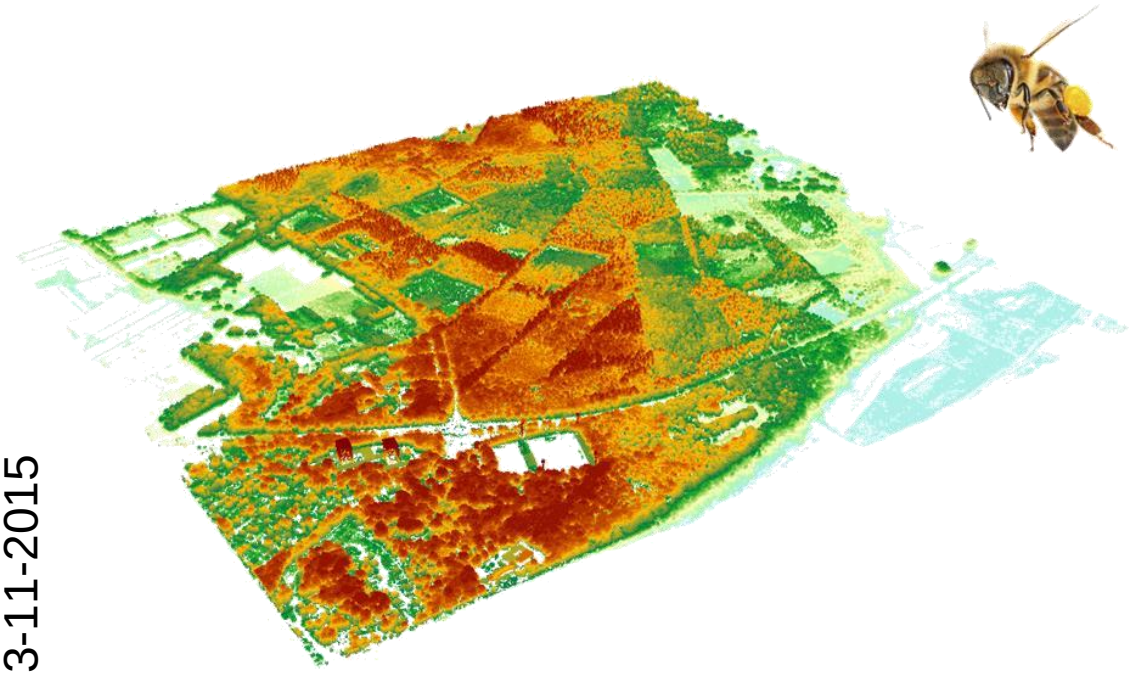
Centre for Geo-Information

Thesis Report GIRS-2015-41

Using LiDAR Point Clouds for predicting Wild Bee Richness at Sub-National Scale

A new methodology for landscape quantification with XYZ points only

Bastiaen Boekelo



23-11-2015

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Using LiDAR Point Clouds for predicting Wild Bee Richness at Sub-National Scale

A new methodology for landscape quantification with XYZ points only

Bastiaen Boekelo

Registration number 90 12 12 078 130

Supervisors

Dr. H.M. (Harm) Bartholomeus
Dr. J. (Jesús) Aguirre-Gutiérrez
Prof. dr. J.C. (Koos) Biesmeijer

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

November 23rd, 2015
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2015-41
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Revision History

Date	Version	Description
November 23rd , 2015	1.0	Graded thesis report
December 15th , 2015	1.1	Minor textual and lay-out improvements

Foreword

For my master GIS and Remote Sensing I was seeking a subject that would cover my interests for ecology. There were open options to work on deforestation topics, but I rather wanted to do something with animals, if possible insects. It was not easy to find something, since the combination of insects and GIS and/or remote sensing appeared to be a small academic niche. Nevertheless, the quest for a suitable topic found its end in Naturalis. It was suggested here that I could investigate the relationship between cuckoo bees and their hosts by means of geo-tagged wild bee records. Though this appealed to me, I liked to go broader, focusing for example on the diversity of wild bees. Possible topics about wild bees and LiDAR had been mentioned before, but studies with this combination had barely been done for insects and only at a small scale. However, after many thoughts and conversations with bee experts I took on the challenge to proof that LiDAR data can help to determine wild bee richness, which resulted in this thesis report.

This thesis combines knowledge from ecology, remote sensing and GIS and I highly enjoyed to work on this integration. It is my hope that these scientific disciplines will find each other more and more in the future.

Abstract

Throughout human history, people have been dependent on nature for the pollination of their cultivated crops. Today, the global economic value of this ecosystem service is estimated at 153 billion euro. Wild bees are major contributors to the world's crop pollination. Not only do they act as a buffer for possible declining honey bee populations, they increase the pollination quality as well, resulting in higher fruit or crop yields. Knowing what environmental conditions are important drivers for species richness is vital for conservation biologists and decision makers. Species Distribution Models (SDMs) can provide this information by correlating species' presence with their associated environmental conditions. From the predictions of multiple models species richness maps can be created. The performance and the output of the SDMs depend on the (quality of the) variables describing these environmental conditions. It is common to quantify environmental conditions with (remote-sensing based) land cover variables. In this study, vegetation structure has been quantified from the AHN2 point cloud with a voxel-based classification method for the Southern half of the Netherlands. The predictive performance of SDMs from land variables is compared with SDMs based on vegetation structure variables, using observations of 60 different wild bee species. Area Under the ROC Curve (AUC) evaluation values provide indications that vegetation structure based landscape variables are explaining single wild bee distribution better than land use based variables. Furthermore, wild bee richness is predicted more precisely by landscape variables derived from vegetation structure than from land use variables. In general, the province of Zeeland and the 'Green Heart' area of The Netherlands are predicted to be species poor, while the Veluwe, Utrechtse Heuvelrug and the East of The Netherlands are predicted to encompass more wild bee species. Results indicate that the certainty of the prediction is related to the spatial distribution of wild bee observation records. Simple methodological implementations, like the use of different SDM algorithms and inclusion of topographical, climatic or other variables might improve SDM performance considerably. Nevertheless, without these adjustments it is shown that point cloud data acquired by airborne LiDAR can contribute significantly to the predictive power of the SDMs as well. Further research is needed to refine and validate the vegetation structure classification and to assess the applicability of this vegetation structure for other (invertebrate) species.

Keywords:
Wild bee richness, LiDAR, SDM, ENM, Voxel, AHN2, Vegetation Structure

INDEX

1 Introduction

This research is integrating methods and principles derived from GIS, ecology and remote sensing. The basic theoretical concepts will be introduced in three paragraphs.

1.1 Wild bees & Landscape Heterogeneity

Wild bees are contributing to the pollination of most plant species worldwide (Winfree 2010; Winfree et al. 2007). Not only do they pollinate wild plants, they also play an important role in the pollination of agricultural crops (Kleijn et al. 2015; Bretagnolle and Gaba, 2015; Park et al. 2015). Together with butterflies, hoverflies and other pollinating species, wild bees provide an ecosystem service that has an approximate economic value of 153 billion euro (Gallai et al. 2009) worldwide and around 22 billion euro at a European scale (Potts et al. 2011). Despite their economic potential, relative little attention is going out to wild insect communities compared to honey bee species (*Apis mellifera* or sometimes *Apis cerana*).

Several studies emphasize that an unilateral focus on honey bees could be controversial (e.g. Winfree 2010; Kremen et al. 2002), also given the unstable population dynamics of the species in recent years (Park et al. 2015; van Engelsdorp et al. 2009; Kremen et al. 2002). Natural occurring wild bee communities could act as a pollination buffer at situations where sufficient artificial pollination seems to become unattainable. Therefore, it is stressed that the colonization of, especially agricultural areas, should be catalysed by creating habitats that are believed to be suitable for wild pollinators (Kremen et al. 2002). A proven way to do this is by creating patches of native flowers close to agricultural fields (Carvalheiro et al. 2012). The main preconditions for a good habitat of a bee species are the availability of sufficient food resources and suitable nesting locations (Gilgert and Vaughan 2011; Westrich 1996). Every bee species has its own preferences for both criterions. Some wild bees make use of various food resources and are tolerant for various nesting locations (e.g. *Bombus terrestris*). Such species are called generalists and are the counterpart of other wild bee species that are dependent on the presence of e.g. a certain soil type and / or plant species. These species are considered specialists and often only occur at certain, geographical areas that meet these requirements. For example, *Colletes herderae* is for its food completely dependent on the presence of the flowering of many ivy (*Hedera* sp.) plants, while it prefers to nest in loess or sandy soils (Peeters et al. 2012). Another example is *Andrena florea* that is completely dependent on the presence of *Bryonia dioica* for its food delivery. A location that fits such requirements is called the *ecological niche* of the species. Hirzel and Le Lay (2008) describe the concept of ecological niche theory as the "function that links the fitness of individuals to their environment". It can be assumed that a diverse landscape increases the chance of more locations with a suitable habitat (or ecological niche that provides high species fitness) than a homogeneous landscape. It is therefore believed that a heterogeneous landscape is positively correlated with wild bee diversity (Hopfenmuller et al. 2014). One could wonder what relevant habitat heterogeneity for wild bees means. Gilgert and Vaughan (2011) mention that "a diversity and abundance of plants that produce nectar and pollen used by insects, combined with a variety of standing or downed dead wood, bare ground, and overgrown vegetation, are the hallmarks of rich heterogeneous pollinator habitat". For bumblebee queens, it has been

suggested that they are depending on complex vegetation structures for their nesting sites (Lye et al. 2009).

It has been mentioned that wild bees could act as a 'pollination buffer' when the size of honey bee populations is low. This might suggest that the contribution of wild bee species to current crop pollination is currently limited, which is unjustified. Research has been done to the relative influence of wild bees compared to honey bees for crop pollination worldwide (Garibaldi et al. 2013). They found in crop systems which are pollinated by both wild insects and honey bees that honey bees account for only 40% till 62% (95%-Confidence Interval) of the crop flower visitation. Pollination quality (like cross-pollination) appears to increase as well when flowers are pollinated by wild insects compared to pollination by wild bees. Furthermore, fruit set increases significantly at flowers visited by wild bees, even if those flowers were regularly visited by honey bees as well (Garibaldi et al. 2013). Recent findings (de Groot et al. 2015) support that both the quality and the quantity of blueberries and apples are positively correlated with the pollination by insects. That wild bees are important contributors to pollination services is emphasized by Kleijn et al. (2015). However, they also emphasize that many wild bee species do not play an important role in crop pollination. Moral arguments should therefore play a pivotal role as well for the conservation of biodiversity.

1.2 Species Distribution Modelling

Mapping species distribution at a national, subnational or regional scale is a commonly applied practice which enables researchers or decision makers to estimate where certain species are believed to occur. To overcome the problem of incomplete species record data, one often uses a widespread technique called species distribution modelling, which is also known as Ecological Niche Modelling (ENM). Elith and Leathwick (2009) define a species distribution model (SDM) as "a numerical tool that combines observations of species occurrence or abundance with environmental estimates". These environmental estimates usually cover the whole area of interest, enabling the possibility to make predictions about the chance of occurrence for the species from the determined relationship. This chance of occurrence is also often called "habitat suitability", which terminology might be more appropriate. After all, geographical barriers (possibly caused by habitat fragmentation) might prevent a species from living somewhere, even though the habitat fulfils the species' requirements. A common evaluation metric is the Area Under the ROC (Receiver Operational Characteristic) Curve (AUC), which provides insight in the SDM performance. Section 2.6.2 will elaborate upon this evaluation metric.

SDMs are widely used to map all kinds of species, like invasive plant species (Chunyuan Diao 2014), bats (Lundy et al. 2012), frogs (Puschendorf et al. 2013) or dragonflies (Jaeschke et al. 2013). Most of the SDMs focus on (larger) vertebrate species. Often remote sensing based land use variables are used as environmental descriptors. For these models, the landscape is often described with variables based on (derived) information inside a particular cell only, like percentage coverage of a certain land use class (Ficetola et al. 2014). Less, though not very few, studies have been dedicated to the mapping invertebrates. For these animals, variables that summarize spatial arrangements of intrinsic landscape elements could be very suitable for invertebrate species (Kumar et al. 2009). Many studies

have been dedicated to the use of these kinds of landscape indices. Relevant variables can be 'distance to certain landscape elements' (e.g. Zulka et al. 2014; Wagner and Fortin 2012), patch size (e.g. Baeza and Estades 2010; Fagan et al. 2009) or edges in the landscape (e.g. Marshall et al. 2006; Aguirre-Gutierrez et al. 2015).

A few studies have been dedicated to the mapping of *Apidae sp.* using SDMs. (e.g. Giannini et al. 2013; Polce et al. 2013). The last study has assessed, by means of SDMs, the potential service provision for field beans of wild and managed pollinators in Great Britain. This was believed to be one of the first studies with SDMs focused on pollinators on such large scale. Here they have used topographical, land cover, climate and pesticides data. For bumblebees, the mean patch area has shown to be an important explanatory variable, while for butterflies this seems to be the edge density (Aguirre-Gutierrez et al. 2015). For wild bees in general, it has been found that woody edges (Kleijn et al. 2004) are positively correlated with wild bee diversity.

The mapping of species richness or diversity is a common practice (Ferrier and Guisan 2006). In a SDM context, there are mainly two methods to express how rich a location is. The first one adds all habitat suitability predictions of the species, while the second method is the summation of the binary species prediction transformations (Dubuis et al. 2011). In the end every location has a value that should indicate how well it facilitates species richness.

1.3 LiDAR & SDM

LiDAR (Light Detection And Ranging) is a laser technique that uses the reflection of light in order to detect the location physical features. There are three main acquisition techniques. For the first one the environment is scanned by a LiDAR device that is attached to e.g. drones or airplanes, called *Airborne LiDAR Scanning (ALS*). During the flight data is collected. The second technique is similar, but for *terrestrial LiDAR Scanning (TLS)* the device is situated on the ground. The third technique is spaceborne LiDAR. Airborne laser scanners are known to have lower spatial resolution than terrestrial LiDAR (but higher than spaceborne), but it can cover a bigger area (Jaboyedoff et al. 2012). The LiDAR device can be a waveform recording device or a discrete-return device. The difference is shown in figure 1 (from Lefsky et al. 2002) that shows a LiDAR beam that assumes an airborne LiDAR system.

One of the research applications of LiDAR technology is in the quantification of vegetation parameters. Examples of these are tree/shrub density, foliage height, mean or maximum vegetation height, variation of plant height, number of vegetation contacts or coarse wood debris (Simonson et al. 2014). Evaluations of the ability of LiDAR to describe these metrics were often positive. These findings underscore why it is believed that this ability of LiDAR to quantify the 3D structure of the natural environment can be used as a tool to map habitat structure (Simonson et al. 2014; Vierling et al. 2008). An interesting vegetation metric is *vegetation structure*. This vegetation parameter is mentioned in many studies, but it appears to have various meanings. Simonson et al. (2014) make a distinction between horizontal and vertical vegetation structure. Examples of vertical vegetation structure metrics are tree/shrub cover, mean/maximum vegetation height, coarse or fine woody debris or variation in plant height. Horizontal vegetation structure can be diversity of land covers,

percentage vegetation cover, patch size/density and edge length / density. An example of a quantification of vegetation structure can be found in a study of Schut et al. (2014). Here they quantified vegetation structure with a voxel based method. A voxel is 3D pixel or volumetric pixel. Airborne derived point cloud data has been used for the creation of point densities in voxels. These point densities have been used as an input for an unsupervised classification of vegetation structure.

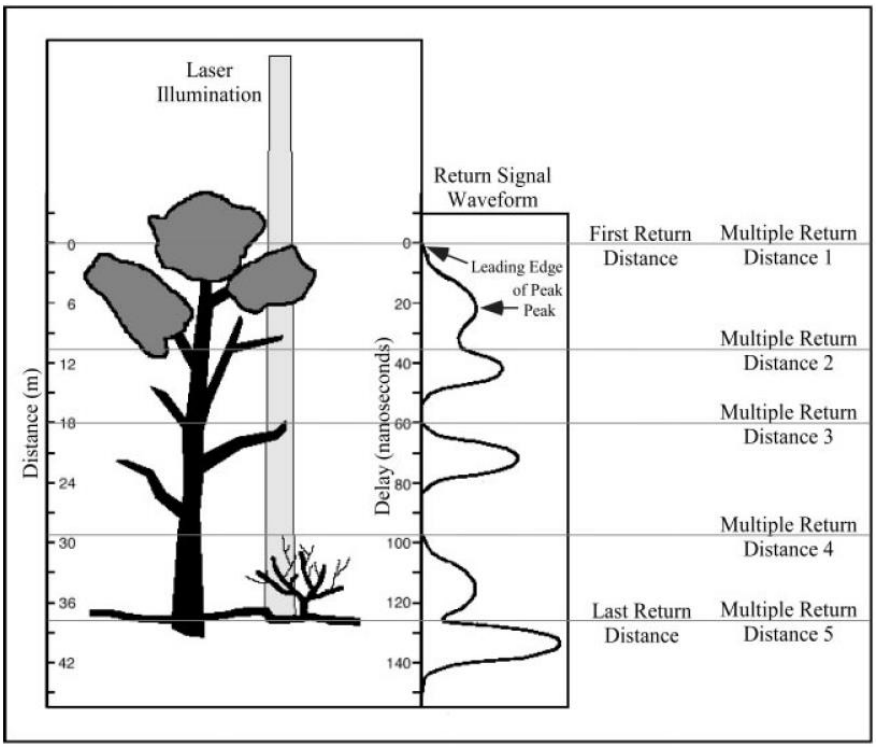


Figure 1: Difference between discrete-return LiDAR vs waveform recording LiDAR. In the left box hypothetical vegetation is drawn. The box connected at the right of it shows the data stored in waveform recording LiDAR system, while the horizontal lines show indicate the returns of a discrete-return LiDAR system. The latter can return multiple (mostly up to five) points per laser beam. Figure copied from Lefsky et al. (2002)

Several studies already indicate that LiDAR derived metrics can be used in SDMs. For example, it has been shown that LiDAR based vegetation can be correlated to northern spotted owls (Ackers et al. 2015) with good model performance (AUC ≈ 0.8). Also, the golden-cheeked warbler and the black-capped vireo have been modelled before (AUC = 0.864 and 0.746 respectively) using LiDAR derived vegetation metrics (Farrell et al. 2013). The results are emphasized by a third study that uses LiDAR to predict the distribution of nine bird species (Ficetola et al. 2014). LiDAR based SDMs showed for all species an AUC higher than 0.72. They also show that SDMs with only LiDAR variables are in general explaining the diversity of birds better than SDMs based on land use variables or combined SDMs. All three studies have in common that they are matching *local vegetation characteristic* (in a pixel of e.g. 30m) to the presence of *birds*. There area (a limited number of) other local studies which focus on the relationship between point clouds and invertebrates. Small-scale studies have been performed that suggest that LiDAR derived metrics can predict distribution of beetles (Müller and Brandl 2009) or spiders (Vierling et al. 2011). These studies are exceptional in the sense that they correlate LiDAR to the occurrence of invertebrate species. However, for both studies this was the direct result of a clear experimental set-up. Upscaling this to a (sub-)national area would be unfeasible.

1.4 Research Motive

The following list summarizes what has been done before and what is lacking in current research papers:

- LiDAR data have been used in SDMs, but those are aimed at birds (e.g. Ackers et al. 2015; Farrell et al. 2013; Ficetola et al. 2014), not invertebrate species and only at a regional scale.
- LiDAR data have been correlated with invertebrates (Müller and Brandl 2009; Vierling et al. 2011), but not in a SDM context and only at a regional scale.
- Wild bee SDMs have been applied at national levels (Aguirre-Gutierrez et al. 2013; Polce et al. 2013), but never in combination with LiDAR.

In this study it will be investigated if vegetation structure information derived from the AHN2 point cloud data can improve SDMs for wild bees in The Netherlands. With that, this research is an effort to fill the research gaps mentioned above. The EIS (European Invertebrate Survey) has maintained a dataset of wild bees in The Netherlands. Next to this, an airborne LiDAR derived point cloud dataset (XYZ only, spatial resolution approximately 11 points/m^2) is available for area of The Netherlands. In this study it is hypothesized that (the addition of) LiDAR derived landscape variables can explain wild bee richness better than land use only landscape variables. These two datasets will be used in order to investigate this.

If the results are promising, the wild bee richness map might be useful for decision makers or conservation biologists. It will also mean that it could be possible that LiDAR derived variables could help to depict which landscape elements are enhancing wild bee richness.

One might wonder why the AHN2 point cloud would be more valuable than LGN ("Landgebruik Nederland") data. As (Ficetola et al. 2014) already suggested, it could be that LiDAR can explain species occurrence better than land use. Inspecting the LGN6 also supports this (figure 2). The aerial photograph shows that field edges are often characterized by higher vegetation. This higher vegetation is barely visible in the LGN6 dataset (dark green pixels). Next to this, the land use classes do not provide information about the vertical structure of the vegetation. It is hypothesized that LiDAR derived vegetation structure can significantly improve this.

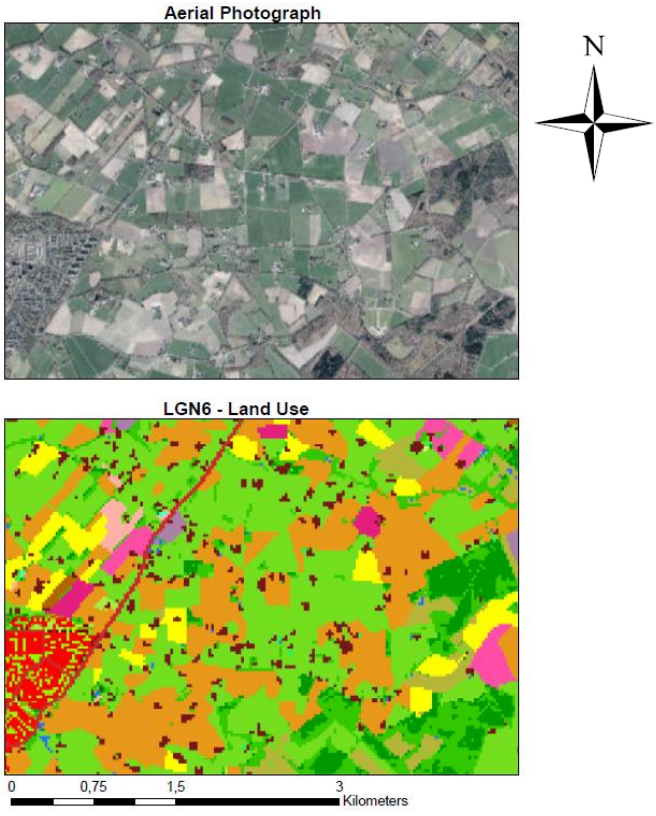


Figure 2: Comparison of aerial photograph (upper) and a land use map based on LGN6 (under) of the area east of Doetinchem. Darker green pixels are classified as 'forest'.

1.5 Research Aims

With this study I have tried to attain the general objective by answering the research questions described below.

General objective

The main objective of this study is to investigate if *airborne discrete-return LiDAR technology* can be used to predict *wild bee richness* at a *large scale* with the aid of *SDM techniques* and to compare the results with, more classical, land cover approach.

Research questions

Question 1
How accurate do the following species distribution models predict, on average, wild bee distribution: 1) LiDAR-only, 2) Land Use-only or 3) a combination of both LiDAR and Land Use data?

Question 2
According to the models, which Dutch areas are (not) facilitating wild bee richness?

Question 3
Which explanatory variables are important for wild bee distribution according to the constructed models?

2 Materials and Methods

2.1 Data & Study Area

Study Area
This study will focus on the Southern part of The Netherlands. In this part of The Netherlands more bee observations have been recorded and the landscapes are differing, ensuring varying environmental conditions. The reason to leave out the Northern part of The Netherlands was to reduce the total processing time of the enormous amount of data. The upper bound (maximum Y coordinate) of the area is 475000m ('Rijksdriehoeksstelsel'). It is assumed that this part of The Netherlands will cover enough variation in order to construct reliable SDMs.

AHN2 point cloud
The AHN2 point cloud is a free dataset that can be downloaded from the national spatial data portal www.PDOK.nl in .laz format (compressed .las). Because of the size, the point cloud is split in tiles of 5000*6250m (width*height). Per tile two datasets are available: one containing the points indicating the surface (NL: 'gefilterde puntenwolk') and one containing all points above (NL: 'uitgefilterde puntenwolk'). Though the points are separated in two different files, the identity of the points is not classified within the laz files.

During the point cloud acquisition, all point records contained regular LiDAR information regarding RGB, intensity, number of return etc.. Nevertheless, only XYZ coordinates are publically available without any other attribute information. Reading the data with various software programmes gives the impression that only first returns are stored in the file. However, this is the program's interpretation of data that lacks the attribute information 'number of return'. This means that essentially all points are available for download.

The initial split in ground and non-ground points is not straightforward and performed by different suppliers. Niels van der Zon, project leader of the AHN2, describes the following regarding this matter:

"The filtering (or classification) is a complex process, where both automated classification algorithms are used, as well as many manual corrections. For the AHN2 there were 4 or 5 suppliers. The method to split the data is different for every supplier. Though the end product should fulfil the criteria of the AHN2, suppliers can design their own route to that. Possibly, a supplier has improved the process during the time of AHN2 acquisition, which means that data could have been treated at e.g. 10 different ways. It is important to emphasize that the classification is not purely a product based on algorithms and laser data. Many additional sources like aerial images, BAG, topographical maps and panorama pictures are used in this process".

The systematic height error is 5 cm and the standard deviation of the height is 5 cm as well. On average, the point resolution is about 11 points/m^2, but it varies and can reach up to more than 20 points/m^2 or only a few points/m^2.

The main goal of the acquisition is to construct a DEM or DSM of The Netherlands. For this goal, vegetation is mostly considered noise. To reduce the effect of vegetation the acquisition has mainly been performed in winter time, ideally between December 1st and March 31st.

The resolution of points located in MIVD ('Militaire Inlichtingen- en Veiligheidsdienst') areas is reduced and height information of non-ground objects is removed in these areas.

For more information one should read the (Dutch) quality assurance document of the AHN2 (Van der Zon 2013).

Bee dataset

The dataset indicating bee presence locations origins from the EIS. Point locations are stored in an excel file, together with the spatial resolution, year of discovery, the record's data source and the species names. Spatial resolution is varying, mostly because the data records have different origins, which are listed below:

- Literature
- Collection
- Field observations from www.waarneming.nl
- Field observations submitted by observer and directly submitted to EIS
- Field observation derived from a city name list

Only validated species records from www.waarneming.nl are stored in the dataset. If a photo of a species has unambiguously proved the identity of the species, the observation was included.

Some records are the result of complete field inventories by professionals, while others are coming from single amateur observations. This is an important aspect of the dataset and should be taken into account in order to interpret the data well.

Records between the beginning of 2003 and the end of 2014 are used. Data of presence records in the North of The Netherlands, outside the study area, have been excluded. Around 50% of the data has a spatial resolution of 1 km^2. The remaining coordinates are rounded down to a spatial resolution of 1 km^2. This way, all records have similar character and double records can easily be detected. A coordinate (e.g. 75000:376000) refers to the *South-West corner* of a km^2. After removal of 'double records' of the same species on the same location but at a different time (this study will not focus on temporal aspects), the species with more than 100 records have been selected for the study. In total there are *60 species* belonging to 17 different genera. In table 1 all species that are selected are listed.

Table 1: Wild bee species used for the study. 'Counts' refers to the number of unique locations in which the species is found.

Species	Counts	Species	Counts
Andrena barbilabris	188	*Andrena chrysosceles*	181
Andrena bicolor	162	*Andrena cineraria*	126
Andrena carantonica	208	*Andrena clarkella*	108
		Andrena dorsata	213

Andrena flavipes	491
Andrena fulva	233
Andrena haemorrhoa	466
Andrena minutula	144
Andrena nitida	181
Andrena subopaca	181
Andrena vaga	247
Andrena ventralis	116
Anthidium manicatum	111
Anthophora plumipes	127
Apis mellifera	429
Bombus campestris	133
Bombus hortorum	184
Bombus hypnorum	270
Bombus lapidarius	561
Bombus lucorum	233
Bombus pascuorum	914
Bombus pratorum	474
Bombus terrestris	594
Colletes cunicularius	140
Colletes daviesanus	133
Colletes fodiens	105
Dasypoda hirtipes	220
Halictus rubicundus	129
Halictus tumulorum	237
Heriades truncorum	133
Hylaeus communis	223

Hylaeus confusus	134
Hylaeus hyalinatus	118
Lasioglossum calceatum	388
Lasioglossum leucozonium	237
Lasioglossum morio	231
Lasioglossum pauxillum	123
Lasioglossum exstrigatum	198
Lasioglossum villosulum	129
Macropis europaea	143
Megachile centuncularis	124
Megachile willughbiella	151
Nomada alboguttata	141
Nomada fabriciana	145
Nomada flava	228
Nomada flavoguttata	167
Nomada fucata	216
Nomada goodeniana	109
Nomada lathburiana	153
Nomada marshamella	112
Nomada ruficornis	212
Nomada succincta	113
Osmia rufa	254
Panurgus calcaratus	107
Sphecodes albilabris	131
Sphecodes monilicornis	169
Sphecodes pellucidus	116

In figure 3 the spatial distribution of the observation is visualized.

BAG - Buildings

The dataset “Basisadmistratie Adressen en Gebouwen” (BAG) is a vector dataset with all buildings of The Netherlands. It contains several features but only the buildings are used for this study. The version released in March 2015 was used.

LGN6

The LGN6 (Land Gebruik Nederland 6) is a raster dataset, constructed in 2007 and 2008 and it contains 39 land use classes. The spatial resolution is 25m. These classes have been reclassified by Jesús Aguirre Gutíerrez into classes which are more relevant for bees with the R tool ClassStat [SDMTools]. The new classes for this dataset are ‘Grassland’, ‘Cultivated / Bare ground’, ‘Moors / Peat’, ‘Forest Mixed’, ‘Forest Deciduous’, ‘Forest Coniferous’, ‘Buildup / Roads’, ‘Water’, ‘Swamps’ and ‘Sandy Soils’.

Overlay datasets

For the vegetation structure raster that has been created several overlays have been used to correct for non-vegetation features or locations where LiDAR point data has been manually manipulated. This has been done for water areas, highways and military areas. The water and the highway dataset are vector datasets from the TOP10NL, which is maintained by the Dutch cadastre. The vector

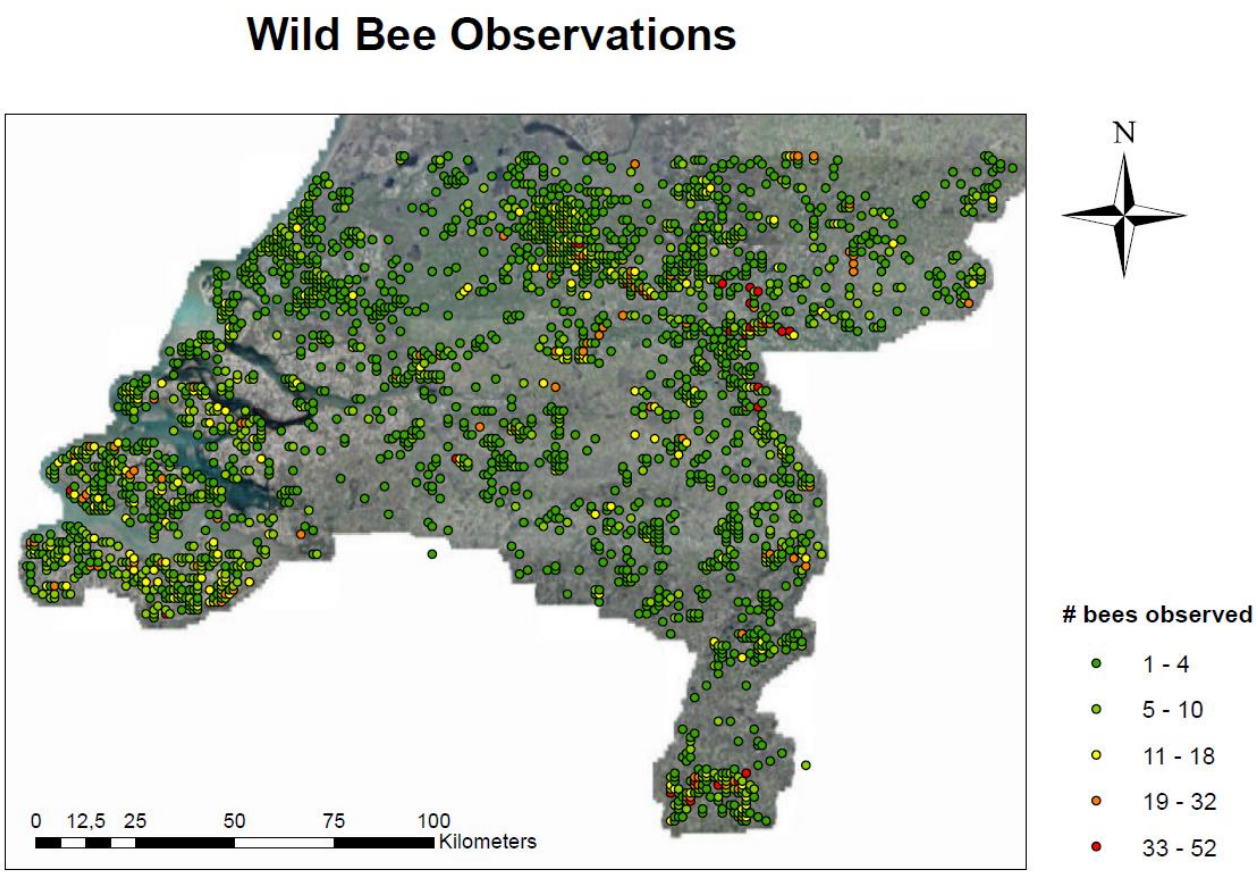


Figure 3: Sample locations of the bee data. Points are coloured after the number of bee species that have been observed from 2003 till 2014.

dataset indicating the location of military areas has been provided by the MIVD ('Militaire Inlichtingen en Veiligheidsdienst).

Soil type

A soil type dataset with a resolution of 1km is used as well. It has 10 classes: Peat soils, Marine Clay soils, Riverine Clay soils, Dune and Marine Sandy soils, Sandy soils, Other Clay and Loam soils, Abroad, Anthropogenic soils and Water. The classes have been derived from tables BODEMGT and CEL in the LKN database for the period 1985-1995.

Food resources availability

A dataset describing the availability of food resources relevant for bees has been constructed by Jolien Morren, in cooperation with Naturalis Biodiversity Centre. It is a result of a combination of CBS data and the BRP (Basis Registratie Percelen). Spatial resolution of this dataset is 1 km and it covers the years 2005-2014. This dataset describes food resource availability by human land use. Next to this dataset, she also constructed a dataset which contains information about possible food resources from wild plants. This information has been derived from the FLORON dataset. A selection has

been made of 100 plants that have been proven to be relevant for honey bees. It was assumed that these flowers are relevant for wild bees as well.

2.2 Analysis overview

The goal of the research is to get insight in the potential of LiDAR technology for large-scale SDMs. Several analyses steps have been undertaken to come to a conclusion. It is a major task to convert point cloud data to relevant landscape variables. This chapter will explain what has been done and what choices have been made. The methodology can be split up into four parts, which are visualized in figure 4.[1]

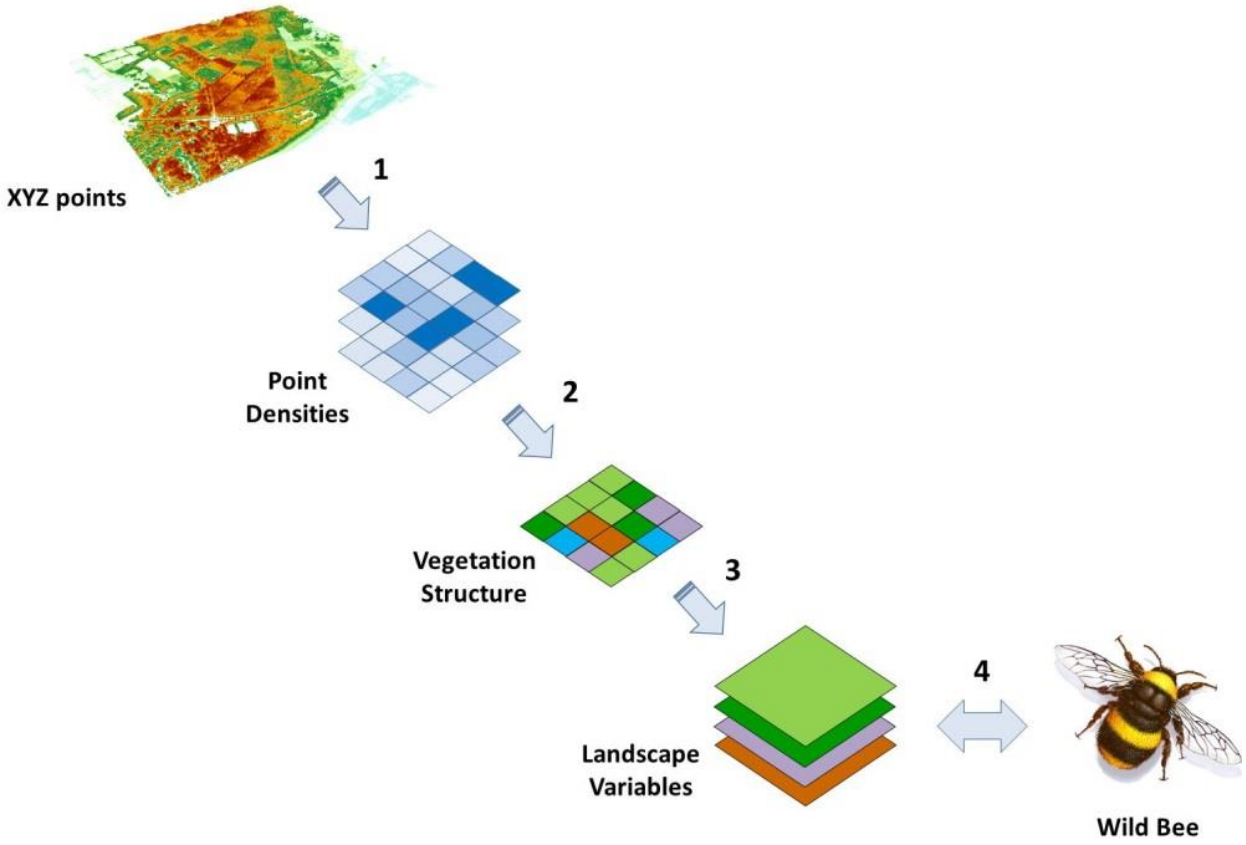
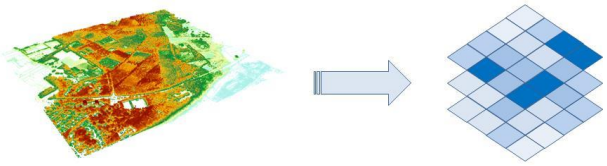


Figure 4: Schematic overview of the methodology in 4 parts. 1- Conversion from point cloud to point density rasters; 2- Classification of point densities to vegetation structure; 3- From vegetation structure to landscape variables; 4- Species Distribution Modelling.

The first processing step has been the transformation of the point cloud into point density rasters. These rasters have subsequently been used for classification to obtain a vegetation structure raster. The third step is to derive landscape variables from this data and the fourth step focuses on the modelling of the species' distribution. All steps are elaborated upon in the following paragraphs in this chapter. The next paragraph will introduce the input data and the area of interest first.

[1] Bumblebee picture copied from http://www.bioquicknews.com/node/2599

2.3 Step 1: AHN2 point cloud to point density



Theory behind methodology

In order to derive relevant information from the massive LiDAR point cloud, data need to be converted. It is chosen to convert the point cloud data into raster pixels with a 25m spatial resolution. This spatial resolution has been chosen since it has the same resolution as the land use dataset and it will contain enough points needed for further calculations later in the process (assuming 8 points/m^2, there will be around 25*25*8 = 5000 points inside this area). The volume in which these points are situated will be referred to as a *voxel* (volumetric pixel). Every point in a voxel is reflecting the presence of an object (e.g. a dog, trees or a traffic light) situated in the XYZ space. This means that, since Z values are available, these points can be indicative of the height of those objects. If a voxel contains many points in the first meter above ground, but no points situated higher, it can be deduced that only objects smaller than 1 meter high are present. In contrast, if the voxel contains only some points in the first meter above ground but relatively much more between 9 and 10 meters high, it can be assumed that big objects are present here, like high trees. This simple reasoning forms the starting point of the methodology. If the number of points inside a certain height layer (e.g. between 2 and 3 meter) of a voxel can be converted into one value, it will summarize to which extent objects are present in this 3D layer. Every voxel will be split into several height layers. This way, every pixel will contain as much values as the number of height layers chosen. The height layers will be used for the development of the vegetation structure raster.

Implementation

It is assumed that low height layers are the most relevant for bees. The length of the intervals between height break points will therefore increase when going higher. Table 2 shows which height layers are chosen initially.

Table 2: Initial proposition for height breakpoints of the voxels

Layer	Height Breakpoints (m)
1	0.05 – 0.20
2	0.20 – 0.50
3	0.50 – 1.00
4	1.00 – 2.00
5	2.00 – 5.00
6	5.00 – 10.0
7	10.0 – 20.0
8	20.0 – 80.0

Ideally, the values obtained are only representing vegetation. However, points can be reflected by many other objects, especially in urban areas. It is not or barely possible to remove all non-vegetation reflection sources, but for buildings it is. Points reflected on houses can be clipped out using the BAG. Other, smaller objects will be included in the analysis. Another complicating matter for the

analyses is the Z-value of the points, which are stored in height above NAP ('Normaal Amsterdams Peil'), while heights above ground would be needed. Lastly, the size of the data and the maximum PC memory space should be taken into account in the methodology. The methodology scheme is visualized in figure 5.

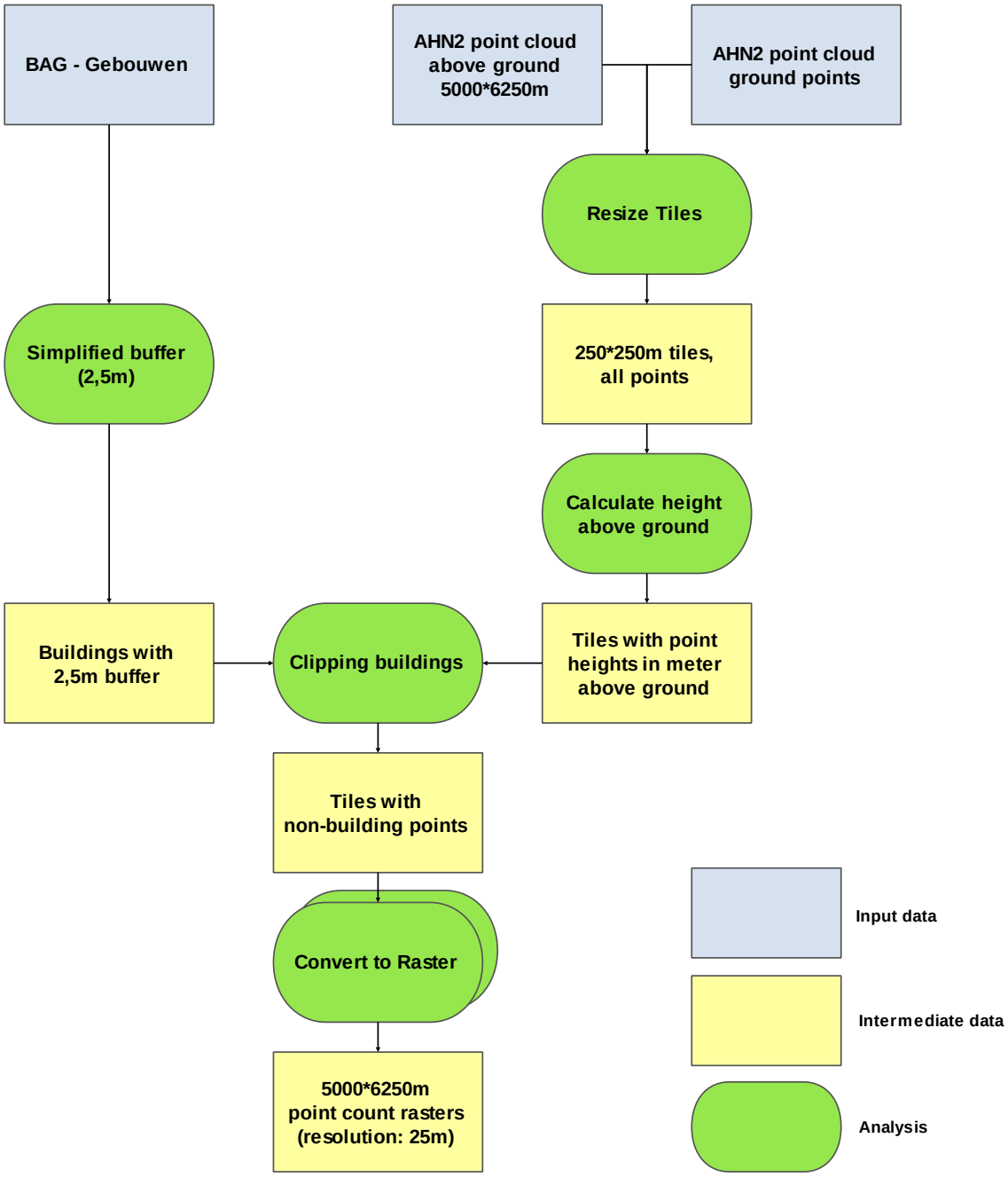


Figure 5: Methodology scheme of the LiDAR processing.

All steps handling point cloud data are performed with the, licensed, LiDAR software '*Rapidlasso - LAStools*',[2] developed by Martin Isenburg. Firstly, all *.laz files, both ground points and non-ground points data are downloaded. These are downloaded and unzipped with an R script. For all ground point datasets a polygon bigger than The Netherlands was overlaid and every point inside this polygon would be classified as a ground point. For PC memory reasons, every 5000*6250m tile (both ground and non-ground) were retiled into 250*250m tiles before proceeding to the next step. Every tile created that contained less than 2 Kb of data was deleted in order to prevent lastools from crashing later on. These were empty files that were only touching the edge of the original 5000*6250m tile. Then the height of the points was transformed from 'height above NAP' to 'height above ground' for every tile. A triangular network (TIN) was made out of the ground points, and Z-values were changed into height of point above this TIN. The Z-values of the ground points itself were set to zero. For some areas in The Netherlands the height conversion failed, mostly because of 'point gaps' in the data, which are areas without sufficient amount of points to be able to create a TIN. Typically, these were (250*250m) areas covered entirely by water or locations with military activity. Points reflecting on buildings in the latter were removed here earlier and point density was reduced by the organisation of the AHN2 (Van der Zon 2013).When large buildings were removed, height conversion sometimes failed. All failed tiles were not included for further processing.

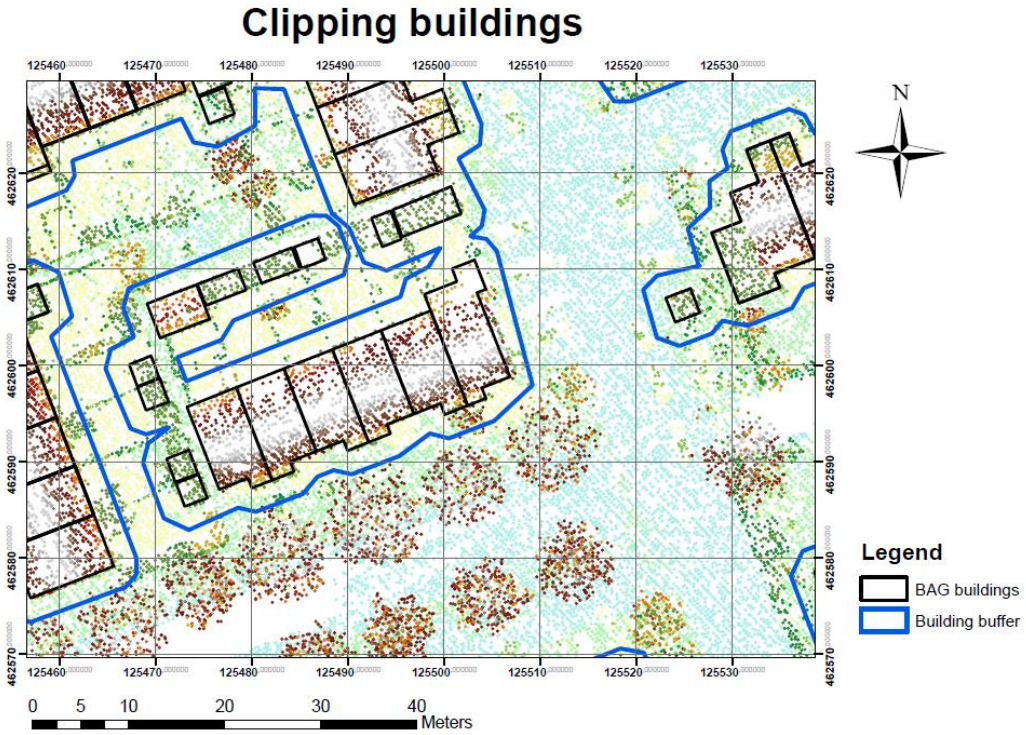


Figure 6: Visualization of the area around building points that have been clipped. The black line delineates the original BAG buildings, blue line the simplified 250cm buffer. Points are coloured after their height. Note that some building points are falling outside de BAG polygons. Buildings are located in Kockengen, The Netherlands. Grid lines are according to 'rijksdriehoeksstelsel'.

Points of buildings were removed using the 'buildings' polygon shapefile of the BAG. Because this is a large dataset, all polygons were first split by municipality. Inspection of the data showed that often some points clearly reflecting on a building were falling a few centimetres outside the BAG polygons.

[2] http://www.cs.unc.edu/~isenburg/lastools/

For high buildings it was seen that, in rare cases, points were dislocated about 150 cm from the BAG polygons. For this research it is important to be certain that points with a high Z-value can be interpreted as a point reflecting on a tree and not a house. Therefore, it has been chosen to use a wide buffer for every polygon with 250 cm. Simplifying the buffered polygons reduced the amount of data with an acceptable reduction of precision (tolerance was set to 50 cm) and speeded up the clipping process significantly.

In figure 6 it is visible which points are deleted. In this image it is visible that the size of the building buffer is very large, thereby also deleting some vegetation. However, it is assumed that the loss of some vegetation around buildings weighs less than the inclusion of building points. The latter might cause a vegetation structure misclassification in a later stage of the analysis, due to the (potentially high) building points.

Lastly, the point cloud was converted to point count rasters, using the proposed height breakpoints. This step also merged the 250*250m tiles into the original AHN2 tile size of 5000*6250m.

Using R the rasters were normalized for 1) volume of the height layers and 2) total number of points.

Normalization by height or volume
The first normalization is performed because in a later phase of the analysis it is important to be able to visualize and identify homogeneous areas with a typical vegetation structure. A disadvantage of using the raw point density output is that the number of points relative to the total amount of points is highly dependent on the vertical length of the height intervals. This means that the height point density in e.g. the 0.20-0.50m layer will almost always be lower than the 10.0-20.0m layer, just because the latter counts the points in a volume which is (20-10)/(0.5-0.2) = 33.3 times greater than the first layer. In a later stage of the analysis homogeneous areas should be identified with a unique vegetation structure type and this difference in 'layer weighting by height' is making the visualization for identification of these areas hard. Furthermore, all layers should have equal weights for a correct and unweighted calculation of the point densities. Therefore, the layers were normalized for their corresponding heights as well by using the next formula for every pixel:

$$PM_x = \frac{Counts_x}{HB_{y+1} - HB_y}$$

Where PM_x is the number of points per 625m^3 (25m*25m*1m), $Counts_x$ the number of points in layer x. HB_y and HB_{y+1} are the lower height breakpoint and the higher height breakpoint respectively.

All rasters together will represent an imaginary voxel of 25*25m and 6m high, equally divided into voxels of 1m high.

Normalization by total number of points
Overlapping flight lines during point cloud acquisition have caused some areas to have more points than others, irrespective of (the character of) the objects present in these areas. To correct for this the *point density* was calculated, using the next formula for every pixel:

$$PD_x = \frac{PM_x}{\sum_{i=1}^{n} PM_i}$$

Where PD_x refers to the point density in layer x relative to the total points found in the entire, newly created, voxel. The units of the PD_x are *point point*$^{-1}$ m^{-1} = % m^{-1}. The denominator refers to the sum of all, normalized, point counts in the voxel and *n* is the number of layers.

A visualization of both corrections is shown in figure 7.

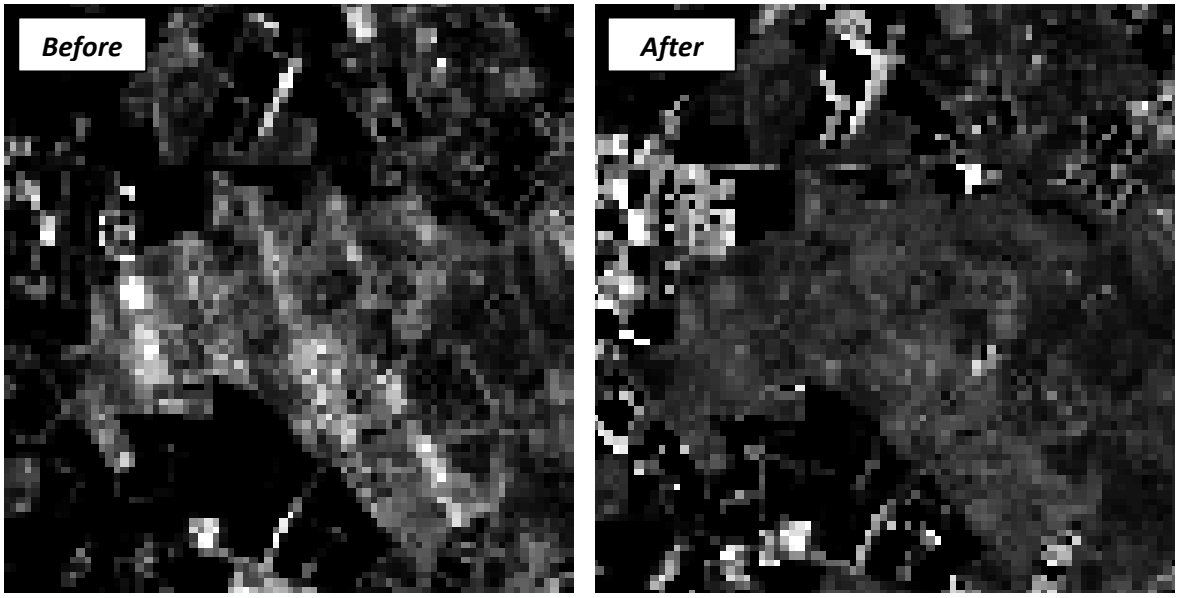


Figure 7: Left figure shows a raster image before the height and density corrections in the 0,1-0,25m layer. The right figures shows the raster after. Note the left area of the picture that changed from relative low (dark) to relative high values (light) and the disappearance of the peculiar stripes caused by overlapping flight lines. Note that the 0.10-0.25 layer is created after the revision of the height breakpoints.

Thresholding

For grasslands or other - typically very flat – areas, the total point counts hardly ever exceeded 100. To prevent that only a few points in the lowest layer would be translated into a value of 100% in the lowest layer ($Counts_1$), thereby suggesting the presence of low vegetation, a threshold has been implemented; if there are less than 100 points between 0.05 and 10 meters the pixel value will be set to NA. If points above 10 meters would be taken into account, the presence of power lines would often cause the total point counts to exceed the threshold of 100 points. In grassland areas that results in the inclusion of pixels which only indicate power line objects. Since these objects are considered non-relevant for this study, only the number of points lower than 10 meters is taken into account for the NA threshold.

Changing the height breakpoints

After inspection of the point density rasters it has been chosen to change the height breakpoints. It appeared that the 1.00-2.00m layer was almost always very similar to the 2.00-5.00m layer. Therefore, it has been chosen to merge these two layers. Because the accuracy of the ground point classification is containing a certain level of noise, the lower height breakpoint is changed to 0.1 m. Because a 0.10-0.20 would be very small, the second breakpoint was set to 0.25. The 0.25-0.50m layer is merged with the 0.50-1.00 m as well, because they also appeared to give very similar point

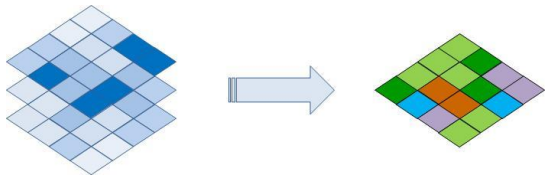
density values. Lastly, to cover more information it has been chosen to include a 20-30 m layer as well. All these mutations resulted in the new, final point density height layers, listed in table 3.

Table 3: Height breakpoints of the voxels used for the classification of the vegetation structure. Note that the last layer (30-80m) has not been used for the classification.

Layer	Height Breakpoints (m)
1	0.10 – 0.25
2	0.25 – 1.00
3	1.00 – 5.00
4	5.00 – 10.0
5	10.0 – 20.0
6	20.0 – 30.0
7	30.0 – 80.0

Summing the first six layer values of a single 25*25m pixel should mostly result in 100 (%). However, for some areas with high vegetation some points will be higher than 30 meter. In these cases, the sum of the first 6 layers will be less than 100%.

2.4 Step 2: Create vegetation structure



Classification

Since the 30-80m layer is not used as an input for the classification point density rasters into vegetation structure, there are six point density layers in total. A supervised classification is performed, by performing a *Maximum Likelihood Estimation*. This method has been chosen because it is a parametric classification algorithm, which makes it fast (instead of *Random Forest*) and not sensitive to outliers (instead of the *k-Nearest Neighbour algorithm*). For the input of this algorithm zones needed to be identified which were used as a raw input for the classification.

The next criteria were used for the identification of the classes:

- The classes should differ in:
  - Most prevalent height layer
  - Maximum height
  - The equality of point density spread amongst the height density layers
- A class should not be too similar to another class

It should be noted that the choice which classes to define and the delineation of the zones are not completely objective. Alternatively, one can do an unsupervised classification. However, one could wonder if the resulting differences between the vegetation structure classes are relevant for bees. While performing supervised classification, the classes meet at least the criteria mentioned above (which are believed to be relevant for bees).

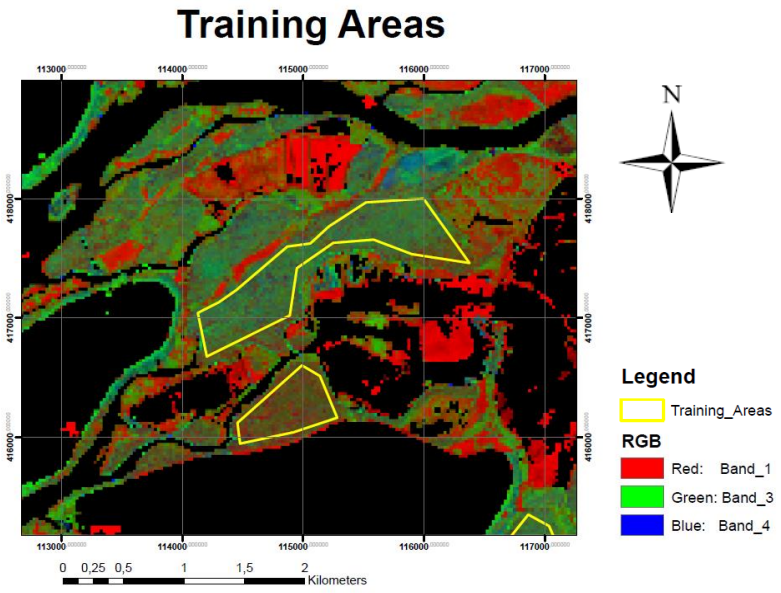


Figure 8: Delineation example of vegetation classes. Bands are the normalized point density values. Band_1 = 10-25cm, Band_3 = 25-100cm, Band_4 = 5-10m. Areas inside the delineation zones are expected to be geospatially homogeneous. Grid lines are according to 'rijksdriehoeksstelsel'

For the delineation of the zones the next criteria were used:

- Pixels inside the polygon should have similar values per layer
- Delineation of pixels should match homogeneous areas visible from aerial images
- Multiple polygons can be used to describe one vegetation structure class

In the end different vegetation classes have been defined. An example of a zonal delineation is shown in figure 8.

Vegetation Profiles

Eventually, nine vegetation profiles were constructed. However, by mistake the second layer has become 0.10-1.00 instead of 0.25-1.00. This means that this layer consists of the layer 0.10-0.25 meter and 0.25-1.00 meter. Therefore, it could occur that the classes exceed the 100%. The classes are shown in figure 9. In table 4 the point density values are listed. Cells are coloured according to it, giving a more intuitive representation of the vegetation structure. Based on the values of the point densities, the classes have been given a name, listed in table 5.

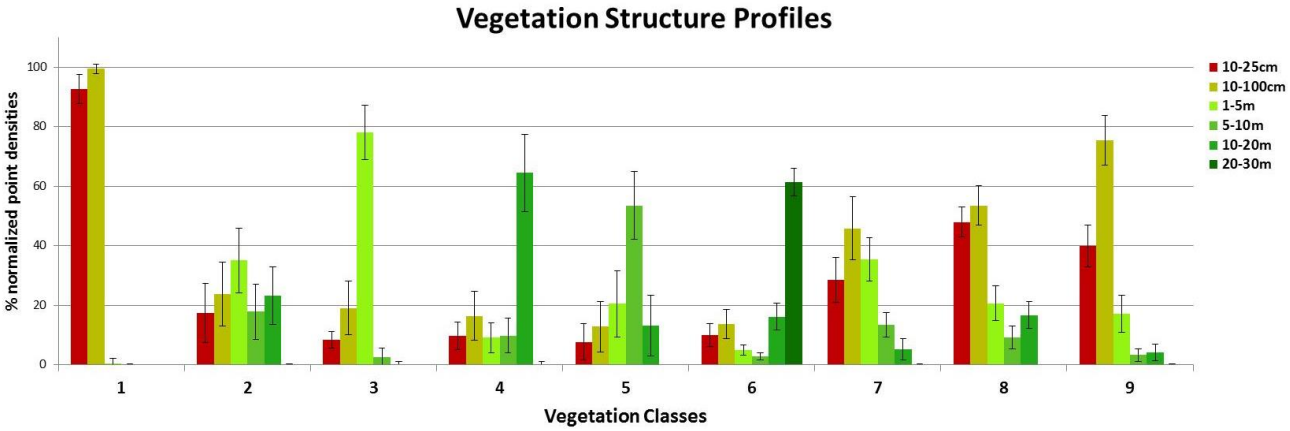


Figure 9: Vegetation profiles of the vegetation classes. Error bars indicate standard deviation of the pixel values in the training area delineation zone.

Table 4: Vertical representation of the point density values for the different vegetation classes. The darker the cell, the higher the point density.

Height Layer (m)	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
20 - 30	*0.0*	*0.1*	*0.0*	*0.3*	*0.0*	*61.3*	*0.1*	*0.0*	*0.1*
10 -20	*0.0*	*23.2*	*0.2*	*64.4*	*13.2*	*16.2*	*5.3*	*16.6*	*4.1*
5.0 - 10	*0.0*	*17.9*	*2.7*	*9.8*	*53.5*	*2.9*	*13.4*	*9.2*	*3.3*
1.0 - 5.0	*0.6*	*35.0*	*78.0*	*9.1*	*20.5*	*5.0*	*35.4*	*20.7*	*17.2*
0.1 - 1.0	*99.4*	*23.8*	*19.1*	*16.4*	*12.8*	*13.6*	*45.9*	*53.5*	*75.3*
0.1 - 0.25	*92.6*	*17.3*	*8.4*	*9.7*	*7.6*	*10.0*	*28.5*	*47.9*	*39.9*

Table 5: Description of the vegetation structure classes. Note that voxels containing less than 100 points between 0,1 and 10 meter have been assigned the class 600.

Vegetation Structures	Description
Class 1	*Low vegetation*
Class 2	*Mixed vegetation, high and low*
Class 3	*Small trees, with some vegetation*
Class 4	*High trees with some understory*
Class 5	*Middle trees with some understory*
Class 6	*Very high trees*
Class 7	*Bushy, mid*
Class 8	*Bushy, low and high*
Class 9	*Bush, much low*
Class 600	*No or very low vegetation*

Vegetation Structure - Ground Truth

It is hard to validate the vegetation structure classes. However, two field trips have been executed in order to get more insight how the vegetation classes relate to the ground truth. The procedure and results are described in a report (appendix). Pictures of the classes are shown there. Here a summary of the report will be given. It should be emphasized that conclusions are indicative, giving the limited reference photos been made.

Summary report

Compared to the other classes, class 1, 6 and 9 can be hosted under the low vegetation classes. Class 6 is referring to none or only very low vegetation, class 1 is low vegetation only and class 9 is again a few decimetre higher than class 1.

The other classes seem to be more similar to each other. It seems that pixels with trees with a relatively large contribution for the lowest height class (0.1-0.25m) are assigned the vegetation class 8. In contrary, trees which barely seem to have understory will be classified as class 5. In both class 5 and 8 not much intermediate (several meters high) vegetation is present. Class 4 is very comparable to class 5, but the height of the trees seems to make the difference here. In both classes coniferous trees seem to dominate, where the green biomass seems to be situated exclusively in the top layer of the trees, thereby increasing the relative importance of the high point density layers.

Class 2 and 7 seem to be very similar. Looking at the vegetation profiles this can be understood. In both cases there is information in all point density layers (except the 20-30m layer), but class 2 seems to have in general higher (intermediate) trees. It is hard to determine the difference between the classes from the photos. It is clear though that both class 2 and 7 have a very mixed vegetation type.

Class 3 is very distinct of all other classes. Big trees were absent here and here vegetation is in between forest-like and bushy. However, only one location with this vegetation type has been visited. This place characterized itself by the high openness of the location, but more locations should be visited in able to define this as characteristic to this class.

Overlays

During the LiDAR processing, 'building points' have already been removed. Nonetheless, the remaining point cloud still contains many other non-vegetation features. Some of them are known and can be corrected for by overlaying other datasets representing the locations of 1) highways, 2) water areas and 3) military areas. Together with NoData areas (caused by unprocessed LiDAR data) these raster layers were placed on top of the vegetation structure raster. The methodology scheme of how these layers are created is visualized in figure 10.

Figure 10:: Methodology scheme of pre-processing of the vegetation structure overlays. Rounded boxes are vector datasets, parallelograms are raster datasets. Blue box = raw input data; yellow box = processed data. Green boxes describe data analyses.

The scheme is the product of several choices that have been made. Water has been chosen to mask the 'NoData areas'. These areas were mainly caused by the failed point height conversion because point density was too low (or points were absent). Therefore, these 250*250m areas were left out the process, resulting in NoData areas, which should be masked.

Highways were also masked since it is assumed that the 'vegetation structure' pixel that touches a road will be more influenced by objects (cars, guardrails, road signs etc..) than by true vegetation. The pixel resolution is 25*25m. If a part of this area overlaps with the highway, the entire pixel value is influenced and unreliable. To compensate for this effect, a buffer of 10 meter has been made around the highways in order to assign more pixels as 'road' instead of having more pixels with wrong vegetation classes.

For many areas with military activity the point density was reduced. Because of the thresholding (100 points) many pixels were even assigned NoData in these areas. Therefore, pixels inside a polygon were inspected on atypical (very low or NA) values. If this was case, the point density was not reduced in this area. Since the polygons are used later to overlay edited areas, polygons were removed when point density values seemed to be unedited. An example of an area with unexpected NoData values is shown in figure 11.

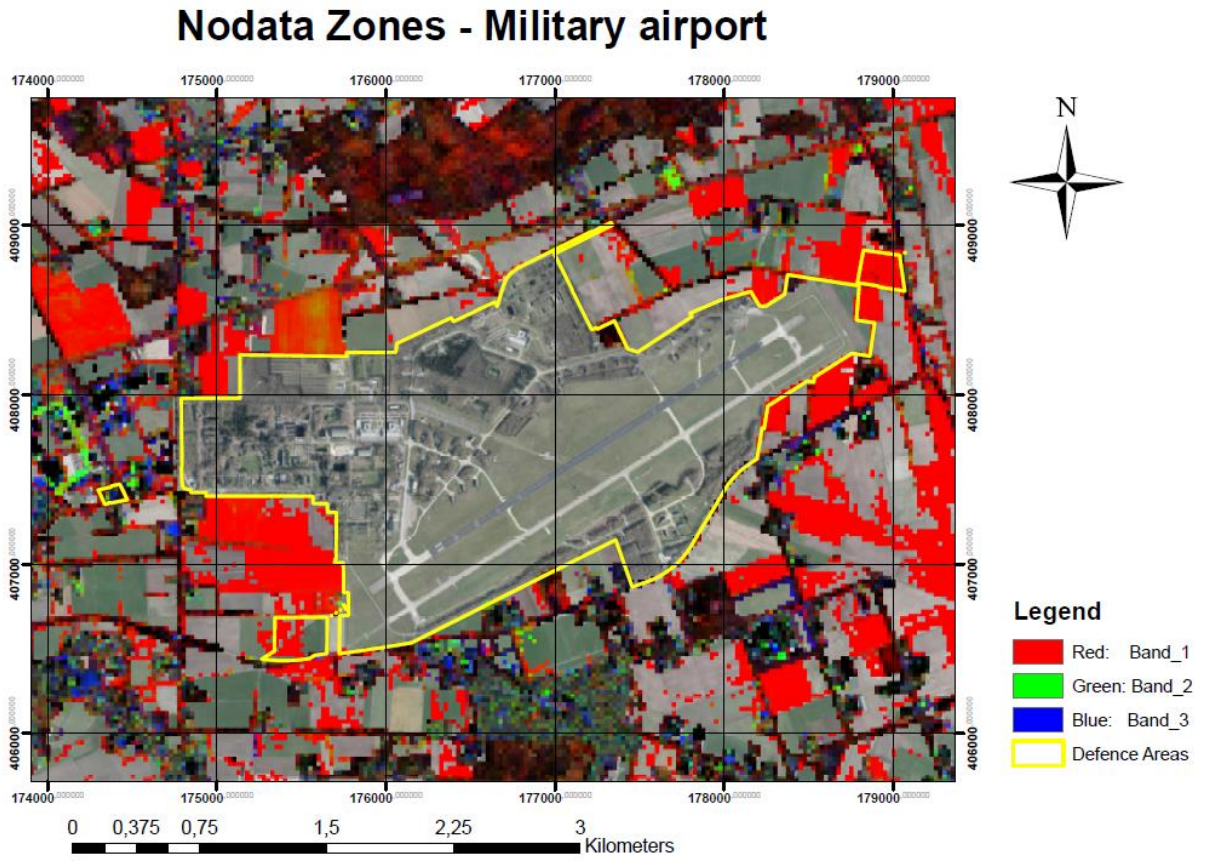


Figure 11: Military flying base Volkel, close to Uden. Point density information is missing in this area, due to removal or reduction of points by the AHN2 organisation. Colours are after the normalized point density values: Band_1 = 10-25cm, Band_2 = 25-100cm, Band_3 = 1-5m. Grid lines are according to 'rijksdriehoeksstelsel'.

Lastly, all areas that clearly contained no information as a result of failed LiDAR point processing were masked as well.

The overlay order was (top-down):
- NoData
- Military Areas
- Water Areas
- Highways
- Vegetation structure raster

In the end a vegetation structure raster was created including water areas or highways. Figure 12 shows an example of how the final vegetation structure raster with overlays looks like. By means of visual inspection of the vegetation structure raster it appeared that, especially in urban areas, some vegetation classes were spatially co-occurring. This was very much the case for vegetation class 7 and 9. Looking to their vegetation profile it indeed seems that these classes are similar. Since such scattering highly influences the mean patch area and edge density later in the process (paragraph 2.5), it has been chosen to merge these two classes. The same was observed for the classes 4 and 6. Since these classes both describe high vegetation, it was chosen to merge these classes as well.

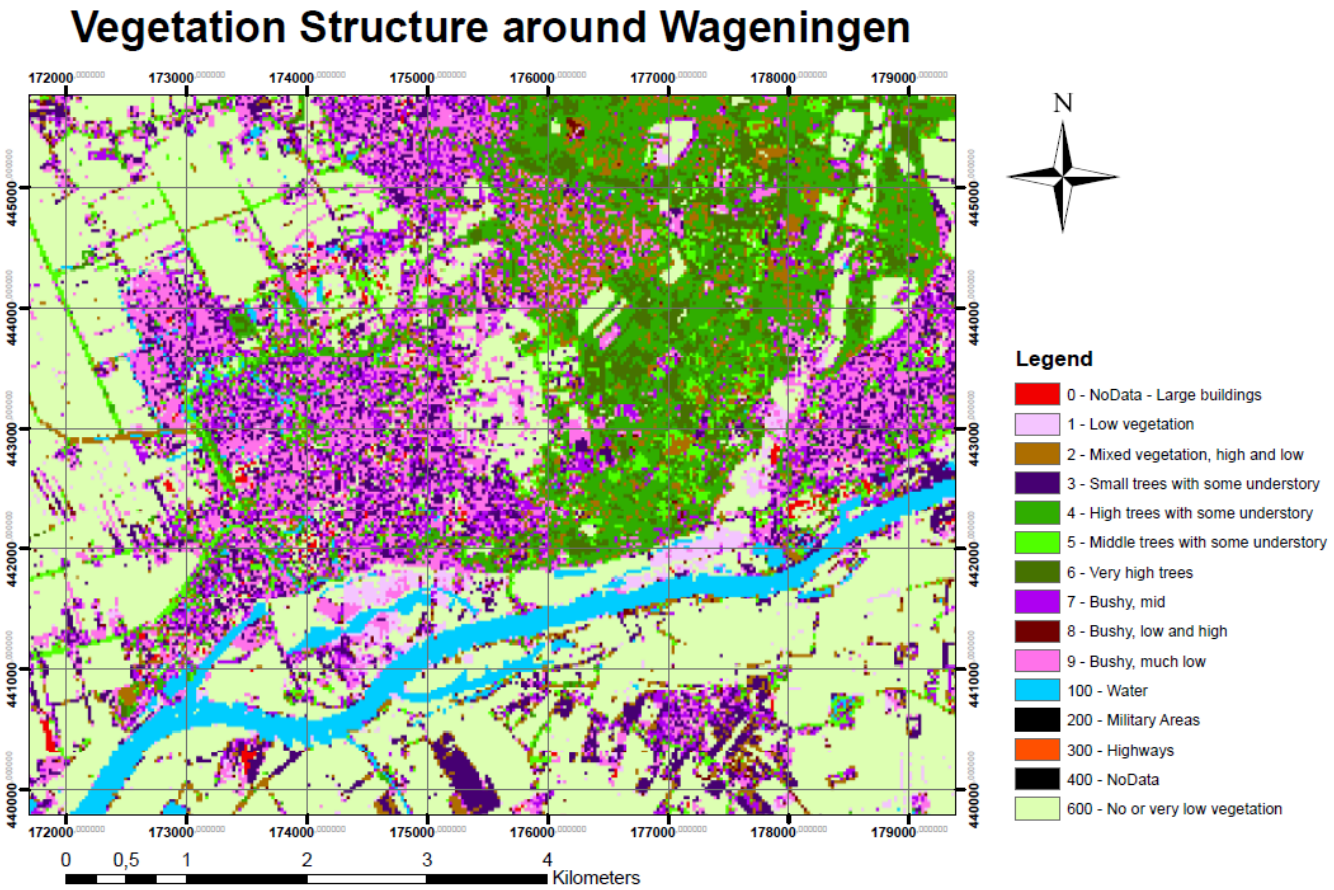
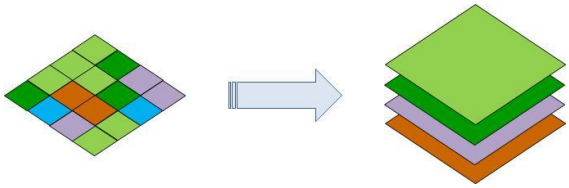


Figure 12: Visualization of the vegetation structure dataset in the surroundings of Wageningen. Grid lines are according to the 'rijksdriehoeksstelsel'.

2.5 Step 3: Creating landscape variables



The spatial resolution of the land use and vegetation classes is 25m, while the bee dataset has mainly a spatial resolution of 1 km. Therefore, landscape variables need to be constructed which typify the landscape in a way which is ecologically relevant for bees. In this chapter it is explained how the landscape variables have been created. All these variables are the input variables for the SDMs.

2.5.1 Variables SDM1 - Land use only

In a study of Aguirre-Gutierrez et al. (2015) LGN landscape variables were reclassified into classes that are similar to land use classes found in historical land use datasets of The Netherlands. The variables are believed to be relevant for wild bees. This land use dataset contains 10 classes, including 3 forest classes. For the present study these forest classes were merged into one class. Two further binary transformations have been performed with this dataset. Firstly, a 'bee habitat suitability raster' has been derived by dividing the classes into suitable and non-suitable. Secondly, a 'Managed-Natural' raster has been created by dividing the classes into 'Managed' and 'Natural'. Table 6 shows how the land use classes have been reclassified.

Table 6: Reclassification scheme of the land use classes. Left column is the same dataset used in Aguirre-Gutierrez et al. (2015). Other columns are new, reclassified, datasets derived from the first.

Reclassified LGN classes	LGN classes with 1 forest type	Suitable	Managed - Natural
Grassland	Grassland	Yes	Managed
Cultivated / Bare ground	Cultivated / Bare ground	No	Managed
Moors / Peat	Moors / Peat	Yes	Natural
Forest Mixed	Forest	Yes	Natural
Forest Deciduous		Yes	Natural
Forest Coniferous		No	Natural
Buildup / Roads	Buildup / Roads	No	Natural
Water	Water	No	NA
Swamps	Swamps	No	Natural
Sandy Soils	Sandy Soils	Yes	Natural

After the reclassifications of the land use dataset five final landscape variables are created with one value for every km^2. Several steps have been performed for this. First, I used the software 'Geospatial Modelling Environment[3] which split the raster datasets into 1*1 km tiff files (consisting out of max. 40*40=1600 pixels). From these files the variables were calculated with help of the R tool 'ClassStat' [SDMTools]. The variables are listed here with a short description of the meaning and origin.

1 - Percentage unsuitable habitat (PUH)

[3] http://www.spatialecology.com/gme/

This data has been derived from the suitability raster. It is the area of unsuitable habitat divided by the total area in a km^2. By accident it is not describing suitability, but unsuitability.

2 - Mean patch area of suitable habitat (MPA_SH)
This variable is a derivative of the suitability raster as well. It calculates how big all individual patches are in a km^2 and consequently calculates the mean.

3 - Mean edge density all classes (ED_LU)
Here the 'LGN classes with 1 forest type' dataset is used. The help of ClassStat [SDMTools] states that edged density is the "*edge length on a per unit area basis that facilitates comparison among landscapes of varying size*" [4]. It further refers to the 'Fragstats' help[5] which contains more theoretical background. First the total length of the borders of a pixel belonging to a certain class that touches another class is calculated. Then this number is divided by the length of the potential total edge (= the maximum edge length possible) in the km^2. This division is performed to account for possible differences in area size in order to make values comparable (e.g. a km^2 positioned across the coast line will contain fewer pixels). For every km^2 tile the edge density was calculated for every class. Then the mean edge density of all classes is stored as the value for that area in this variable.

4 - Edge Density of Managed-Natural (ED_MN)
This variable makes use of the managed / natural dataset. The mean patch area of suitable habitat and the mean edge density of all classes are measures for the general configuration of the landscape. The variable described here has been included as well because these edges between managed and natural systems are believed to be a measure for the connectivity of the (agricultural) landscape (Aguirre-Gutierrez et al. 2015). For this variable the edge density between the managed and natural pixels will be calculated (see table 6).

5 - Number of classes (NumClass)
This variable uses the 'LGN classes with 1 forest type' dataset as well. It counts the number of unique classes that are present in the km^2.

It has been considered to include the Simpson's index of diversity as well (with varying focal lengths) as a measure of landscape heterogeneity at various spatial levels, but after visual inspection of the datasets it appeared that high and low values were highly similar to high and low values of edge density. Therefore, it has been chosen to not include this variable in the SDM part.

For an elaborate description of the calculation of landscape variables one should read the help of FragStats.

[4] https://cran.r-project.org/web/packages/SDMTools/SDMTools.pdf
[5] http://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf

2.5.2 Variables SDM2 - Vegetation structure only

The approach followed for the construction of vegetation variables was similar as for the land use variables. For some of them a reclassification of vegetation classes has first been performed. In table 7 the reclassification scheme is showed.

Table 7: Reclassification scheme of the vegetation structure raster. It should be noted that the overlays (NoData, Water and Roads) are not listed here.

Original Vegetation class	Merged Vegetation Classes	4 Vegetation Classes
600 – No or very low vegetation	*6 – No or very low vegetation*	*1 – No or very low Veg*
1 – Low veg	*1 – Low Veg*	*2 – Low Veg*
3 – Small trees + understory	*3 – Small trees + understory*	
7 – Bushy: mid	*7 – Bushy: lower*	
9 – Bushy: much low		
2 – Mixed Veg	*2 – Mixed Veg*	*3 – Mid Veg*
8 – Bushy: low/high	*8 – Bushy: higher*	
4 – Higher trees	*4 – Higher trees*	*4 – High Veg*
6 – Very high trees		
5 – Middle trees	*5 – Middle trees*	

The first column of table 7 shows the 10 original vegetation classes constructed. The second column shows the new vegetation classes after the merging of the original classes 7 & 9 and 4 & 6, as explained earlier. This column has been used for the calculation of the vegetation structure based landscape variables (except for the edge density). The third column shows which conglomeration of vegetation classes are used to reclassify the vegetation into four height classes. It should be noted that the classes from the overlays (highways and water) are not included here.

Ten vegetation variables have been constructed, which are listed and explained below.

1 - Edge density of 4 vegetation classes (ED_VEG)
This is the mean edge density of the 4 vegetation classes. It has been chosen to merge some classes for similarity reasons. Though they are differing in structure, it does not mean that a border between every vegetation structure class combination should be considered an edge relevant for bees (e.g. class 4 and 5 or class 1, 3 and 7). Assuming vegetation height is an important aspect of vegetation edges a distinction has been made for height, resulting in the four defined vegetation height classes. In this dataset the overlays (military areas, water, highways, NoData) were present as well, but not included in the edge analysis.

2 - Mean patch Area (MPA_VEG)
This is the mean of the average patch areas of every vegetation class in a km^2.

3-10 - Vegetation abundance (VEG_1, VEG_2, … ,VEG_7, VEG_8)
For every class, the percentage coverage over all non-NA pixels inside every km^2 was calculated. This value has been used as vegetation abundance indicator of the vegetation classes.

It has been decided to exclude another variable described in Aguirre-Gutierrez et al. (2015), which is the 'number of unique classes in a km^2', because this was almost always 10 and would therefore barely add information to the prospective SDMs.

2.5.3 Variables SDM3 - Land use and Vegetation structure

For this SDM some variables earlier constructed are used, but others are made as well. From the land use dataset ED_MN and PUH are used again here. The other variables are derivations of a new overlay. This overlay is a combination of the vegetation structure raster and some land use classes. In table 8 the overlay order is listed.

Table 8: Overlay scheme of land use and vegetation structure variables. Classes listed higher in the table are on top of lower classes. 'VEG' = Merged vegetation structure raster, 'LU' = Land Use classes with 1 forest type, 'LUVEG' = Raster dataset consisting of land use and vegetation classes.

Original Class name	Origin	Original Class Value	New Class Value	LUVEG Class Name
NoData	VEG	0	0	NoData
Highways	VEG	9	60	Roads/Build-up
Build-up	LU	60	60	
Water	VEG	10	30	Water
Water	LU	30	30	
Mixed Vegetation	VEG	2	2	Mixed Vegetation
Higher trees	VEG	4	4	Higher trees
Middle trees	VEG	5	5	Middle trees
Bushy, low/high	VEG	8	8	Bushy, low/high
Cultivated/Bare ground	LU	20	20	Cultivated/Bare ground
Small trees with some understory	VEG	3	3	Small trees with some understory
Bushy, low/mid	VEG	7	7	Bushy, low/mid
Grassland	LU	10	6	Grassland
Grassland	VEG	6	6	
Low Vegetation	VEG	1	1	Low Vegetation

For this order several choices and assumptions have been made:
- Land use class 10 (Grassland) will overrule vegetation class 1 (Low Vegetation). It is assumed that LGN6 has in general classified grassland correctly. The pixel area assigned as vegetation class 1 will therefore probably be dominated by high grass. 'Grassland' will then be more accurate than 'Low Vegetation', because the latter can also mean other types of vegetation like moors.
- Vegetation class 3 and 7 are below land use class 20. This is done to be able to distinguish naturally occurring vegetation from vegetation of agriculture. Agriculture is assumed to be a better way to describe these areas, since the diversity and arrangement of the vegetation differs significantly from natural vegetation. Therefore, it has been chosen to let agriculture overrule low vegetation structure classes. Vegetation structure classes with higher vegetation are assumed to be more important than agricultural soils.

- Water from the land use dataset and water from the overlay of the vegetation structure dataset have been merged into one water class.
- Highways (vegetation structure raster) and the Roads/Buildup (land use raster) have been merged into one Roads/Buildup class. When roads are present it is assumed that vegetation classes will be more erroneous, because it can be influenced by many road objects, which is actually no vegetation.

From this new dataset several variables have been derived.

1 – Mean patch Area (MPA_LUVEG)
From the new overlay the average patch area is calculated. This is the mean of the average patch area of every vegetation class in a km^2.

2 - Edge Density Managed-Natural (ED_MN)
This is the same variable used before in the land use only SDM.

3 - Percentage unsuitable habitat (PUH)
This is the same variable used before in the land use only SDM.

4 – Edge density of 6 classes (ED_LUVEG)
From the new overlay a reclassification to 6 classes has been made, which is shown in table 9. For the same reasons as described earlier it has been chosen to merge vegetation structure classes. Note that water and roads/build-up are included for edges here.

Table 9: Reclassification scheme. Class values are listed before a qualitative description of the class. Note that these descriptions are indicative.

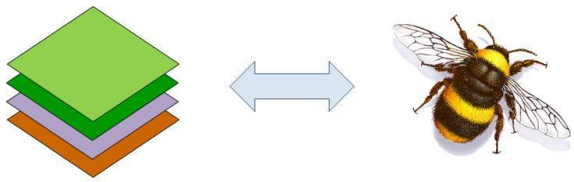
LUVEG Class Name	New Name for ED_LUVEG
0 - NoData	0 - NoData
1 - Low Vegetation	1 - Lower Vegetation
7 - Bushy, low/mid	
3 - Small trees with some understory	2 - Higher Vegetation
4 - Higher trees	
5 - Middle trees	
2 - Mixed Vegetation	3 - Mid Vegetation
8 - Bushy, low/high	
6 - Grassland	4 - Managed areas
20 - Cultivated/Bare ground	
30 - Water	5 - Water
60 - Roads/Buildup	6 - Roads/Buildup

From those 6 classes the mean edge density is calculated at the same way as describe before. It is assumed that edges between those classes are beneficial for wild bees. NoData pixels have not been taken into account.

5-15 – Class abundance (LUVEG_1, LUVEG_2, … , LUVEG_30, LUVEG_60)
For every class (except NoData pixels), the percentage coverage over all non-NA pixels inside every km^2 was calculated. This value has been used as vegetation abundance indicator of the vegetation structure classes. Note that the LUVEG_1 until LUVEG_8 are describing the same vegetation structure classes as the vegetation structure variables. However, because of the overlays values do not have to be similar. LUVEG_20, LUVEG_30 and LUVEG_60 are land use classes.

Also for this SDM, 'number of unique classes' inside a km^2 has not been included, because all km^2 showed similar results.

2.6 Step 4: Modelling Species Distribution



2.6.1 Single Species Modelling

Once the landscape predictor variables are constructed the SDM procedure can start. The modelling is a complex process in which many choices need to be made. This paragraph will explain what the single species SDMs do and elaborate on the used parameter settings.

For every individual wild bee species a habitat suitability map and a binary absence-presence map will be made (projected). The performance of the model is, amongst others, dependent on 1) the chosen model algorithm, 2) the allocation of the pseudo-absence data and 3) the selection of the testing data. Therefore, an *ensemble modelling approach* is chosen, which will be explained later in this chapter.

The SDM approach will be repeated three times, with different variables. The previous paragraph explained how variables have been created using a land use dataset and the new vegetation structure raster. Next to these constructed variables, two external predictor variables have been added to the models in order to increase their predictive value. This is *food availability (FA)* and *presence of sandy soils (SAND)*. The latter is a reclassification of the soil type dataset and has been performed by personal judgment. Peat soils and clay soils are reclassified as 'non-sandy', while sandy soil and loamy soils are classified as 'sandy'. Anthropogenic soils have been classified as 'sandy' as well, because most of the build-up areas contain some degree of sand (e.g. between pavement tiles, at which some bees could nestle as well, like *Dasypoda hirtipes* (Peeters et al. 2012). However, it should be noted that the classification is subjective to a certain extent. In total 3 different SDMs will be performed (figure 13). It should be noted that, by mistake, *the sandy soils dataset has not been in included in SDM2*. This could have an influence on the final outcome of the predictions.

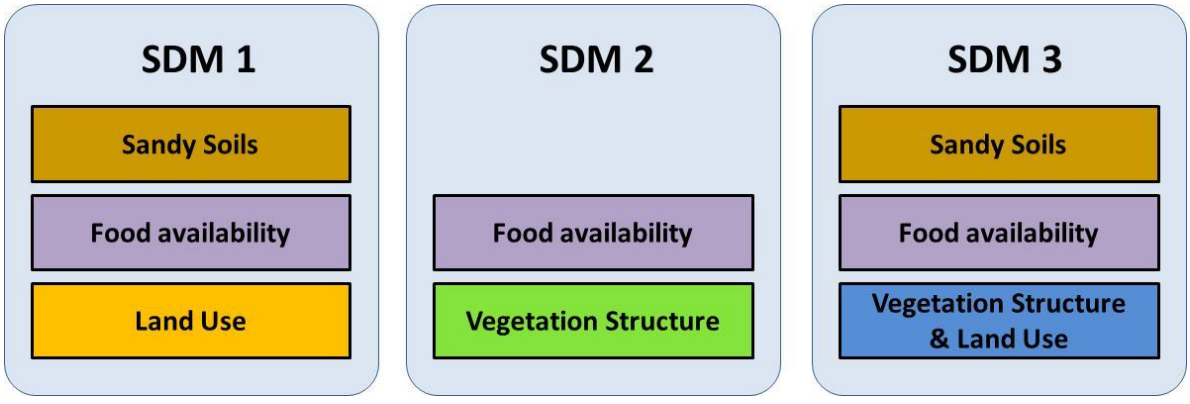


Figure 13: Three Species Distribution Models based on different variables. All models have 'Food availability' as input variable. Not that the 'Sandy Soils' variable is, by accident, excluded from SDM 2.

The whole procedure can be summarized in four steps:

- Run of multiple SDMs per species
- Ensemble modelling
- Project ensemble model
- Stack maps

These steps will be explained one by one in the following paragraphs.

2.6.2 Multispecies Modelling

Create pseudo-absences

In order to be able to run SDMs, there need to be absence data. Since the wild bee species dataset does not have absence data included, pseudo-absences (PAs) will be created. These PAs will be randomly allocated across locations (km^2) where the target species is not found but one or more other species were. This is called the target group approach (Polce et al. 2013). Though this is not as ideal as true absence data, it is believed that PAs can be used for this in order to create suitable species input for the SDMs (Barbet-Massin et al. 2012). The maximum number of records of a species in the study area is 914 (*Bombus Pascuorom*) and the minimum is 105 (*Colletes fodiens*). In total there are 2643 locations where at least 1 species is found. Barbet-Massin et al. (2012) evaluate how much PAs should be selected and that the answer depends on the algorithm used. Only GLM (General Linear Model) algorithms will be performed in this study (see next section). For this algorithm, it has been found that assigning as many PAs (though not covering the entire study area) as possible is best (Barbet-Massin et al. 2012). A possibility is to use 10 times more PAs as the number of presences for a species. However, for *Bombus pascuorum* this would mean that every non-presence cell will become a PA, leaving no room for randomization. Therefore, I have chosen to use 1250 PAs for all species. This ensures that three important criteria are met:

- For the vast majority of the species there are about 5-10 times more PAs than presences;
- In the case of *Bombus Pascuorum* there are still 2643 - 913 = 1730 locations available for the random allocation of 1250 PAs. Ideally this would be more, but a reasonable variation of PA allocation is still possible;
- In the cases of all other species there are sufficient PA allocation possibilities.

The random allocation of PAs could - by chance - favouring higher or lower values of certain environmental variables. To reduce the possible effect of this, the *PA allocation process is repeated three times*, resulting in three different absence-presence datasets for a species of interest.

Run GLMs

In this step the presence-absence data is compared to the predictor / landscape variables. In this study the following settings are used:

- Data split: This is the percentage of presence data that will be used for model construction. The other part is used for model evaluation. Value is set to 75%.

- Model evaluation runs: 10 times. This means that (for every absence-presence dataset!) the data will 10 times split up into different data selections using the 75% data split. This again means that – per species - in total 3*10 = 30 models will be constructed.
- Model algorithms: Only GLMs have been used. This has mainly been chosen for practical reasons. It has been chosen to perform a quadratic model without interaction. This means that the algorithm remains linear, but variables will be quadrated in order to correct for a possible non-linear response. No interaction is chosen because no hypothesis about interactions is made and results would become unnecessary difficult to interpret.
- Model evaluation method: Based on the model, habitat suitability (or chance of occurrence) is calculated for the 25% of the presence data which was set aside in the beginning. There are several methods that can be used to determine the model performance. For this research, the Area Under the Curve (AUC) has been chosen. The AUC is a model evaluation method that is threshold independent. How this method works exactly can be read in e.g. the study of Jiménez-Valverde (2012). Here the metric has also been criticized, but it is still widely used for SDM purposes. Besides, this study focuses on the difference in performance between vegetation structure based SDMs and land use based SDMs, rather than the creation of SDMs with high predictive power. Advantages of the AUC method are that it is threshold independent and it seems not to be affected by prevalence: the proportion of the data representing presence (Raes and Ter Steege 2007). An AUC value of 0.5 indicates that the model has no predictive value and a value of 1 is the highest score possible.
- Projecting the model: In the research the outcome of the model projection will be referred to as the 'model prediction'. When the relationship between species occurrence and predictor variables is established, the models need to be projected for the whole study area. This is done using the equation generated from the quadratic GLM. The primary output of the SDMs consists of values between 0 and 1 (or 0 and 1000, dependent on the preference of the user). If preferred, these data can be transformed - with the aid of an evaluation metric - into binary predictions of species occurrence. For this study the threshold of the suitability values for the binary transformation is chosen by the AUC method. This method optimizes the sensitivity and specificity (Jiménez-Valverde 2012).

Ensemble modelling

At this stage, the models constructed in the previous step need to be 'summarized' in order to come to one species distribution map of the wild bee of interest. This is done using an ensemble method, which uses the 30 suitability score maps. To derive the final suitability map, the median value of every cell (km^2) is chosen which results in the final habitat suitability value map. It is possible to choose the mean value as well, but to reduce the effect of possible outliers this method has not been chosen. AUC scores are stored of this final habitat suitability map. Lastly, the map will be transformed into a presence-absence map based on the AUC method again.

Variable Importance

AUC values indicate how well the model performs. However, this gives no insight yet in how and which variables influence the prediction of the suitability score. Therefore, the variable importance is calculated.

The calculation of the variable importance goes in several steps. The steps describe the situation of a variable importance calculation procedure for a single species with only 1 (pseudo-)absence - presence dataset.
1 - The values of a variable of interest (e.g. edge density) are shuffled. This means that the values of the cells are randomly reassigned over the study area;
2 - A GLM is performed taking all variables, including the shuffled one, into account;
3 - Based on the constructed GLM, suitability values are projected over the entire study area;
4 - A Pearson's correlation coefficient is calculated between the newly calculated suitability values and the suitability values originating from the prediction of the non-shuffled variables;
5 - The variable importance is calculated by subtracting the correlation coefficient from 1;
6 - Step 1-5 are repeated X times, depending on the user settings;
7 - Step 1-6 are repeated for every explanatory variable.

This procedure gives X number of variable importance values for every variable. However, for this study an ensemble approach has been chosen, which makes the entire procedure more complex. The variable importances calculated are the result of the next processing steps:

1 - For all 30 (because of the following parameter settings: 3 different PA selection and 10 data splits) GLM-based SDMs values of an explanatory variable (e.g. mean patch area) are shuffled, before the model is established.
2 - For every model the values are projected over the entire study area.
3 - A median ensemble model is created.
4 - The Pearson's correlation coefficient is calculated between the values of the predictions of this new median ensemble model (with a shuffled variable) and the original median ensemble prediction values.
5 - Step 1-4 are repeated 4 times for the same explanatory variable, resulting in 4 variable importance values of the ensemble model.
6 - Step 1-5 are repeated for every explanatory variable, resulting in 4 variable importance values for every variable used in the model.

3 Results

The outcomes of the SDMs are summarized in this chapter. The constructed richness maps will be shown, as well as the AUC and variable importance values. The last paragraph will zoom in on habitat suitability predictions of three species. The first paragraph will show how the land use data compare with the vegetation structure data.

3.1 Land Use vs Vegetation Classes

It can be hypothesized that the vegetation structure is representing similar information as the land use data. In order to investigate this, the datasets have been compared. For every pixel (resolution 25m) it has been checked which vegetation structure class and what land use class was assigned. The number of pixels for every unique combination of vegetation structure and land use has been counted and listed in the tables 10 and 11. This should give an impression how the vegetation classes relate themselves to the land use classes. In total there are 61290006 cells.

Inspecting the tables, one can make several interesting observations. Next to the expected results ('Grassland' relates to vegetation class 6, Vegetation class 9 to 'Build-up / Roads', 'water' is correlated with water), the tables provide interesting additional information. One can see that many vegetation classes are more or less equally spread over different land use classes. Vegetation class 2 seems to be related to grasslands, cultivated soils / bare ground and forest. A similar pattern can be found for vegetation classes 3, 4 and 5. Vegetation class 1 is, next to its relation with grasslands, cultivated soils and moors / peat, also the most prominent vegetation type in swamps. Vegetation class 7 seems to be highly correlated to grasslands and build-up regions, which makes sense considering the many low objects in cities. Vegetation class 8 is more spread over different classes (grassland, cultivated soils, moors / peat, forest and build-up / roads).
It's noteworthy that the land use class 'forest' seems to be an aggregation of many vegetation types. Vegetation class 2 and 4 are most prominent in this land use class, but all other classes (except for vegetation class 1, 3, 6, 9 and 10) make a considerable contribution as well.

Summarizing these results it is clear that land use dataset describing the landscape at a different manner than the vegetation structure dataset.

Table 10 and 11: The upper table (10) shows what the distribution is of vegetation structure classes over the land use classes (in percentage). The upper left cell value indicates that 35,3 % from all pixels belonging to the vegetation structure class 'Low Vegetation' (total 1245144) is overlapping with the land use class 'Grassland'. The bottom table (11) is showing the same, but uses the land use classes as starting point. Darker cell colours indicate higher values.

	Vegetation Structure Classes (%)									
Land Use Class	1 - Low Veg	2 - Mixed Veg	3 - Small trees with Understory	4 - Higher trees	5 - Middle trees	6 - No or very low veg	7 - Bushy: low/mid	8 - Bushy: low/high	9 - Highways	10 - Water
Grassland	**35.3**	**26.4**	**37.3**	**18.8**	**36.0**	**51.3**	**35.1**	**32.1**	**4.9**	**2.7**
Cultivated BareGround	**36.2**	**12.5**	**17.0**	**4.8**	**13.4**	**43.8**	**12.0**	**17.7**	**1.1**	**0.5**
Moors/Peat	**13.2**	**2.1**	**0.5**	**0.2**	**1.1**	**0.4**	**0.7**	**8.5**	**0.0**	**0.1**
Forest	**1.7**	**48.6**	**5.1**	**66.2**	**33.7**	**0.2**	**11.2**	**27.9**	**0.9**	**0.2**
Build-up Roads	**4.6**	**9.5**	**38.9**	**9.6**	**15.1**	**3.2**	**38.4**	**10.8**	**93.0**	**0.4**
Water	**1.3**	**0.3**	**0.5**	**0.3**	**0.4**	**0.3**	**0.7**	**0.6**	**0.1**	**95.5**
Swamps	**5.6**	**0.4**	**0.2**	**0.1**	**0.1**	**0.3**	**1.0**	**1.9**	**0.0**	**0.4**
SandySoils	**2.2**	**0.2**	**0.5**	**0.0**	**0.3**	**0.3**	**0.9**	**0.4**	**0.0**	**0.2**
Total Cell Counts	**1245144**	**2550431**	**1543013**	**1699094**	**1389596**	**13162683**	**4767686**	**585371**	**168825**	**3533160**

	Land Use Classes (%)							
Vegetation Class	Grassland	Cultivated / BareGround	Moors / Peat	Forest	Build-up Roads	Water	Swamps	SandySoils
1 - Low Veg	**3.9**	**5.8**	**42.5**	**0.6**	**1.5**	**0.5**	**34.2**	**19.6**
2 - Mixed Veg	**6.0**	**4.1**	**14.1**	**33.7**	**6.4**	**0.2**	**4.7**	**3.8**
3 - Small trees with understory	**5.1**	**3.4**	**1.9**	**2.1**	**15.9**	**0.2**	**1.7**	**6.1**
4 - Higher trees	**2.8**	**1.0**	**1.1**	**30.6**	**4.3**	**0.1**	**0.5**	**0.5**
5 - Middle trees	**4.5**	**2.4**	**3.9**	**12.7**	**5.6**	**0.1**	**0.9**	**2.9**
6 - No or very low veg	**60.1**	**74.3**	**14.1**	**0.8**	**11.4**	**1.2**	**21.5**	**30.4**
7 - Bushy: low/mid	**14.9**	**7.4**	**9.0**	**14.6**	**48.6**	**0.9**	**23.6**	**31.3**
8 - Bushy: low/high	**1.7**	**1.3**	**12.9**	**4.5**	**1.7**	**0.1**	**5.4**	**1.5**
9 - Highways	**0.1**	**0.0**	**0.0**	**0.0**	**4.2**	**0.0**	**0.0**	**0.0**
10 - Water	**0.8**	**0.2**	**0.5**	**0.2**	**0.4**	**96.7**	**7.5**	**3.9**
Total Cell Counts	**11228545**	**7764671**	**387581**	**3671892**	**3760648**	**3492424**	**202149**	**137093**

Spatial Comparison

The tables in previous paragraph give a general idea how the land use classes are relating to the vegetation structure classes and vice versa, but not how this is expressed spatially. This is shown in figure 14. Here it is clearly shown what differences exist between the land use and vegetation structure raster. It's clear that the land use raster differentiates better in agricultural practices (light green, brown, pink, purple, yellow), but the field edges are hardly visible. The opposite is true for the vegetation structure raster: there is no differentiation in agricultural practice, but field edges with trees (darker green or brown) are clearly visible. This is a clear difference between both datasets.

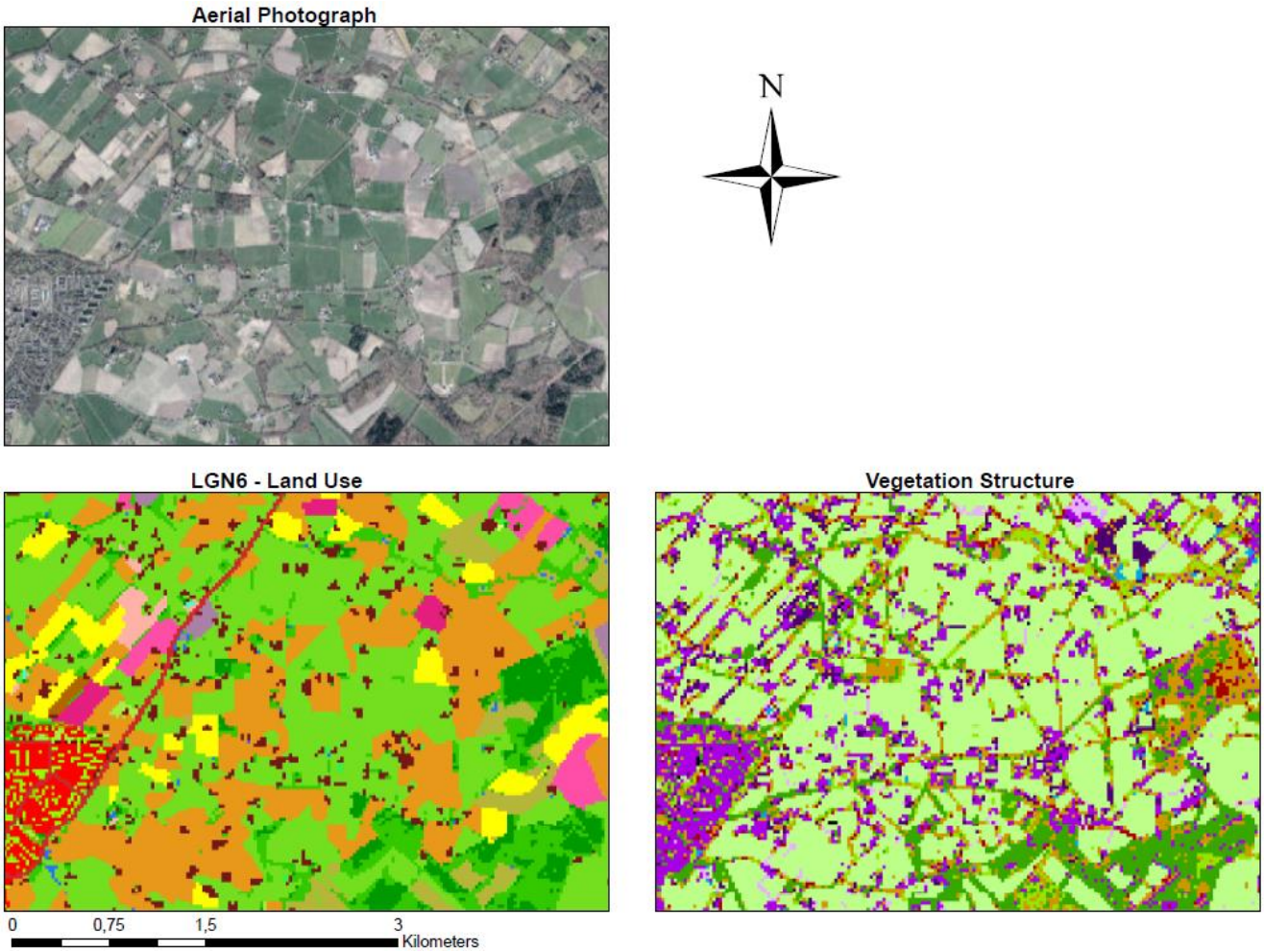


Figure 14: Three maps (aerial photograph, LGN6 and Vegetation Structure) showing the area east of Doetinchem, The Netherlands.

3.2 Model Performance

To be able to interpret the wild bee richness maps in a proper manner, the model performance should be analysed first. This will be done according to the average AUC values (figure 15). Here it can be seen that the LUVEG SDM performs best, followed by the VEG SDM and the LU SDM seems to perform worst. The standard deviations seem to indicate that the variation is too high to let the differences be significant. However, the three possible pairwise (per species) t-tests all show significant differences (p-value < 0.001 in all cases). This means that the LUVEG SDM performs systemically better than the other models. This effect can also be seen in figure 16. Green dots (LUVEG SDMs) are in almost all cases the highest, followed by VEG and LU respectively. Another typical effect is visible here, which is that the AUC values decrease with the number of observations, as known from, amongst others, Aguirre-Gutierrez et al. (2013)

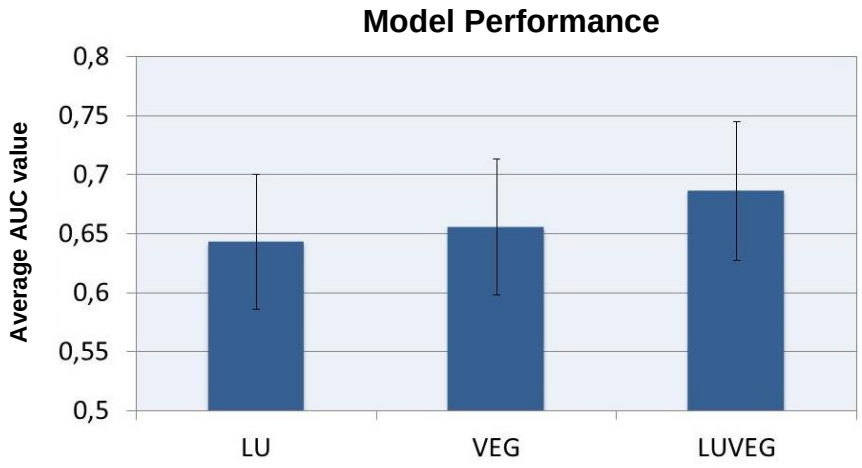


Figure 16: Average AUC values over the 60 SDMs of Land Use (LU), Vegetation Structure (VEG) and both variable (LUVEG). Error bars represent standard deviation.

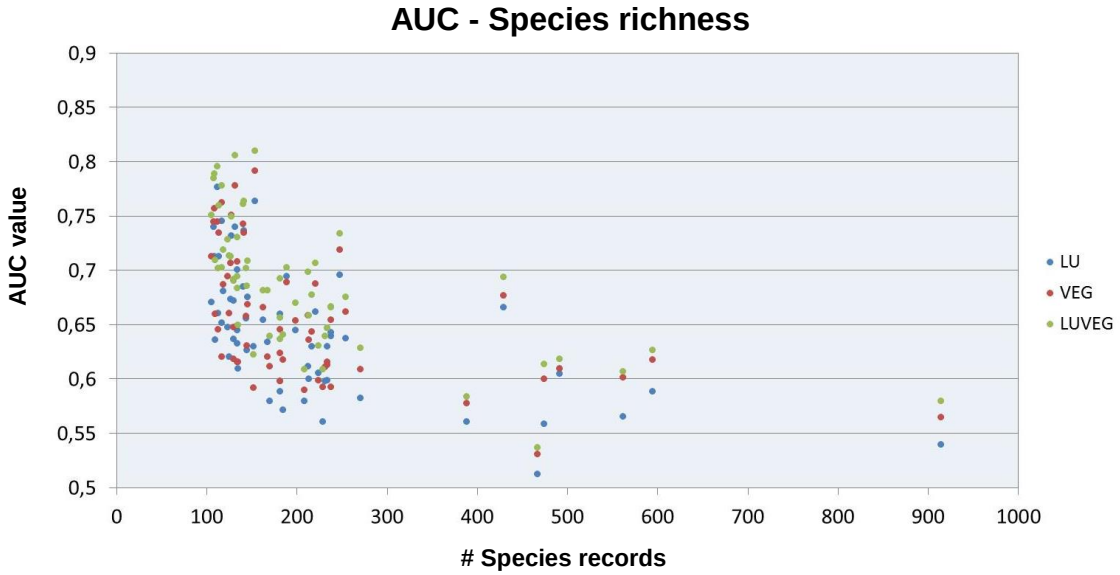


Figure 16: Scatter plots of AUC values against the number of records. Blue dots represent the land use (LU) SDM, red dots are the vegetation structure (VEG) SDM, green dots the combined (LU + VEG) SDM.

3.3 Species Richness Maps

This paragraph summarizes the predictions of the SDMs by creating wild bee richness maps. As mentioned in the introduction there are mainly two methods to create a map of species richness. One can add the binary predictions or the habitat suitability values of all species. In this research the first method has been chosen. A disadvantage of the second method is that the final value is multi-interpretable: a moderate value could indicate high suitability of some species and low suitability for others, while at the same time it could mean that it is in general moderately suitable for all species. This problem is also recognized by Dubuis et al. (2011). The transformation of habitat suitability values to binary presence-absence predictions is a threshold-based method. Different methods are possible. As explained earlier, the threshold is chosen that optimizes the sensitivity (proportion of observed presences predicted correctly) and the specificity (proportion of - randomly allocated - PAs correctly predicted). The exact method can be read in Jiménez-Valverde (2012).

All binary predictions of species occurrences (based on the ensemble modelling approach) are stacked (summed). This results in the creation of a *species richness map*, which indicates the number of unique species at a location. For every SDM type (land use, vegetation structure or both), a richness map has been made.

The richness maps are very useful for getting insight in the spatial patterns of the wild bee richness. However, it would be interesting as well to add information about the model consistency or variation in prediction. After all, the procedure of the randomization of PA allocations and differing model testing areas allowed the SDM to predict 30 different habitat suitability values per location. If the 30 habitat suitability values are in general close to each other, model predictions can be considered consistent. To get more insight in the model prediction consistency '90% certainty maps' are shown as well. For these maps, only the cells where a presence is predicted 27 or more out of 30 are considered a presence for every species. This does not directly visualize model prediction consistency (or variation), but combined with the original map it does. For example, if an area that is in general predicted to be species rich in the original richness map but species poor in the 90% certainty map (relative to the entire study area), it can be concluded that the predictions were not consistently high there. Secondly, the map gives an impression of areas that are believed to be species rich with higher certainty, which would not have been visible in a regular map that indicates the variation of the predictions of the species.

All maps are visualized in figure 17.

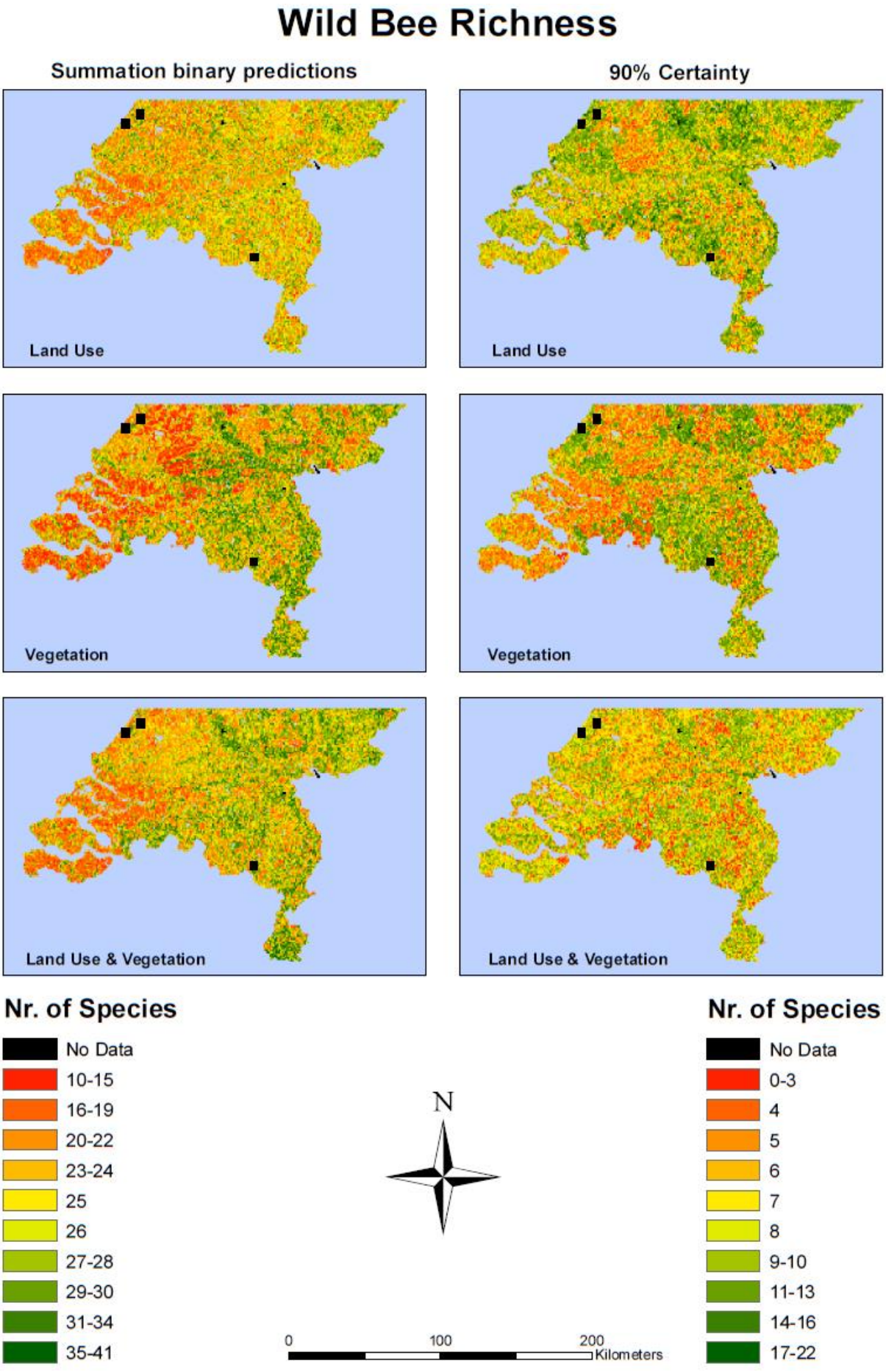


Figure 17: The Left column shows the expected wild bee richness according to the binary predictions of the individual SDMs. The right column shows the prediction of the richness with 90% certainty.

The maps show some striking similarities and differences. In this section there will be first a comparison between the ensemble model predictions and subsequently an elaboration upon the significance of the 90% maps.

Comparing ensemble model predictions

The first clear similarity is that the Zeeland province does not seem to support many bee species for all of the models. Secondly, the 'Green Heart' of The Netherlands (more or less the area between Rotterdam, Amsterdam and Utrecht) seems to be species poor, though this is most clear in the VEG SDM prediction.

The land use model prediction seems to be very scattered in the rest of The Netherlands, no very clear patterns arise. In contrast, the VEG map shows several patterns. According to this map it seems that the eastern border of The Netherlands is in general species rich. Also the 'Veluwe' and 'Utrechtse Heuvelrug' seem to be species rich. In the LUVEG map patterns are similar, though in general less obvious than the VEG map. This map shows high species richness in the South of Limburg, more than the other maps. The differences between the predictions are visualized in figure 18. Differences are high in the Green Heart and the Veluwe / Utrechtse Heuvelrug. Also, the Eastern part of Limburg the land use map seems to predict less species compared to the vegetation structure map.

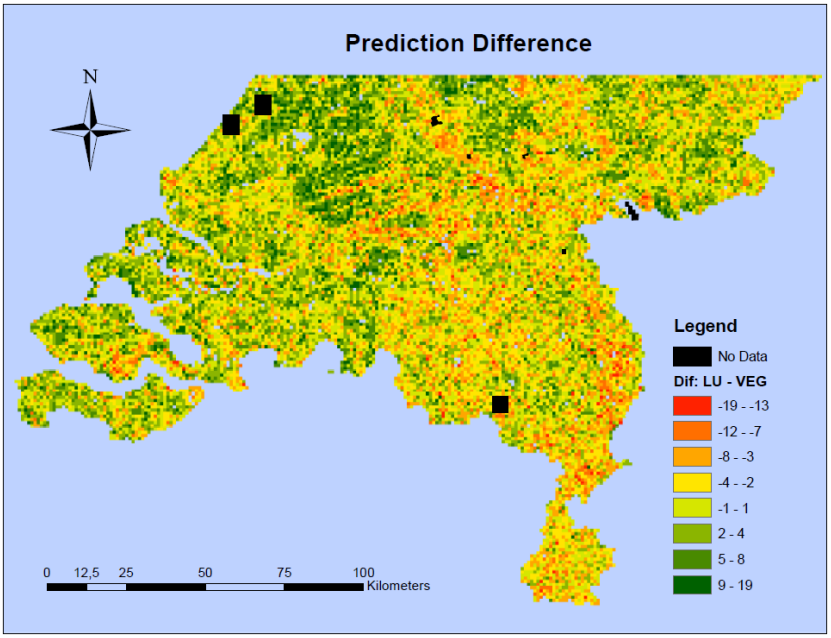


Figure 18: Prediction differences between the LU model and the VEG SDM. Vegetation structure prediction values have been subtracted from the land use values. Red locations indicate higher species richness predictions for the vegetation structure raster.

90% maps

In general the 90% certainty maps show clearer patterns than the ensemble models. The general species richness is lower than the ensemble maps, but this is due to the preselection of locations of at least 27 presence records. In all these maps it is clear that Zeeland and the 'Green Heart' are species poor. Furthermore, the 'Randstad' area (West of the Netherlands, dominated by urban areas) seems to be more species rich than other areas, which is more visible in the 90% maps than the

ensemble maps. The model predictions of both vegetation structure and land use show something interesting. Zeeland has on average relatively higher wild bee richness than the Green Heart on the 90% map, while this is vice versa according to the ensemble modelling map.

For all models, the Utrechtse Heuvelrug and the Veluwe are, here as well, predicted to be species rich. In general, the east of The Netherlands and Limburg seem to be species richer than other areas.

3.4 Variable Importance

This paragraph zooms in on the importance of the variables that determine the predictions of the species occurrence, and with that, species richness. The first section will elaborate on which variables seem (not) to be important for the prediction of a single bee species, whereas the second section will relate the variables of the different models to the species richness found.

3.4.1 Landscape variables & Individual Species

Land use

It appears that the average importance of the variables differs are very differing (figure 19). It seems that food availability and edge densities play on average an important role for species suitability scores. Also the sandy soils scores generally high. Striking is the low importance of the 'number of unique classes' and the 'mean patch area'.

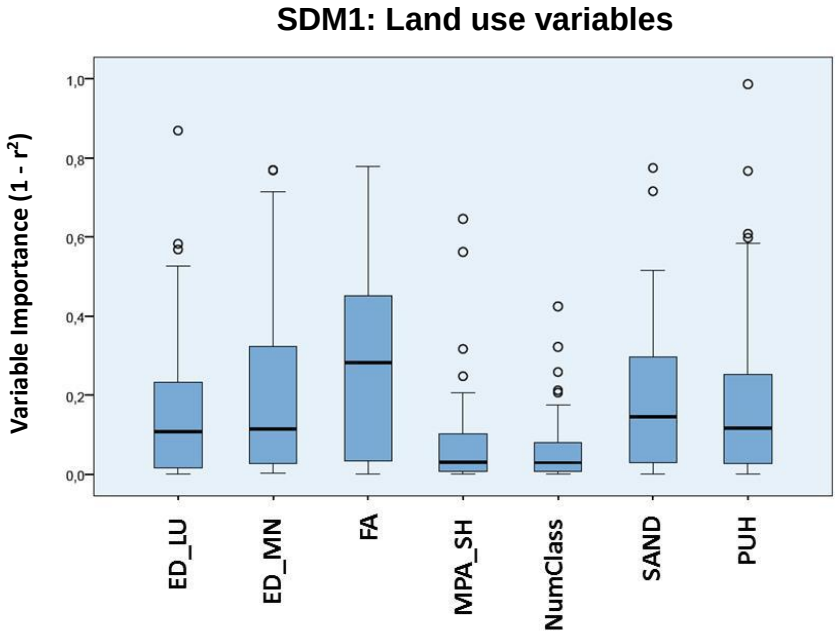


Figure 19: Boxplot showing the distribution of variable importance values for the predictors of the 60 wild bee species of the Land Use SDM

Vegetation Structure

There is one striking observation in figure 20. The food availability seems to be relatively the most important variable. It further seems that vegetation class 2 (mixed vegetation) and class 6 (no or very low vegetation) are in general important. Also here, edge density and mean patch area seem to exert little influence on the SDMs.

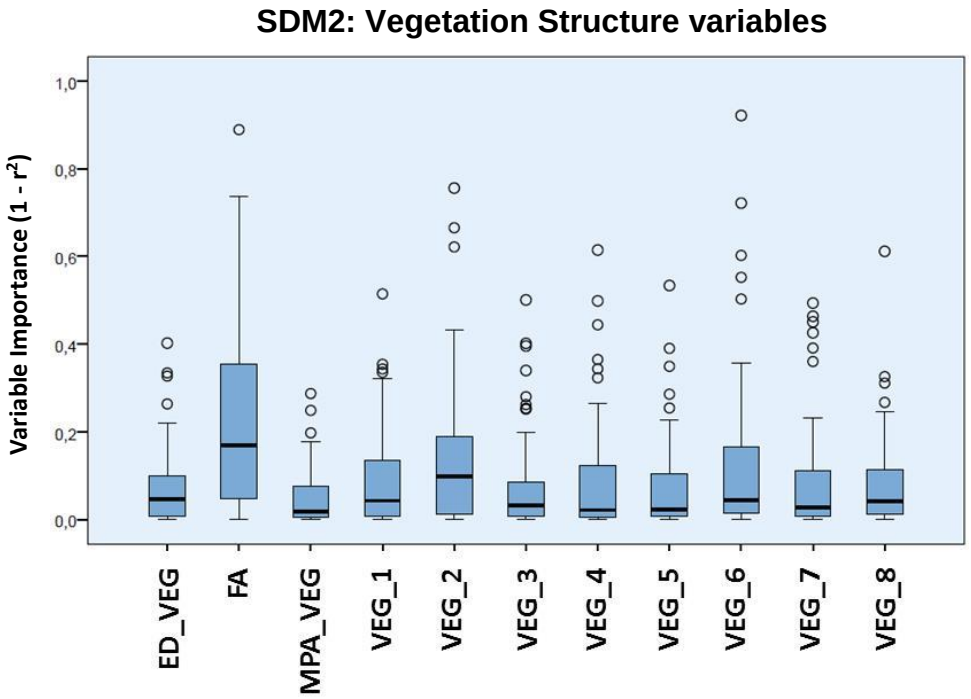


Figure 20: Boxplot showing the distribution of variable importance values for the predictors of the 60 wild bee species of the Vegetation Structure SDM.

Land use and Vegetation Structure

For the LUVEG model there are many variables, which differ a lot in importance (figure 21). First of all, it is visible that LUVEG_20 (cultivated / bare ground) seems to be the most determining variable, followed by LUVEG_60 (roads / build-up). Furthermore, PUH, sandy soils and food availability are important here as well. On average, the vegetation classes seem to have less influence on the model, though LUVEG_2 (Mixed vegetation), LUVEG_4 (higher trees) and LUVEG_6 (No or very low vegetation) are more important than others. The edge densities used here seem to have not much influence as well.

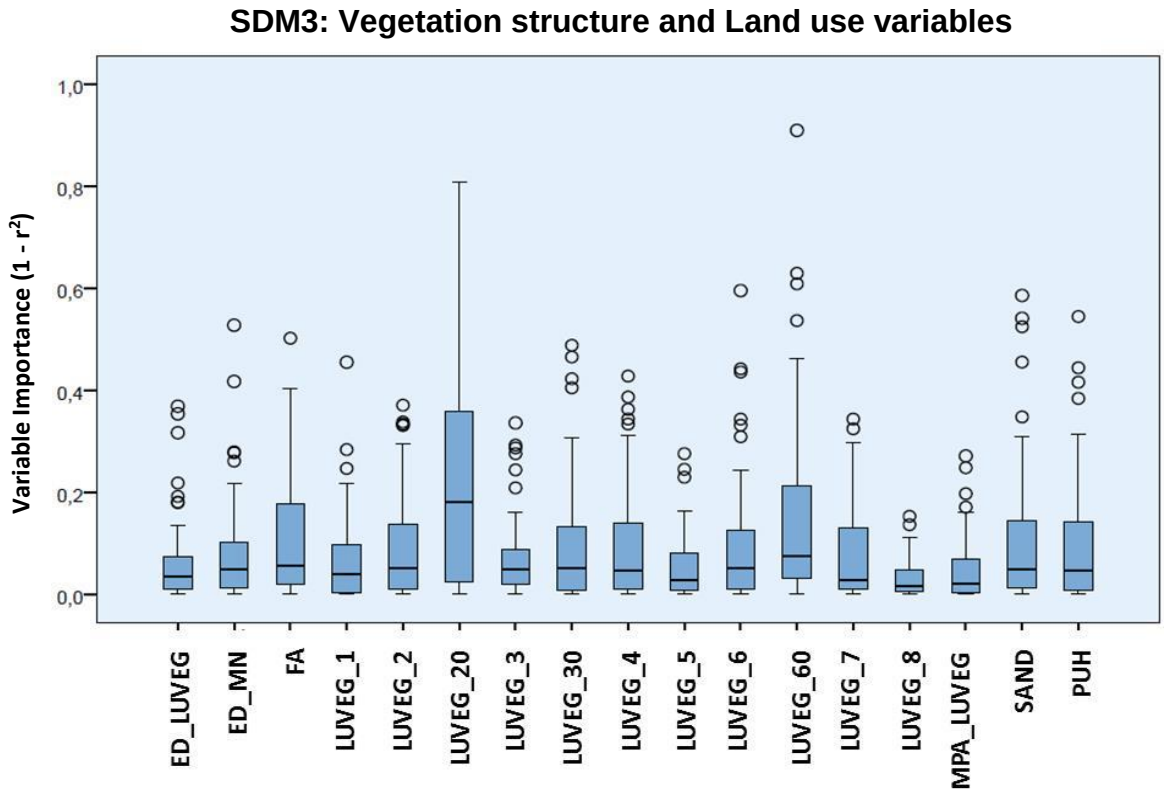


Figure 21: Boxplot showing the distribution of variable importance values for the predictors of the 60 wild bee species of the Combined SDM.

Overall boxplot observations

The importance of edge density seems to be in general low, whereas food availability seems to be one of the most important variables. The importance of the % abundance of the vegetation structure classes is in general relatively low, though with exceptions. Grasslands, Roads and mixed vegetation score relatively high importance.

3.4.2 Landscape variables & Species Richness

GLM for Richness

In section 3.4.1 it is discussed how important a variable seems to be for the prediction of single bee species. Here the correlation between the *landscape variables* and the final *species richness* map will be shown. Three new (non-quadratic, no interaction) GLMs have been constructed that use the same landscape variables, but the dependent variable is now the species richness. The slope coefficients of the variables and the r^2 (coefficient of determination) of the three models are presented in this section. This r^2 should be an indication about how well the variables are explaining *wild bee richness*, rather than indicating what the predictive power is for individual species (like the AUC earlier). Next to this, the variable importances will be calculated again as well. These have been calculated the same way as for the individual species SDMs, but now it indicates how important variables are for explaining the *richness* of the species, rather than their average importance for the prediction of single wild bee species. The calculation has been performed at the same way as in the SDMs earlier. First the values of one of the landscape variable are shuffled . All explanatory variables (of e.g. the land use SDM, including the shuffled one) are used as independent variable in the GLM. The slope coefficients of this new model are then used to make a new species richness prediction. The new species richness values are plotted against the original species richness values, resulting in a r^2. This r^2 is subtracted from 1 to obtain the importance of the (shuffled) variable. This whole procedure is repeated 5 times for every variable of all three SDMs. Again, higher values indicate higher importance. Coefficients and variable importances are listed in table 13. This procedure has been done for all three SDM types (LU, VEG and LUVEG).

Collinearity

The slope coefficients of the landscape variables can only make sense once it is certain that they have no collinearity with each other. Therefore, pairwise correlations between the independent (landscape) variables have been performed in order to conclude, based on their r^2, collinear or not. When to exclude a variable can be trivial. For this study, it has been chosen to leave out variables that have a r^2 of 0.75 or higher with another variable. However, sometimes a variable was collinear with another one with a r^2 between 0.7 and 0.75. In these cases, it has also been investigated how collinear the variable was with others. If it showed collinearity to a certain extent (> 0.5) with other variables, it had been chosen to still exclude this variable. Following this procedure, the reliability of the slope coefficients of the variables is enhanced.

<u>SDM1 - Land use (LU SDM)</u>

Variables were in general not very collinear. However, ED_LU appeared to be collinear with ED_MN (r^2 = 0.70) and also, to a less extent, with the mean patch area of all classes (r^2 = 0.57). Therefore, ED_LU has been excluded.

<u>SDM2 - Vegetation structure (VEG SDM)</u>

For the vegetation structure no variable has been removed. The highest correlation found was between VEG_3 and VEG_7 (abundance of 'Small trees with some understory' and 'Bushy, low/mid' respectively), which had an r^2 of 0.66.

SDM3 - Land use and Vegetation structure (LUVEG SDM)

There were several variables that were correlated to other variables (overview table 12). The ED_MN showed high correlation with build-up areas. This can be explained, because build-up areas often show a very fragmented pattern (for the LGN6 raster and the vegetation structure raster as well), resulting in higher edge densities. Grasslands are classified suitable for bees and it can be assumed that this is causing the high correlation between PUH and grasslands. ED_LUVEG shows high correlation with the mean patch area and LUVEG_7. The latter can probably be due to the many scattered LUVEG_7 pixels in urban areas, which automatically increases the edge densities. Lastly, mixed vegetation (LUVEG_2) was correlated with FA.

Table 12: Collinear predictor variables of the LUVEG SDM

Variable 1 (included)	**Variable 2 (excluded)**	**R^2**
LUVEG_60 (Build-up / Roads)	ED_MN	0.79
LUVEG_6 (No or very low vegetation; Grassland)	PUH	0.76
MPA_LUVEG	ED_LUVEG	0.72
LUVEG_7 (Bushy, low/mid)	ED_LUVEG	0.67
LUVEG_2 (Mixed Veg., high/low)	FA	0.67

It has been chosen to exclude ED_LUVEG, because it was also, to a lesser extent, correlated with other variables (r^2 around 0.40-0.50). The variables FA and PUH are also removed together with ED_MN to include mostly vegetation structure variables. This leaves one landscape heterogeneity variable (MPA_LUVEG), which is also completely based on vegetation structure (thus point cloud information).

The importances of the variables are visualized in bar graphs in figure 22.

Table 13: Overview of the importance of the variables, together with their corresponding slope coefficient, of the species richness prediction for the three SDMs. In the first column the r^2 is derived from the correlation of the original richness predictions with the predictions based on the linear model. 'LU' means 'land use model', 'VEG' means ' vegetation structure model' and 'LUVEG' means 'combined model'.

Model	Variable	Variable Importance ($1-r^2$)	Coefficients
LU $r^2 = 0.4598$	*MPA_SH*	**0.4093**	**-6.8067**
	PUH	**0.2285**	**-0.0531**
	FA	**0.0204**	**0.0153**
	NumClass	**0.0048**	**0.2103**
	SAND	**0.0013**	**0.1748**
	ED_MN	**0.0010**	**-3.5726**
VEG $r^2 = 0.4834$	*ED_VEG*	**0.1143**	**121.6583**
	MPA_VEG	**0.1088**	**-11.2294**
	VEG_5	**0.0290**	**-0.1743**
	VEG_3	**0.0209**	**0.1581**
	VEG_4	**0.0087**	**0.1006**
	FA	**0.0053**	**0.0195**
	VEG_2	**0.0047**	**0.0714**
	VEG_1	**0.0015**	**0.0407**
	VEG_6	**0.0001**	**0.0109**
	VEG_7	**0.0000**	**0.0019**
	VEG_8	**0.0000**	**NA**
LUVEG $r^2 = 0.4287$	*MPA_LUVEG*	**0.1901**	**-7.4727**
	SAND	**0.0963**	**2.4934**
	LUVEG_3	**0.0845**	**0.4676**
	LUVEG_60	**0.0584**	**-0.1054**
	LUVEG_1	**0.0573**	**-0.1692**
	LUVEG_20	**0.0222**	**-0.0547**
	LUVEG_5	**0.0215**	**-0.1374**
	LUVEG_4	**0.0120**	**0.0547**
	LUVEG_8	**0.0099**	**-0.0037**
	LUVEG_2	**0.0072**	**-0.0495**
	LUVEG_7	**0.0048**	**-0.0508**
	LUVEG_6	**0.0016**	**-0.0147**
	LUVEG_30	**0.0006**	**-0.0089**

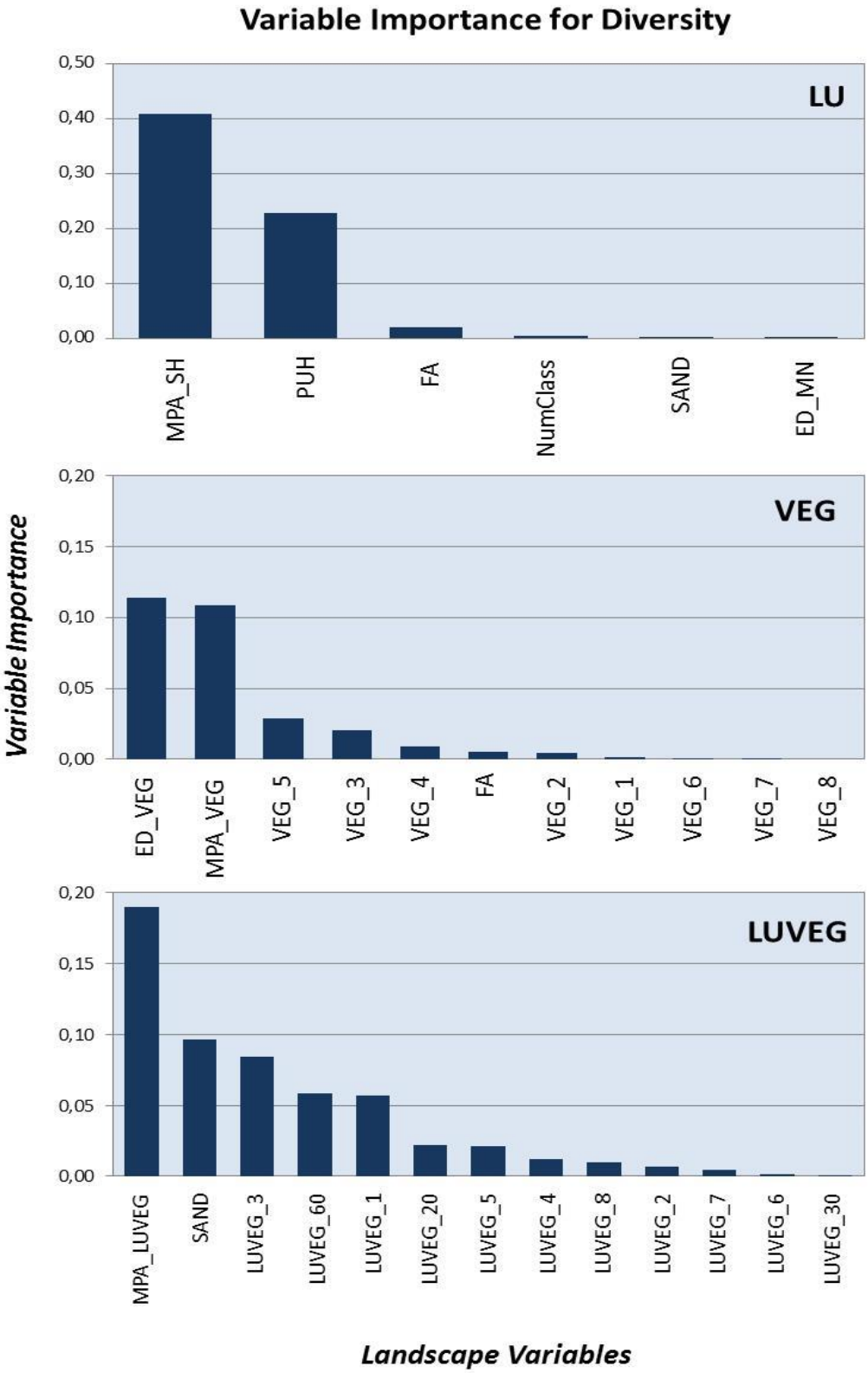


Figure 22: Variable importances for every model type. Note that some variables of the total are excluded because of collinearity. Note that the bars are averages of 5 shuffle repetitions. The standard deviation has not been shown, because the values were only differing a bit (coefficients of variation were in general < 0.01)

Variable Importance

Table 13 provides a clear overview on how the landscape variables exert influence on the prediction of wild bee richness for the different SDMs. For the LU, VEG and LUVEG model it seems that mean patch area is (one of) the most important variable, as well as edge density. In the vegetation structure model it is the most important variable, while in the other two models the total edge density was collinear with the mean patch area.

There are other interesting results. Higher unsuitability decreases the expected bee richness in the land use model, which can be expected. Sandy soils seem to be important for the combined model, while it has much less influence on the land use model, though it has a positive influence for both. Vegetation class 3 (positive), 4 (positive) and 5 (negative) seem to be important as well for both the LUVEG model and the LU model. It is interesting to note that all other vegetation classes are negatively correlated with species richness in the combined model, while positively in the VEG model.

Coefficient of determination

It can be seen that the vegetation structure variables can explain species richness best ($r^2 = 0.48$), while the land use ($r^2 = 0.46$) and the combined model ($r^2 = 0.43$) seem to be less precise. It should be noted that these coefficients indicate how well the models can explain the *species richness*. This is a key difference between the AUC values, which have given an impression on how the SDMs perform for the prediction of *single species*. Therefore, the results found here are not one-to-one comparable with the results found earlier. In line with this, it should be noted that the AUC values give an indication about the accuracy of the model predictions, while this is not the case for the three coefficients of determination found here. The accuracy describes how much your actual predictions are deviating from the truth, which was tested by the data set aside for testing the model performance (resulting in AUC values). The coefficients of determination are derived from the correlation between the original richness maps with the predicted richness maps. Since the predicted species richness is a product from the correlation with the original species richness map (stack of binary values), the accuracy is playing no role here. After all, no comparison has been made with species richness data that is considered to be (higher likely to be) true. The original richness map is a prediction itself (made by the SDMs). Therefore, the r^2 is indicating the *precision*, which indicates how much the prediction deviates from the original richness map.

3.4.3 Linear Model predictions

The landscape variables (raster format) have been multiplied with their corresponding (derived) slope coefficient. The summation of this new data, together with the and the model intercepts resulted in three new species richness maps. Here, spatial patterns can be analysed and compared to the original richness maps (figure 23). It indeed seems that the maps are showing the same pattern (high species richness in Veluwe / Utrechtse Heuvelrug and low species richness in Zeeland). Furthermore, it can be seen that the vegetation structure map shows the highest contrast between species rich and species poor regions. The land use seems still rather noisy and with limited spatial differences. Also, the combined model seems to be more species poor, while the vegetation structure map seems to be the most species rich.

Predicted Species Richness Maps

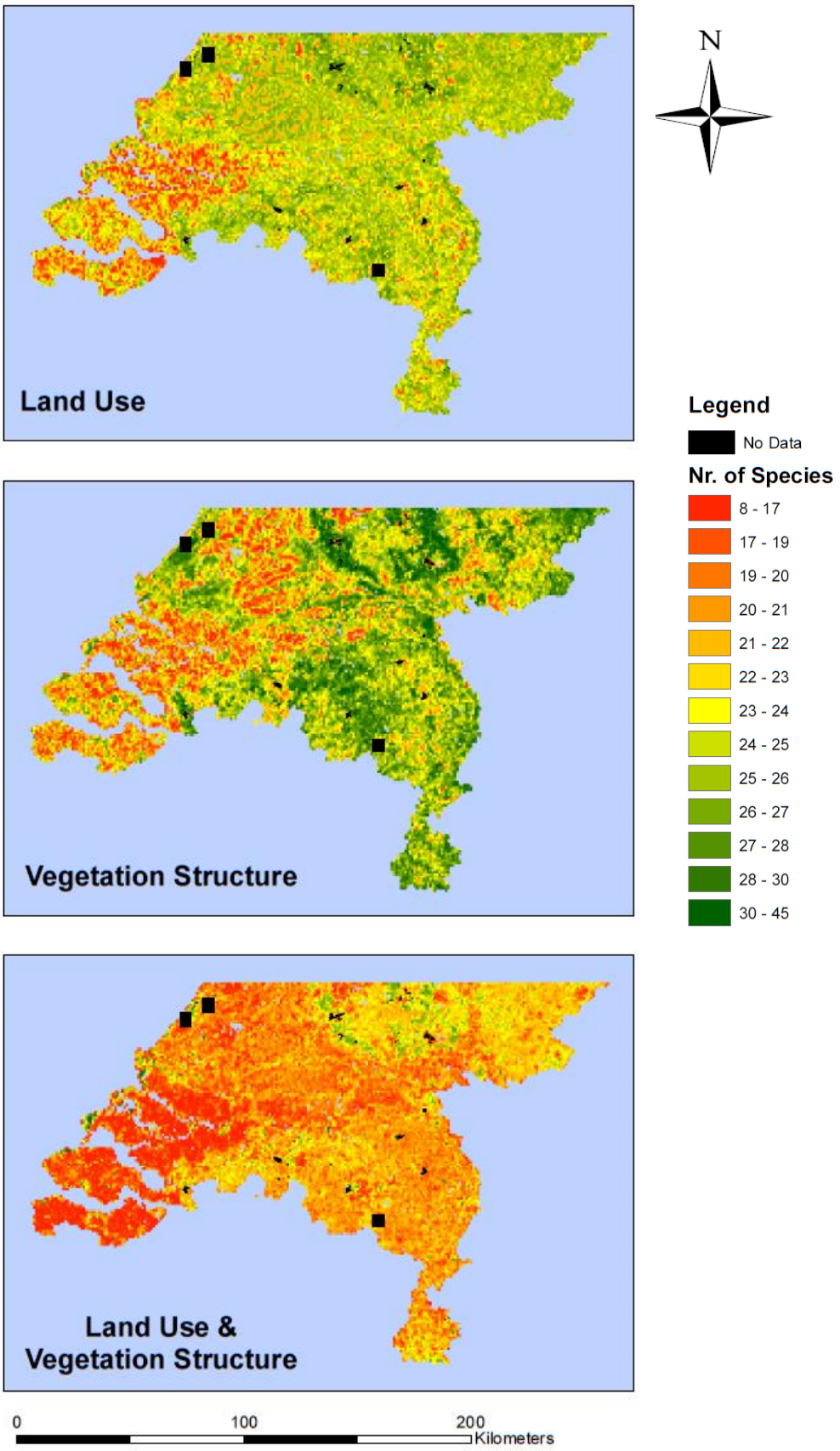


Figure 23: Species richness predictions according to the derived slope coefficients of the variables. The maps are named according to the type of input variables of the SDMs. Note that the last map seems to be very different, but this is mainly caused by a general prediction of about 3 till 4 species less per location (see legend) compared to the other two maps.

3.5 Single Species Predictions

It is hard to validate the richness maps, since independent data is barely available. However, it is possible to inspect the habitat suitability maps of some single species for peculiarities in order to see if model predictions make sense. Figure 24 shows three examples of single species' suitability predictions.

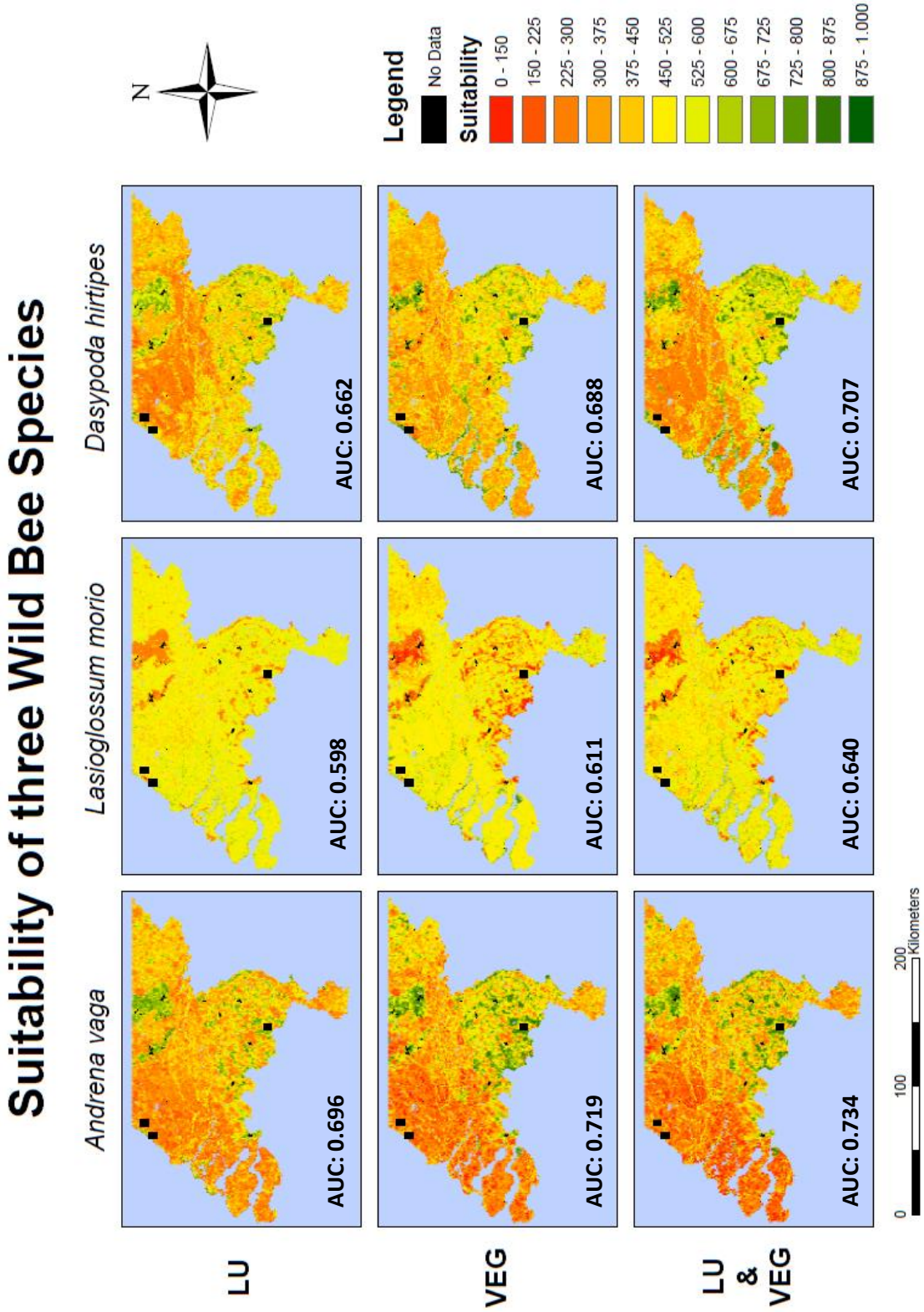


Figure 24: Habitat suitability maps for three different species predicted by land use (LU) variables, vegetation structure (VEG) variables and variables combining the two (LUVEG).

These three species have been chosen in consultation with Menno Reemer (EIS), because he had a good understanding of the biology and the rough distribution of the species. Both the personal conversation with Menno and the book "De Nederlandse Bijen" (Peeters et al. 2012) have provided information about the ecology of the species. Wild bees will be discussed per species.[6]

Andrena vaga (NL: 'grijze zandbij')

This species is believed to occur in sandy areas and open till half-open habitats. It likes open till half-open habitats, parcel edges, moors and dikes. The species has only rarely been spotted across the coast line though. The maps show that this species is mostly found in the Veluwe area and (mostly according to the vegetation structure map) in the South-East region of The Netherlands. As visible from figure 24 the model types are not differing lot in their prediction. These areas are indeed known to be dominated by sandy soils. Corresponding with the biology, habitat suitability is not increased across the coast line. Also, parcel boundaries are more often segregated by vegetation than in the West of The Netherlands (there agricultural parcels are mostly separated from each other with ditches).

Lasioglossum morio (NL: 'langkopsmaragdgroefbij')

This species is believed to be very common in The Netherlands. It is often found in villages, cities and warm areas. It likes to nestle in decaying walls and in rock gardens. There it can dig tunnels sometimes longer than 20 cm. It can be found in various loamy or sandy soil types. From the maps it indeed seems that the species is very general and only few areas show to be less suitable (Utrechtse Heuvelrug and Veluwe). This prediction seems to make sense as well.

Dasypoda hirtipes (NL: 'pluimvoetbij')

This species likes to nestle in dry and sandy biotopes, like sandy paths, road edges or dikes. In The Netherlands it is known to occur at the coast line and the higher sandy soils. This is in line with the predictions, with the exception that the land use map does not show the higher suitability at the coast line, while the vegetation structure and combined map do.

Though the accuracy of the maps are not extremely high (AUC between 0.598 and 0.734), the spatial patterns seem to match the expected pattern to a certain extent according to the biology of the species. This indicates that the maps are far from random. It should be noted that only 5% of all habitat suitability maps are shown here, which is a small subset. No extrapolations should be made for the quality of all SDMs, but after inspection of these maps it is less likely to assume that (many) other maps show unexpected or random habitat suitability patterns.

[6] All pictures are copied from www.wildebijen.nl

4 Discussion

The discussion consists of three parts. The first three paragraphs will try to answer the research questions with the results found in the study. The fourth paragraph the flaws in the methodology will be discussed. The last two paragraphs elaborate on what the results mean in the context of this study on the overall scientific value.

4.1 Model Performance - AUC

A widely chosen model evaluation method is the AUC (Area Under the Receiver Operational Characteristic curve). Ackers et al. (2015) refer to a AUC of 0.717 a score "within acceptable limits, whereas a score of 0.809 is considered as a good predictive power. In this perspective, it seems that the average AUC values found for the three SDMs are not very high (AUC = 0.64, 0.66 and 0.69 for LU, VEG and LUVEG SDM respectively). It should be remembered however, that it is not the main goal of the study to provide the highest accurate maps. The main goal is to investigate how LiDAR derived landscape variables perform in comparison with land use landscape variables. Therefore, given the context of the study, low AUC values should not be considered unfortunate. To increase the prediction accuracy, one could implement three actions: *1) take more algorithms, 2) Include rarer wild bee species, and 3) use more environmental variables*.

1 - Include more algorithms

It is very likely that the choice to use an ensemble approach with only GLM algorithms has reduced the average AUC values as well. In a paper of Aguirre-Gutierrez et al. (2013) it is shown that other algorithms (e.g. Maximum Entropy) result in, on average, an increase of the AUC values. That Maximum Entropy (MaxEnt) generally predicts presences more accurate can be seen in figure 25. This figure shows that algorithms respond differently to the number of species presence records. It can be expected that implementing multiple algorithms and species with fewer records will result in higher AUC values. Another possibility is to implement an ensemble approach with multiple algorithms. Using AUC as evaluation metric it can be assumed that the consensus approach provides best results according to Aguirre-Gutierrez et al. (2013).

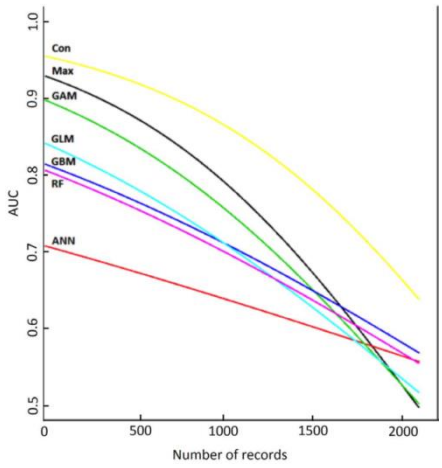


Figure 25: Figure showing the response of various algorithms to the increase of the number of species presence records. Figure edited from Aguirre-Gutierrez et al. (2013).

2 - Include rarer wild bee species

Figure 16 already shows that the AUC value tends to decrease with an increase of the number of records for a species of interest. Figure 25 (Aguirre-Gutierrez et al. 2013) underlines this. This observation can also be explained by the species' biology. It is likely that wild bee species with many records are more widespread than species that has only a few observations. The dispersal of a more widespread species will be less constricting by the environmental conditions than a specialized

species (e.g. species dependent on very specific nesting locations or food sources), since they have a higher likelihood to fulfil their resource requirements in a greater number of habitat patches (Tscheulin et al. 2011). Vice versa, this means that bee species with more records will be present in a wider range of landscape types, with higher variation in environmental conditions, than rare species. As a result, it can only be that the model prediction accuracy will tend to decrease for general or widespread species. After all, it is hard to define exactly which areas are typical for a species that does not have a clear landscape preference. The choice to only include species that have more than 100 observations at a unique location therefore automatically decreases the average AUC values.

3 - Include more environmental variables
Together with sandy soil and food availability, only vegetation structure and/or land use variables are used. Since it is assumed that many other factors are determining the presence of the species, other variables are often included in SDMs as well, like relief, climate data as temperature, precipitation or the occurrence of certain plant species (Peeters et al. 2012; Polce et al. 2013). Although the variables used in this studied are assumed to be important, other variables could significantly improve the prediction as well.

A technique that helps to evaluate how well the SDMs are performing is to create null-model (Raes and Ter Steege 2007). This approach has been applied before in another wild bee SDM study (Polce et al. 2013); it shows to which extent the SDMs are performing better than purely by chance. Though the SDM types can be compared within this study, this technique ensures that they actually are explaining some of the variation found as well.

Paired t-tests have shown that all models differ significantly from each other, where the LUVEG SDM seems to perform best, followed by the VEG and LU SDM. This is indicating that the vegetation structure variables are a better input for the SDMs than land use variables. A side-note to this is that less land use variables are used than vegetation structure variables. This means that it is possible that the vegetation structure variables have more ability to cover the variation in the landscape of the species, just by having more variables. However, this is not necessarily the case and increasing the number of variables also increases the chance of overfitting the data (though to a limited extent given the number of variables and wild bee observations), so it does not mean that more variables will result in better predictions per definition. Therefore, it is concluded here that SDMs based on vegetation structure variables provide comparable and, if not, better prediction results than land use variables.

Single Species prediction maps
Next to the AUC values, a visual inspection of the wild bee richness maps also confirms that in general predictions seem to be logical. In general, patterns seem to be non-random and to a certain extent follow the biology of the species. These results are only indicative. For a good validation of the maps one should visit random, yet unexplored field locations in order to investigate how habitat suitability values relate to the presence (or number of) species.

Linear Richness Models

Lastly, the linear models that correlate the found species richness (sum of binary values) with the original variables again show different results. The coefficients of determination now indicate that the combined model predicts the worse (r^2 = 0.429) than the vegetation structure model (r^2 = 0.483), even though the 'sandy soils' parameter was not included in the latter. This suggests that the variables derived indirectly from point cloud data is better able to (or at least comparable) predict the variation in species richness than the variables derived from land use.

4.2 Wild bees in The Netherlands

The constructed maps can be discussed at various ways. In this paragraph there will be an elaboration of the analysis of 1) the spatial patterns in the models, 2) the certainty of prediction.

4.2.1 Spatial patterns & Tenability

Zeeland and the Green Heart

All predictions of the three SDMs show that the species richness in Zeeland is low. This is in line with the expectation for both land use and vegetation structure variables. The area is characterized by intensive agriculture and visual analysis of aerial photographs indeed shows a monotonous landscape, dominated by agricultural parcels. Tree patches are very rare and vegetation rows between parcels are only at a limited number of parcels borders. These patterns are also translated into the LGN6 and the vegetation structure raster. For the LGN6, Zeeland is clearly dominated by agricultural fields of potatoes, grains, maize, beets and other crops. The vegetation structure shows mostly "no or very low vegetation" in this area, sometimes interrupted by rare long strokes of different kinds of higher vegetation. However, these strokes are rare. It can be assumed that this broad-scale cultivation system and the scarcity of natural field edges decrease the species richness at such landscapes (Benjamin et al. 2014). Another species-low area is the Green Heart of The Netherlands. This can very well be observed in the map of the VEG SDM prediction. Here it is shown clearly that species richness is low. This effect is less pronounced in the land use and combined model predictions. An important cause for this can be found by 1) the distribution of the species' observations and 2) the suitability classification of the LGN6 according to Vogiatzakis et al. (2014).

Many land use classes in Zeeland are classified as unsuitable. In this dataset this province is therefore one of the most unsuitable areas in The Netherlands. This can also explain the negative coefficient of the PUH landscape variable in the land use SDM, since only few species have been found there (which is consequently correlated with high PUH). However, though the species richness is expected to be low, there are many species records in Zeeland. This can be due to the execution of a former project of the EIS to enrich the data in this province, which resulted in high observation densities of wild bees for this area. Since the target group approach is used for the allocation of PAs, this sampling bias is causing the models to assign relatively many PAs here, while other areas can contain fewer PAs. This explains the prediction difference in the Green Heart region between the SDMs. For the land use SDMs, grassland is classified as suitable, and because of the negative slope coefficient for the PUH variable (which is included in the combined SDM and the land use SDM), the models are more prone to predict species to be present in the Green Heart. Since the Green Heart does not

contain many observations, PAs are only rarely allocated here, so even though the Green Heart might be species poor, the PUH variable makes it species richer. The vegetation structure SDM does not have this problem, since the PUH variable is not included here and both the Green Heart and Zeeland are dominated by high values of the same variable: “No or very low vegetation”. It can therefore be expected that predictions in the Green Heart will show similar results as predictions in Zeeland. The validity of the suitability classification of Dutch grasslands (according to Vogiatzakis et al. 2014) can also be argued, given that many grasslands are managed and therefore often homogeneous with few food resources and limited nesting locations for many species. Classifying this area as unsuitable might provide results that are more similar to the vegetation structure SDM.

Utrechtse Heuvelrug and the Veluwe

A big similarity between the models can be found in the region of the Utrechtse Heuvelrug and the Veluwe. It appears that in for all SDM types species richness is predicted to be high in these areas. The landscapes are here dominated by (different types of) forest, moors and other natural areas. Aerial imagery shows that, especially in the Veluwe, there are many open spots. Though the LGN6 shows mainly the presence of pine forest and deciduous forest, the vegetation structure raster shows a more heterogeneous pattern of vegetation structure types. The implying monotony of these landscapes in the LGN6 raster might clarify why the Veluwe and the Utrechtse Heuvelrug do not distinctly show higher bee richness than elsewhere in the land use map, while this is clearly the case for the other maps. Because of the diversity of the landscape, it can also be expected that it can encompass suitable conditions for many bee species.

Other areas

There are several other areas where wild bee richness appears to be high; in the mid-South of The Netherlands, Limburg and across the Eastern border of The Netherlands. All these are having sandy soils, which positively affects wild bee richness for the combined model (variable coefficient = 2.49, indicating that there are on average 2.5 wild bees more in sandy soils). In general, these are also the areas which are characterized by smaller scale agricultural compared to the West of The Netherlands. Further, aerial imagery and the vegetation structure raster (but not the LGN6) show that parcel borders are very often characterized by higher vegetation, thereby creating beneficial edges for wild bees in that landscapes. In the mid-East of the area (West-Brabant, Noord-Limburg), relative species richness decreases in general a bit compared to the mid-South (mid-Brabant), which is probably due to the increase of cultivation of crop types that grow in Zeeland. Except for the big difference in the Green Heart of The Netherlands, the three models are more or less in line with respect to the spatial patterns. In general it seems that the habitat suitability prediction of the land use SDM gives a noisier impression than the other two SDMs. This might be a sign that the vegetation structure and combined SDM are better able to capture the environmental variation than the land use SDM.

The South of Limburg is known for its high biodiversity for several reasons. There are relief differences with fast warming sandy soils, which provide suitable nesting locations for many bees. Furthermore, agriculture mostly consists of small-scale practices and field edges are often characterized by higher vegetation strokes suitable for wild bees (Peeters et al. 2012). However, this higher spot of wild bee richness is not visible on the maps, while one would expect this (pers. conv. with Menno Reemer and Peeters et al. 2012). The reason for this is that many species that are prevailing in South-

Limburg are mostly more rare and exclusively occurring in this area. Therefore, they have not reached the threshold of 100 observations. Including these species as well, might provide a map that is more representative for the total wild bee richness of The Netherlands.

4.2.2 Prediction Certainty

Here the 90% map will be compared with the stacked binary predictions map for the three SDMs. For all SDMs it is visible that the expected number of unique species per location is, obviously, less than the original map. However, there are more differences that can be pointed out. Comparisons will be made per SDM.

Land use SDM

There are several differences between the 90% map and the original species richness map. It appears that the predictions of single wild bee species in the Veluwe and Utrechtse Heuvelrug are more certain than in other areas. This is pointed out by the more pronounced high wild bee richness of these areas in the certainty map compared to the spatial patterns in the original species richness map. In contrast, the Green Heart of The Netherlands shows less species richness compared to the entire study area than the original map, indicating predictions were less certain here. This could be explained by the distribution of the wild bee observations again. Because the number of species observations is low in the Green Heart, the environmental conditions of the Green Heart are less well covered than the environmental conditions in Zeeland (because the high sampling density there). A similar effect is visible for the Randstad area. Here the species richness is predicted to be high and even comparable to the species richness predictions in the Utrechtse Heuvelrug and Veluwe in the certainty map. In the original map this is not the case. Here, in the Randstad mediocre species richness is predicted compared to e.g. areas as the Veluwe. This indicates that the certainty of the prediction is higher in the Randstad compared to other areas, which can be due to the high sampling density again.

Vegetation structure SDM

In general, the two maps of the VEG SDM are more similar to each other than the maps of the LU SDM. As expected from the observation in the land use SDM, the difference in the relative species richness prediction of the Randstad is clearly visible here as well. However, the difference between the Green Heart and Zeeland seems to be non-existent here. This can probably be due to the different variables used for the models. As explained earlier, for the LU SDM the Green Heart has a different landscape composition as Zeeland, while for the VEG SDM these areas appear to be similar. Therefore, it is not surprising that the difference in prediction certainty for the VEG SDM is not as pronounced as in the LU SDM, since environmental conditions are similar. Because of this similarity, the higher sampling density in Zeeland compared to the Green Heart can assert only little influence on the prediction certainty of the two areas.

Combined SDM

Since these maps are constructed from both vegetation structure and land use variables, it is not surprising that similar patterns as described earlier in this paragraph are coming forward. Again there is a difference in Zeeland and the Green Heart between the certainty and the original map, as well as for the Randstad. From all the maps it seems that the distribution of the wild bee observations plays a pivotal role in the certainty of the prediction of certain regions in the study area.

4.3 Variable importance

There are two types of results concerning the importance of the variables: 1) the average importance of a variable for the prediction for all single wild bee species and 2) the importance of the variable for the prediction of wild bee richness. It appears that these are not necessarily similar. The two types will both be discussed in this paragraph.

4.3.1 Single Species SDMs

From the boxplots, describing the distribution of the importance of the variables for 60 species, it appears that several variables are believed to be important. First of all, the food availability is the most important variable for the land use only and the vegetation structure only model, and it seems to be the third important variable for the combined model. That this variable appears to be important can be reasonable, since the presence of wild bees is believed to be dependent on 1) suitable nesting locations and 2) the availability of sufficient food resources (Westrich 1996). This variable is the only one describing where the availability of wild bee food is high or low, whereas the others can be considered to be more related to nesting locations. Section 4.3.2 will elaborate upon why food availability seems to be important for the distribution of single species and how this relates to the findings of the species richness.
In general, it seems that (for the vegetation structure and combined model) the landscape variables describing the % abundance of a vegetation class exert little influence on the model, while the edge density values and the mean patch area seem to be even less important. This observation is not the case for the land use model, because the ED_MN seems to be the second most important variable.

4.3.2 Species Richness

That the mean patch area and edge density appear to play a minor role for the species distribution should be placed into context. Section 4.3.1 seems to contradict literature, while this is not necessarily the case. It could very well be the case that the predictions of many bees are more dependent upon certain vegetation or land use types than landscape diversity variables (edge density and mean patch area). This way, it could still be that these variables are important, but less compared to the presence of certain beneficial vegetation structure types. However, for species richness it can be assumed that this effect will occur less. After all, a landscape that supports and/or predicts many wild bee species to be present will probably have a high diversity of vegetation / land use types, which subsequently increases the chance of suitable ecological niches for wild bees. Therefore, if the landscape is heterogeneous, the mean patch area and edge density would decrease and increase respectively. This theory could explain why the mean patch area and edge density are the most important variables for the prediction of wild bee richness. It is known that some wild bee species are very sensitive for the presence of suitable edges or the mean patch area and habitat fragmentation can significantly reduce the population size of species requiring a minimum suitable area (Tscharntke et al. 2012). However, it is also noted here that this does not have to be the case for every species. This paper also exemplifies that the relationships between biodiversity patterns and the landscape are often not straightforward. These results are also not in line with Aguirre-Gutierrez et al. (2015), where edge density and mean patch area seem not to have much influence on the

presence of the species (except for bumblebees, they were positively affected by the edge density between managed and natural areas).

There are other variables, though to a less extent, that seem to be important for the prediction of wild bee richness. Vegetation class 3 (small trees with some understory) and 5 (middle trees with little understory) are relatively important as well. A visual interpretation of the photographs shows why these results are reasonable. Vegetation class 3 seems to contain very mixed vegetation with only smaller trees and much understory. This landscape is varied in itself and with understory present it is arguable that food resources are available. This vegetation class is also, in contrast to vegetation class 5, positively correlated with species richness. That species richness is lower in landscapes that are more dominated by vegetation class 5 can be understood by the photographs (shown in the appendix: figure 35) as well. Almost no understory is present under the pine trees and this vegetation class seems to be very monotonous. It can be expected that this class offers limited food availability and only few different kinds of nesting locations, hence the ecological niche diversity is low.

It is also interesting that food availability seems to be, on average, very important for the prediction of a single wild bee species, while it only seems to play a marginal role for species richness. It is hard to explain what is causing this discrepancy. A hypothesis is suggested here. The food availability is in general very high in the Utrechtse Heuvelrug and - to a less extent - in the Veluwe, while it is very low in the West of The Netherlands and South-Limburg (figure 26). However, there are many records in all of these regions. Assuming a spatial correlation of observations, it can be deducted that some species are occurring more in areas of high food availability, while others are more present in areas of low food availability. In both cases, it will be an important explanatory variable for the presence prediction of the species. However, for the total richness of the species it might be less correlated, because species with both a negative and a positive slope coefficient are included in the model.

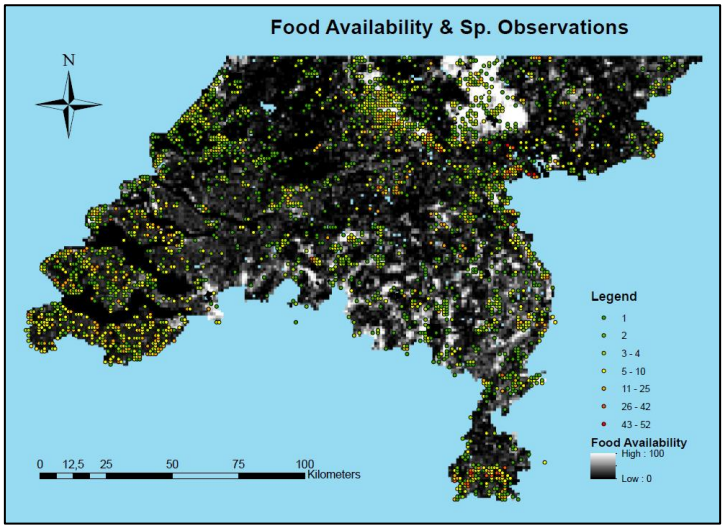


Figure 26: Food availability raster in relation with the wild bee observation locations.

4.4 Flaws

There are several analysis steps that have not been executed perfectly, which may have influenced the final findings. Here I discuss what aspects of the methodology could have been done better and how these aspects might have affected the final results per category.

Point density rasters

The height breakpoints of the point density rasters were changed later, but the values of the height weight correction were not changed correctly. The original height breakpoints (later these have changed by the addition of the second and third layer) in the lastools script were [0.1, 0.25, 0.5, 1, 5, 10, 20, 30, 80] and the corresponding weight correction should have been [6.67, 4, 2, 1/4, 1/5, 1/10, 50]. However, the following weight correction has been used: [6.67, 3.33, 2, 1, 1/3, 1/5, 1/10, 1/60]. This means that the second layer got a bit lower weight, the fourth layer 4 times too much, the fifth layer 5/3 times too much, the sixth layer 2 times too much and the highest a bit too little. In general, it can be said that the middle layers have received more weight than they should have had. Using the right correction values gives the lower layers more relative weight for the classification later. Following the assumption that wild bees are most dependent on the situation on ground level it can be assumed that applying the right weights will improve the models, since the lower layers will relatively gain more weight.

The highest layer (30-80m) has been used for the calculation of the total points in a voxel. This means that, in case there are trees higher than 30 m in an area, the total point density values of the layers will not add up to 100%. Therefore, it might be better to exclude the highest layer for the summation of the points in a voxel since it contains hardly information and point density values are easier to interpret. Implementing this will only have a marginal effect on the classification and further outcomes.

Pixels with less than 100 points between 0.15m and 10m were automatically classified as the vegetation structure class 'no or very low vegetation'. However, for some (probably accuracy) reason there was an area at the north edge of the study area that clearly should have been classified as grassland (based on aerial images), but here the points systematically exceeded the threshold. This has influenced the quality of the vegetation structure raster, but since the area is relative small (ca. 1 km^2) it can be assumed that the effect on the outcome of the quality parameters of the SDMs is negligible.

There were two tiles where the ground point classification failed (black areas in West of The Netherlands). This might be solved by using different tools from the lastools package. For time reasons this has not been done yet. Adding still missing data will mainly change the visual appearance of the map, rather than create big changes in final outcomes.

Species Distribution Modelling

An important analysis step has been forgotten before the species distribution models were started: no collinearity check between the landscape variables has been performed. This probably did not influence the species' presence-absence prediction, since collinear variables correct for each other in GLMs. However, without this check it becomes difficult to interpret the variable importance values, since it is hard to establish if a variable is not important because it has no relation with the bee species or because another variable is already accounting for the, related, variation in the landscape. For example, food availability seems to have some degree of correlation between vegetation class 2 (r^2 = 0.65), class 4 (r^2 = 0.59) and class 6 (r^2 = 0.51). Now though these three classes seem still to be relatively important for the outcome of the suitability values, they will probably be more important if food availability were excluded from the model.

Lastly, because of a mistake, the PUH variable should have been the inverse, which is '% suitability habitat' (PSH). Changing this variable will not change the final outcome, but rather enhance the interpretability and intuitiveness of its values and slope coefficient.

4.5 Further Research

Already in 2002 it has been hypothesized that LiDAR technology would become an important tool in ecology, because of its ability to quantify the 3D vegetation structure (Lefsky et al. 2002). The authors state that it is hard to predict how widely applied the technology is in ecology in 5 years. The findings are suggesting that point cloud data derived from LiDAR technology improves the prediction of wild bee richness for the Southern part of The Netherlands compared to a land use-only approach. However, measures could be taken to increase the model quality parameters (r^2, AUC). Suggestions will be made in the following sections. In any case, further research should include proper validation of 1) the vegetation structure and 2) the species richness map.

Validation
Going to the field is essential for a proper validation of the vegetation classes and to understand their true 3D structure. Here a possible validation procedure will be described. Random locations should be selected for a field visitation. The vegetation class should be unknown before the visit. Expert opinion (e.g. the classifier of the point densities) should decide what class it is. Every vegetation class should be visited multiple times (the more visits, the better). This way, a confusion matrix can be constructed that indicates the accuracy of the class estimation and the outcome of the maximum likelihood classification. It would be good if the field expert makes quantitative estimations of some aspects of the vegetation as well, like the estimated vegetation height or vegetation density close to the ground and at higher levels. This would also provide useful information about the character of the vegetation class for a more thorough description and understanding of the vegetation class. It is hypothesized that if this 'ground-truthing' is performed at least 10 times for a vegetation class, at randomized locations, one could have general insight in the in-between variation of the vegetation class as well.

Species presence predictions should be validated with independent sample data as well. Some species poor and rich km^2 cells should be selected where sampling has not been carried out before. Then different accuracy measures for the different species can be derived as well (from false positive, true positive, etc..). This will give a general overview of the prediction accuracy of the species richness map.

Wild bees & SDMs
At the start of the research bee species that had more than 100 records in the time span of 2003-2014 were selected. This way, one can be sure that the species distribution models (with no more than 13 variables) were not tending towards overpredicting bee suitability and presence-absence. However, assuming a correlation between the number of observations and the generality of a wild bee species, the sum of binary predictions of the selection of wild bee species (the species richness map) might not be a representative map for the total richness found. For example, the richness of

wild bees in The Netherlands is assumed to be the highest in South-Limburg, because many bee species are only observed and believed to occur there (Peeters et al. 2012) and personal conversation with Menno Reemer). However, this is not very clear in the species richness map, simply because most of the species were too rare to be included in the primary selection of bees. Together with this, one could say that the selection of more generic species is causing a 'smoothening' of the species richness map.

Another point of improvement is related to the selection of the km^2 that are used to construct the species models. In this study a PA could be allocated to a location where one or more other bees *of the 60 pre-selected bees* were found. However, it might be fairer to include km^2 cells where at least one other bee is found, *regardless of the selection*. For example, at a location 7 bee species are found, but none of those 7 bees were common enough to be included in the SDMs. However, these cells are very suitable for the allocation of PAs in the first phase of the modelling. Inclusion of these areas might broaden the range of landscape variation at which PAs could be allocated, thereby enabling the model to better distinguish the 'typical landscape' of the wild bee of interest. Although it is hard to state to which extent, it can be expected that this will translate into improved model prediction, hence higher AUC values.

The PA allocation method used in general and for this study in particular is fair (but not necessarily the best) given the nature of the wild bee dataset. Records come from random amateur inputs as well as from professional field inventories for certain locations. When more species are found at a certain location, it can be hypothesized that there is a higher possibility that this location was better searched for wild bee species than a location where only one or a few wild bee species are found. This assumption advocates making the chance of PA allocation dependent on the number of other bee species found. Though this assumption is not true per definition, it can be justified that implementing such a method would be fairer than continue with the present method, which assumes equal possibility of an absence for every location. For example, a place where only once *Bombus terrestris* is found has the same probability to get a PA as a location where 30 other bee species have been recorded. This way, many single-species observations (mostly of a common species) are mainly adding noise to the models. Some research has already been performed to different PA allocation methodologies (Senay et al. 2013), but as far of my knowledge a chance-based PA allocation method has not yet been investigated yet. Further research could be dedicated to the possible SDM improvements that can be made by such a method.
Irrespective of the method used, the presence records could have a bias towards certain geographic regions. Also in this study, the density of wild bee observations seems to be higher in ecologically attractive areas (with the province of Zeeland as exception). This should be taken into consideration as well in the interpretation of habitat suitability predictions of the species. After all, there is a chance that higher habitat suitability is predicted in ecologically attractive areas, partly because the visitation rate was higher there compared to the visitation rate in less attracting areas. More coordinated field visits to unattractive areas (like Zeeland) could reduce this problem, which might improve the predictive power of the SDMs.

Vegetation Structure Classification
The validation of the vegetation structure has already been discussed. Here other points for improvement of the vegetation structure raster will be discussed. First of all, the final vegetation class of a

25m*25m pixel does not provide information about vegetation density. This means that the data are able to provide information about the relative vegetation presence ratios in different heights, but not the difference between e.g. 2 and 12 trees. Two possible improvement methods are suggested here. The first one is straightforward: increase the spatial resolution of the data. It is suggested to use a pixel size of 5 meter. This should be fine enough to cutback the lack of vegetation density problem, since such an area can vary much less in vegetation quantity than the area that is 25 times as large (25m resolution). At the same time it is assumed to be coarse enough to have enough points (ca. 200-300) that can be used to create the different point density rasters.
The second method to overcome the vegetation density problem is more complicated, but could be worth investigating. Once the spatial resolution of the raster is increased, a degree of *vegetation openness* could be introduced. Rubene et al. (2015) have also found that wild bee and wasp diversity can be supported by the openness of the landscape, because more flowers can grow at open spots. Openness should be calculated over a 25*25m area again, based on the percentage of pixels that have a maximum vegetation height of e.g. 50 cm. Including this variable in the vegetation structure classification process might significantly improve the informativeness of the classes for wild bees.
For both methods it can be argued as well that information about the average NDVI (Normalized Difference Vegetation Index) over the spring and summer (most important time of insect activities would improve the classification as well). This way, it is easier to distinguish between e.g. grasslands and bare ground or urban zones and lower / mixed vegetation.

The point counts of the lowest layer (0.1 - 0.25 m) are a result of a combination of point height precision (possible noise) and the real presence of vegetation. These can even be assumed to be correlated: the presence of vegetation might cause higher imprecision, because it makes the determination of the true ground level more difficult. Still, it could be that some geographic areas are less precisely measured than others. Also, the different algorithms used for the split between ground and non-ground data by the different data suppliers of the AHN2 might cause spatial differences in the precision or accuracy of points in the lowest height layer. Therefore, it is suggested to shift the lowest layer up a bit. A new proposition for the height breakpoints (in meter) would be [0.15 - 0.4 - 1.0 - 5.0 - 10 - 20 - 30]. For a successful classification of vegetation structure of the landscape it is pivotal that the existing variation in vegetation structure is sufficiently covered in the defined classes. The importance of this can be illustrated by an example where only 2 classes are defined: high vegetation without understory and low vegetation. If the area of a - still unclassified - pixel would be characterized by high vegetation with understory, both classes seem to be unsuitable for this pixel. Therefore, more classes can be defined to prevent such 'misclassifications'. The height layer 20-30m is therefore valuable, since it helps to define high forests. It can be argued that high forests with canopies in the 20-30m layer have a significant different meaning for bees than high forests with the canopy in the 10-20m layer. However, having these two layers for the classification helps to prevent the earlier described 'misclassification'. Besides, these classes can be merged later (like in this study the original vegetation class 4 and 6). However, class difference should not be too small in order to prevent overcomplicating the (interpretation of) the vegetation structure classes. Expert opinion about relevant differences in vegetation structure for wild bees, in combination with a clear understanding of classification algorithms should provide more insight in what vegetation classes to define and, if needed, merged for a successful vegetation structure classification.

Landscape Indices

This research is combining horizontal and vertical vegetation metrics as presented in Simonson et al. (2014). The primary vegetation classes have been created based on the vertical structure of the vegetation. Horizontal vegetation structure metrics (referred to as the landscape variables) have subsequently been calculated (edge density, mean patch area) from these vertical vegetation structure classes, creating variables that tell something about the 3D structure of the green landscape. For the classification of the vegetation structure a voxel-based method has been chosen similar to the one described in the paper of Schut et al. (2014). However, other metrics can be chosen as well. Here the LiDAR data has been transformed to nominal data, similar to land use classes, before it was transformed to a quantitative landscape variable again. Perhaps other metrics are suitable as well. 'Variation in plant height' or the 'mean / maximum vegetation height, has been proposed as well (Simonson et al. 2014). With LiDAR derived point clouds perhaps new metrics can be developed. An example could perhaps be 'degree of edginess'. In this study, calculation of the edge density is a direct derivation of a binary (either suitable or not suitable) determination between adjacent pixels. However, with LiDAR derived metrics, this might also be approached quantitatively. Using the maximum or mean vegetation height of a 25*25m (or 5*5m) pixel, it can be assumed that certain height differences are describing a different degree of suitability. If suitability of single edges can be defined, an overall edge suitability score could be derived for an entire km^2. As far as I know this has not yet been suggested before.

Probably more vegetation metrics can be thought of, but the core message is that LiDAR point clouds may be suitable for the creation of such variables describing the landscapes. Creating a relevant and correct representation of the 3D architecture of the landscape is challenging. However, since many invertebrate data are often not available at high spatial resolution, making steps here might be pivotal to derive high quality environmental SDM predictors.

Overlays

The highways and the water areas used as an overlay of the vegetation structure raster have been derived from the TOP10NL. For a better comparison between the vegetation structure raster and the land use model one should just extract water bodies and roads from the LGN6. With this, other main roads would be included as well.

And beyond

The research motive was to integrate current knowledge in 1) LiDAR for ecology, 2) SDM for pollinators and 3) LiDAR for SDMs, here for the first applied at subnational scale. AUC values show that LiDAR data can improve SDMs. However, the effect size between the different SDM types could be stronger. Now only XYZ LiDAR data have been used, while other more advanced vegetation parameters, like leaf area index (Forzieri et al. 2011) or canopy profile (Zhao et al. 2015) could be derived as well with the use of the intensity values, number of return or RGB data. More sophisticated vegetation structure classification algorithms might be developed this way. Therefore, points with more attributes, acquired in summer time could significantly improve the characterization of vegetation structure. Perhaps the follow-up dataset (point cloud of the AHN3) can improve the quality of the classification.

With or without additional point data, it can be assumed that the vegetation structure can be used for other animal species as well. Future studies should be conducted to examine this hypothesis.

4.6 Scientific Relevance

This study for the first time provides evidence that high-resolution point clouds can be used for the prediction of richness of invertebrates at a large, subnational scale. In this report, a new methodology that creates variables which describe (indirectly) the 3D structure of the landscape is applied. This has been done by combining existing methodologies to characterize vertical vegetation structure first and subsequently horizontal vegetation structure (Simonson et al. 2014). The vegetation structure at a small scale (25*25m patches) was defined by the vertical distribution of the points. A similar methodology for vegetation structure classification was used by Schut et al. (2014). Because of the coarse resolution (1 km^2) of the bee dataset it was necessary to create landscape variables, which take the horizontal distribution of vertical based vegetation structure classes into account. This way, large scale XYZ data have for the first time quantified the landscape in such manner that it could be used for wild bee SDMs.

The results of the research indicate that the distribution of the wild bee records influences the species richness prediction pattern. This is strongly linked to the PA allocation method used. A new method of PA allocation is suggested in this study, which will let the allocation probability of a PA depend on the number of other wild bee species found at a certain location. It is hypothesized that such a method may improve multi-species SDMs based on presence-only datasets.

5 Conclusion

The aim of the study was to investigate if point cloud data can be used in species distribution models at subnational scale. The findings presented in this study support this hypothesis. Vegetation structure has been derived from basic XYZ point information from the AHN2 dataset. From this vegetation structure, landscape variables have subsequently been created and used as SDM input variables.

For the prediction of single wild bee species it is shown that the model performance of vegetation structure SDMs is systematically higher than land use SDMs. Combining land use and vegetation structure data provides even more accurate predictions. The habitat suitability predictions of three single species maps seem to be in line with the biology of the species, which supports the assumption that the SDMs provide maps with at least reasonable quality. Depicting which variables are most important is hard because of collinearity. However, for all SDMs it seems that the availability of food resources is playing an important role for the prediction of single species.

Wild bee richness is predicted more precisely by landscape variables derived from vegetation structure than from land use or combined landscape variables. For all SDMs the mean patch area and edge density are important landscape variables, which is in line with literature. That this is also true for the landscape variable derived from point clouds underlines the ecological relevance of LiDAR data.

In general, Zeeland and the 'Green Heart' area of The Netherlands are predicted to be species poor, while the Veluwe, Utrechtse Heuvelrug and the East of The Netherlands are predicted to be richer in wild bee species. One should be aware that the spatial distribution of the predicted wild bee richness depends partly on the combination of the spatial distribution of the wild bee observations and the configuration of the different landscape types. Prediction certainty also appears to be higher in regions with many wild bee records.

Simple methodological implementations might improve the results considerably. Nevertheless, even without these adjustments this study shows that point clouds acquired by airborne LiDAR can contribute significantly to (an improvement of) species distribution models. Further research should be dedicated to the validation and refinement of the vegetation structure classification and to the applicability of this vegetation structure for other (invertebrate) species.

Acknowledgements

For this thesis I had to integrate knowledge of multiple scientific disciplines. Luckily, I was privileged with three competent supervisors with different backgrounds who guided me through this process. They have supported me at different ways. I want to thank Harm for keeping me from getting trapped in the infinity of possible analyses. Suggestions like "try to keep in mind what you really want to proof…" or "perhaps you should make this a discussion point…" really helped me to focus on my research goals. I want to thank Koos for the opportunity he gave me at Naturalis. During my research he supported me by being innovative and by suggesting new ideas for analyses that have enriched my thesis. I want to thank Jesùs for him being my supervisor next door at which I could regularly ask questions about anything. Jesùs, though you were very busy with finishing-up your PhD, you always had time to help me with various issues, mostly about the SDMs.

There were other people who helped me during my thesis period which I would like to mention. Maarten van 't Zelfde (CML, Naturalis) has contributed a lot with the clipping and buffering of the BAG buildings. Menno Reemer (EIS) has supported me with the bee dataset and he helped me interpret the wild bee (richness) maps. Joost Michael (Ministry of Defence) helped me allocating NoData areas, which would have been a very time-consuming job otherwise. I would also like to thank my father who helped me increase the readability of this thesis report. Lastly, I thank the lastools forum community for explaining to me why my scripts were failing.

There is also something like moral support. The Cambridge dictionary describes this as follows: "If you give someone moral support, you encourage that person and show that you approve of what he is doing, rather than giving practical help". Anna, thank you for standing next to me during the whole course of this thesis.

References

Ackers, S.H., Davis, R.J., Olsen, K.A., & Dugger, K.M. (2015). The evolution of mapping habitat for northern spotted owls (Strix occidentalis caurina): A comparison of photo-interpreted, Landsat-based, and lidar-based habitat maps. *Remote Sensing of environment, 156*, 361-373

Aguirre-Gutierrez, J., Biesmeijer, J.C., Van Loon, E.E., Reemer, M., WallisDeVries, M., & Carvalheiro, L.G. (2015). Susceptibility of pollinators to ongoing landscape changes depends on landscape history. *Diversity and Distributions, 21*, 1129-1140

Aguirre-Gutierrez, J., Carvalheiro, L.G., Polce, C., van Loon, E.E., Raes, N., Reemer, M., & Biesmeijer, J.C. (2013). Fit-for-purpose: species distribution model performance depends on evaluation criteria - Dutch Hoverflies as a case study. *PLoS One, 8*, e63708

Baeza, A., & Estades, C.F. (2010). Effect of the landscape context on the density and persistence of a predator population in a protected area subject to environmental variability. *Biological Conservation, 143*, 94-101

Barbet-Massin, M., Jiguet, F., Albert, C.H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution, 3*, 327-338

Benjamin, F.E., Reilly, J.R., & Winfree, R. (2014). Pollinator body size mediates the scale at which land use drives crop pollination services. *Journal of Applied Ecology, 51*, 440-449

Bretagnolle, V., & Gaba, S. (2015). Weeds for bees? A review. *Agronomy for Sustainable Development, 35*, 891-909

Carvalheiro, L.G., Seymour, C.L., Nicolson, S.W., & Veldtman, R. (2012). Creating patches of native flowers facilitates crop pollination in large agricultural fields: Mango as a case study *Journal of Applied Ecology, 49*, 1373-1383

Chunyuan Diao, L.W. (2014). Development of an invasive species distribution model with fine-resolution remote sensing. *International Journal of Applied Earth Observation and Geo-information, 30*, 65-75

de Groot, G.A., van Kats, R., Reemer, M., van der Sterren, D., Biesmeijer, J.C., & Kleijn, D. (2015). Kwantificering van ecosysteemdiensten in Nederland. In W.U. Alterra (Ed.). Wageningen: Alterra, Wageningn UR, EIS Nederland, Naturalis Biodiversity Centre

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions, 17*, 1122–1131

Elith, J., & Leathwick, J.R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics, 40*, 677-697

Fagan, W.F., Cantrell, R.S., Cosner, C., & Ramakrishnan, S. (2009). Interspecific variation in critical patch size and gap-crossing ability as determinants of geographic range size distributions. *Am Nat, 173*, 363-375

Farrell, S.L., Collier, B.A., Skow, K.L., Long, A.M., Campomizzi, A.J., Morrison, M.L., Hays, K.B., & Wilkins, R.N. (2013). Using LiDAR-derived vegetation metrics for high-resolution, species distribution models for conservation planning. *Ecosphere, 4*, art42

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology, 43*, 393-404

Ficetola, G.F., Bonardi, A., Mücher, C.A., Gilissen, N.L.M., & Padoa-Schioppa, E. (2014). How many predictors in species distribution models at the landscape scale? Land use versus LiDAR-derived canopy height. *International Journal of Geographical Information Science, 28*, 1723-1739

Forzieri, G., Guarnieri, L., Vivoni, E.R., Castelli, F., & Preti, F. (2011). Spectral-ALS data fusion for different roughness parameterizations of forested floodplains. *River Research and Applications, 27*, 826–840

Gallai, N., Salles, J.-M., Settele, J., & Vaissière, B.E. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics, 68*, 810-821

Garibaldi, L.A., Steffan-Dewenter, I., Winfree, R., Aizen, M.A., Bommarco, R., Cunningham, S.A., Kremen, C., Carvalheiro, L.G., Harder, L.D., Afik, O., Bartomeus, I., Benjamin, F., Boreux, V., Cariveau, D., Chacoff, N.P., Dudenhoffer, J.H., Freitas, B.M., Ghazoul, J., Greenleaf, S., Hipolito, J., Holzschuh, A., Howlett, B., Isaacs, R., Javorek, S.K., Kennedy, C.M., Krewenka, K.M., Krishnan, S., Mandelik, Y., Mayfield, M.M., Motzke, I., Munyuli, T., Nault, B.A., Otieno, M., Petersen, J., Pisanty, G., Potts, S.G., Rader, R., Ricketts, T.H., Rundlof, M., Seymour, C.L., Schuepp, C., Szentgyorgyi, H., Taki, H., Tscharntke, T., Vergara, C.H., Viana, B.F., Wanger, T.C., Westphal, C., Williams, N., & Klein, A.M. (2013). Wild pollinators enhance fruit set of crops regardless of honey bee abundance. *Science, 339*, 1608-1611

Giannini, T.C., Chapman, D.S., Saraiva, A.M., Alves-dos-Santos, I., & Biesmeijer, J.C. (2013). Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. *Ecography, 36*, 649-656

Gilgert, W., & Vaughan, M. (2011). The value of pollinators and pollinator habitat to Rangelands: Connections among pollinators, insects, plant communities, fish, and wildlife. *Rangelands, 33*, 14-19

Hirzel, A.H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology, 45*, 1372–1381

Hopfenmuller, S., Steffan-Dewenter, I., & Holzschuh, A. (2014). Trait-specific responses of wild bee communities to landscape composition, configuration and local factors. *PLoS One, 9*, e104439

Jaboyedoff, M., Oppikofer, T., Abellan, A., Derron, M.-H., Loye, A., Metzger, R., & Pedrazzini, A. (2012). Use of LiDAR in landslide investigations: a review. *Natural Hazards, 61*, 5-28

Jaeschke, A., Bittner, T., Reineking, B., & Beierkuhnlein, C. (2013). Can they keep up with climate change? – Integrating specific dispersal abilities of protected Odonata in species distribution modelling. *Insect Conservation and Diversity, 6*, 93-103

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography, 21*, 498-507

Kleijn, D., Berendse, F., Smit, R., Gilissen, N.L.M., Smit, J., Brak, B., & Groeneveld, R. (2004). Ecological Effectiveness of Agri-Environment Schemes in Different Agricultural Landscapes in The Netherlands. *Conservation Biology, 18*, 775-786

Kleijn, D., Winfree, R., Bartomeus, I., Carvalheiro, L.G., Henry, M., Isaacs, R., Klein, A.M., Kremen, C., M'Gonigle, L.K., Rader, R., Ricketts, T.H., Williams, N.M., Lee Adamson, N., Ascher, J.S., Baldi, A., Batary, P., Benjamin, F., Biesmeijer, J.C., Blitzer, E.J., Bommarco, R., Brand, M.R., Bretagnolle, V., Button, L., Cariveau, D.P., Chifflet, R., Colville, J.F., Danforth, B.N., Elle, E., Garratt, M.P., Herzog, F., Holzschuh, A., Howlett, B.G., Jauker, F., Jha, S., Knop, E., Krewenka, K.M., Le Feon, V., Mandelik, Y., May, E.A., Park, M.G., Pisanty, G., Reemer, M., Riedinger, V., Rollin, O., Rundlof, M., Sardinas, H.S., Scheper, J., Sciligo, A.R., Smith, H.G., Steffan-Dewenter, I., Thorp, R., Tscharntke, T., Verhulst, J., Viana, B.F., Vaissiere, B.E., Veldtman, R., Westphal, C., & Potts, S.G. (2015). Delivery of crop pollination services is an insufficient argument for wild pollinator conservation. *Nat Commun, 6*, 7414

Kremen, C., Williams, N.M., & Thorp, R.W. (2002). Crop pollination from native bees at risk from agricultural intensification. *Proc Natl Acad Sci U S A, 99*, 16812-16816

Kumar, S., Simonson, S.E., & Stohlgren, T.J. (2009). Effects of spatial heterogeneity on butterfly species richness in Rocky Mountain National Park, CO, USA. *Biodiversity and Conservation, 18*, 739-763

Lefsky, M.A., Cohen, W.B., Parker, G.G., & Harding, D.J. (2002). Lidar remote sensing for ecosystem studies. *BioScience, 52*, 19-30

Lundy, M.G., Buckley, D.J., Boston, E.S.M., Scott, D.D., Prodöhl, P.A., Marnell, F., Teeling, E.C., & Montgomery, W.I. (2012). Behavioural context of multi-scale species distribution models assessed by radio-tracking. *Basic and Applied Ecology, 13*, 188-195

Lye, G., Park, K., Osborne, J., Holland, J., & Goulson, D. (2009). Assessing the value of Rural Stewardship schemes for providing foraging resources and nesting habitat for bumblebee queens (Hymenoptera: Apidae). *Biological Conservation, 142*, 2023-2032

Marshall, S.D., Walker, S.E., & Rypstra, A.L. (2006). Two ecologically-divergent generalist predators have different responses to landscape fragmentation. *OIKOS, 114*, 241-248

Müller, J., & Brandl, R. (2009). Assessing biodiversity by remote sensing in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. *Journal of Applied Ecology, 46*, 897-905

Park, M.G., Blitzer, E.J., Gibbs, J., Losey, J.E., & Danforth, B.N. (2015). Negative effects of pesticides on wild bee communities can be buffered by landscape context. *Proc. R. Soc. B, 282*, 20150299

Peeters, T.M.J., Nieuwenhuizen, H., Smit, J., van der Meer, F., Raemakers, I.P., Heitmans, W.R.B., van Achterberg, K., Kwak, M., Loonstra, A.J., de Rond, J., Roos, M., & Reemer, M. (2012). *De Nederlandse Bijen*. Nederland, Leiden: Naturalis Biodiversity Center & European Invertebrate Survey

Polce, C., Termansen, M., Aguirre-Gutierrez, J., Boatman, N.D., Budge, G.E., Crowe, A., Garratt, M.P., Pietravalle, S., Potts, S.G., Ramirez, J.A., Somerwill, K.E., & Biesmeijer, J.C. (2013). Species distribution models for crop pollination: a modelling framework applied to Great Britain. *PLoS One, 8*, e76308

Potts, S.G., Biesmeijer, J.C., Bommarco, R., Felicioli, A., Fischer, M., Jokinen, P., Kleijn, D., Klein, A.-M., Kunin, W.E., Neumann, P., Penev, L.D., Petanidou, T., Rasmont, P., Roberts, S.P.M., Smith, H.G., Sørensen, P.B., Steffan-Dewenter, I., Vaissière, B.E., Vilà, M., Vujić, A., Woyciechowski, M., Zobel, M., Settele, J., & Schweiger, O. (2011). Developing European conservation and mitigation tools for pollination services: approaches of the STEP (Status and Trends of European Pollinators) project. *Journal of Apicultural Research, 50*, 152-164

Puschendorf, R., Hodgson, L., Alford, R.A., Skerratt, L.F., & VanDerWal, J. (2013). Underestimated ranges and overlooked refuges from amphibian chytridiomycosis. *Diversity and Distributions, 19*, 1313-1321

Raes, N., & Ter Steege, H. (2007). A null-model for significance testing of presence-only species distribution models. *Ecography, 30*, 727-736

Rubene, D., Schroeder, M., & Ranius, T. (2015). Diversity patterns of wild bees and wasps in managed boreal forests: Effects of spatial structure, local habitat and surrounding landscape. *Biological Conservation, 184*, 201-208

Senay, S.D., Worner S.P., & Ikeda T. (2013. Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PloS one*, 8.8

Schut, A.G., Wardell-Johnson, G.W., Yates, C.J., Keppel, G., Baran, I., Franklin, S.E., Hopper, S.D., Van Niel, K.P., Mucina, L., & Byrne, M. (2014). Rapid characterisation of vegetation structure to predict refugia and climate change impacts across a global biodiversity hotspot. *PLoS One, 9*, e82778

Simonson, W.D., Allen, H.D., Coomes, D.A., & Tatem, A. (2014). Applications of airborne lidar for the assessment of animal species diversity. *Methods in Ecology and Evolution, 5*, 719-729

Tscharntke, T., Tylianakis, J.M., Rand, T.A., Didham, R.K., Fahrig, L., Batary, P., Bengtsson, J., Clough, Y., Crist, T.O., Dormann, C.F., Ewers, R.M., Frund, J., Holt, R.D., Holzschuh, A., Klein, A.M., Kleijn, D., Kremen, C., Landis, D.A., Laurance, W., Lindenmayer, D., Scherber, C., Sodhi, N., Steffan-Dewenter, I., Thies, C., van der Putten, W.H., & Westphal, C. (2012). Landscape moderation of biodiversity patterns and processes - eight hypotheses. *Biol Rev Camb Philos Soc, 87*, 661-685

Tscheulin, T., Neokosmidis, L., Petanidou, T., & Settele, J. (2011). Influence of landscape context on the abundance and diversity of bees in Mediterranean olive groves. *Bull Entomol Res, 101*, 557-564

Van der Zon, N. (2013). Kwaliteitsdocument AHN versie 1.3. In N.v.d. Zon (Ed.)

van Engelsdorp, D., Evans, J.D., Saegerman, D., C., M., Haubruge, E., Kim Nguyen, B., Frazier, M., Frazier, J., Cox-Foster, D., Chen, Y., Underwoord, R., Tarpy, D.R., & Pettis, J.S. (2009). Colony collapse disorder: a descriptive study. *PLoS One, 4*, doi:10.1371/journal.pone.0006481

Vierling, K.T., Bässler, C., Brandl, R., Vierling, L.A., WeiB, I., & Müller, J. (2011). Spinning a laser web: predicting spider distributions using LiDAR. *Ecological Applications, 21*, 577-588

Vierling, K.T., Vierling, L.A., Gould, W.A., Martinuzzi, S., & Clawges, R.M. (2008). Lidar: shedding new light on habitat characterization and modeling. *Frontiers in Ecology, 6*, 90-98

Vogiatzakis, I.N., Stirpe, M.T., Rickebusch, S., Metzger, M.J., Xu, G., A., R.M.D., Bommarco, R., & Potts, S.G. (2014). Rapid assessment of historic, future and current habitat quality for biodiversity around UK Natura 2000 sites. *Environmental Conservation, 42*, 31-40

Wagner, H.H., & Fortin, M.-J. (2012). A conceptual framework for the spatial analysis of landscape genetic data. *Conservation Genetics, 14*, 253-261

Westrich, P. (1996). *Habitat requirements of central European bees and the problems of partial habitats*. The Linnean Society of London and The International Bee Research Association

Winfree, R. (2010). The conservation and restoration of wild bees. *Ann N Y Acad Sci, 1195*, 169-197

Winfree, R., Williams, N.M., Dushoff, J., & Kremen, C. (2007). Native bees provide insurance against ongoing honey bee losses. *Ecol Lett, 10*, 1105-1113

Zhao, K., García, M., Liu, S., Guo, Q., Chen, G., Zhang, X., Zhou, Y., & Meng, X. (2015). Terrestrial lidar remote sensing of forests: Maximum likelihood estimates of canopy profile, leaf area index, and leaf angle distribution. *Agricultural and Forest Meteorology, 209-210*, 100–113

Zulka, K.P., Abensperg-Traun, M., Milasowszky, N., Bieringer, G., Gereben-Krenn, B.-A., Holzinger, W., Hölzler, G., Rabitsch, W., Reischütz, A., Querner, P., Sauberer, N., Schmitzberger, I., Willner, W., Wrbka, T., & Zechmeister, H. (2014). Species richness in dry grassland patches of eastern Austria: A multi-taxon study on the role of local, landscape and habitat quality variables. *Agriculture, Ecosystems & Environment, 182*, 25-36

Appendix: Photos Vegetation Structure

In order to get more insight in the true 3D configuration of the different vegetation structure classes, two field trips have been organized. The first field trip was North of Rozendaal in the National park 'the Veluwezoom'. The second field trip was located East of Wageningen. The locations where the photos have been made for both validation routes are indicated in figure 27. As a preparation, points have been selected which were located inside a patch (consisting of at least four agglomerated pixels) of the same vegetation class. These points of interest (POIs) were loaded into a handheld GPS. Geotagged photos were made at the locations of the points. The fact that the patch was bigger than one pixel ensures that GPS imprecision would cause no confusion about the identity of the vegetation structure class could arise. For this reason, no random sampling technique has been chosen. Later, the photos were analysed to come to a qualitative description of every vegetation class. It should be noted that the photos made have not captured every possible variation inside the vegetation structure classes and should therefore not be used as absolute reference. Also, the photos could be shot several years after the point cloud acquisition, which might cause differences between the expected and observed vegetation. However, it is assumed that these photos give a proper indication about the general structure of the vegetation classes.

In this section I will try to identify the characteristics of the vegetation structure type by photo interpretations. For some photos a fish eye lens was used in order to catch the understory and the canopy as well in one photo. Sometimes multiple photos were stitched together into one photo using ICE (Image Composite Editor)[7]. The name of the vegetation class mentioned are the ones given before the field visitation and described in paragraph 2.4. However, one should always be aware that the names are indicative.

[7] http://research.microsoft.com/en-us/um/redmond/projects/ice/

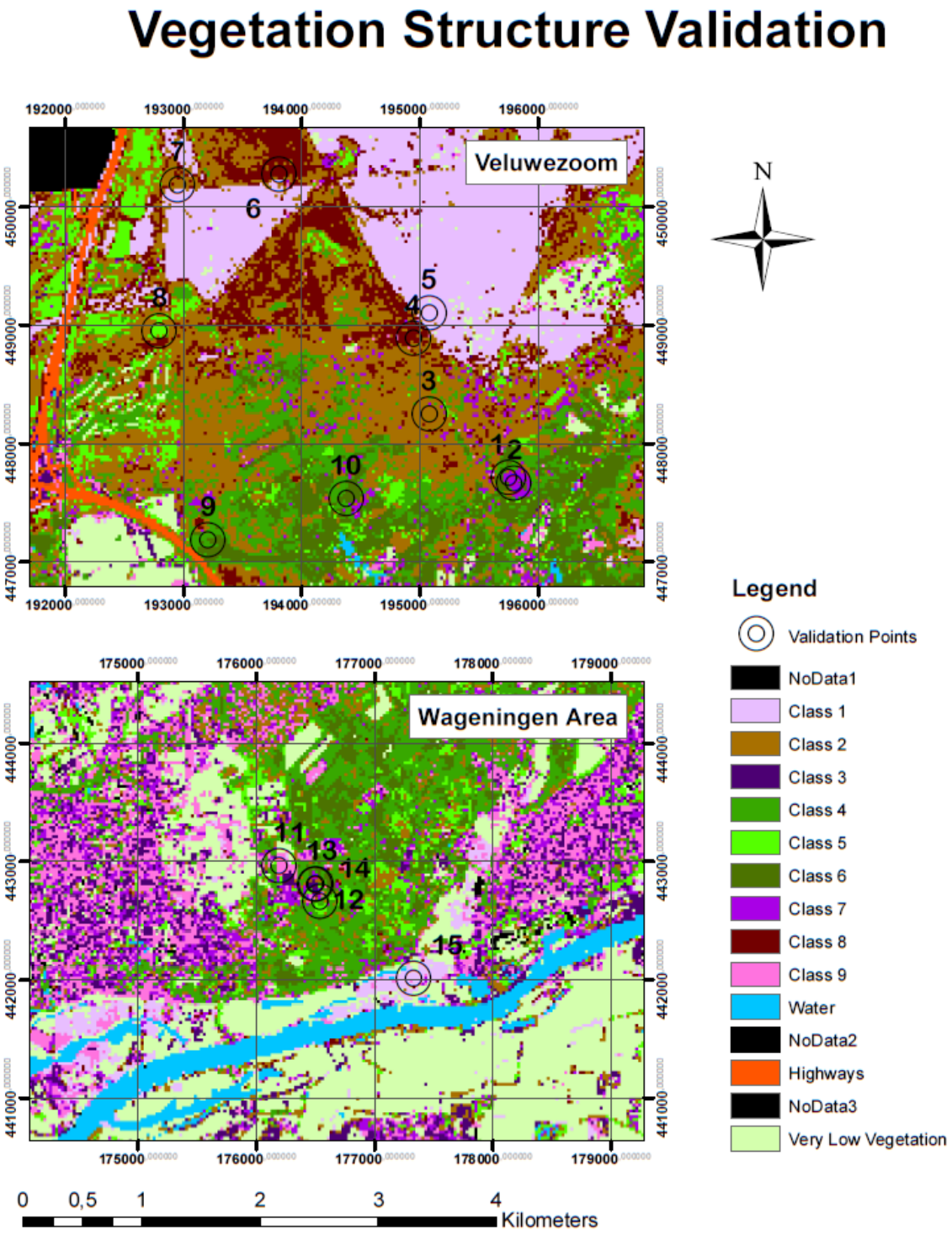


Figure 27: Validation areas: Veluwezoom (upper) and Wageningen area (lower). Point labels are the names of the photos. Grid lines are according to 'rijksdriehoeksstelsel'

Class 1 - '*Low vegetation*'

This vegetation type is characterized by the presence of only very low vegetation. On figure 28 and 29 it is shown that this can be grass landscapes or moors.

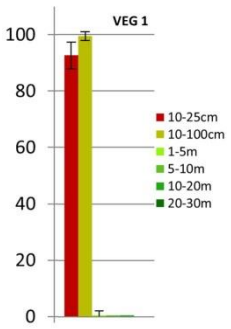
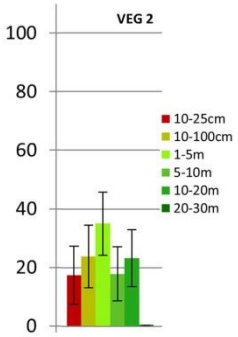


Figure 28: Point W15

Figure 29: Point VZ5

Class 2 - '*Mixed vegetation, high and low*'

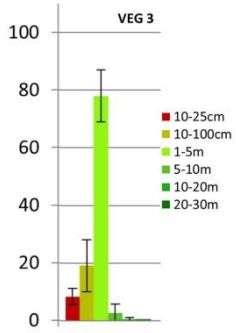


This vegetation class is showing much variation in vegetation structure. Several photos are made of this class and all show much variation in vegetation heights. All locations though have to a certain extent vegetation in the lower layers. At figure 30 there are many blue berries and other low bushes, while figure 31 is showing much grass.

Figure 30: Point V3

Figure 31: Point V7

Class 3 - '*Small trees with some understory*'



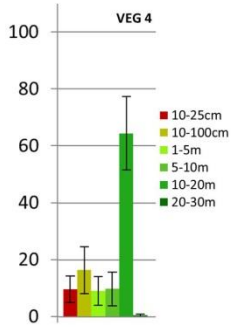
This class is rare and patches of this class are in general small. This class is also unique in its appearance compared to the other classes. The vegetation height seems to, in general, not exceed the 6 meter, and at this particular spot there is a lot of bramble (figures 32 and 33). This would result in relative high point density values in some lower layers. According to the vegetation profile, this would be the 1-5 meter layer, which seems to fit. At this particular spot the vegetation had a very open structure and many insects and flowers could be found as well. It should be noted though that the openness at this spot cannot be extrapolated to all areas with this vegetation class, since the method used is openness insensitive.

Figure 32: Point W13

Figure 33: Point W12

Class 4 - '*High trees with little understory*'

Figure 34 shows high pine trees with mainly leaves in the top of the high canopy, which explains the relative high contribution of the 10-20 point density layer compared to the other layers. At other locations this class gave similar photographs.



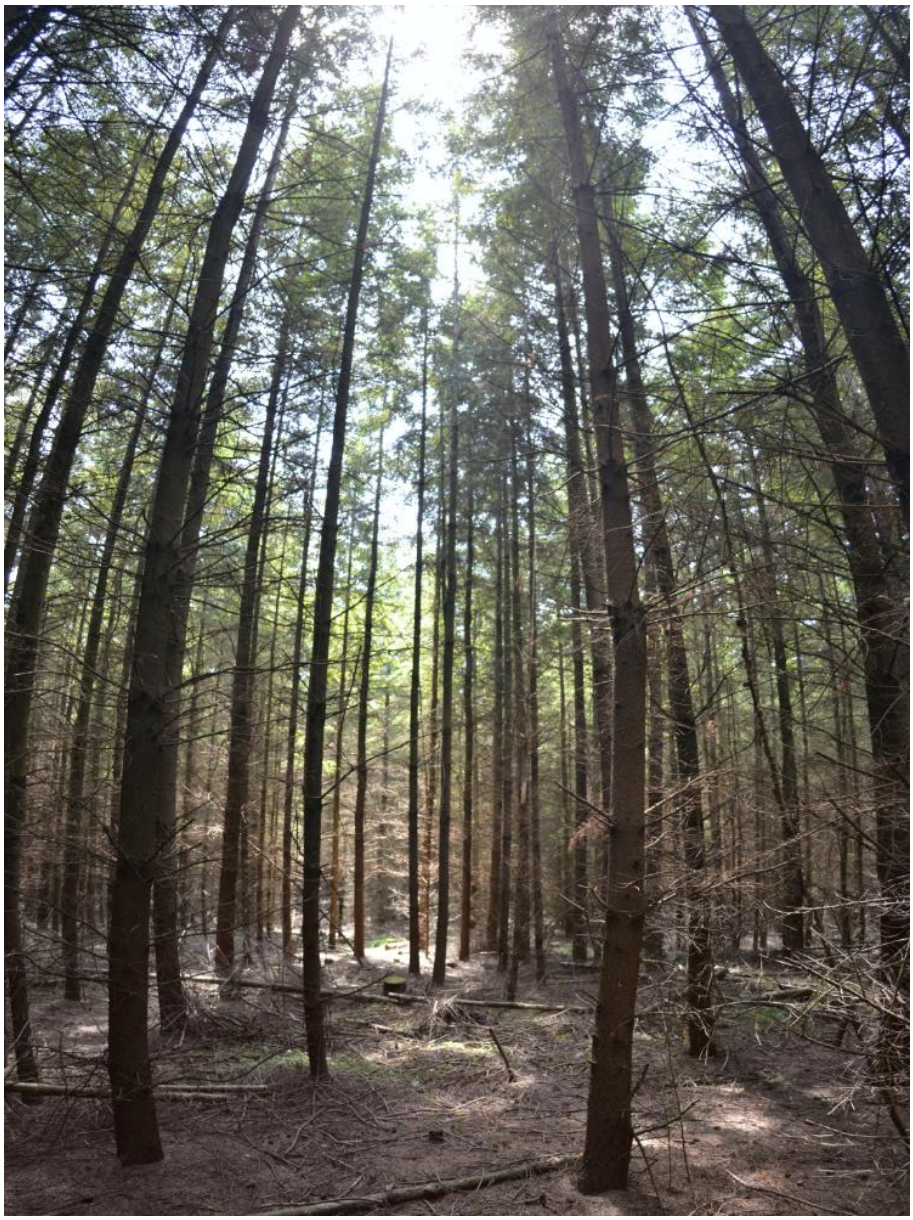


Figure 34: Point W14

Class 5 - '*Middle trees with little understory*'

All photos of this class were consisting of small pine trees (generally not much higher than 10 m). Understory was almost not present at this location (figure 35).

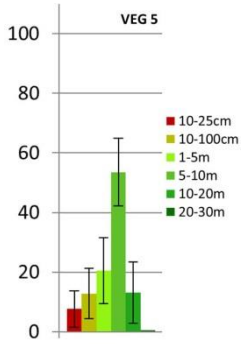


Figure 35: Point V8

Class 600 - '*No or very low vegetation*'

The character of this class can easily be defined by looking at aerial pictures. Figure 36 shows that grasslands and agricultural soils are in general defines as 'no or very low vegetation'. If no object or vegetation is present, a pixel will have become this vegetation class.

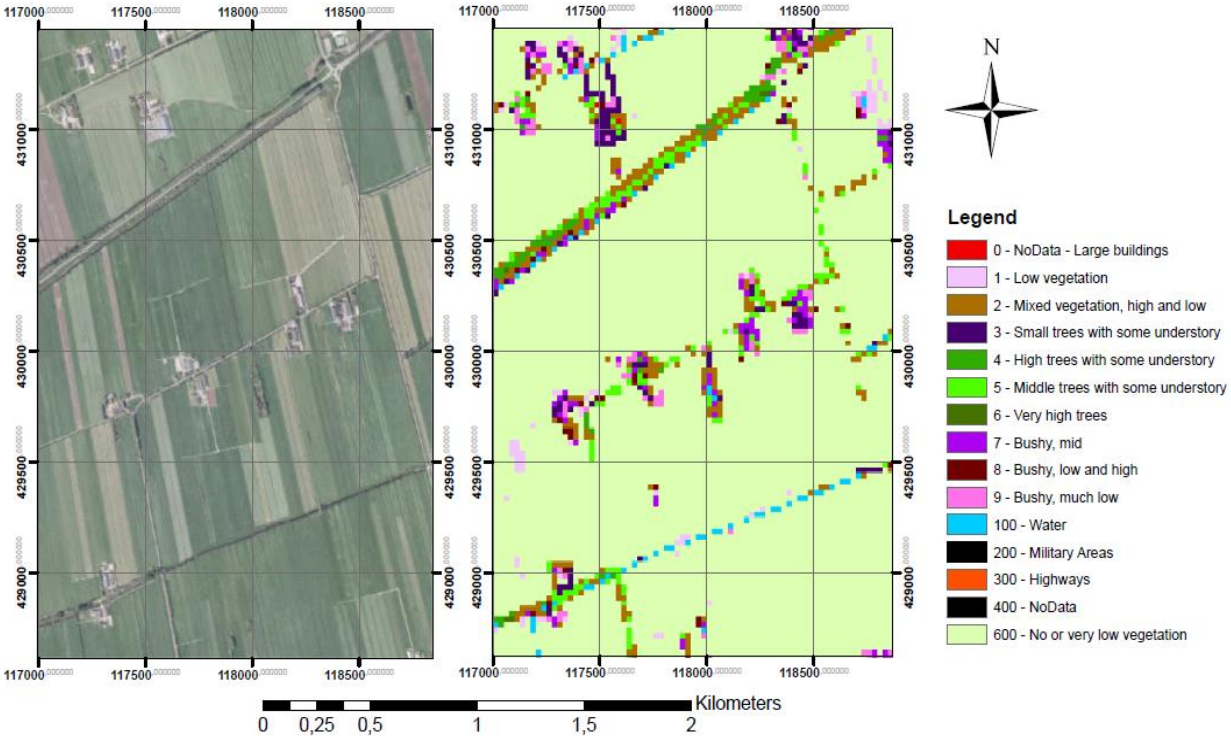


Figure 36: Two maps showing the same area. The left map origins from the ArcGIS base map. The right map is the vegetation structure raster. Grid lines are according to the 'rijksdriehoeksstelsel'. The area is west of Giessenburg, The Netherlands

Class 6 - '*Very high trees*' (merged with class 4 later)

This class has high trees which contains sometimes some smaller vegetation. It seems to be coniferous in all cases. Low vegetation was generally barely present (figure 37), but in figure 38 low ferns are present.

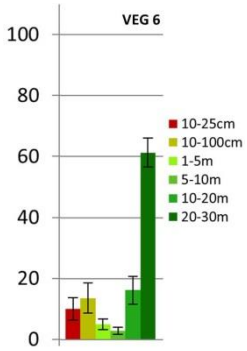


Figure 37: Point V9

Figure 38: Point V10

Class 7 - '*Bushy, mid*'

High trees were present in this class, but the lower vegetation / understory was much more dominant. Seedlings, low trees and a lot of ferns sometimes were present as well (figures 39 and 40). This is also in line with the vegetation profile of this class.

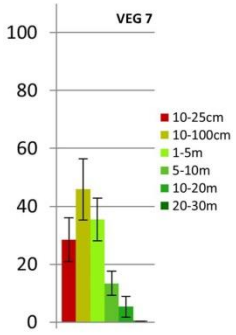


Figure 39: Point V1

Figure 40: Point V2

Class 8 - '*Bushy, low and high*'

This class had always much understory (figures 41 and 42). At the locations of the photos there were always blue berries bushes present which probably explain the high point density in the first, and perhaps the second, layer. Furthermore, there seemed to be more leafs/vegetation in the height layer between 1 and 8m, compared to class 4 or 5. A better name of this class would probably be 'trees with much understory'.

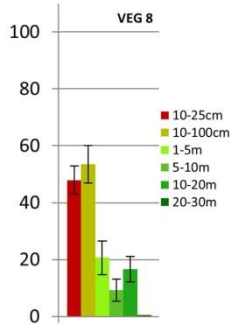


Figure 41: Point V6

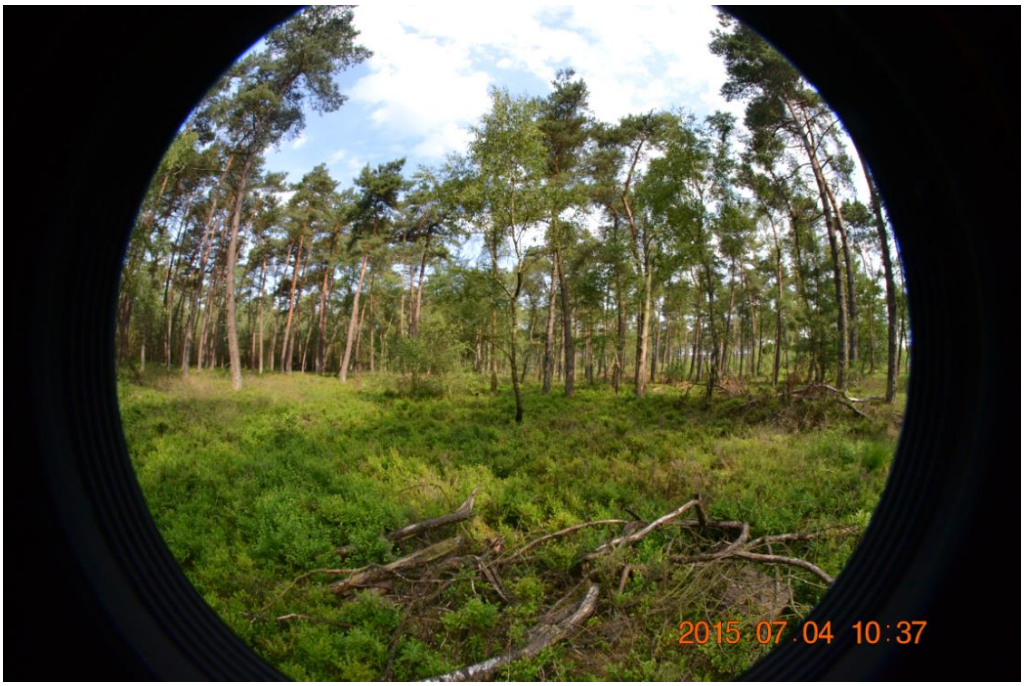


Figure 42: Point V4

Class 9 - '*Bushy, low*' (merged with class 7 later)

Only the photo at figure 43 was made of this class. It appeared to be a vine yard. Unfortunately no other picture of this class is available. It can probably be said though that this class will mostly contain low vegetation, though in general higher than class 1.

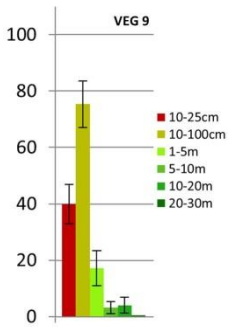


Figure 43: Point W11

For a summary of the vegetation class validation, one should read paragraph 2.4.