relevant IVs for the study area are moisture regime (mF), nutrient availability (mN) and acidity (mR). Remote sensing can be used to map these IVs based on the spectral signature of vegetation. The objective is to model the spatial variation of IVs by using hyperspectral and LiDAR data. Current estimations of IVs are too inaccurate for subsequent use models that predict vegetation types, so adding LiDAR variables offers potential to improve the model accuracies. For the LiDAR variables, AHN-2 data is used. The hyperspectral predictors are derived from APEX images. Furthermore, a database of 40 sampled vegetation plots are available for which the average IVs were calculated. Partial Least Squares (PLS) regression was used to find the relation between these IVs and hyperspectral and LiDAR variables. The accuracy was determined based on the R^2 of Leave One Out (LOO) validation. The R^2 of validation is 0.654, 0.515 and 0.822 for mF, mN and mR respectively. The mN and mR models were based on spectral predictors, since no LiDAR predictors remained significant in the model. For the mN model, many wavelengths are significant including a large region from 800 to 1300 nm. The mR model results in fewer significant wavelengths including the region from 1035 to 1100 nm. The LiDAR data did improve the model prediction for mF, since a R^2 of 0.491 was found when only spectral predictors were used and 0.721 was found when LiDAR data is added. Elevation and vegetation height were significant LiDAR predictors for the mF model. The significant wavelengths for mF are usually located in higher wavelengths then for mN and mR and include the regions from 1485 to 1635 nm and 2285 to 2370 nm.

Inaccuracies occur in many steps, including the averaging of IVs over a vegetation plot and the inaccuracy of the relation between the spectral signature of vegetation and the IVs. Moreover, PLSR brings some inaccuracies, since it is sensitive for some settings and the input variables used. Nonetheless, relatively accurate estimations of IVs were found which can be used for nature management and vegetation type predictions. The relation found in this research can be used to predict IVs in other areas where the same vegetation types occur; otherwise new samples and analysis have to be performed.

Keywords: Herbaceous vegetation | Plant Indicator Values | PLSR | Hyperspectral images | LiDAR

Table of Contents

1. Introduction

1.1 Background

Remote sensing offers opportunities for vegetation mapping, which can support nature conservation. Airborne photography was used to map vegetation in natural areas, but is time and costs intensive (Schmidtlein and Sassin 2004). Verrelst et al. (2012) used remote sensing to map vegetation density by deriving LAI in floodplains. This is important, since the vegetation hinders the water flow into the floodplains, which increase flood risk. Colgan et al. (2012) used remote sensing to predict tree species by deriving multiple variables from hyperspectral and LiDAR data in savannahs to gain insight in these ecosystems. Furthermore, biochemical properties of plants have been successfully mapped using remote sensing, like nitrogen, water and cellulose (Kokaly, Asner et al. 2009). Remote sensing can also be used to estimate soil properties from bare ground (Mulder, de Bruin et al. 2011). Therefore, remote sensing offers many opportunities to derive valuable information for nature management. Though, when information about abiotic factors is required and the soil is covered by vegetation, additional data is needed. This can be field samples, but is time and costs intensive. Another approach is to use plant Indicator Values (IVs) to derive information about soil properties. Individual plant species root and grow optimally within a certain intensity range of several abiotic conditions (e.g., moisture, pH, nutrient availability, salinity). The optimum of a species towards an abiotic factor is marked by IVs, which are measures for the preference of plant species (Diekmann 2003). By using the relation of IVs as an indication of soil properties with spectral signatures of plant species, there is an opportunity of using remote sensing to map soil properties. In this way, field measurements can be replaced and data are obtained faster and cheaper. Maps of IVs contain important ecological information and are also valuable for nature conservation and management. Management strategies can be improved when the underlying abiotic factors that influence the occurrence of plant species, are known. Also, managers can assess their management based on monitoring IVs and they may adjust their management strategy. Furthermore, IV maps can be used as input for models that predict occurrence of vegetation types. For example, IVs are used as input data for the PRObability Based Ecological target assessment (PROBE) model. This model calculates the occurrence probability of plant communities; specific and a priori defined collection of plant species (Witte, De Haan et al. 2007). Witte et al. (2007) finds that on average, the PROBE model classifies 85% of the plots to the correct vegetation type (Witte, De Haan et al. 2007). Altogether, maps of IV estimates can be valuable for modellers and managers.

1.2 Problem definition

Full coverage maps of IVs can be made by using remote sensing. Partial Least Squares Regression (PLSR) has been used in various studies (Schmidtlein and Sassin 2004; Schmidtlein 2005; Witte, Wojcik et al. 2007)) as a method to determine the relation between hyperspectral reflectance data and IVs. This method is often used in applications that deal with numerous linearly dependent predictor variables. Roelofsen et al. (submitted) mapped IVs from hyperspectral airborne data and used the maps to predict the occurrence of vegetation associations. The accuracy of the resulting maps was not sufficient, because both steps contained uncertainties (Roelofsen, Kooistra et al. Submitted). The prediction of moisture (mF) was the weakest and may be better predicted by adding terrain variables to the hyperspectral predictors.

This research focuses on the issue that current estimations of IVs and resulting IV maps from hyperspectral data are too inaccurate for subsequent use in models. Therefore, there is a need for more accurate IV predictions. The occurrence of plant species might also be influenced by terrain features, which are often overlooked in studies for IV predictions. The reflectance of vegetation is dependent on the health status and the physical- and chemical structure of vegetation (Schmidtlein and Sassin 2004). These properties are only partly correlated with plant species. When elevation and terrain data from Light Detection And Ranging (LiDAR) is added, it may improve the accuracy of the prediction models. For example, higher moisture content is expected in lower elevated areas compared to higher elevated areas. The study of Schmidt et al. (2004) employed terrain variables to predict vegetation types in coastal wetlands. Height, slope, aspect and terrain position (gully, midslope, ridge or flat) were added to the hyperspectral predictors. The resulting vegetation map had an accuracy of 66%, compared to 40% when only hyperspectral variables were used. Similarly, Ecker et al. (2010) combined 70 optical spectral predictors (400-900 nm), which were based on only four spectral bands from where derivatives were calculated, with LiDAR variables to predict gradients of IVs in wetlands by using PLSR. About 30 different topographic variables were added, including altitude, slope, curvature, eastness, northness and several measures for sediment transport, erosion and wetness. Adding LiDAR data improved the prediction of all IVs. Moreover, the LiDAR variables performed better than spectral variables for predicting mF. This was determined by using the squared Pearson's product–moment coefficient (Ecker, Waser et al. 2010). The possibilities for using LiDAR data will be analysed in this research. Such data has to be prepared in a different way, since the amount, units and meaning of the variables are different compared to the hyperspectral data.

1.3 Objective and Research Questions

The objective of this research is to model the spatial variation of plant indicator values by using hyperspectral and LiDAR data.

The following research questions are based on the objective:

1. How can the IVs for moisture, nutrients and pH be mapped using hyperspectral and LiDAR data?
2. Which prediction variables can be extracted from LiDAR data that may improve the prediction of IVs?
3. Does the use of LiDAR data improve the accuracy of IV predictions compared to predictions using only hyperspectral data?
4. Which LiDAR and hyperspectral variables are significant when predicting moisture, nutrients, and pH?
5. Are the maps of IVs representative for the area?

The hypothesis is that hyperspectral remote sensing data can be used to predict plant indicator values and adding LiDAR data will improve the accuracy of the prediction models.

1.4 Thesis outline

Chapter 2 provides a literature review about the IVs and introduces the method applied for this research. Furthermore, it outlines what has been done in previous studies to compare the results in the end. Next, the study area, data and data pre-processing are described in chapter 3. Subsequently, the methodology of the analysis is explained (chapter 4), followed by the results and discussion in chapter 5 and 6 respectively. Finally chapter 7 contains the conclusions of this research and some recommendations for further studies are given in chapter 8.

2. Literature Review

2.1 Indicator Values

Plant indicator values of Central Europe are defined by Ellenberg (1991), who distinguished seven IVs: light, temperature, continentality, nutrients, soil moisture, pH and salinity. Ellenberg IVs are plant functional response types. Species are ranked according to their optima of occurrence in environmental factors. In addition to Ellenberg, other IV systems exist, which focus on a specific geographic area. A specific IV system for the Netherlands is designed by Witte et al. (2008) and is based on the ecotypes system of the Netherlands (Runhaar, van Landuyt et al. 2004). The relevant IVs for the Netherlands are salinity (mS), moisture regime (mF), nutrient availability (mN) and acidity (mR) (Witte and Kooistra 2008). IVs from plant species are correlated with soil properties, though they are vegetation attributes and not environmental attributes. The average IVs of a location will reflect the conditions better than the values of individual species, because plant species also occur in areas that deviate from their optimum (Diekmann 2003; Kafer and Witte 2004; Zelený and Schaffers 2012). These averaged IVs are mapped in this research. When a plant species occurs in a wide IV range, the average IV becomes less meaningful. To illustrate the concept of IVs, the IVs for common heather *(Calluna vulgari)* are given in Table 1. From these values we deduce that this species occurs in dry, nutrient poor and acidic soils. The system of Witte et al. (2008) bases the calculation IVs of a plant species on the percentage of occurrence in the different factor classes. A large database of vegetation plots is used to determine the IVs per plant species. For moisture regime, the indicator values for the classes are: 1 = open water, 2 = wet, 3 = moist, 4 = dry. The mF indicator value of for example Kleine Bevernel *(Pimpinella saxifraga)* was calculated as the sum of the product of the class-values weighted by the fraction of species occurrence in each class: 1×0.000 + 2×0.000 + 3×0.578 + 4×0.432 = 3.42 (Witte and Kooistra 2008). In general terms, mF is calculated as in formula 1, where F indicates the fraction of occurrence for the specific value of mF. IVs for mN and mR are calculated similarly, but with only three classes.

$$mF = 1 * F_{water} + 2 * F_{wet} + 3 * F_{moist} + 4 * F_{dry} \qquad (1)$$

Cover-weighted averaging of IVs in a vegetation plot can be performed to test if the accuracy improves when using the cover percentage of the species as a weight for calculating the average IVs (Diekmann 2003). This method did not lead to a statistically better result compared to averaging the IVs independent of their cover percentage (Schaffers and Sýkora 2000; Kafer and Witte 2004; Klaus, Kleinebecker et al. 2012). Moreover, many requirements of the sample plots are needed in order to use the cover-weighted averaging method (Diekmann 2003).

Table 1. Plant Indicator values for common heather *(Calluna vulgari)*. The arrows indicate the values for the IVs for this plant species (Witte and Kooistra 2008). This species occurs in dry, nutrient poor and acidic soils.

mF	Open water	1	2	3	↓ 4	Dry soils
mN	Nutrient Poor	↓ 1	2		3	Nutrient Rich
mR	Acid	↓ 1	2		3	Alkaline

IVs are an indication of the environmental conditions, so they relate with soil properties. The indicator values of mF and mN were compared to field measured variables related to moisture and nutrients (Schaffers and Sýkora 2000). mF contained high correlation with average groundwater level and the average lowest moisture content (correlation between 0.84 and 0.88). mN shows highest correlation with N-accumulation and biomass production (0.83 and 0.82 respectively) (Schaffers and Sýkora 2000). Figure 1 shows the relation between the calculated average mR (based on 40 vegetation plots) and the field measured pH for which mR is an indication in the same 40 plots. The Pearson's correlation coefficient (r) is 0.76. The deviations are dependent on the different vegetation types, since they are clustered in the scatterplot. In general, more wet vegetation types have relatively low mR values in relation with field measured pH. Furthermore, a RMSE of 0.81 pH units is found by (Wamelink, Goedhart et al. 2005) when predicting the soil pH based on mR. These results indicate that the relation between plant IVs and actual soil properties contain uncertainties, but is reasonable.

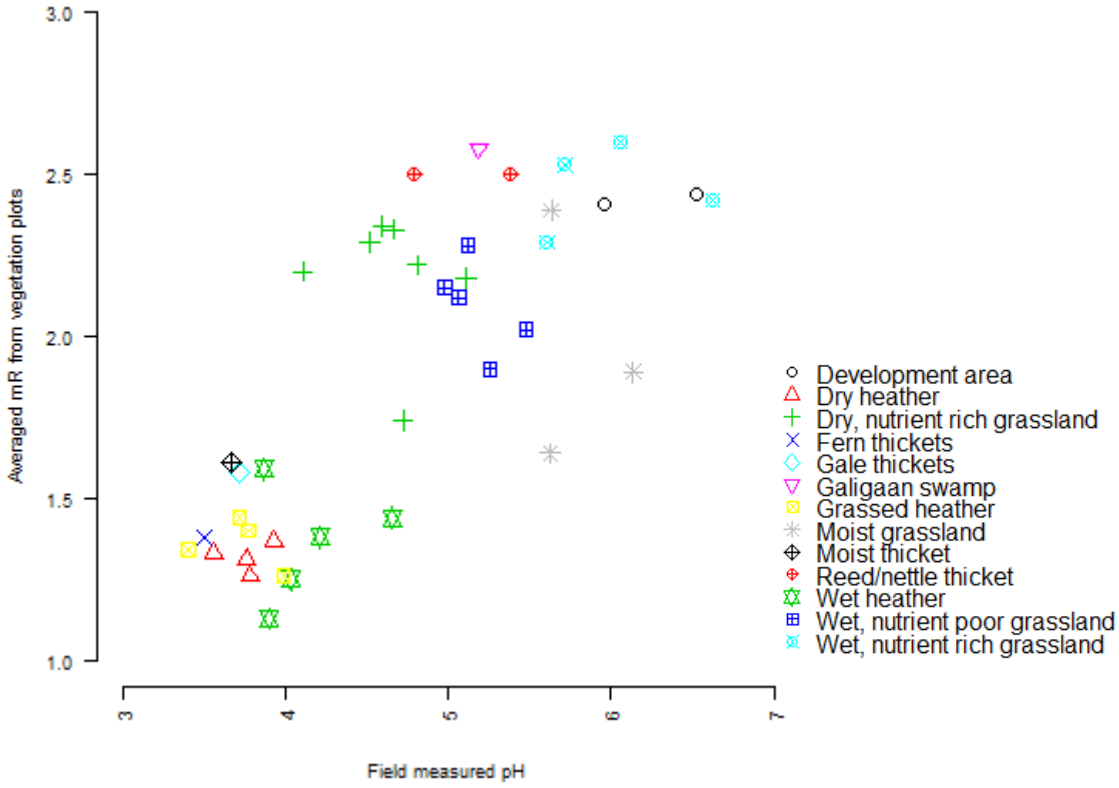


Figure 1. Relation between the averaged mR calculated from 40 sampled vegetation plots and the field measured pH of the same plots. The colours and symbols indicate different vegetation types.

2.2 Partial Least Squares Regression

Remote sensing offers several possibilities to map vegetation based on different methods. PLSR is used in various studies (Schmidtlein and Sassin 2004; Schmidtlein 2005) to determine the relation between spectral reflectance and vegetation or IVs. PLSR finds a linear relation to get a single equation by decomposing the predictors X and dependent variable Y into several principal components, where the orthogonal score T of X is correlated with Y (formula 1) (Wold, Eriksson et al. 2004).

$$Y = XB - E = XW_a^*C + E = TC + E \qquad (2)$$
$$W_a^* = W_a\,(P^TW_a)^{-1}$$

Where B is the vector of regression coefficients; E is a residual error matrix; W_a is the PLS weights; a is the number of latent variables (LVs); P and C are loadings for X and Y, respectively. Consequently, PLSR vectors depend on the response variables and not only the maximum variance of the predictor, like in Principal Component Analysis (PCA) (Cho, Skidmore et al. 2007). PCA is used to visualize large amounts of data. A first component is made by making linear combinations of the original variables. In this way, the axes are changed to new axes in the direction of the largest variation. PC1 is the vector of the largest variation in the data; PC2 is the second largest etc. This does not have to be the case for the PLSR used in this research. PLSR avoids the assumption that the observations have to follow a specific (normal) distribution and must be independent. PLSR is often used because it can deal with many, correlated explanatory variables with relatively few samples (Mevik and Wehrens 2007). When the number of predictor variables is equal to or exceeds the number of observations, multiple linear regression cannot be used. Moreover, highly correlated variables need to be removed when using multiple regression. When using PLSR, significant information of the variables contributes to a few latent variables. Therefore, related variables can together explain the variance in one component, so the variables do not have to be independent. The PLSR method has some limitations, since the method is sensitive for outliers and based on the assumption that the data has a linear relation. When the relation is more complex, different methods may perform better. For example non-linear PLS (NPLS) (Eriksson, Andersson et al. 2006) or Bayesian Model Averaging can be used for non-linear relations between predictor and response variables (Witte, Wojcik et al. 2007).

Selecting the number of latent variables is a critical point in PLS regression, which is usually based on minimizing the Root Mean Squared Error (RMSE) of the cross-validation (Feng, Elmasry et al. 2013). An alternative is to choose the amount of latent vectors based on the Predicted REsidual Sums of Squares (PRESS) (Cho, Skidmore et al. 2007).

Jackknifing can be used with PLSR to test if the predictors contribute significantly to the prediction. Jackknifing is the approximate t-test of regression coefficients based on jackknife variance estimates. It can be used to remove predictors, which do not contribute to the prediction significantly, so the model becomes more reliable (Martens and Martens 2000).

Furthermore, brightness normalized PLSR can be used to correct the spectral variables for the orientation of the vegetation and therefore minimize shade effects for reflectance data acquired under heterogeneous illumination conditions. In this way, all brightness effects are removed which may improve the accuracy of the PLSR model (Feilhauer, Asner et al. 2010). In the study of Feilhauer (2010), the R^2 increased up to 0.36 for predicting leaf chemistry, but not for all predictions gave a

significant improvement. Moreover, the spectral signatures were based on simulated canopy reflectance and not on actual hyperspectral images.

2.3 Mapping IVs

Mapping of IVs has been done in previous studies using various methods, input variables and vegetation types. Schmindtlein (2005) used PLSR to predict IVs with hyperspectral data in the northern Alps near Salzburg. The soil consists of moraine deposits and the vegetation type is mainly grasslands. Witte and Kooistra (2008) used PLSR to predict IVs based on hyperspectral images for grasslands at the Doode Bemde in Vlaanderen. Schmidtlein and Sassin (2004) used hyperspectral data to predict floristic gradients by using Ellenberg IVs. The study was performed for grasslands in southern Germany. Roelofsen et al. (submitted) used Gaussian Process Regression to predict IVs of the island Ameland in the Netherlands. The accuracies of the IV predictions for the different studies vary (Table 2).

Table 2. Accuracy (R^2 of validation) of IV predictions performed by different studies.

			R^2 validation		
Research	**Veg. type**	**Method**	**mF**	**mN**	**mR**
Schmidtlein, 2004	Grasslands	PLSR	0.66	0.75	0.76
Schmidtlein, 2005	Grasslands	PLSR	0.58	0.66	0.68
Witte, 2008	Various	PLSR	0.73	0.51	0.40
Roelofsen, 2013	Coastal	Gaussian Process Regression	0.65	0.76	-

3. Study Area and Data Preparation

3.1 Study Area

3.1.1 Study Area Selection

The study area is part of the natural area Kampina in the province of Noord-Brabant, the Netherlands (Figure 1). The area, 51°34'N 5°17'E, is approximately 1544 ha and is owned and managed by Natuurmonumenten; an organisation that buys, protects and manages natural areas in the Netherlands (Natuurmonumenten 2012). Several habitats that are important for various plant and animal species are found in the Kampina. The number of endangered plant species, such as round sundew *(Drosera rotundifolia)* and wire sedge *(Carex lasiocarpa),* has increased to 60 different species the latest years (Aptroot 2009). It is an area that contains many plant species occurring in stream valleys, fens, dry and wet heathlands, dry pine forest in the north and more wet alder forests in the south (Natuurmonumenten 2012). A lot of management has been carried out in recent years. Excavations of agricultural fields were performed to convert the fields to natural areas between 2000 and 2005 in the south-eastern part of the study area (Bruijn, Voorn et al. 2010). In these areas the groundwater table was raised, causing many plant species to reappear. Some fields in the middle of the study area are still active agricultural fields, whereas a field in the southeast is now part of the natural area of Kampina, but is not excavated. Furthermore, a dike was located around the Beerze river valley in the south to keep water in the area.

This area is suitable for this research, because it contains many vegetation types over a short distance. In this way, a lot of variation occurs in the ranges of the three IVs. Therefore this method can be tested well and be used as a showcase for the province of Noord-Brabant. The study area is drawn by selecting the areas with low vegetation types consisting of mainly grassland, herbaceous and small shrub species and excluding forest and agriculture, which results in a study area of approximately 1050 ha (Figure 2).

3.1.2 History of the Landscape and Heathland Ecosystems

The cover-sand landscapes in the Netherlands originated from the last ice age (18-14 ka BP), when wind covered large areas with sand (Berendsen 2008). Higher areas still contain sandy soils from aeolian cover sand. Nowadays, streams flow through these sandy soils (Bruijn, Voorn et al. 2010) and fens were created when wind exposed the underlying loamy soils. Rainwater cannot infiltrate through the loam, so water is collected in fens. Over time, many fens became nutrient rich and acid and finally silted.

Heather vegetation is an important vegetation type for the Dutch history of cultural landscapes. Heathlands were a more common vegetation type in the history of the sand covered landscape. It originated because of deforestation, intensive grazing by sheep and stabbing of turf for the fertilization of agricultural fields (Berendsen 2008). From the second half of the nineteenth century, forestation and agricultural reclaiming of areas resulted in a decrease of heathlands. Fertilisation and removal of sheep resulted in grass encroachment and forestation. Heather occurs at an early stage of succession, so without active management, heather species will compete with other plant species. Especially moor grass (*Molinia caerulea*) is competing with heather in the study area. Dry heather occurs mostly in the dry sand covered areas, whereas wet heather is most pronounced around the stream valleys. Additionally a lot of other plant and moss species occur in heathlands.

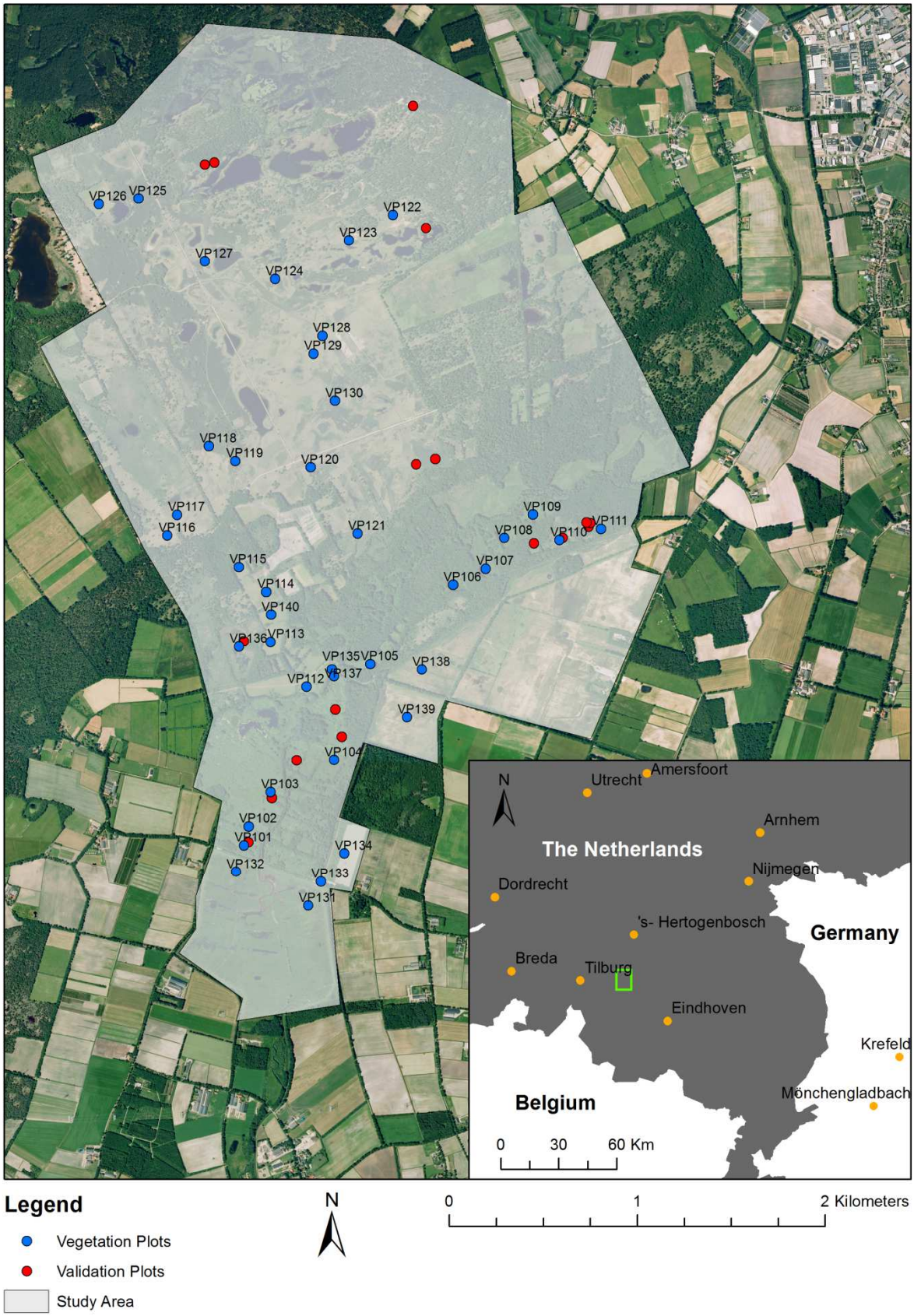


Figure 2. General location of the study area in the south of the Netherlands (indicated by a green box). Indicated in grey is the study area, overlaid on an aerial photograph. The blue points indicate vegetation plots used for the IV estimation. Red points are vegetation plots used for external validation.

Heather usually occurs in acid and poor nutrient environments. Also soil moisture is important to maintain the different heather species. When biomass accumulates in the area, nutrients and pH will increase and heather occurrence decreases. Therefore, conserving heather vegetation requires a combination of grazing, removal of trees and removal of the topsoil to take away biomass and nutrients for preventing other vegetation species like grasses and trees to take over (Bruijn, Voorn et al. 2010).

3.2 Field Sampled Vegetation Plots

A database of 40 vegetation plots located in the study area was available (Figure 2). The names of the plots were defined as VP1xx, where xx represent the number of the plot without a specific order. The plots measure 2x2 meters and the vegetation was sampled between 12 and 14 August 2012, whereas additional measurements were taken during July and August. A random-stratified sampling method was used to determine the locations of the plots (Figure 1), where the strata represented broad vegetation classes from the vegetation map of Natuurmonumenten. The following information has been acquired per vegetation plot: location and elevation (measured with RTK GPS, accuracy usually < 1 cm), date and time of sampling, broad vegetation type, species composition including moss species, vegetation height (herb layer) and total vegetation coverage divided into herb, and moss layer, litter and bare soil. Additional measurements for each vegetation plot include: soil pH (5 measurements at 20 cm below surface level distributed equally over the plot), leaf area index (measured with LAI2000 instrument), chlorophyll content, leaf nitrogen content,, leaf carbon content, leaf dry matter content, phenolic content, tannin content and lignin content. Additionally, 17 plots are available from Kampina vegetation mapping (hereafter: validation plots), of which the location (Figure 2) and the occurring plant species are known (Bruijn, Voorn et al. 2010).

3.3 Field Reflectance Data

For each vegetation plot, canopy reflectance was measured using an ASD FieldSpec Pro FR fieldspectrometer (Analytical Spectral Devices Inc.). Approximately 10 spectral measurements were taken for each plot, using a white spectral disc as reference. The measurements were acquired during clear sky conditions between 12 and 26 July 2012. Additionally, 11 reference points of water, bare soil and concrete were acquired simultaneously with the overflight of the APEX sensor on 30 June 2012.

3.4 Pre-processing LiDAR Data

3.4.1 Available LiDAR Data

Elevation data of the Netherlands was available from the second version of Actueel Hoogtebestand Nederland (Up-to-date Height Model of The Netherlands) (AHN-2) and provided topographical data of the study area. AHN is based on airborne LiDAR, where the return time of a laser pulse is used to determine the distance between the surface and the pulse source, which translates to surface elevation. Two products of AHN-2 are available, the Digital Elevation Model (DEM) and Digital Terrain Model (DTM). The DEM is an elevation model that includes all objects on the surface, while the DTM only contains the actual earth surface. The spatial resolution of this data is 0.5 m and the height error is less than 5 cm. The flights were carried out between 11 February and 20 March 2009 (Waterschapshuis 2012).

3.4.2 Prepare LiDAR Data

The AHN-2 data tiles that covered the study site were mosaicked together and subsequently clipped to the extent of the study area in ArcGIS. Overlaying the vegetation points with the elevation data revealed that no height information is available for six vegetation plots. Here, water or tree shadows prevented the laser pulse to return to the airplane. To be able to use these plots, AHN-2 pixels were resampled to 2m in order to expand the extent of available data. This reduced the number of plots without data to two. The AHN-2 was interpolated to generate data for the two remaining plots, meaning that when the DTM has no data for a specific cell, the value of the mean of the circle with a radius of 12 cells is acquired. A radius of 12 cells is the minimum to use for VP132 to get an elevation value, while a radius of two cells was sufficient for VP102. Finally, the surface rasters were resampled to exactly match the pixel size of the hyperspectral images (approximately 2.12m), so there is no unnecessary loss of spatial resolution. In this way, a matching overlay can be made, in which all pixels correspond to the exact same location in all layers.

3.4.3 Quality Check LiDAR Data

Field measured elevation data was available from GPS measurements. When comparing this elevation with the elevation obtained from LiDAR data, an r of 0.89 is found (Figure 3), which is reasonable considering that inaccuracies of the different measurements are included. A large difference of 1.6 m was observed for plot VP104. This difference cannot be explained spatially, since elevations nearby are rather homogenous. Therefore, the GPS measurement of this plot can be considered as an outlier, which is confirmed by the high inaccuracy of 4.6 m from this GPS measurement.

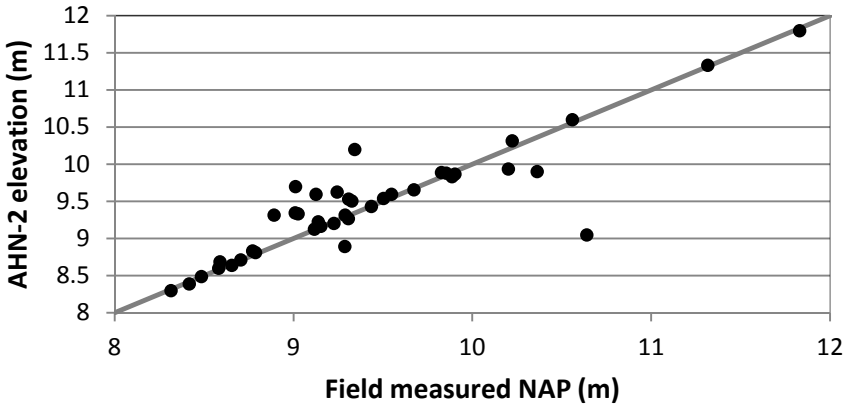


Figure 3. Elevation from AHN-2 data vs. field measured elevation for the 40 vegetation plots. The line indicates the 1:1 line.

3.5 Pre-processing Hyperspectral Images

3.5.1 Available Hyperspectral Data

Hyperspectral images from the Airborne Prism EXperiment (APEX) sensor were available. The APEX sensor covered the study area on 30 June 2012 in three flightlines (Figure 4). Spectral reflectance data was acquired in 288 bands covering wavelengths from 413 to 2453 nm (APEX 2012). The spatial resolution was 2.12 m. The images were atmospherically and geometrically corrected by VITO using the AHN-2 elevation data. Some known issues are indicated in the delivery report: residual striping

occurs in the lower wavelengths, some minimal across track striping may occur, natural spectral directional effects occur in shaded and sloped areas (VITO, 2012).

3.5.2 Preparation of Hyperspectral Images

The images of the three flightlines were delivered in four images, which were mosaicked and clipped to the study area in ENVI. For the mosaicking, the order of overlapping images is indicated in Figure 4. Flightline 38 was used to substitute for areas obscured by clouds in the other flightlines. The extent of clouds and cloud shadows was manually determined and removed from the images. Furthermore, the first 10 and last 8 bands were removed for analysis due to a low signal/noise ratio. Additionally, bands measured in water absorption wavelengths were also removed (Table 3). The spectral regions of 518 -1334, 1440-1792 and 1962-2402 nm remained for analysis, covered by 240 spectral bands.

Table 3. APEX bands that were removed from analysis.

Removed bands	**Reason of removal**	**Wavelength lower band (nm)**	**Wavelength upper band (nm)**
1 – 10	Striping and noise	411.2	507.5
147 - 156	Water absorption	1347.2	1433.8
197 - 216	Water absorption	1803.5	1956.8
281 - 288	Striping and noise	2410.4	2454.0

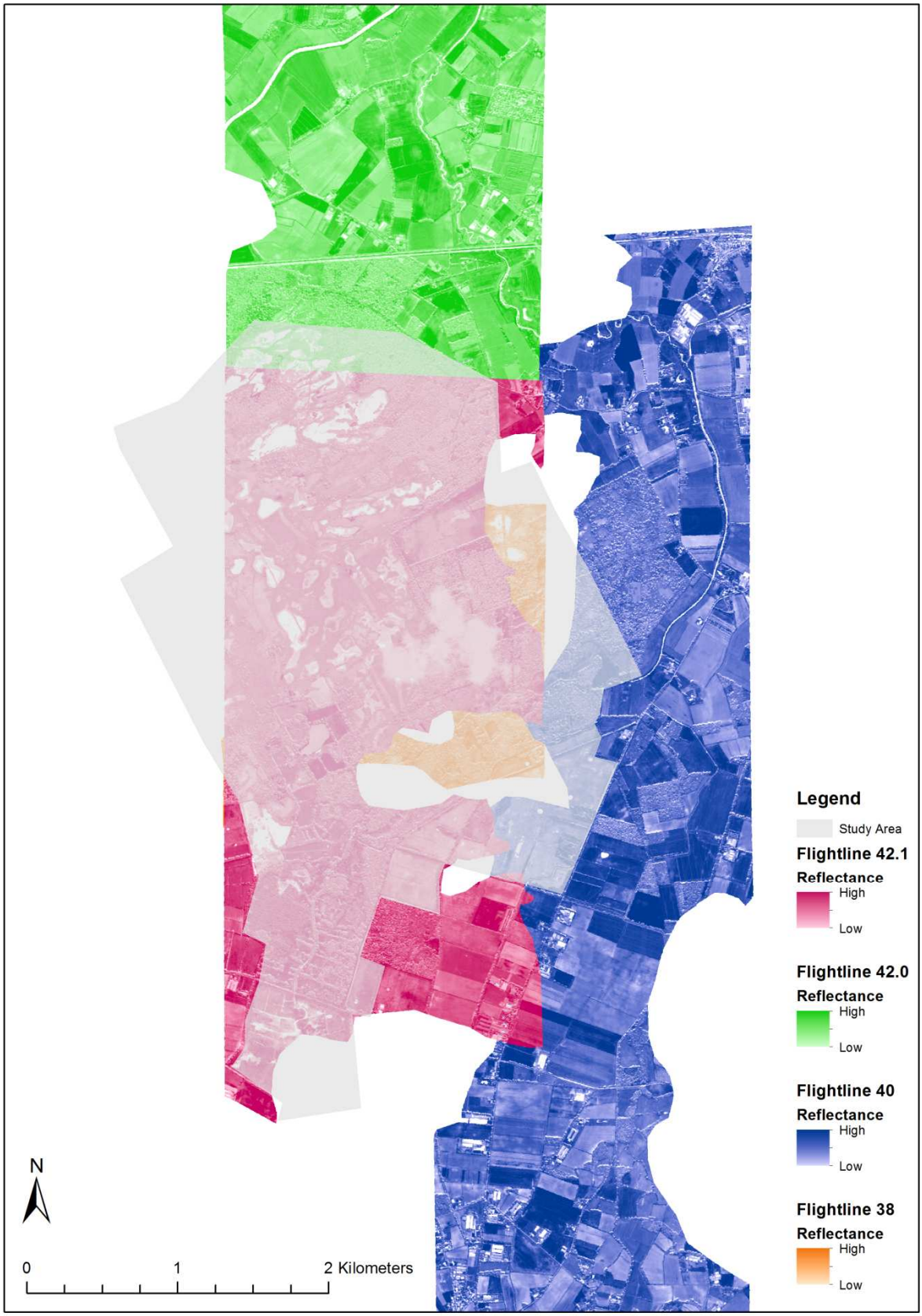


Figure 4. Locations of the three different flightlines indicated by different colours. Note that flightline 42.1 and 42.0 are part of one flight, but were separated later. Also note that clouds and cloud shadows are already removed from the flightlines.

3.5.3 Quality Check Hyperspectral Data

First a comparison was made between APEX reflectance for different flightlines at the same location. Flightline 38 and 42.0 can be compared, because they overlap at some vegetation plots where no clouds or cloud shadows occur. The reflectance was sampled for both flightlines for VP114, VP115, VP136 and VP140 for the selected bands that are used for analysis (Figure 5). Flightline 38 overestimates the reflectance systematically compared to flightline 42. This may be the result of the different moment of data acquisition and thus, different atmospheric effects and weather conditions. Since the reflectance is structural different between flightlines, values may be over or underestimated in certain areas. The actual effect of this difference is not expected to be problematic for this research, because the difference does not seem to be large.

Next, a quality check of the APEX images was done by comparing APEX reflectance with field spectrometer data. Both field spectrometer measurements for each vegetation plot and for some reference points were available. Appendix I shows the spectral signature of the APEX images and the spectral signature of the field spectrometer for the vegetation plots. A comparison between the flightlines reflectance and the field spectrometer reflectance was made for two vegetation plots. A plot with a relatively low and a relatively high reflectance in the NIR wavelengths were used for comparison (Figure 6). Also in this case, flightline 38 overestimates the reflectance compared to flightline 42. The deviation from the field measured reflectance is not systematically and differs throughout the spectrum. In the visible (VIS) and near infra-red (NIR) part of the spectrum, both flightlines overestimate the reflectance compared to the field spectrometer. This deviation may be due to the atmospheric conditions during the flights, which were not optimal. The sky was not clear, which lead to higher reflectance values. Although the images were atmospherically corrected, some effects still remain in the images.

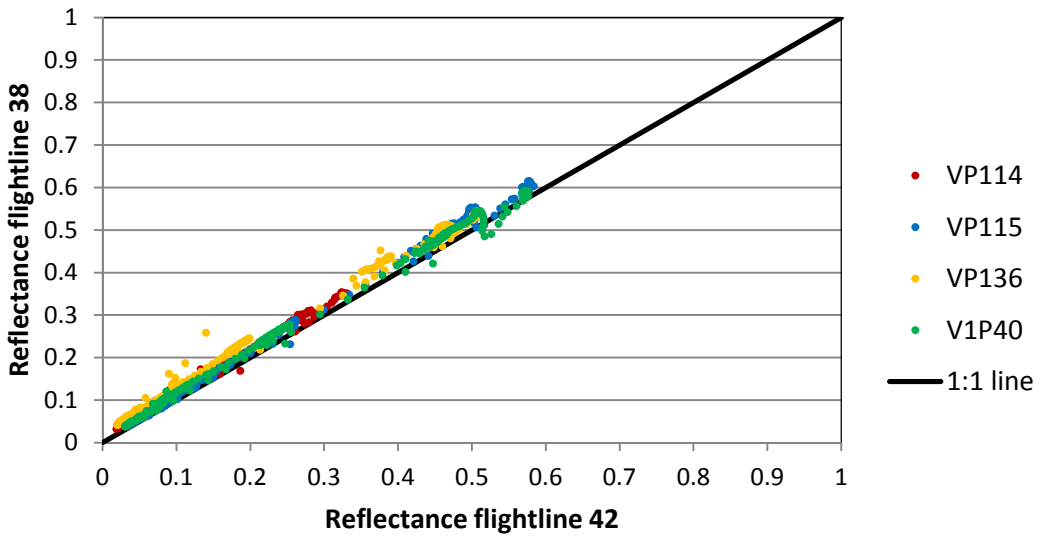


Figure 5. Reflectance of flightline 38 versus 42 for four vegetation plots. Flightline 38 systematically overestimates the reflectance compared to flightline 42.

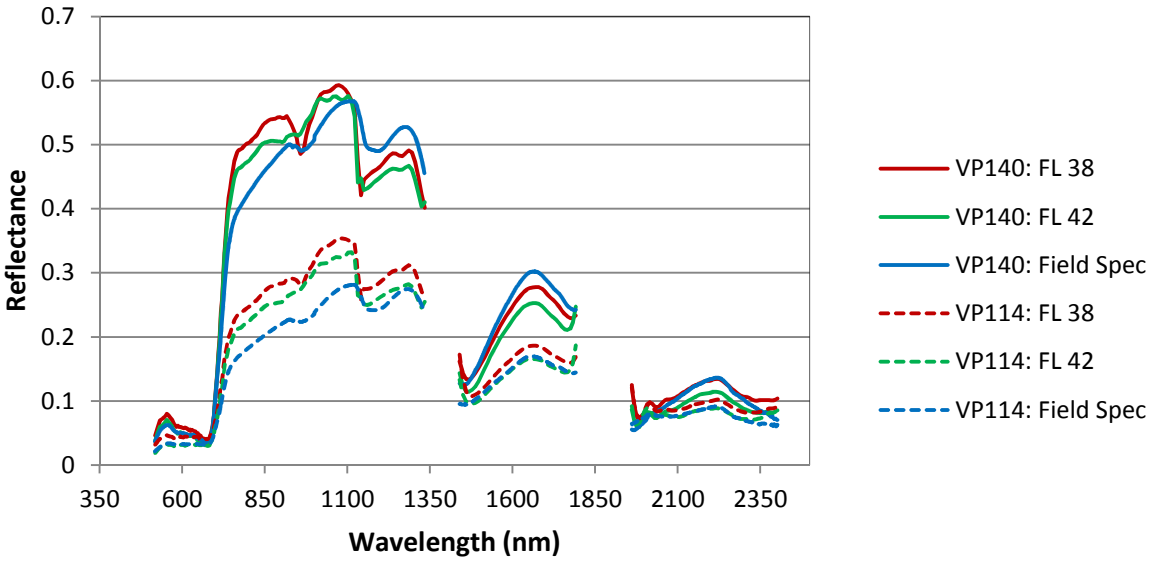


Figure 6. Spectral signatures of VP114 and VP140 for flightline 38, 42 and the field spectrometer, all acquired on June 30, 2012.

Finally, five field spectrometer reference points were used for a quality check, but their locations are uncertain because of inaccurate GPS measurements. Therefore, the locations were manually determined based on the description of the sample points, though the exact location may deviate. The results of the comparisons are given in Appendix II. The two spectral signatures for open water correspond well in general. The last water reference point shows some underestimation of the APEX reflectance, but this can be explained because of the reference point may be relocated too far from the water bank, so reflectance becomes lower. The reflectance of sand shows some more deviation, the APEX reflectance is overestimated compared to the field spectrometer. However, no predictions of IVs have to be made for bare soil areas.

3.6 Overlay of AHN-2 and APEX Data

The definitive study area consists of the areas where both APEX and AHN-2 provide data minus non-vegetation objects such as water, roads, houses, agricultural fields, forests and solitary trees. The removal of the objects indicated in Figure 7 is described below.

3.6.1 Clouds

Clouds and cloud shadows were manually delineated and removed from the APEX data. In this way, the study area remains continuous and edge effects do not occur. Clouds can also be removed with a classification algorithm in e.g. ENVI. This classification is reasonable accurate, but it is difficult to get the right balance of pixels that have to be removed. The clouded areas need to be removed in a way that no edge effects occur, this means that no pixels at the edges are still influenced by clouds, which can alter the results. Therefore, a generous bounding of clouds is preferred, but then pixels in the study area will also be classified as clouds and the area becomes pixelated. All in all, it was decided that manual removal was most efficient and accurate.

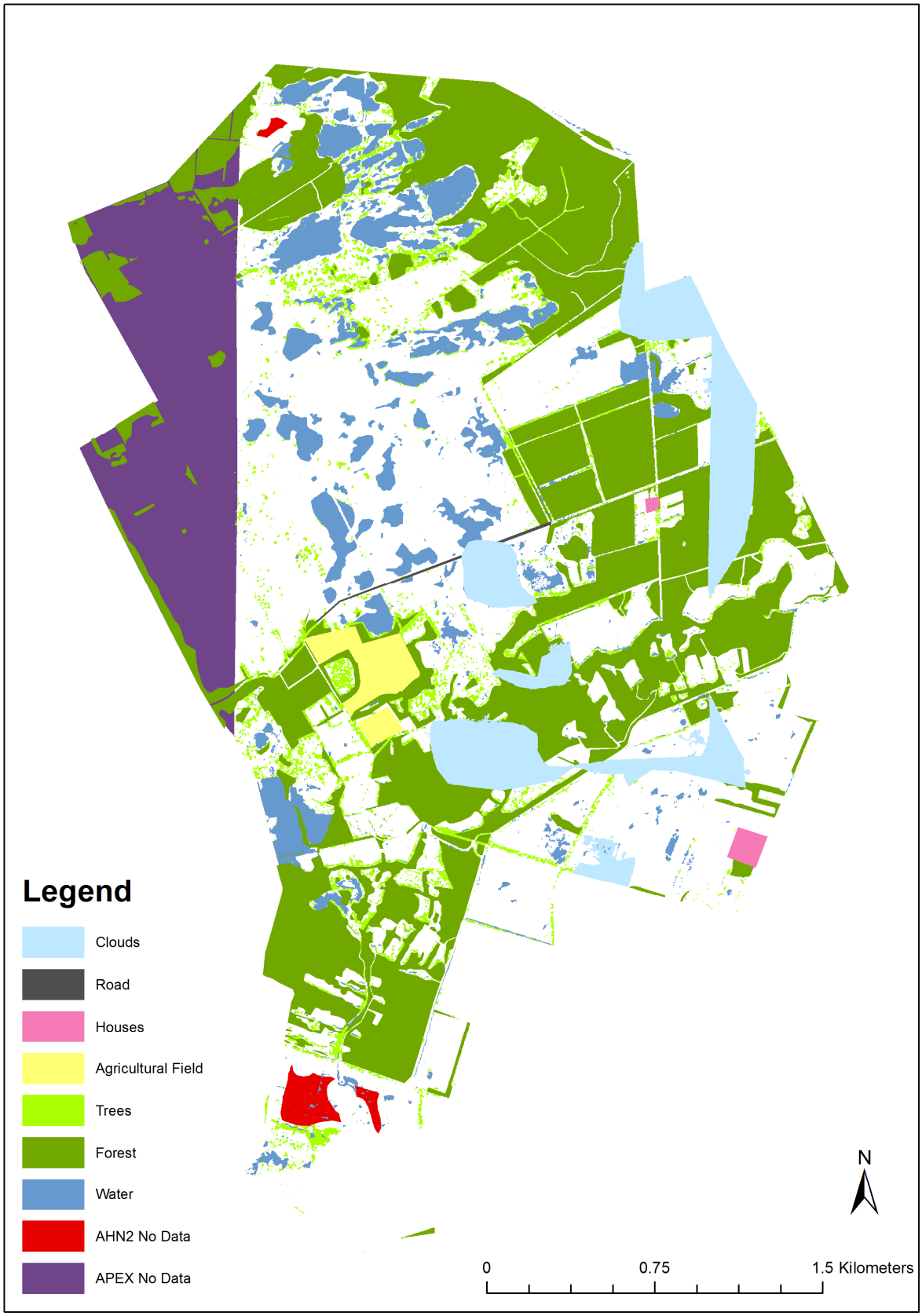


Figure 7. Areas removed from the study area.

3.6.2 Trees

Trees were removed from the area, because they are not suitable for the prediction model. The relation between the spectral signature of trees and their preference of environmental conditions is different than that for low vegetation. Moreover, trees were not present in the vegetation plots. The trees were removed using the 1:25.000 topographic map from 2009. All pixels that are classified as coniferous, deciduous or mixed forest were selected. In this way, complete fields were removed, but some tree lines and solitary trees remained. These trees were identified by exploiting the difference of spectral signature between trees and herbaceous vegetation in a classification. The images were classified into two classes: tree and non-tree areas. Classification was based on a subset of ten bands randomly chosen in the range from the remaining bands from the APEX images, so classification is faster and easily repeatable. The maximum likelihood classifier was used based on Regions Of Interest (ROI) (Figure 8). A solitary tree was selected indicated by the red polygon, so only solitaire trees are selected and no other vegetation and the green polygon is selected to classify the rest of the study area. The maximum likelihood classification calculates the probability that a given pixel belongs to a specific class and each pixel is assigned to the class with the highest probability (ENVI 2012).

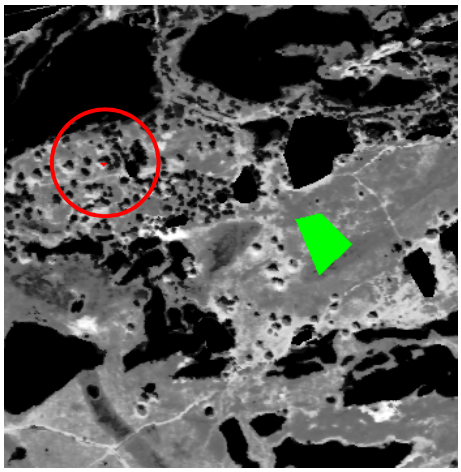


Figure 8. Tree classification training ROIs, a solitaire tree (red polygon in the centre of the red circle) and land (green polygon) are selected for the maximum likelihood classification.

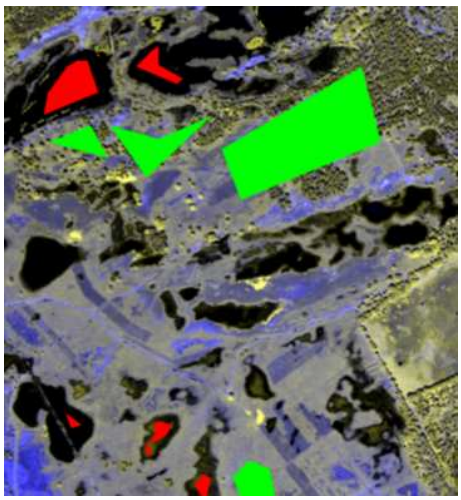


Figure 9. Water classification training ROIs, water (red polygons) and land (green polygons) are selected for the maximum likelihood classification.

3.6.3 Water

Comparable to tree identification, open water was classified by selecting ten bands from the range of remaining APEX bands. Again, the maximum likelihood classifier was used based on ROIs. Water classification also includes tree shadows, since their spectral signature is comparable and these areas also had to be removed. Water has a distinct spectral signature, so classification goes well. Relatively large ROIs were selected (Figure 9) because some areas contain water plants. Again two classes were created to be able to check the accuracy of the classification clearly. The classification was done separately from the tree classification, but could be done together.

3.6.4 Other Removed Areas

The agricultural fields in the middle of the study area were removed, because these fields are still used for extensive agriculture. Estimating IVs for these areas makes little ecological sense, because the species composition of agricultural fields is determined by external factors like agricultural management, rather than by the abiotic conditions. The fields were manually delineated and removed, as were two houses and a concrete road. Additionally, nodata areas from the AHN-2 data were removed. These areas are mainly located in the southern part of the study area where open water occurs. Finally, a large area located in the west is removed from the study area, because the APEX flightlines do not cover this area.

In the end, all masked areas from Figure 7 were removed from the study area. The resulting polygon was used to clip the topographical variables and hyperspectral images to get an overlay needed for later steps. The final study area containing only the areas with low herbaceous vegetation is presented in Figure 10.

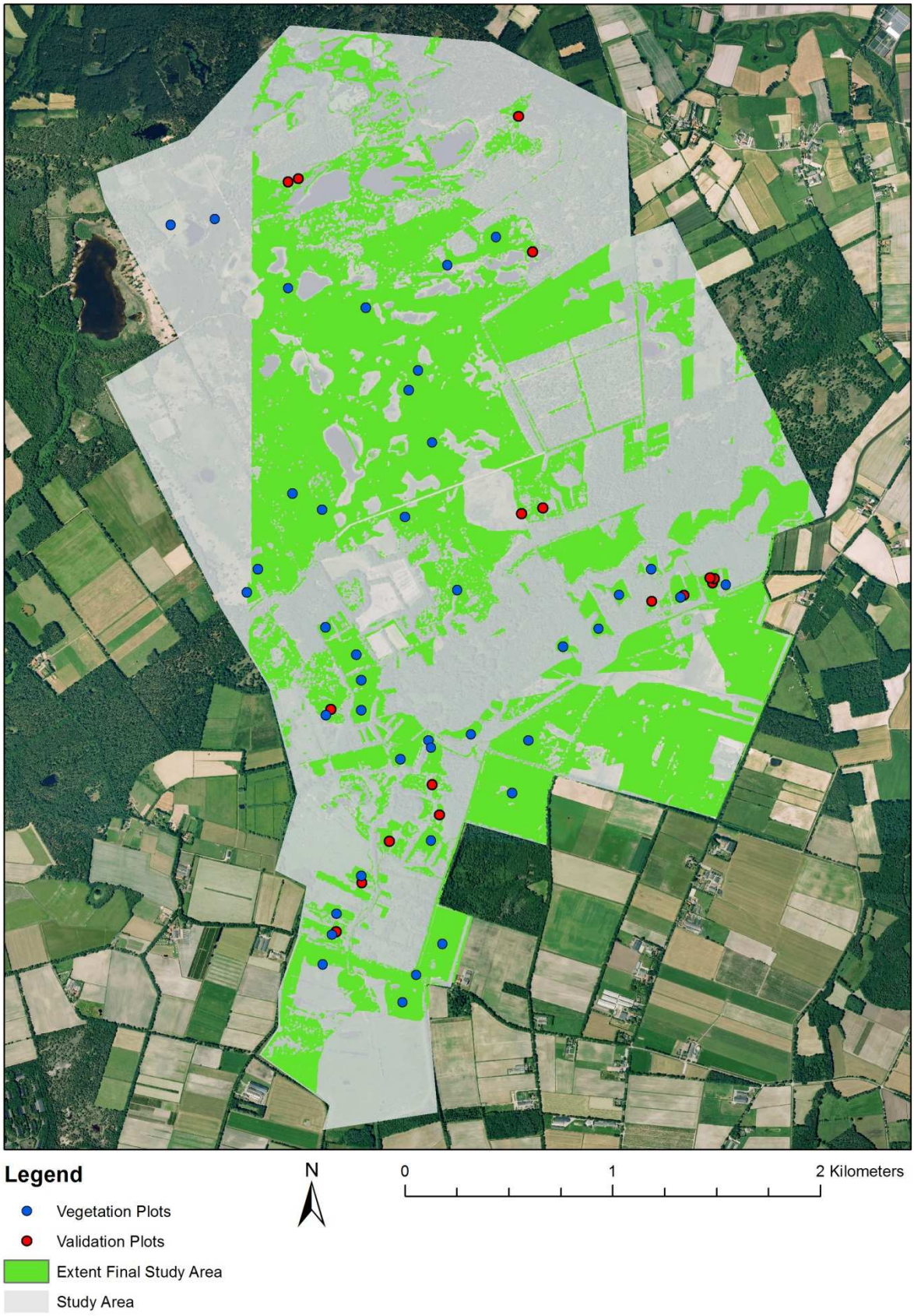


Figure 10. The extent of the final study area after removal of the masked areas. The blue points indicate the vegetation plots used for this study and the red points are vegetation plots for external validation.

4. Methodology

4.1 General Methodology

The general methodology is based on the following working steps, which correspond to the numbers of the boxes in the flowchart as presented in Figure 11:

1. Calculate the average IVs per plot;
2. Prepare the DEM and DTM. Select, calculate and sample the LiDAR prediction variables at the location of each plot;
3. Prepare the hyperspectral images and sample the spectral prediction variables;
4. Apply PLSR by using the average IVs as response and LiDAR and hyperspectral variables as predictors to model the IVs;
5. Assess the results and accuracy of the PLSR model and adjust inputs and model settings;
6. Select the best models to estimate full coverage IV maps;
7. Evaluate the estimated IV maps.

The subsequent sections will describe each working step in closer detail. The data preparation is already described in chapter 3.

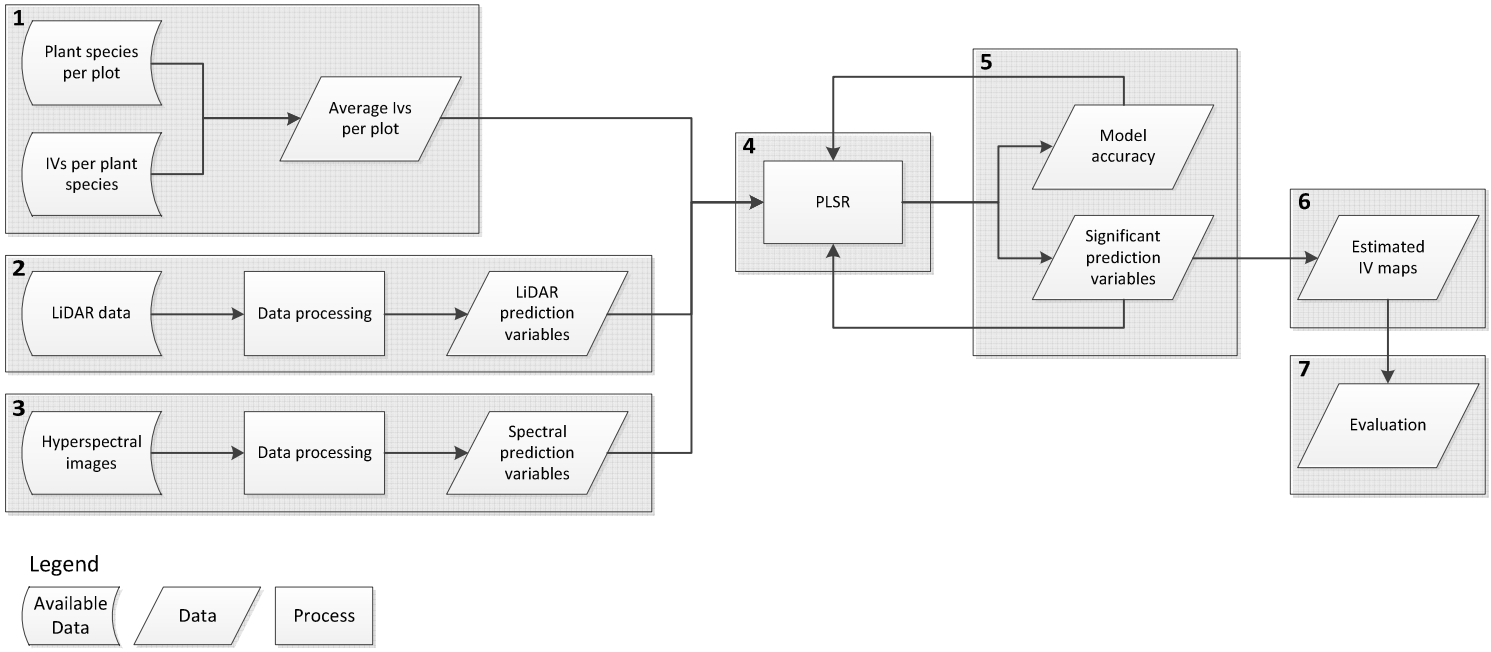


Figure 11. Flowchart of the general methodology. The large black boxes enclosing one or more small boxes are a single working step. The numbers in the boxes correspond to the numbers of the working steps.

4.2 Averaging IVs

A database with the IVs per plant species from Witte et al. (2007) was used for this research. The most relevant IVs for this study area are mF (moisture regime from 1 = open water to 4 = dry soil), mN (nutrient availability from 1 = poor to 3 = rich) and mR (acidity from 1 = acid to 3 = alkaline) (Runhaar, van Landuyt et al. 2004). First, the plant species that occur in each plot need to get their corresponding plant species number for which the IVs are defined. Second, the negative IVs were removed from the database, since these plant species are insensitive for the particular IVs and do not contain information about the preference of environmental conditions. The relative coverage was not taken into account, since it will not improve the analysis (Kafer and Witte 2004). Instead a qualitative approach was used, so all plant species in the plot received equal weights. Therefore, the average IV was calculated as the sum of the values divided by the number of plant species occurring in the plot. This was done for each IV separately and resulted in a table with 40 average values per IV corresponding to the 40 sample plots. This table was used as an input for PCA to create a biplot in order to check for outliers. Finally, this table was imported to ArcGIS using the GPS recorded coordinates. The indicator values for the external validation plots were calculated in a similar fashion.

4.3 Deriving LiDAR Predictors

The DTM of Kampina (Appendix III) was used to derive the following topographical variables: elevation, slope, aspect, northness and eastness, flow direction, flow accumulation, curvature (planofile and profile) and vegetation height. In this way, the elevation data is used in an optimal way to derive more information that may contribute to the IV predictions. For each variable, a fixed window of 3 by 3 pixels (6.35x6.35m) was used to calculate the centre cell. These calculations were performed in ArcGIS and each derivative is described below (ESRI 2011):

- Elevation is derived from the DTM and indicates the height in meters above NAP. The elevation in the study area ranges from 6.8 to 18.3 meters. The lowest vegetation plots are located in the south-western part. The highest plots are part of the sand covered ridges in the north.

- The slope (Appendix IV) is defined as the gradient of maximum height change in degrees from 0 to 90.

- For the aspect, each pixel receives a value that indicates the direction of the maximum slope. It ranges from 0 to 360 degrees, clockwise from north. Flat areas get a value of -1.

- The northness and eastness are based on the aspect. The values range from -1 to 1 and indicate to which degree the aspect is located to the north or to the east. Northness is calculated as $\cos(\pi * aspect/180)$, whereas eastness is calculated as $\sin(\pi * aspect/180)$.

- For the flow direction, each cell gets the value of the direction of the steepest down flow. The direction values are based on Figure 12, so a pixel will get a value of e.g. 16 when the steepest downward flow is westwards.

Figure 12. Direction values used for creating a flow direction map.

- The flow accumulation is based on the flow direction and is defined as the number of cells that flow to that particular cell from neighbouring cells.
- Curvature is the change in slope. Negative values indicate concave slopes (the slope increases over height) and positive values indicate convex slopes (the slope decreases with increasing height). A curvature of 0 is a flat area. The curvature can be separated in more detail as planofile curvature, which indicates the curvature when looking in the direction perpendicular the direction of the maximum slope. Profile curvature looks in the direction of the maximum slope.
- The vegetation height is calculated by subtracting the DTM from the DEM.

The vegetation height of the 40 plots was also available from field measurements. When comparing this field measured vegetation height with the object height derived from AHN-2 data, it became clear that the object height could not be used for analysis (Figure 13). The Figure shows a systematic lower vegetation height detected by LiDAR compared to the field measured vegetation height. In fact, only one vegetation plot has a vegetation height of more than 10 cm. This result can be explained by the date of acquisition of the AHN-2 data in early spring, when the low vegetation has not enough leaves and branches to be detected by a LiDAR pulse. Consequently, the vegetation height derived from LiDAR data was removed as explanatory variable for modelling the IVs and the field measured vegetation height was included instead. The highest vegetation in the study area was found for ferns and nutrient rich grasslands.

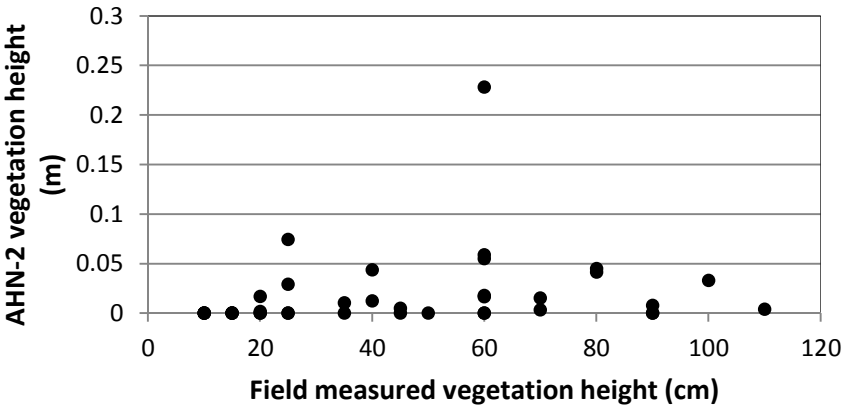


Figure 13. Vegetation height derived from AHN-2 data vs. field measured vegetation height of the 40 vegetation plots.

The values of the topographical variables were sampled for the 40 vegetation plots. Since the pixels of the AHN-2 data and the plot size do not match exactly, a sample method was chosen. Since it was not important to maintain the original values of the topographical variables and the plots were usually not located at edges of different vegetation types, the bilinear sample method was used. The bilinear sampling method assigns the value to the point location based on the distance-weighted average of the four closest neighbours (Lillesand, Kiefer et al. 2007). After sampling, the Pearson's correlation coefficient (r) is calculated to determine if a correlation exists and which topographic variables are likely to contribute to the IV predictions. These coefficients indicate the extent of a linear relation between two variables. The correlation coefficient r ranges from -1 to 1, where -1 indicates a strong negative correlation, +1 a strong positive correlation and 0 indicates no correlation between the two variables. Predictor variables with high positive or negative r values are expected to have a high contribution in the estimation of the IV.

4.4 Deriving Hyperspectral Predictors

The reflectance values were derived from the hyperspectral images for each vegetation plot to use as predictor variables. The pixels of the APEX images and the plot size did not match exactly, so the bilinear sampling method is used in the same way as sampling the topographic variables. Two vegetation plots (VP025 and VP026) were located outside the APEX flightlines. Therefore, the field spectrometer data was used in order to include these plots for the calibration of the prediction model. A reflectance value was given for the corresponding APEX wavelengths by selecting the nearest wavelength reflectance value from the field spectrometer data. Finally, the Pearson's correlation coefficient was calculated between each spectral variable and the IVs to determine if a correlation exists and which variables were likely to contribute to IV predictions.

4.5 Prediction of IVs

4.5.1 PLSR

The average IVs and the derived LiDAR and hyperspectral predictors were organized in a table. The columns contain the three IVs, 11 LiDAR predictors and 240 hyperspectral predictors for each of the 40 vegetation plots in the rows. By using this input data, the prediction of the IVs was done with PLSR by using the pls package in RStudio. PLSR was used as a standard method in this research to compare different models.

4.5.2 Accuracy Assessment

Since the amount of available vegetation plots is small, it was not feasible to split the data of the 40 plots into training and validation datasets. By using Leave One Out (LOO) cross validation, every data point can also be used for the calibration of the regression model. The cross validation was performed to predict the IVs for a certain plot by using all plots except for the specific plot. Next, the predicted and observed IVs were compared to determine the accuracy of the prediction. The accuracy of the model was calculated by the Root Mean Square Error (RMSE) of the prediction (Formula 3). The RMSE is a measure for the difference between the predicted and the observed value in the same units as the IVs (Ott and Longnecker 2010). A lower RMSE indicates a better prediction of the response variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(IVobs_j - IVpred_j\right)^2} \qquad (3)$$

Additionally, the R^2 of the validation was used to assess the accuracy of the models, because it is a measure of the predicted values in relation to the 1:1 line, where observed and predicted values are equal. The R^2 given in the pls package is based on the Nash-Sutcliffe model efficiency coefficient (Formula 4). The IV_{obs} is the observed IV and IV_{pred} is the predicted IV based on the regression model for n=40 vegetation plots.

$$R^2 = 1 - \frac{\sum_{j=1}^{n}\left(IVobs_j - IVpred_j\right)^2}{\sum_{j=1}^{n}\left(IVobs_j - \overline{IVobs}\right)^2} \qquad (4)$$

The R^2 ranges from –infinity to 1, 1 indicates a perfect match when the prediction equals the observations. The R^2 is 0 if the mean of the observations predicts as good as the model, negative values indicate that the mean of the observed values gives a better prediction than using the modelled values.

4.5.3 Prediction Models

The response variables mF, mN and mR were predicted once at a time with the available predictors. To determine if the LiDAR variables improve the prediction, the model with only hyperspectral predictors was compared to the model with added LiDAR predictors. Additionally, brightness normalised PLSR (Feilhauer et al., 2010) was used for spectral predictors only to check if this correction improves the model accuracy. Consequently, models were fitted for each IV by using:

1. Hyperspectral predictors
2. Hyperspectral and LiDAR predictors
3. Hyperspectral predictors with brightness normalization

An available R script with the PLSR function was used. Some different settings had to be specified. First, the amount of latent variables was chosen manually based on a bar plot with the lowest RMSE of the cross-validation (Figure 14). Second, scaling can be requested within the PLSR function, which standardise each variable by dividing it by its standard deviation (Mevik and Wehrens 2007). Therefore, the standard deviations of the prediction variables are standardized to 1 and the regression coefficients become weighted. This is an important step, because the predictor variables have different units, ranges and standard deviations.

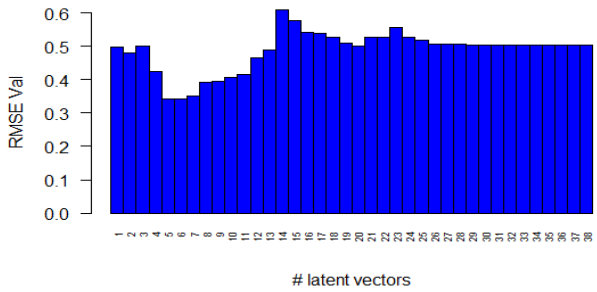


Figure 14. Example plot of the RMSE of LOO validation for different numbers of latent vectors. In this example, 4 latent variables would be selected for the model fit.

4.5.4 Band Selection

The R^2 of validation (based on LOO validation) increases when only significant predictors are used as input, because non-significant variables are difficult to fit. Therefore, these variables were removed to get a more robust model. However, the R^2 of the calibration will decrease, because the non-significant variables will explain some small parts of the variation. Jackknifing was used to determine the stability of the regression coefficients. The jackknifing method uses the variance of a regression coefficient in the model fit to statistically test if the coefficient is significantly different from zero. By using the resulting p-values of the regression coefficients, non-significant predictors were removed and the significant predictors are used as new input for the PLSR. The significance level for removing predictors needs to be chosen carefully. In general, variables with a p-value smaller than 0.05 were removed. This value can be chosen higher or lower based on the amount of significant predictors remaining in the model. The removal of variables was continued until a stable model was formed, in which all variables are significant.

4.6 Mapping IVs

The most accurate regression models for mF, mN and mR were selected based on the highest R^2 of cross-validation. These final PLSR models were applied to the total study area, creating full coverage maps of the IVs by using the intercept and the regression coefficients. The coefficients had to be recalculated first, since they were weighted because of the scaling. Note that each PLSR model required different predictor variables as a result of the band selection. The full coverage raster layers of the significant predictors were loaded in R using the raster package. Every pixel of the study area is located in one row and every value for the different raster layers is present in the columns. This table was used as an input for the prediction function of the pls package (Mevik and Wehrens 2007). In this way, an estimation of the IVs is made for each pixel. The estimations were not restricted to the observed range of IV values, so the produced estimates could have a value beyond the theoretical range of IVs. Next, this table was converted back to a raster in order to get a spatial map by importing the table into ArcGIS.

4.7 Evaluation

In addition to the R^2 and RMSE of the models, the IV estimations were evaluated by a visual check. Some areas of specific vegetation types were selected to compare if the resulting IV maps coincide with the relative expected values. Additionally, an independent external validation is performed by using the validation plots. The observed averaged IVs for these plots were compared with the predicted values in the maps by plotting the values and calculate the R^2.

5. Results

5.1 Average IVs

The data range of the average IVs of the 40 vegetation is given in Table 4 and Figure 15. The three IVs are not normally distributed over the IV ranges. No low mF values were found in the plots, because none were taken in open water. For mN, many plots have low values meaning they occur in nutrient poor areas. Furthermore, values near the maximum value of three do not occur for mN and mR. Also the mR range shows a bimodal structure, since no medium values occur for the vegetation plots. Figure 15 also shows the correlations between the three IVs. There is a medium negative correlation between mF and the other two IVs (r = -0.37 and -0.51 with mN and mR respectively). The correlation between mN and mR is strong and positive (r = 0.89), where high mN values correspond with high mR values. It is expected that this relation is also noticeable in the PLSR. In general, though not valid for the entire study area, dry areas correspond with nutrient poor soils with a low pH. Figure 16 also shows the relation between mN and mR, and some plots deviate from others and may be more difficult to predict, such as VP115 (dry, nutrient rich grassland). However they cannot be removed, because they are not errors and do occur in the study area.

Table 4. Properties of average IVs of the 40 plots.

	Min	Max	Mean	SD
mF	1.82	3.9	2.68	0.554
mN	1	2.73	1.61	0.546
mR	1.13	2.6	1.90	0.482

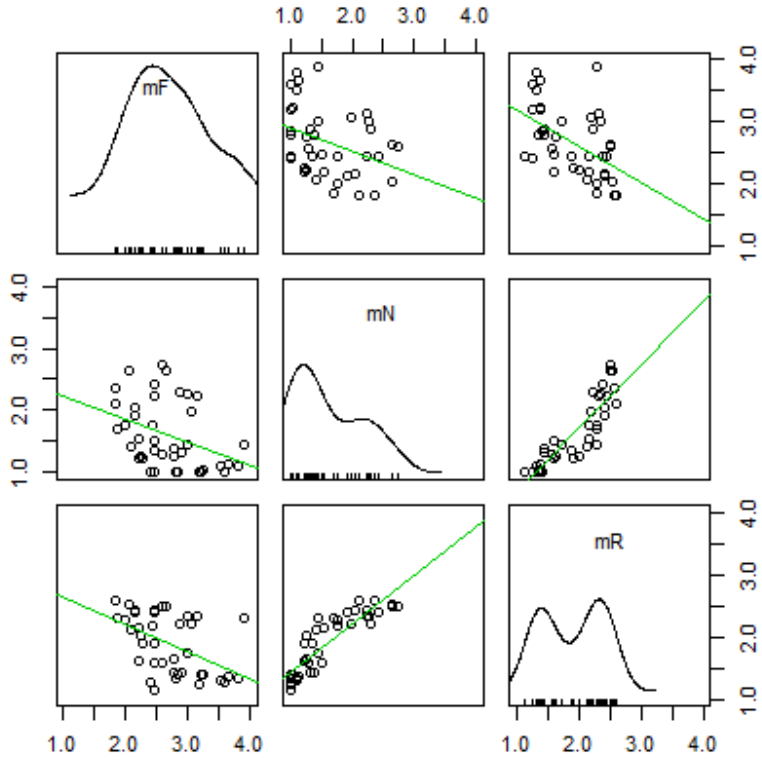


Figure 15. Relations between the average IVs of the 40 plots. The histograms of the occurring IVs are given in the diagonal boxes. Note that mF ranges from 1-4, mN and mR from 1-3.

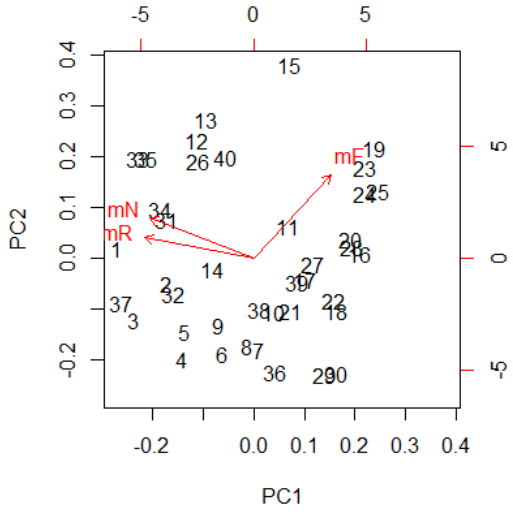


Figure 16. Biplot of the IVs and vegetation plots based on PCA. The numbers correspond with the last two numbers of the vegetation plot names.

5.2 Correlation between IVs and Predictor Variables

The Pearson correlation coefficient is calculated between the three IVs and all predictor variables sampled for the 40 vegetation plots. In general, correlation patterns of the spectral predictors for mF are mirrored with the correlations for mN and mR (Figure 17). mR and mN show higher spectral correlations than mF, so these IVs are expected to be better predictable with spectral predictors only. For large parts of the spectrum, mN and mR behave comparable, which is expected because they are correlated (r= 0.89). In general, correlation coefficients of mR are higher compared to mN. The correlation is especially high for spectral bands around 900nm. The correlation between LiDAR variables and the IVs is in general much lower. Only mF shows a correlation with vegetation height and elevation that is comparable to the correlation with spectral variables. The other LiDAR variables are not expected to contribute to the IV predictions using PLSR.

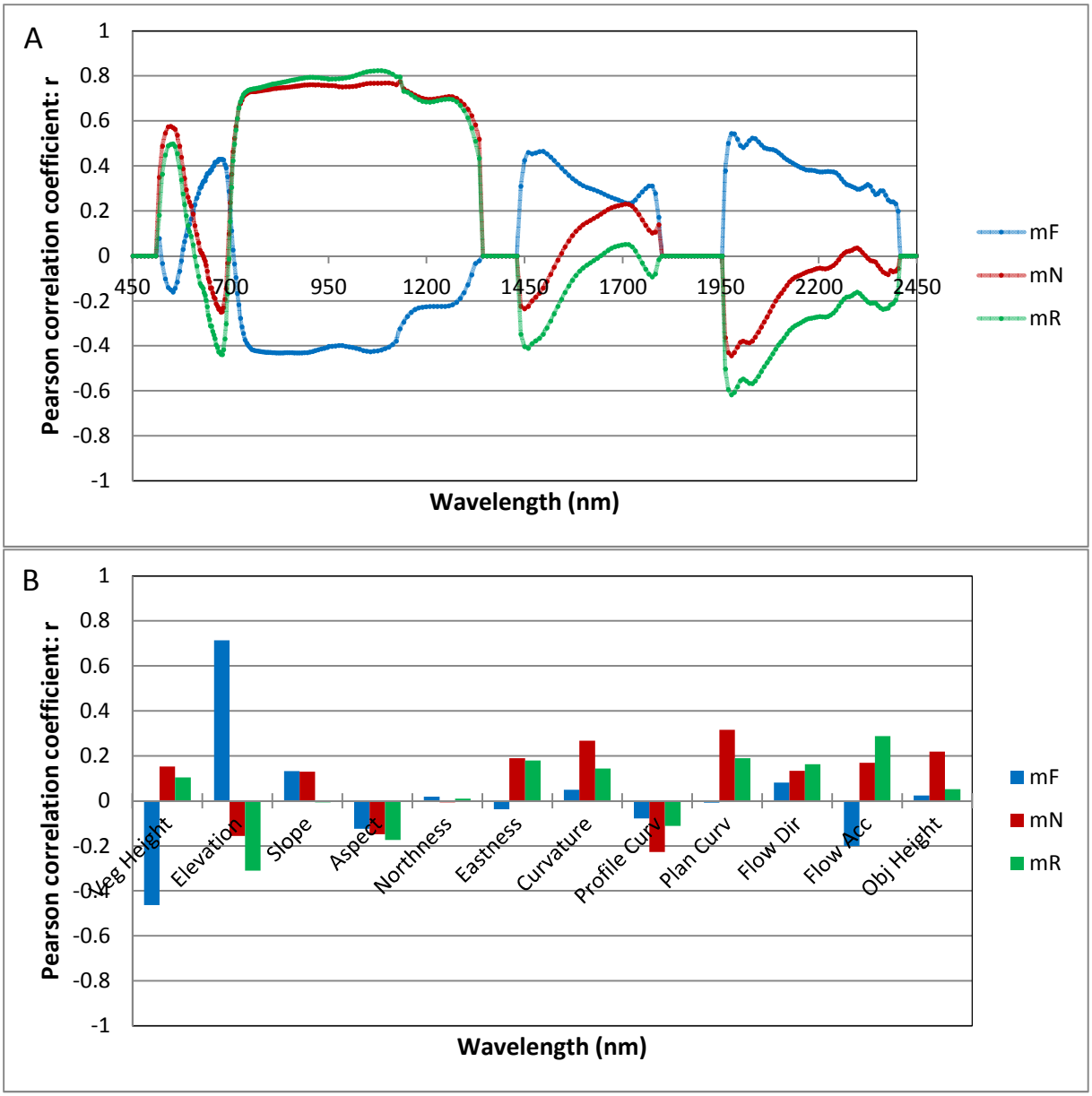


Figure 17. Correlogram based on the Pearson correlation coefficient (r) for hyperspectral predictors (A) and LiDAR predictors (B).

5.3 Prediction of IVs

5.3.1 Hyperspectral Predictors

PLSR is performed to relate observed indicator values to several combinations of spectral and LiDAR predictor variables. The results for the different prediction models are shown in Table 5 and the selected significant wavelengths are given in Figure 18. The p-value of selecting significant predictors (0.05) is changed is some cases, because the model removed too many or too few predictors. To determine the usefulness of LiDAR predictors, first a model is calibrated using only spectral predictors. A high R^2 of 0.822 is found for predicting mR, whereas relatively low accuracies are found for mF and mN (0.491 and 0.515 respectively).

5.3.2 Hyperspectral and LiDAR Predictors

The PLSR with full predictors include hyperspectral and LiDAR predictors. The results of the models do change for all three IVs when the LiDAR variables are added, though no significant LiDAR

predictors were found for mN and mR. In fact, adding LiDAR variables reduced the prediction capability for the models for mN and mR, since non-explanatory LiDAR variables had to be fitted as well in the first run. Probably, the model's robustness is reduced when non-significant variables are added. Also the selected significant wavelengths for mR change, however the ones for mN remain nearly the same. The full mF model also contains elevation and vegetation height as significant predictors. Since vegetation height cannot be used in the final model, an additional model is fitted without this variable, where only elevation remained as a significant LiDAR predictor. For the other two IVs, the vegetation height does not have to be removed separately, because it is no significant predictor.

Table 5. Fitted PLSR models with their properties and accuracies.

IV	Predictors	Method	# LVs	# Pred.	RMSE cal	RMSE val[6]	R^2 cal	R^2 val[6]
mF[1]	Spectral	PLSR	3	12	0.344	0.390	0.606	0.491
mF	Spectral	bnPLSR	3	16	0.372	0.405	0.537	0.452
mF	Spectral+ LiDAR	PLSR	3	50	0.253	0.289	0.786	0.721
mF[2]	Spectral+ LiDAR	PLSR	3	42	0.285	0.322	0.729	0.654
mN	Spectral	PLSR	1	59	0.357	0.375	0.562	0.515
mN[3]	Spectral	bnPLSR	1	10	0.394	0.413	0.465	0.412
mN[4]	Spectral	PLSR	2	42	0.275	0.308	0.763	0.702
mN	Spectral+ LiDAR	PLSR	1	48	0.357	0.376	0.561	0.513
mR	Spectral	PLSR	3	22	0.173	0.201	0.867	0.822
mR	Spectral	bnPLSR	1	7	0.274	0.287	0.668	0.636
mR[5]	Spectral+ LiDAR	PLSR	3	28	0.197	0.224	0.833	0.778

Notes: [1]: Adjusted significance level to 0.1; [2]: Vegetation height is removed as input predictor; [3]: Adjusted significance level to 0.01; [4]: Plots with low mN values are removed; [5]: 6LVs were selected in the first run instead of 2; [6]: Validation is based on LOO cross-validation

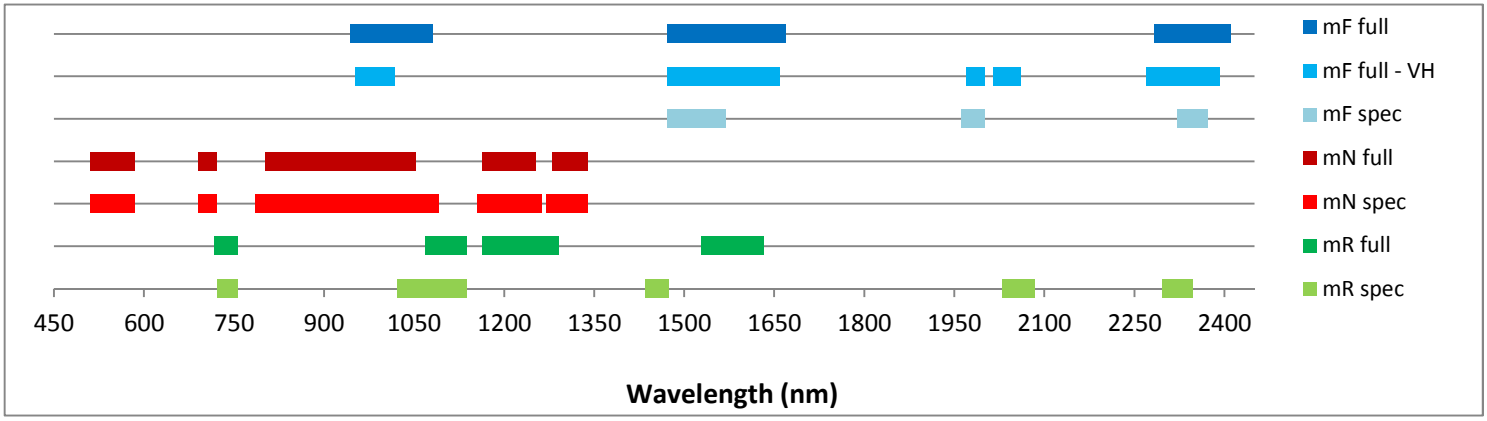


Figure 18. Significant wavelengths selected by the PLSR models. Note that for mF full: elevation and vegetation height are significant LiDAR predictors and for mF full – VH: elevation is a significant predictor too.

5.3.3 Brightness Normalised PLSR

Table 5 shows that brightness normalised PLSR decreases the prediction accuracy for the three IVs. Heterogeneous illumination effects may not occur in the APEX images, in which case the correction removes variances that were actually part of the information that could be used for the prediction. Since the brightness normalization is not used in this research and it is not part of the research questions, the significant wavelengths of these models are not included in Figure 18.

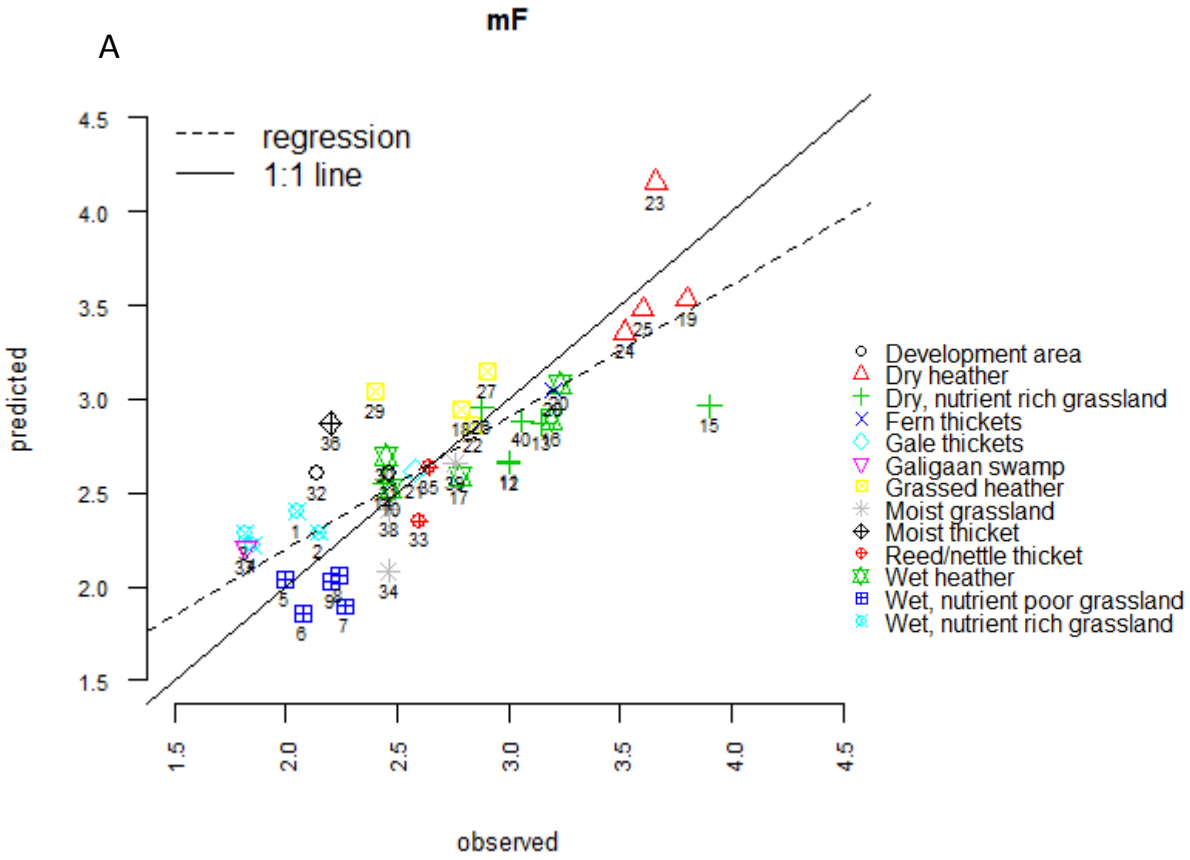
5.3.4 Best Model Selection

By comparing the different prediction models, the best model per IV is selected based on the maximum R^2 of the LOO validation (Table 6). For mF, LiDAR predictors improve the model considerably, but only elevation is used in the final model. For mN and mR, PLSR models with only spectral predictors turn out to be the best and no brightness normalization is used. The scatterplots of the best models show the observed and predicted IVs based on the LOO validation (Figure 19). In general, predicted IVs tend towards mean values resulting in overestimation of low values and underestimation of high values. For mF, the over or underestimation is dependent on the vegetation types, since specific vegetation types are mostly clustered below or under de 1:1 line. VP132 and VP102 are not detected as outliers for the mF prediction, which might have been the case since no elevation data was available and VP132 is based on interpolation over a large area. The mN model overestimated for many low mN observations between 1 and 1.2 except for one outlier, which is dry heather in VP125. It also underestimates slightly for higher values. mR performs well, almost as much vegetation plots are under as well as overestimated. As expected from the biplot, mF and mN are difficult to predict for VP115, which is dry nutrient rich grassland. This vegetation type generally gives bad prediction which could be explained by the fact that the overall relation between mF and mN is different. In general, dry soils correspond with poor nutrients instead of rich. The regression coefficients for the three best models are given in Appendix V. For mF, spectral wavelengths from 967 to 1002 nm give negative regression coefficients. High reflectance in these wavelengths indicates healthy green vegetation (Lillesand, Kiefer et al. 2007) and results in lower mF values, which indicate wet areas. For mN and mR, high reflectance for wavelengths around 1000 nm will result in a higher mN and mR, which indicate nutrient rich and alkaline soils. The other way around occurs at wavelengths from 1488 to 1644 nm, where higher reflectance values increases mF, so drier conditions occur. Furthermore, the wavelength of the red edge (706 nm) is selected for the mN model which is related with the chlorophyll content and therefore the nutrient availability. The correlation between mN and mR is not clearly visible in the significant wavelengths as was expected from the correlogram (Figure 17).

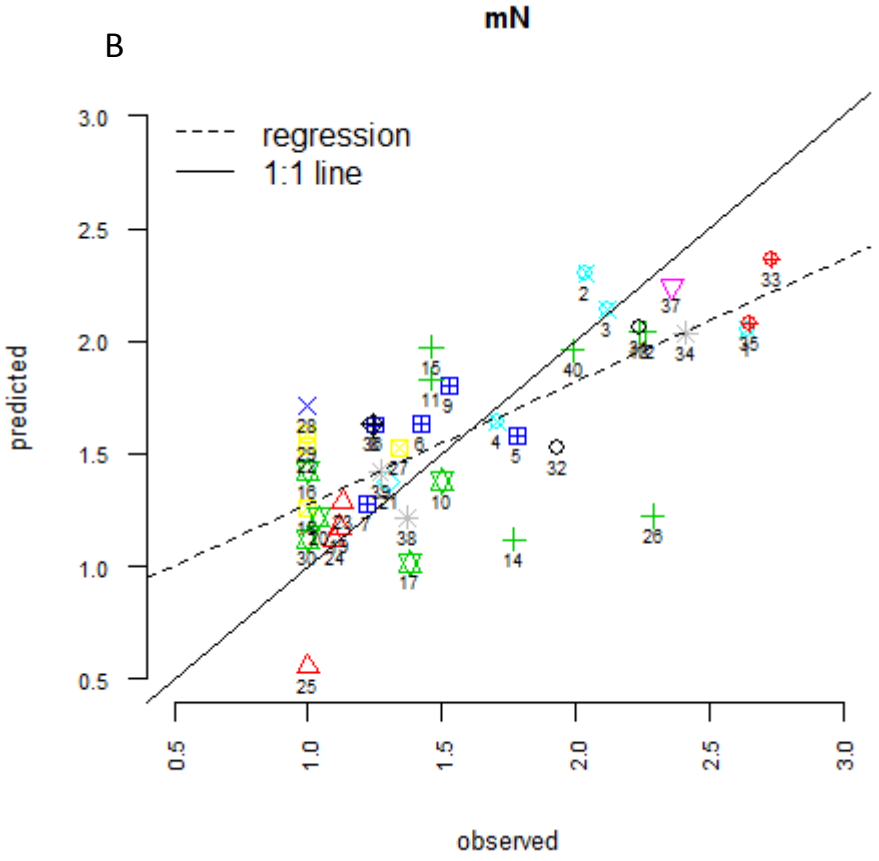
Table 6. The best models selected per IV with their properties.

IV	Method	Predictors	# LVs	# Pred	R^2 cal	R^2 val	RMSE cal	RMSE val
mF	PLSR	LiDAR+Spectral	3	42	0.729	0.654	0.285	0.322
mN	PLSR	Spectral	1	59	0.562	0.515	0.357	0.375
mR	PLSR	Spectral	3	22	0.867	0.822	0.173	0.201

A


mF
regression
1:1 line
predicted
observed
Development area
Dry heather
Dry, nutrient rich grassland
Fern thickets
Gale thickets
Galigaan swamp
Grassed heather
Moist grassland
Moist thicket
Reed/nettle thicket
Wet heather
Wet, nutrient poor grassland
Wet, nutrient rich grassland

B


mN
regression
1:1 line
predicted
observed

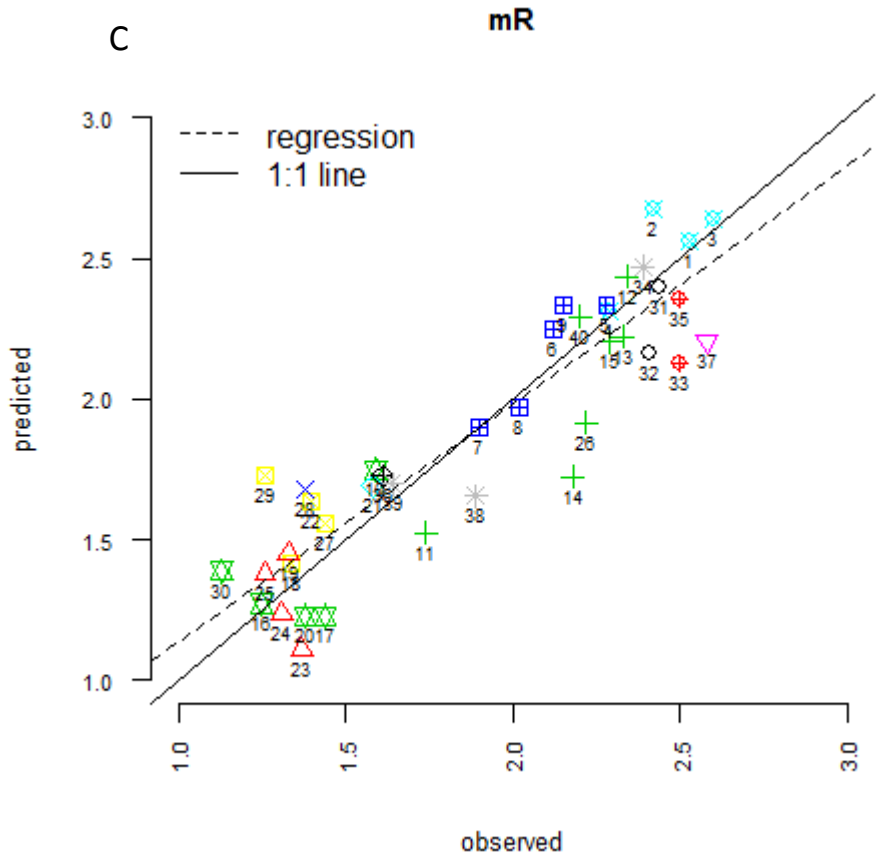


Figure 19. Scatterplots of the observed IVs vs. the predicted IVs by PLSR model for mF (A), mN (B) and mR (C). The symbols indicate the vegetation types and the labels are the numbers of the 40 vegetation plots.

5.3.5 Accuracy Assessment of Best Models

Figure 20A shows that low observed mN values have the largest residuals. In addition, mN also shows the highest residuals for higher observed mN values. Moreover, a (linear) pattern for mN is visible in the residual plots, which suggests that a non-linear model may be more appropriate to use and will give better predictions. There is no structure visible for mR predictions, which indicate independent residuals so the model is robust. mF shows an outlier for the observed mF value near 4, which is predicted much lower. Figure 20B shows a random pattern with values under and above the 0 line for the three IVs. This indicates equal variance of the residuals for the predicted IV range (a normal distribution of the residuals) and no further corrections are needed. Figure 21 shows the relation between the residuals and the vegetation types. The relation shows some dependence, since some vegetation types predict well around the 0 line and others deviate a lot for all three IVs. All IVs are overestimated for grassed heather and all are underestimated for dry, nutrient rich grassland, reed/nettle thicket and moist grassland. The averaged residuals should have been near zero when they are independent from the vegetation types.

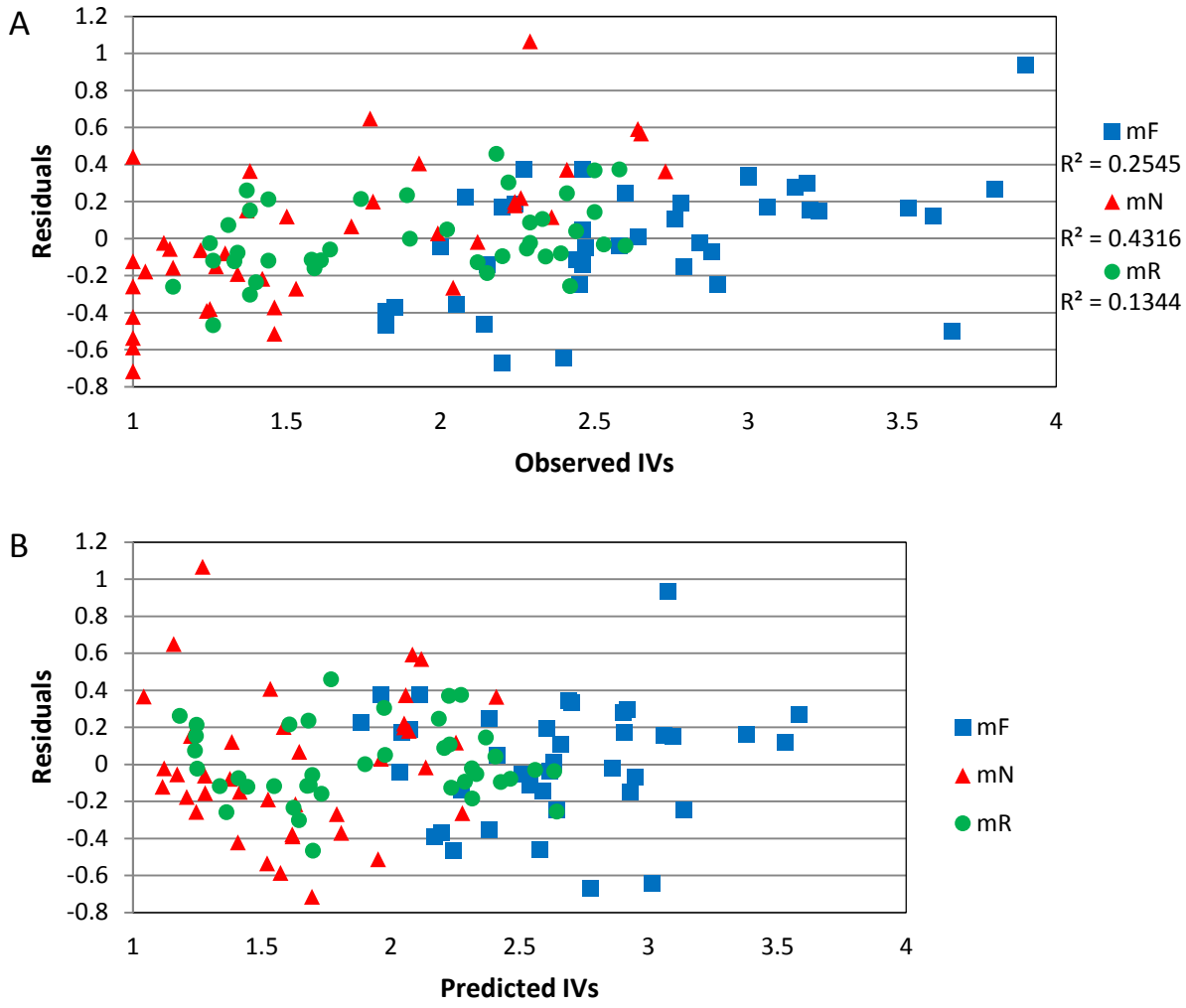


Figure 20. Residuals of the model predictions vs. observed (A) and predicted (B) IVs for the 40 vegetation plots.

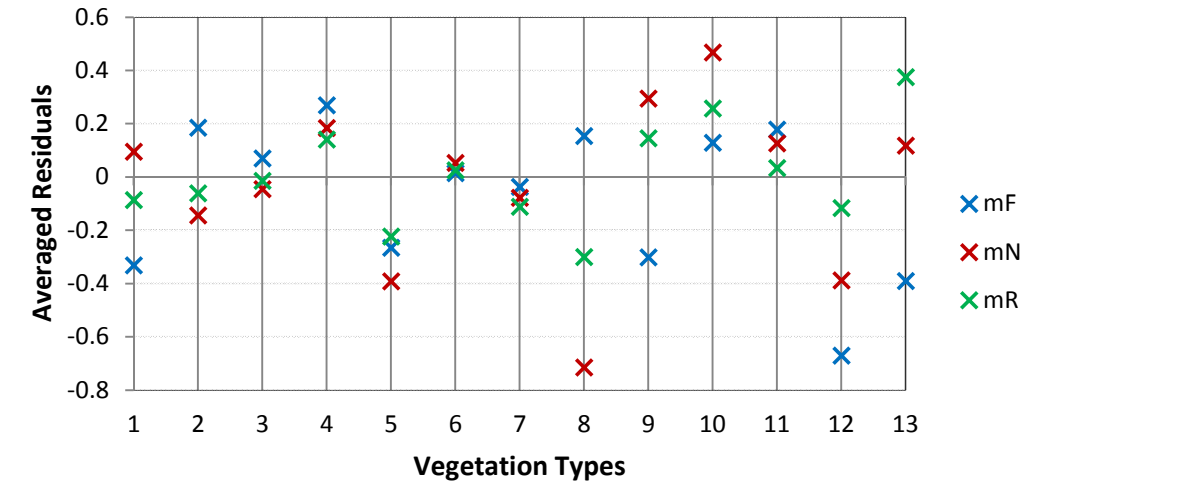


Figure 21. Averaged residuals for each vegetation type. Note that not all vegetation types contain the same amount of plots. The legend for the vegetation types is given in Table 7.

Table 7. Vegetation types correspond to the numbers in Figure 21 with the number of plots that are located in the vegetation types.

# in Figure 25	Vegetation type	Number of plots
1	Wet, nutrient rich grassland	4
2	Wet, nutrient poor grassland	5
3	Wet heather	5
4	Dry, nutrient rich grassland	7
5	Grassed heather	4
6	Dry heather	4
7	Gale thickets	1
8	Fern thickets	1
9	Development area	2
10	Reed/nettle thicket	2
11	Moist grassland	3
12	Moist thicket	1
13	Galigaan swamp	1

The accuracy of the mN model was lowest, which may be the result of different vegetation types occurring in the same mN range. For example, Figure 22 shows the averaged spectral profiles of vegetation types with low mN values between 1 and 1.2. The spectral profiles of these vegetation types deviate especially in the NIR wavelengths, even though they must receive nearly identical mN predictions. Moreover, the selected significant wavelengths for the mN model occur in this region where the reflectance differs. The model is not calibrated with higher wavelengths where the reflectance of the vegetation types is similar, because these wavelengths have a low correlation with mN. Therefore, the mN prediction becomes inaccurate for low mN values. This is confirmed by a model calibration where only vegetation types with mN values smaller than 1.5 (23 plots) are used as input. This model has a R^2 of nearly 0, so mN cannot be predicted accurately and the mean value gives a prediction just as good as the regression model. When these plots are removed from the model, the model improves to a R^2 of 0.70 (Table 5). These results confirm that the model is sensitive for the vegetation plots with a low mN, but those plots cannot be removed. This will only falsely suggest a better prediction, as these vegetation types are an integral part of the study area. For mR, the vegetation types are much more clustered over small regions of the IV range, instead of mixed through the whole range as mN (Figure 19C). mF also has different vegetation types clustered in the range of mF values. For example, the driest mF values only occur for dry heather (Figure 19A). Thus, when the vegetation types can be separated based on their spectral differences, the prediction becomes more accurate.

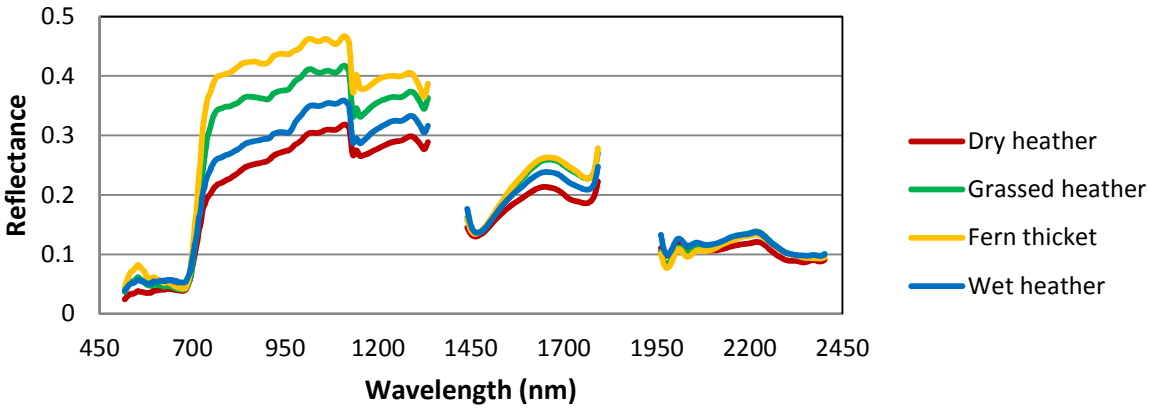


Figure 22. Averaged spectral profiles of vegetation types which occur in nutrient poor areas with mN values between 1 and 1.2. The spectral profiles deviate, even though they must receive nearly identical mN predictions.

5.4 Mapping of IVs

The final models for mN, mR and mF were applied to the whole study area. For mN and mR only hyperspectral bands were used, while for mF also the elevation map was used as input. The resulting map of mF is given in Figure 23; the maps of mN and mR are shown in Appendix VI and VII respectively. The ranges of the estimated values are indicated in Table 8 and the predicted IV ranges are visualised in histograms in Appendix VIII. The histogram patterns match the pattern of the histograms of the averaged IVs from the 40 vegetation plots (Figure 15). Again mR and to a lesser extent, mN show a bimodal structure, which suggest a valid prediction. The mF predictions exceed the theoretical limits of the IV, but only a small amount of pixels falls outside this range (Appendix IX). The large negative values for mF are located at sand paths. This can be explained by the fact that the model is not calibrated on bare soil, but only on the sampled vegetation types. Over 10% of the pixels contain mN estimates outside the IV range, but the values do not greatly exceed those limits since many pixels are predicted just below a mN value of 1. Most of the pixels outside the mN range occur around water bodies. For mR, some fields are predicted below the range especially in grassed heather.

Table 8. Data ranges of the predicted IVs. Note that mF theoretically ranges from 1-4, whereas mN and mR range from 1-3.

	Minimum	**Maximum**	**Mean**	**Standard deviation**	**Fraction of study area with IV estimate exceeding theoretical range**
mF	-3.625	5.412	2.604	0.4367	0.585%
mN	-0.09595	3.524	1.432	0.3689	10.75 %
mR	-0.5029	3.158	1.687	0.3989	3.45%

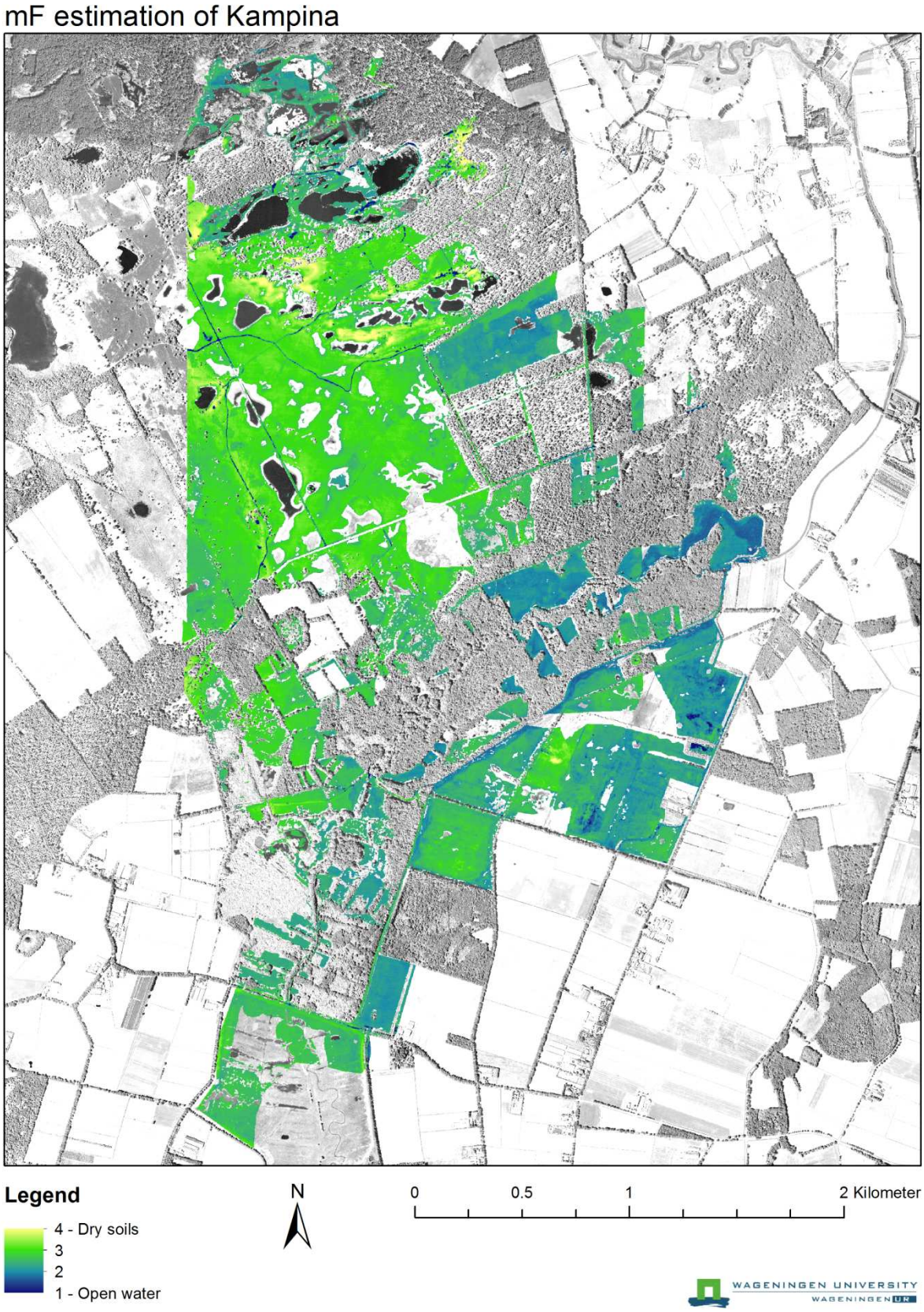
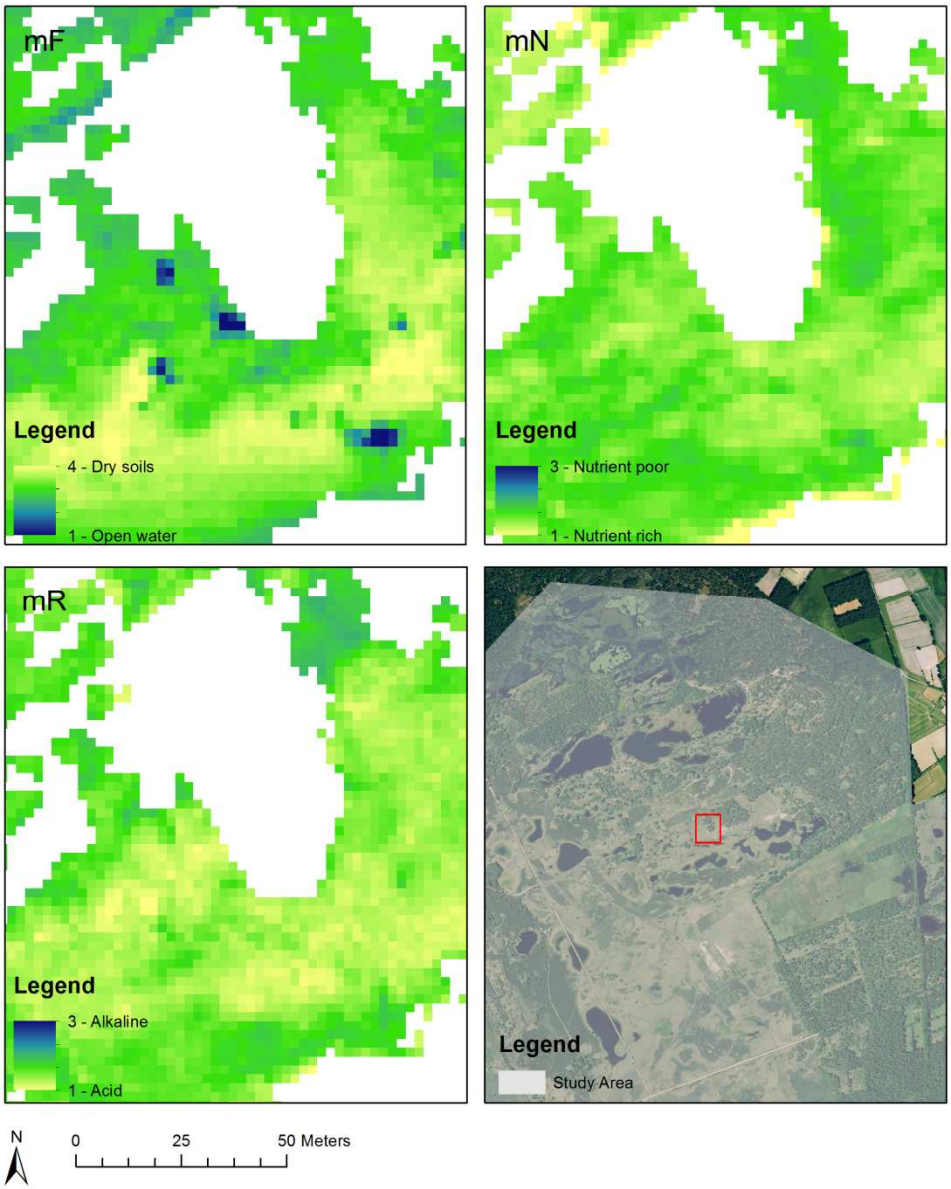


Figure 23. mF estimation of the study area, overlaid on an aerial photograph.

5.5 Evaluation of Spatial Pattern

Figure 24 shows the predictions of IVs for two locations with dry heather and nutrient rich grassland respectively. Dry heather is predicted well, the area receives high values for mF, which indicates dry conditions (Figure 24A). Also the nutrient poor conditions are predicted well, with values lower than 1.5 just like observed in the sampled vegetation plots of similar vegetation type. Also low mR values around 1.3 were found, which was expected. The dry, nutrient rich grasslands (Figure 24B) have a different relation between indicator values than general, but also predict quite well. This field is selected, since VP115 was an outlier and is located in this field. Although, the estimated values may deviate from the observed values, the spatial pattern seems to be valid. The mF is predicted in relatively dry conditions (though, must be lower based on the observed value for VP115); whereas mN values near two indicate a nutrient richer area. This check explained that although the model fit may not be accurate for all vegetation types, the relative spatial patterns still contain a lot of information.

A



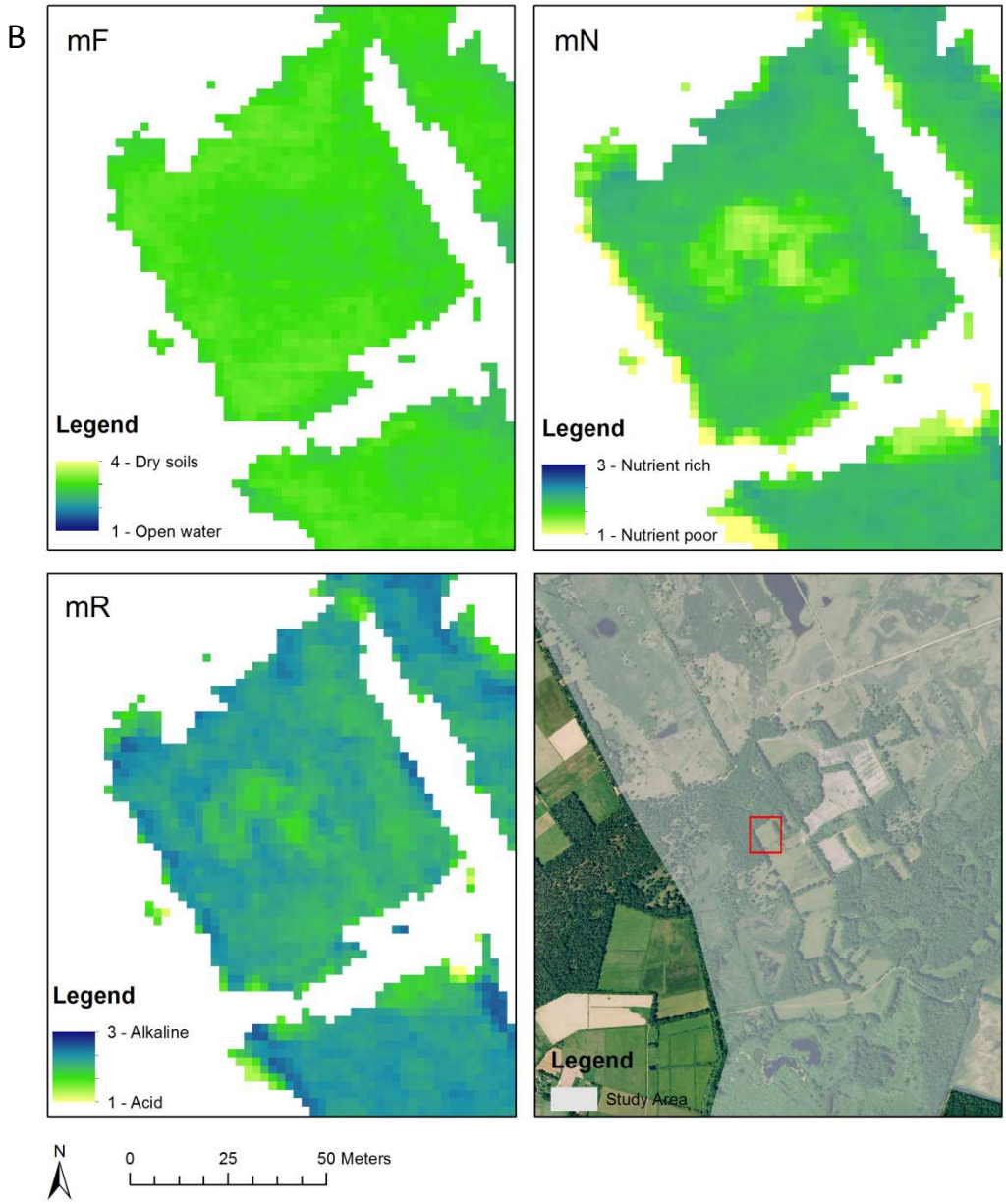


Figure 24. Zoom in areas of the predicted IVs for dry heather (A) and nutrient rich grassland (B).

5.6 External Validation

The results of the external validation for the 17 plots are given in Figure 25. The R^2 and RMSE are calculated as a measure for the accuracy and are given in Table 9. The accuracies of the IV estimations are lower than obtained from the LOO validation of the models. Again, the mN model is proven to be most inaccurate compared to mF and mR.

Table 9. R^2 and RMSE for the prediction of the external validation plots.

	mF	**mN**	**mR**
R^2	0.56	0.44	0.57
RMSE	0.47	0.44	0.36

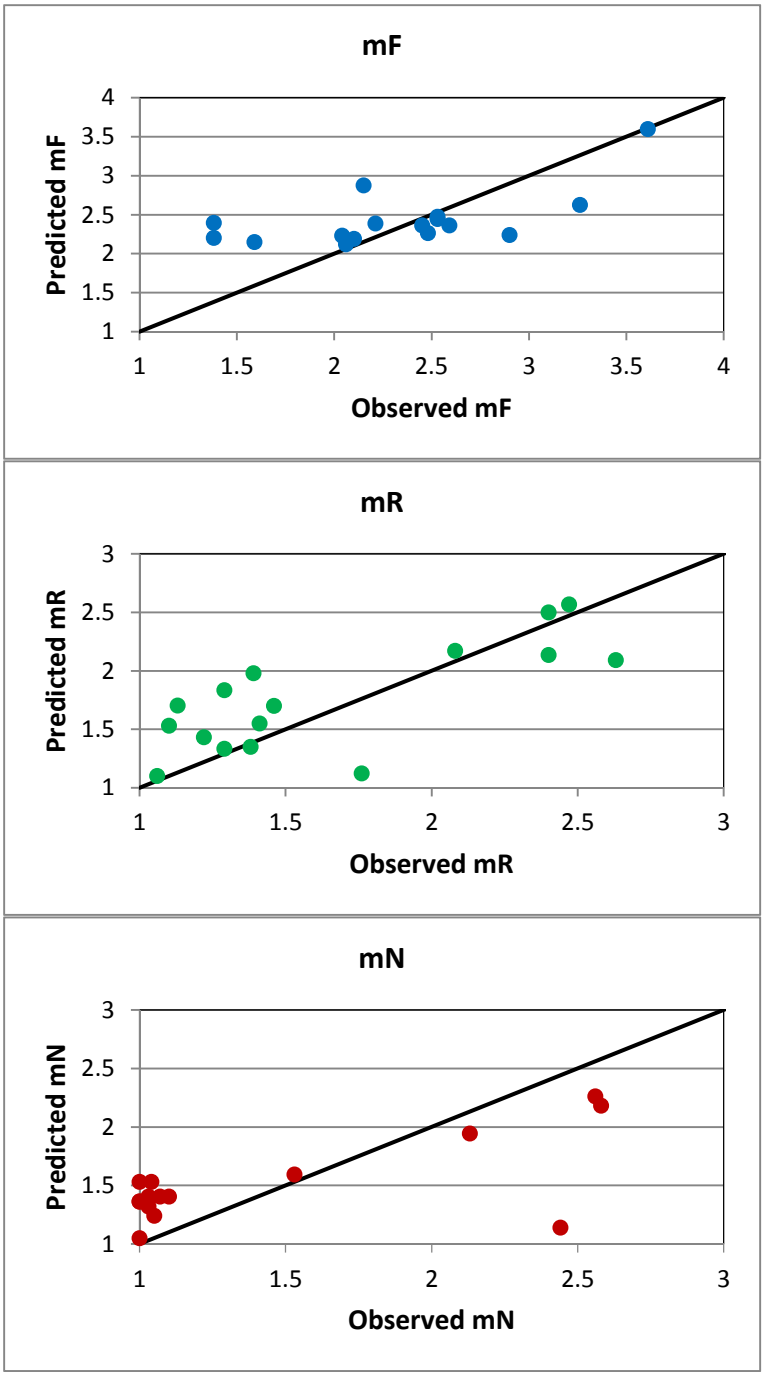


Figure 25. External validation results of 17 validation plots. The observed IVs are plotted against the predicted IVs from the best models. The line indicates the 1:1 line.

6. Discussion

6.1 Averaging IVs

Averaging IVs in vegetation plots contains some uncertainties, because the IVs of the plant species are based on the occurrence in environmental conditions, but plant species do not always grow in their optimum conditions. However, since averaging of multiple plant species is performed, the results should be more reliable instead of only using the occurrence of a single plant species. Additionally, the relation between spectral signature of plant species and IVs is uncertain, because the spectral signature of vegetation relates to other factors. The reflectance relates to chlorophyll content, leaf properties, leaf water content, etc. and not directly to moisture, nutrients and pH of the soil properties. Moreover, the actual soil moisture content is related to soil texture, which cannot be detected by reflectance of vegetation (Lillesand, Kiefer et al. 2007).

6.2 Estimating IVs

6.2.1 Input Data

The DTM derivatives show a weak correlation with the IVs (Figure 17B). They are probably more important in areas with more pronounced elevation differences, where the moisture and nutrient content is more dependent on topographical conditions. Furthermore, the DTM derivatives maps are based on a 3x3 window, which may be too small and more aggregation is needed to prevent pixelated maps. Nevertheless, the reliability of the LiDAR data is sufficient, when comparing with the field measured NAP ($r = 0.89$) (Figure 3). Furthermore, interpolation was performed for VP132 and VP102, but the plots were not detected as outliers for the mF prediction (Figure 19A), in which elevation is part of the regression model. This is evidence for an accurate interpolation method that is sufficient for this research. Additionally, inaccuracies of the derivatives are not important for the prediction, since no derivatives are used in the models.

The APEX images also contain inaccuracies, because they were acquired during suboptimal weather conditions. Clouds and cloud shadows were removed from the area and atmospheric correction was applied, but atmospheric effects are still visible when comparing different flightlines at the same location (Figure 5). Significant wavelengths for which the flightlines and field spectrometer reflectance have large differences may give an extra inaccuracy to the prediction. This is the case for the mN prediction, where the model selects wavelengths around 700 and 900 nm, where the largest deviations occur between flightlines and field spectrometer. Furthermore, no validation is performed to check the accuracy of the classification of trees and water. Edge effects occur in some places when the visual check was not sufficient, so predictions around trees and water can be inaccurate in places where still influences of water and trees occur. The mN prediction indicates some uncertainties, because predictions are lower than the theoretical IV range around water bodies (Appendix IX). The spectral variables for VP125 and VP126 were determined as the nearest wavelength measurement from the field spectrometer. When taking the mean of the wavelength width of APEX bands may give more accurate results. Moreover, the data from the field spectrometer is collected on a different date than the APEX flight was carried out, which can result in extra inaccuracy. This can explain the high residuals for these vegetation plots, for which especially the mN prediction for VP126 is inaccurate.

6.2.2 PLSR

PLSR is a flexible method, because different IVs can be predicted in different areas based on different predictor variables. Any input spectral predictors can be used and any LiDAR predictors can be added. Though, PLSR also contains some limitations. The method is sensitive for specific combinations of predictors that are used to estimate the relation with the response variable. Addition and removal of predictors influences the other predictors as well. For example, the significant spectral predictors change when LiDAR variables are removed, even though these variables do not significantly contribute to the prediction (Figure 18). Even when only one variable is removed, such as vegetation height for the mF model, the significant spectral predictors change as well. This suggests some limitation of PLSR when combining LiDAR and spectral predictors. Furthermore, the PLSR models are sensitive for the amount of latent variables chosen based on the RMSE of validation. When choosing a different number of latent variables, when the difference in RMSE of validation is minimal, the significant predictors in the model changes, as the accuracy. Finally, the variable reduction method used in this study, is not a robust method. It is based on removing the predictors with a p-value bigger than 0.05 from jackknifing, but this level cannot be used for all cases. Jackknifing has some limitations as well. By using a t-test, it assumes that the distribution of the regression coefficients is a normal distribution, but this is not checked by this method. Therefore, the actual p-values should be used with care and are more indicators of the relative stability of the regression coefficients.

Finally, no spatial uncertainty map can be created, since the accuracy per pixel is unknown when using PLSR. Though, this should be possible when more analysis and calculations are done. This is a limitation of using PLSR, while this is part of the output of different methods like Gaussian Process Regression. When the accuracy is known, it can be used as input for vegetation type modelling. In this way, the uncertainty can be added to the uncertainty of the new modelling step, so spatial accuracies can be calculated.

6.2.3 Predicted IVs

mN predictions become inaccurate for low observed mN values, as can be seen from the high residuals in Figure 20A. Different vegetation types can have different strategies to cope with the low amount of nutrients, which result in different spectral signatures. Therefore, spectral predictors cannot fully explain the variation in mN values in the study area. Consequently, adding other input predictors may improve the prediction. Also, water, nutrient, or pH stress for vegetation can result in similar effects in the spectral profile, which makes the prediction more difficult. In this case, mN was difficult to predict, but is dependent on the vegetation types occurring in the area. In contrast, accurate mN predictions can be found in coastal areas, where mF is more difficult to predict accurately (Roelofsen, Kooistra et al. Submitted). The accuracy decreased when vegetation height was removed as predictor variable for the mF models. If a map of vegetation height was available, the mF prediction could be improved and a better prediction could be made for the study area. The accuracy of the predicted IVs is somewhat comparable with other studies (Table 2), but also differs (Schmidtlein and Sassin 2004; Schmidtlein 2005; Witte, Wojcik et al. 2007). Higher accuracies are found for mF than in these studies, which can be explained by the fact that LiDAR predictors did improve the accuracy of the model significantly. Furthermore mR predictions are also more accurate, which may be the result of the simple bimodal distribution of mR values in the study area. mN accuracies are less accurate than found in the other studies, which suggests a more complex relation with vegetation types. Many vegetation types occur in the low range of mN, different from other

studies, where mN is predicted for grasslands. Furthermore, overestimation of low values and underestimation of high values is a similar result that is found in other studies (Schmidtlein and Sassin 2004; Schmidtlein 2005; Roelofsen, Kooistra et al. Submitted).

6.3 External Validation

The independent external validation by using the 17 vegetation plots provided useful information, because the accuracy was reasonable. This was a completely independent external validation unlike other studies which only used LOO cross-validation (Schmidtlein and Sassin 2004; Schmidtlein 2005) or separated the dataset into a test and training dataset (Witte, 2007). The lower accuracies found can be assigned to some factors that add to the inaccuracy of this validation. The dataset is completely independent from the model calibration, only contains 17 plots which were sampled by different people and do not cover the full study area with vegetation types. Furthermore, the samples do not cover the full IV ranges and a small number of plots seem to be outliers (which could be related to vegetation types) and reduce the accuracy significantly. Finally, the acquisition time is different because these plots were sampled in 2009, when the conditions could be different.

6.4 Implication of IVs Maps

The averaged IVs that are mapped are correlated with soil properties, even though they are vegetation attributes and not environmental attributes. A change in the environment when no change in vegetation occurs will not be detected by this method, so the IVs are not directly related to the environmental factors. Additionally, human influence and management may result in more inaccuracy, because this is not used in the models but may adjust the occurrence of vegetation types which were otherwise not occurring based on the soil properties. Uncertainties in the relation between IVs and actual soil properties are irrelevant when IVs are used for vegetation modelling (PROBE). In contrast, it is important when IVs are used for nature management, when information about the soil properties is needed. This brings an extra uncertainty to the already indirect relation between the reflection of plant species and IVs. Though, the relation between the averaged mR of the vegetation plots compared with the actual soil pH is proven to give a strong correlation (Figure 1). Also mF and mN are highly correlated with various field measured variables as shown by (Schaffers and Sýkora 2000). Therefore, the IVs give a worthy indication of the soil properties.

The prediction of IVs can be used for management in Kampina, but repeated image acquisition is required to keep the IV estimates up to date, since the area is expected to change over time, because of management, succession, etc. In the areas that are part of the conserved area, naturally the LiDAR variables will not change much over small periods of time when no drastic management takes place, like excavating the top soil. mR has the longest temporal resolution followed by mN, when biomass is not removed from the area. mF has a small temporal resolution, because of the influence of weather and seasonal conditions. Monitoring problems like dehydration of the soil will be possible when flights are performed about once every five years in the same growing period. The vegetation response of a changing environment is lagged, so it may be difficult to detect changes accurately in a shorter time period (Schmidtlein 2005). The relations that are found in this study can be used again when approximately the same vegetation types still occur in the area. In this way, the relation found should still be valid, since the spectral signatures of vegetation in the same time of the year, do not change.

Since the model is based on the reflectance of specific vegetation types, the prediction model cannot be used in other areas with different vegetation types. However, the method could be used when new sampling and analysis is performed. Based on the results in this research, the amount of samples should be based on the number of vegetation types occurring in the study area. For each vegetation type, a number of samples should be taken. Finally, the hyperspectral images are relatively expensive, so the natural value of the specific area determines if this method is useable for nature conservation.

7. Conclusions

IVs can be successfully mapped from combined hyperspectral and LiDAR data by using PLSR. For the study area in Kampina, the R^2 of the validation is 0.654, 0.515 and 0.822 for mF, mN and mR respectively. Brightness normalisation reduced the accuracy of the models, so for this case, the correction is not used for the IVs predictions. The best mN and mR models are based on spectral predictors only, since no LiDAR predictors are significant for the prediction of these IVs. The LiDAR data did improve the model prediction for mF, where a R^2 of 0.491 is found with only spectral predictors and is improved to 0.721 when LiDAR data is added. However, vegetation height cannot be used, so a R^2 of 0.654 remains for use in this research. For this study area, only elevation and vegetation height turned out to be significant predictors for mF. Many other topographical variables were extracted and may be useful for other study areas. The spatial variation of the estimated IVs reflect the expected spatial patterns well. Moreover, the external validation shows a R^2 of 0.56, 0.44 and 0.57 for mF, mN and mR respectively. Therefore, this research is evidence that accurate IV predictions can be made when combining hyperspectral and LiDAR data. This is important for further research, where IVs can be used as an input for vegetation type predictions.

8. Recommendations

For this research, the LiDAR data obtained in early spring was not suitable to determine the vegetation height for low vegetation. The accuracy of the prediction of mF can be increased by using LiDAR point data of the study area, which is acquired on a date when vegetation cover (leaf-on-period) is still present. In this way, the vegetation height can be derived by selecting the highest point in a defined window. Moreover, more information about the vegetation structure can be derived when using the LiDAR point data. Variables about the vegetation structure can be added to the prediction model and may improve model accuracies. A LiDAR sensor can be mounted next to the hyperspectral sensor, so data can be acquired at the same time.

Relatively accurate estimations of IVs are found with this method, which reduces time and costs compared to manual field sampling. The complex relation between reflectance of vegetation and IVs is difficult to predict. Therefore, it is important to include the different vegetation types that occur in the sampling or remove non-sampled vegetation types from the study area. Otherwise the accuracy of the prediction will be overestimated. When the same vegetation types occur, the relation of IVs with spectral and topographical variables can be used that is found in this research or from other similar studies. In this way, fewer samples are needed. Furthermore, the database of synbiosis can be used, which contains vegetation plots with sampled plant species (Synbiosis 2012). These can be used for the calibration of IV models and will decrease the amount of plots needed. However, the location of the synbiosis plots is inaccurately determined, so when sampling the predictors, sampling over a larger area may be needed.

Although there are uncertainties in all steps of this method, a relatively accurate prediction of IVs can be made. If these predictions are sufficient to use as an input for vegetation prediction models, should be determined in further research.

References

APEX. (2012). "APEX sensor." Retrieved 06-09-2012, from http://www.apex-esa.org/.

Aptroot, A. (2009). Flora- en vegetatiekartering van Kampina in 2009. Rapport Natuurmonumenten. 's-Graveland.

Berendsen, H. J. A. (2008). Landschappelijk Nederland.

Bruijn, L. d., P. Voorn, et al. (2010). De flora en fauna van de Kampina: overzichtsrapportage inventarisaties 2000-2008.

Cho, M. A., A. Skidmore, et al. (2007). "Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression." International Journal of Applied Earth Observation and Geoinformation **9**(4): 414-424.

Colgan, M. S., C. A. Baldeck, et al. (2012). "Mapping savanna tree species at ecosystem scales using support vector machine classification and BRDF correction on airborne hyperspectral and LiDAR data." Remote Sensing **4**(11): 3462-3480.

Diekmann, M. (2003). "Species indicator values as an important tool in applied plant ecology - a review." Basic and Applied Ecology **4**(6): 493-506.

Ecker, K., L. T. Waser, et al. (2010). "Contribution of multi-source remote sensing data to predictive mapping of plant-indicator gradients within Swiss mire habitats." Botanica Helvetica **120**(1): 29-42.

Ellenberg, H. W., H.E.; Duell, R.; Wirth, V.; Werner, W.; Paulissen, D. (1991). "Indicator values of plants in Central Europe." Scripta Geobotanica **18**: 248.

ENVI. (2012). "ENVI Tutorial: Classification Methods." Retrieved 28-11-2012, from http://www.exelisvis.com/portals/0/tutorials/envi/Classification_Methods.pdf.

Eriksson, L., P. L. Andersson, et al. (2006). "Megavariate analysis of environmental QSAR data. Part II - Investigating very complex problem formulations using hierarchical, non-linear and batch-wise extensions of PCA and PLS." Molecular Diversity **10**(2): 187-205.

ESRI. (2011). "ArcGIS Resource Center." Retrieved 19-03-2012, from http://help.arcgis.com/en/arcgisdesktop/10.0.

Feilhauer, H., G. P. Asner, et al. (2010). "Brightness-normalized Partial Least Squares Regression for hyperspectral data." Journal of Quantitative Spectroscopy & Radiative Transfer **111**(12-13): 1947-1957.

Feng, Y. Z., G. Elmasry, et al. (2013). "Near-infrared hyperspectral imaging and partial least squares regression for rapid and reagentless determination of Enterobacteriaceae on chicken fillets." Food Chemistry **138**(2-3): 1829-1836.

Kafer, J. and J. P. M. Witte (2004). "Cover-weighted averaging of indicator values in vegetation analyses." Journal of Vegetation Science **15**(5): 647-652.

Klaus, V. H., T. Kleinebecker, et al. (2012). "NIRS meets Ellenberg's indicator values: Prediction of moisture and nitrogen values of agricultural grassland vegetation by means of near-infrared spectral characteristics." Ecological Indicators **14**(1): 82-86.

Kokaly, R. F., G. P. Asner, et al. (2009). "Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies." Remote Sensing of Environment **113**(SUPPL. 1): S78-S91.

Lillesand, T. M., R. W. Kiefer, et al. (2007). Remote Sensing and Image Interpretation.

Martens, H. and M. Martens (2000). "Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)." Food Quality and Preference **11**(1-2): 5-16.

Mevik, B. H. and R. Wehrens (2007). "The pls package: Principal component and partial least squares regression in R." Journal of Statistical Software **18**(2): 1-23.

Mulder, V. L., S. de Bruin, et al. (2011). "The use of remote sensing in soil and terrain mapping - A review." Geoderma **162**(1-2): 1-19.

Natuurmonumenten. (2012). "Kampina." Retrieved 06-09-2012, from http://www.natuurmonumenten.nl/content/kampina-0.

Ott, L. R. and M. T. Longnecker (2010). An Introduction to Statistical Methods And Data Analysis.
Roelofsen, H. D., L. Kooistra, et al. (Submitted). "Modelling a priori defined floristically based vegetation types in coastal ecosystems using remote sensing derived vegetation characteristics."
Runhaar, J., W. van Landuyt, et al. (2004). "Revision of the ecological species groups for the Netherlands and Flanders." Gorteria: Tijdschrift voor Onderzoek aan de Wilde Flora **30**(1): 12-20.
Schaffers, A. P. and K. V. Sýkora (2000). "Reliability of Ellenberg indicator values for moisture, nitrogen and soil reaction: A comparison with field measurements." Journal of Vegetation Science **11**(2): 225-244.
Schmidt, K. S., A. K. Skidmore, et al. (2004). "Mapping coastal vegetation using an expert system and hyperspectral imagery." Photogrammetric Engineering and Remote Sensing **70**(6): 703-715.
Schmidtlein, S. (2005). "Imaging spectroscopy as a tool for mapping Ellenberg indicator values." Journal of Applied Ecology **42**(5): 966-974.
Schmidtlein, S. and J. Sassin (2004). "Mapping of continuous floristic gradients in grasslands using hyperspectral imagery." Remote Sensing of Environment **92**(1): 126-138.
Synbiosis. (2012). "Database met vegetatieopnames." Retrieved 06-09-2012, from http://www.synbiosys.alterra.nl/natura2000/googlemapslvd.aspx.
Verrelst, J., E. Romijn, et al. (2012). "Mapping vegetation density in a heterogeneous river floodplain ecosystem using pointable CHRIS/PROBA data." Remote Sensing **4**(9): 2866-2889.
Wamelink, G. W. W., P. W. Goedhart, et al. (2005). "Plant species as predictors of soil pH: Replacing expert judgement with measurements." Journal of Vegetation Science **16**(4): 461-470.
Waterschapshuis, H. (2012). "AHN." Retrieved 06-09-2012, from http://www.ahn.nl/.
Witte, F., M. De Haan, et al. (2007). "PROBE: A model for vegetation targets." Landschap **24**(2): 77-87.
Witte, J. M. and L. Kooistra (2008). Vegetatiekartering via remote sensing. Kiwa Water Research. Nieuwegein.
Witte, J. P. M., R. B. Wojcik, et al. (2007). "Bayesian classification of vegetation types with Gaussian mixture density fitting to indicator values." Journal of Vegetation Science **18**(4): 605-612.
Wold, S., L. Eriksson, et al. (2004). "The PLS method -- partial least squares projections to latent structures -- and its applications in industrial RDP (research, development, and production).".
Zelený, D. and A. P. Schaffers (2012). "Too good to be true: Pitfalls of using mean Ellenberg indicator values in vegetation analyses." Journal of Vegetation Science **23**(3): 419-431.

Appendices

I. Spectral Signatures of the 40 Vegetation Plots

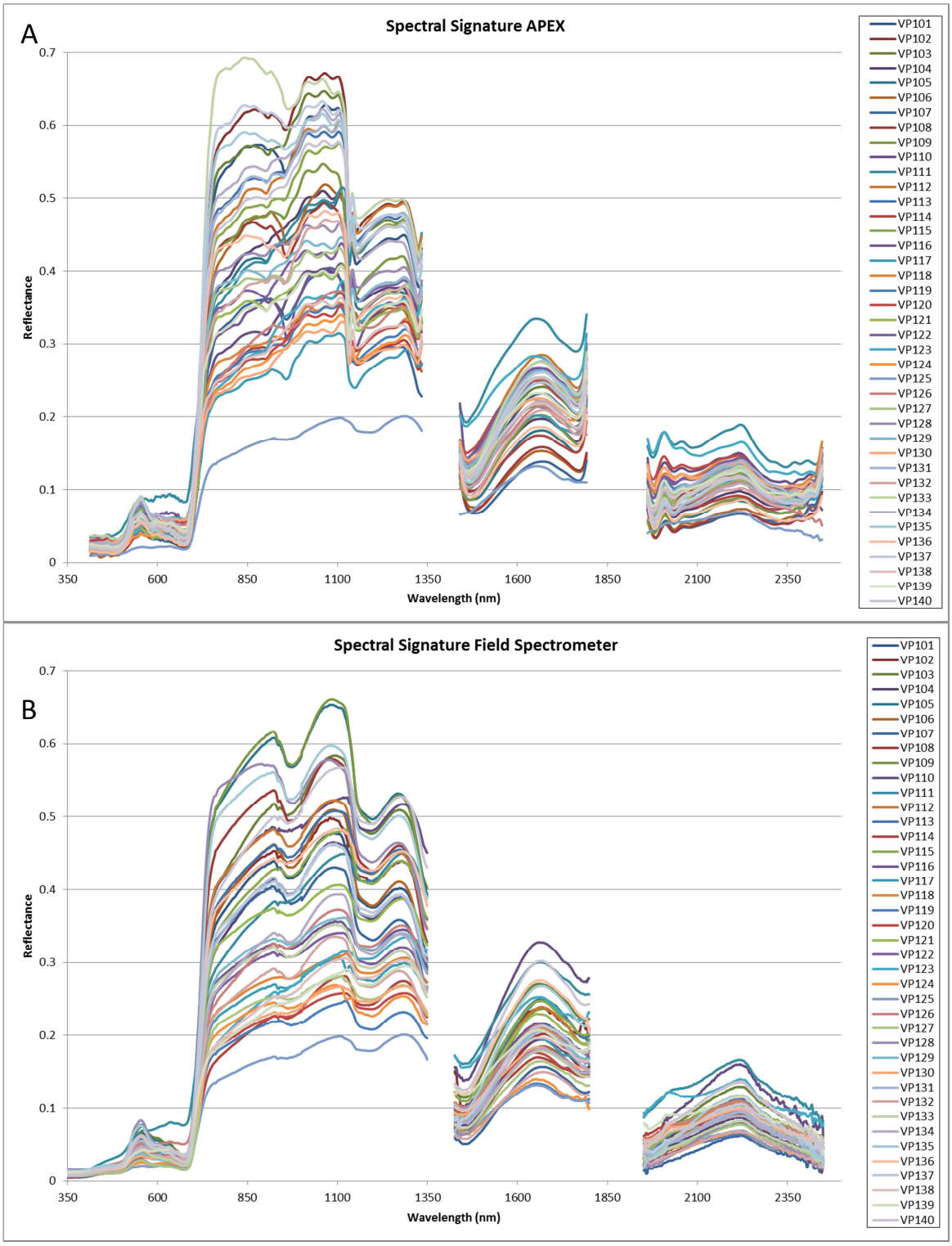


Figure 26. The spectral signatures are derived for the 40 vegetation plots for both the APEX images (A) and the field spectrometer data (B). Note that the data is not acquired on the same day.

II. Reference Spectra from Field Spectrometer Compared to APEX Reflectance

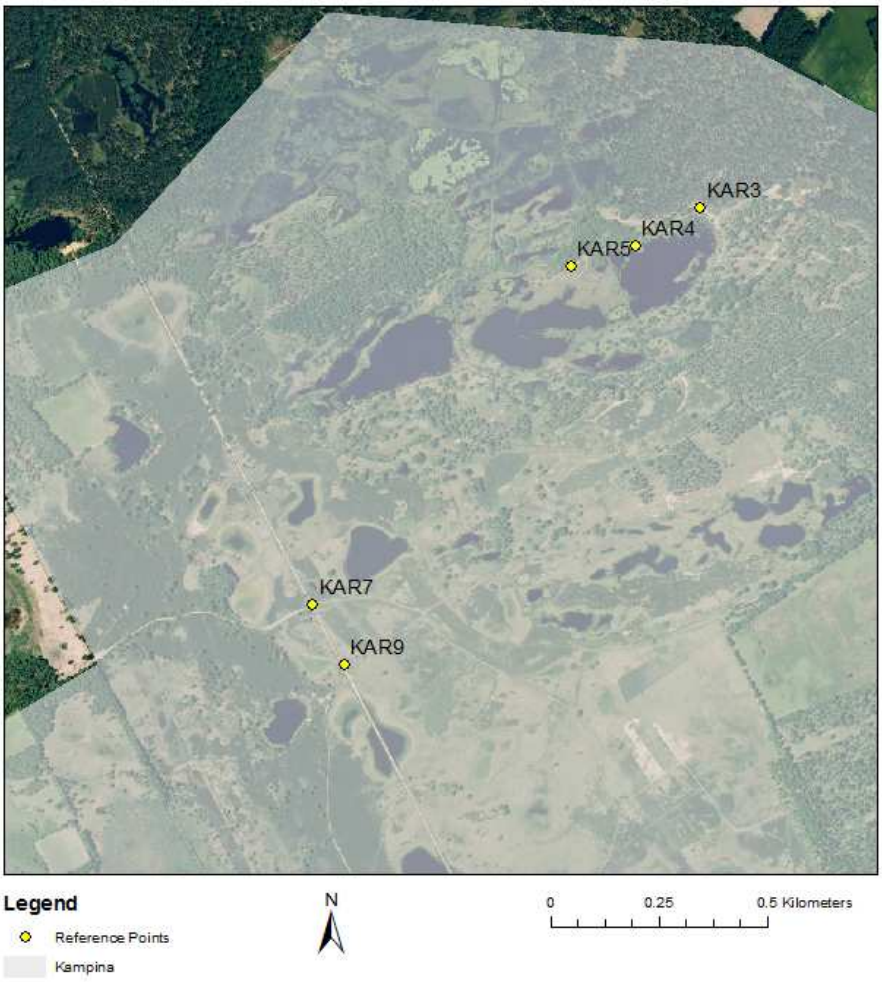
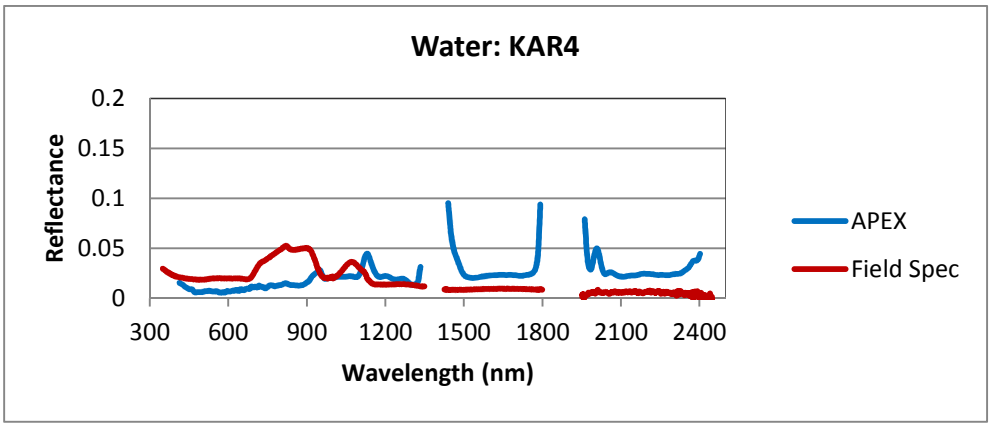


Figure 27. Locations of the five reference spectra overlaid on an aerial photograph. Note that the photograph is taken in 2006 and in a different time of the year, so water points are actually located in water on not in bare sand.



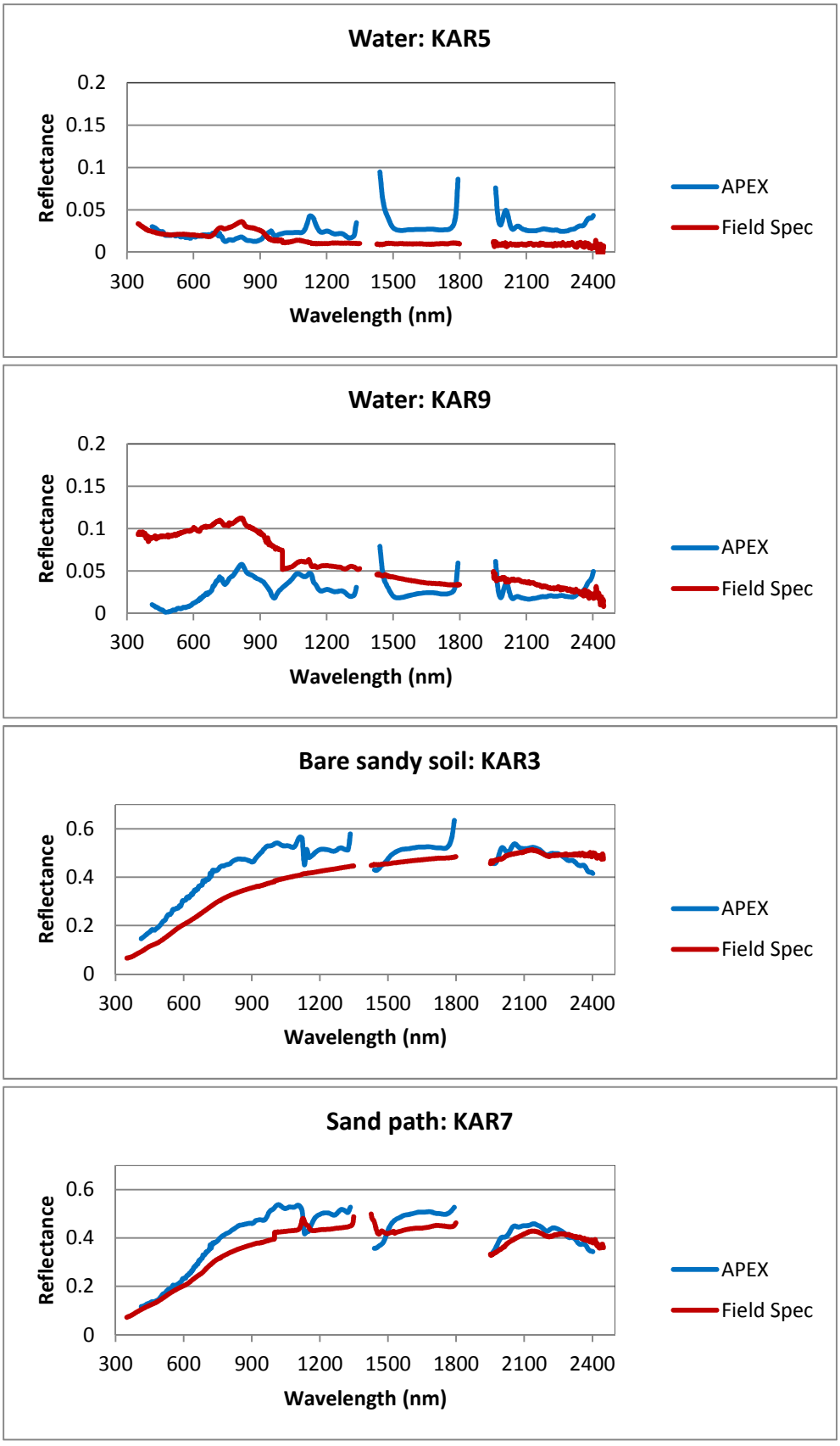


Figure 28. Spectral signatures of the reference points. Five reference spectra from the field spectrometer can be compared to the reflectance in the APEX images.

III. DTM of the Study Area

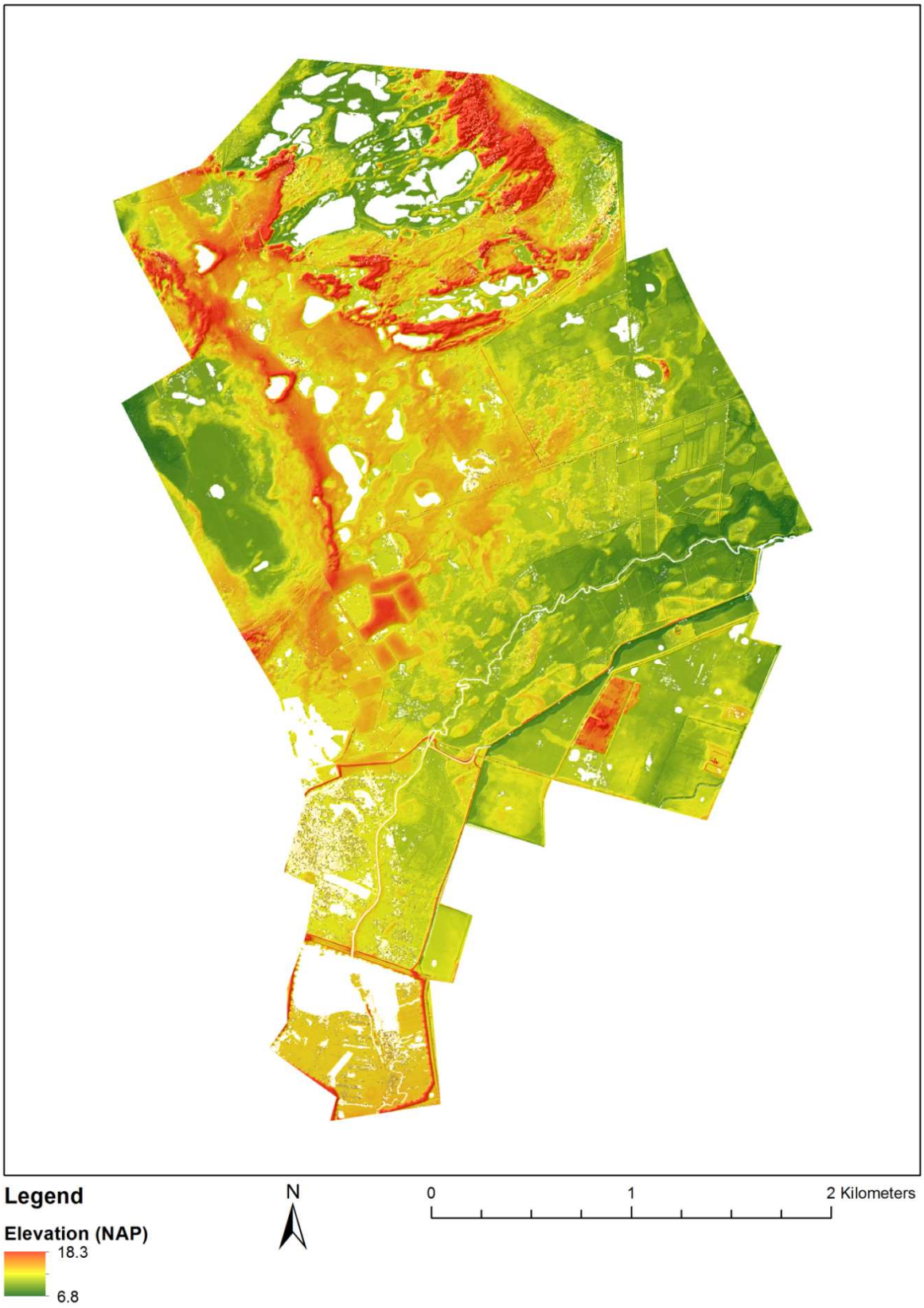


Figure 29. The DTM is based on the AHN-2 data. The white areas are parts where no data is available because of water and tree shadows.

IV. Slope of the Study Area

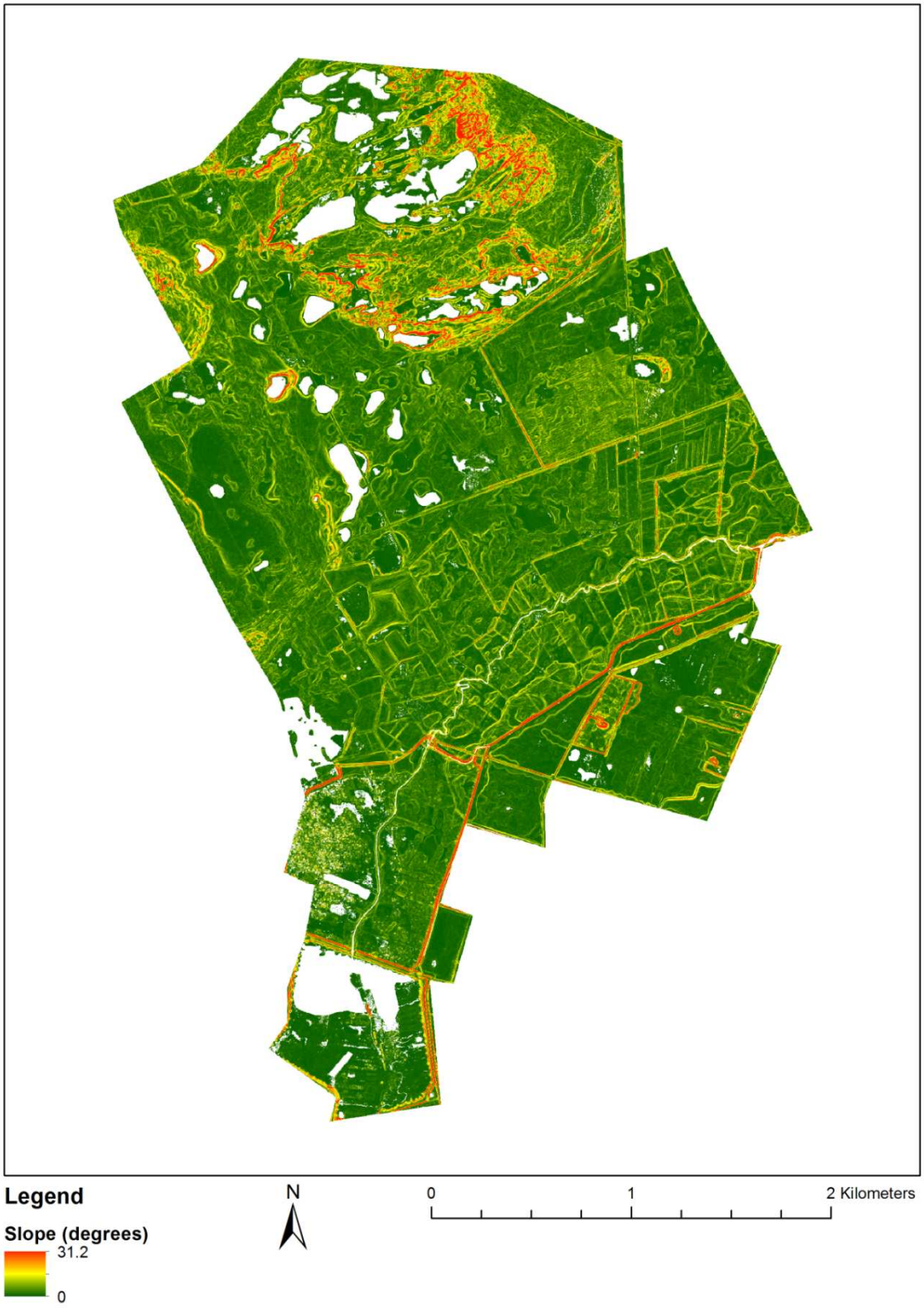


Figure 30. This slope map is based on the DTM of Appendix III. It is calculated based on a 3x3 window.

V. Regression Coefficients of the Three Best Models

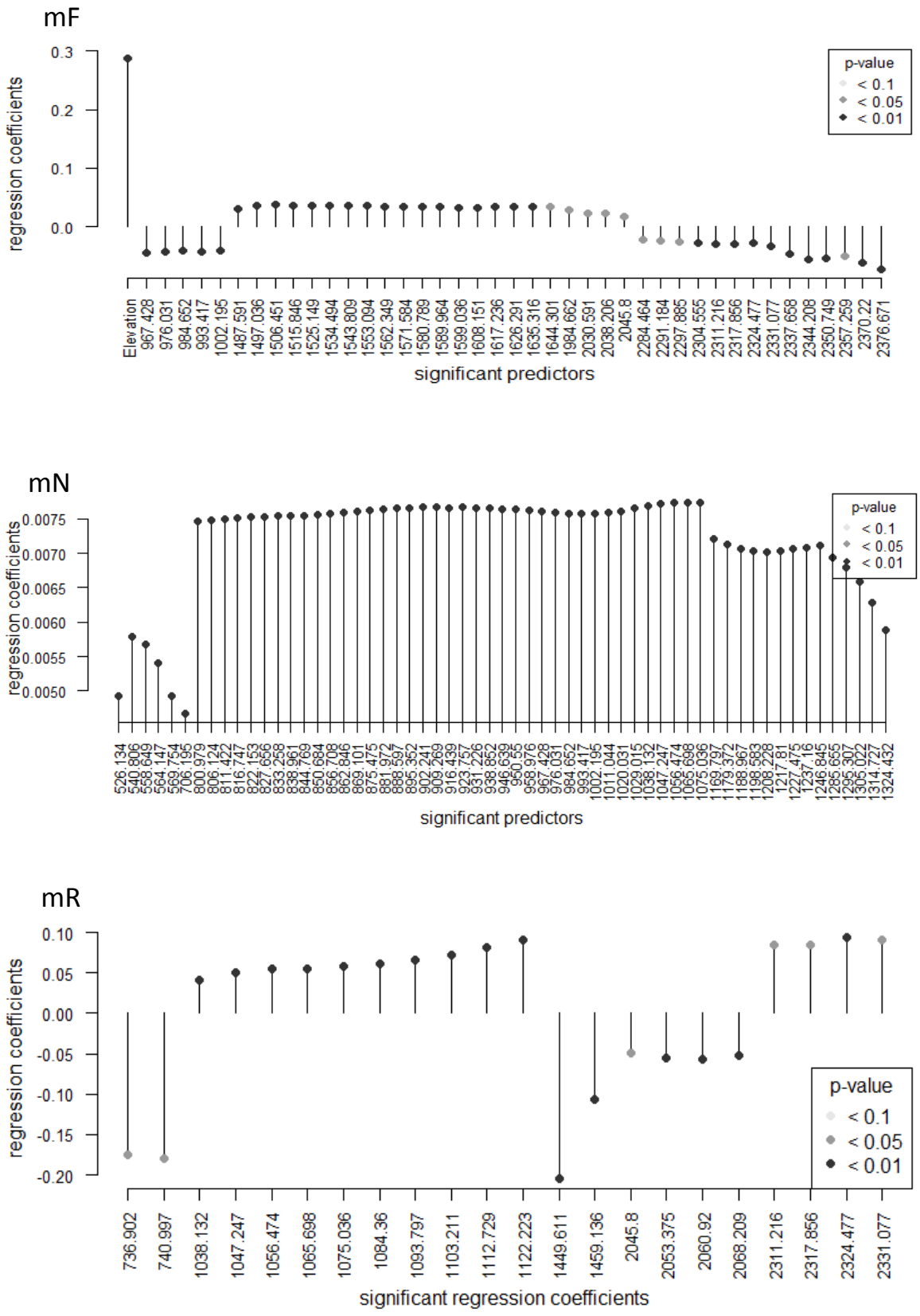


Figure 31. The significant regression coefficients of the three models. The p-value indicates the significance of the coefficients. As can be seen, all predictors have significance levels smaller than 0.05.

VI. mN Prediction

Figure 32. Estimated mN in the study area.

VII. mR Prediction

Figure 33. Estimated mR of the study area

VIII. Histograms of Predicted IVs

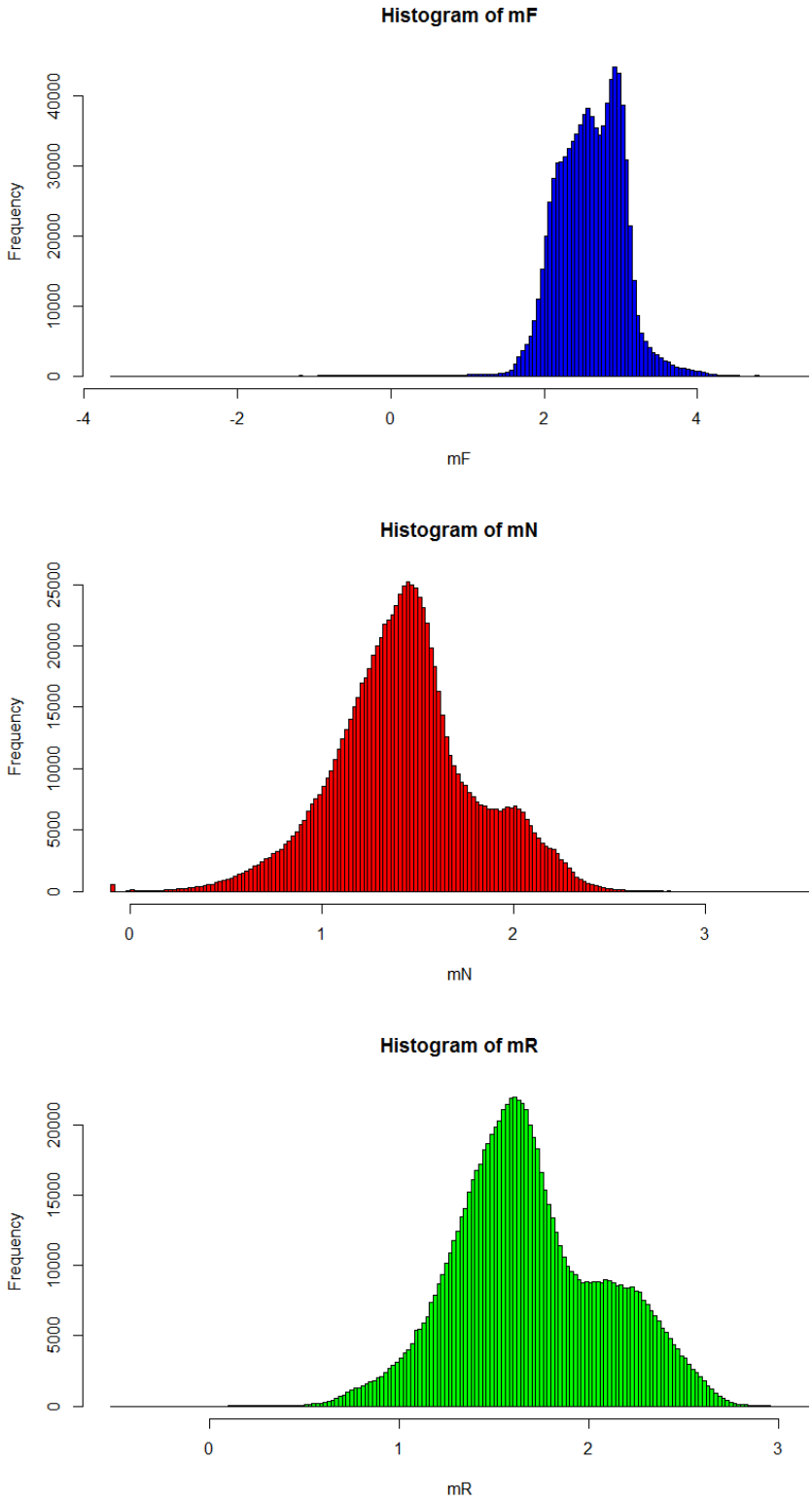


Figure 34. The amount of pixels in the study area is given in the histograms. The frequency also indicates the amount of pixels.

IX. Areas with Predicted IVs Outside Theoretical IV Range

Figure 35. Predicted values lower and higher than the theoretical IV range. The zoom-in areas give special awareness. For mF, the sand paths are given very low values. Especially the edges around water are given small mN values. For mR there are more continous fields that are given low values, whereas the zoom-in area indicates grassed heather.