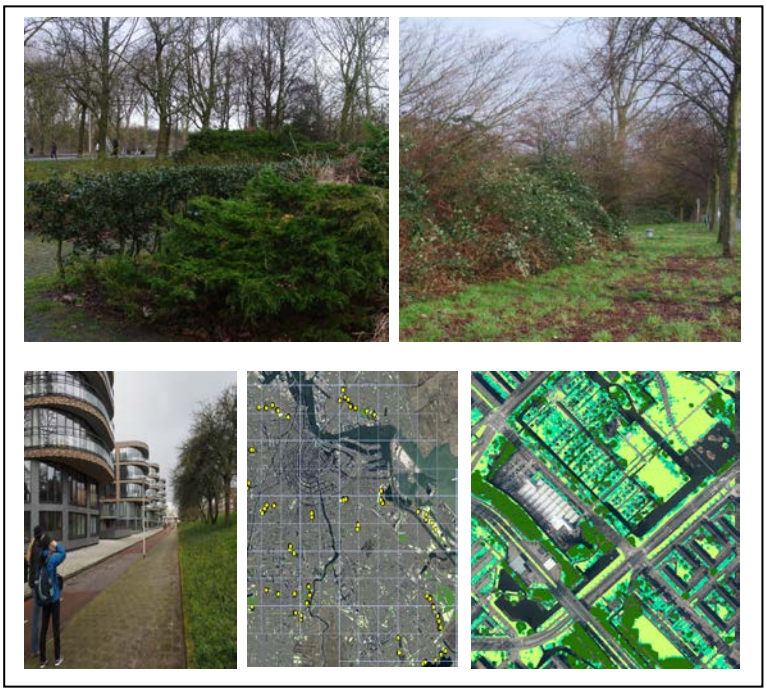
Centre for Geo-Information

Thesis Report GIRS-2015-10

VEGETATION STRUCTURE CLASSES AND THEIR RELATIONSHIP WITH TICK OCCURENCES IN AMSTERDAM

Xiaxia Gao



April 2015

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Vegetation structure classes and their relationship with tick occurrences in Amsterdam

Xiaxia Gao

Registration number 89 09 19 250 130

Supervisor:

Dr Ir Sytze de Bruin

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

April 17, 2015
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2015-10
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Abstract

Tick-borne diseases and tick bites are an increasing concern in the Netherlands. Among all the environmental factors, vegetation plays an important role in tick life circle by providing shelter, water, a place for seeking hosts, etc. Knowledge about relationships between vegetation structure patterns and tick occurrence patterns can help mapping potential tick distribution and contribute to improve local risk management of ticks by the public health service in Amsterdam. It is also required to have vegetation structure map for monitoring and risk management of ticks in Amsterdam.

In this study, firstly grass, shrubs, and trees three vegetation structure categories were defined. Height and height spreads were used as criteria for classification of point cloud data created by Semi Global Matching of aerial photographs. The topographic dataset KBKA10 provided locations of vegetation for filtering the point cloud data. AHN2 terrain grid offered terrain height values for extracting vegetation area objects height. The aerial image of Amsterdam was used for providing ground information and selecting training locations of the three classes.

Then supervised classification with *k*-Nearest Neighbour and Classification And Regression Tree (CART) analysis were applied for mapping vegetation structure classes. Accuracy was assessed by computing confusion matrices. Thirdly, logistic regression was conducted with tick data, vegetation structure classes from the CART result maps, distance to water area and distance to built-up area to assess relationships between these factors and tick occurrence patterns in Amsterdam. The model was finally used for spatial prediction of tick occurrence

The results show CART produced most accurate vegetation structure maps. The overall classification accuracy of CART result was 86%. The grass class had a high accuracy 97% and high reliability 93%. The tree class had a high reliability 99%. The overall classification accuracy of the *k*-Nearest Neighbour result was 77%. The grass class had 99% accuracy. The tree class had a reliability approaching 100%. Distance to water and distance to built-up area were found not to be significantly related to tick absence and presence in Amsterdam. The significant predictor vegetation structure classes influence tick occurrence in Amsterdam to a small extent. I found a weak association between vegetation structure classes and tick presence and absence in Amsterdam. The predicted tick presences appear at sites where shrubs or trees grow.

Keywords: Vegetation structure map, Supervised classification, *k*-Nearest Neighbour, Classification And Regression Tree, Tick occurrence, Logistic regression, Amsterdam

Acknowledgements

It is interesting to map vegetation in Amsterdam urban area with remote sensing datasets, supervised classification method and CART method. In this study I also had new experience to explore vegetation and its relationship with pests distribution and new data point clouds from Semi Global Matching method. Additionally working with open source software R brought lots of fun and difficult tasks. In the period I have understand more about 'dare to try'. I have gained much experience and scientific research skills through this major thesis study.

First of all I want to thank my supervisor Sytze de Bruin for guiding my research. He was always open and nice for discussions and gave me lots of guidance and encouragement when I came across difficulties in methods or scripting.

I would like to thank experts Jan Buijs and Marjolein van Adrichem. They raised up this thesis topic and offered different datasets. They offered nice suggestions for this study and helped me with understanding terms of ticks and vegetation. We conducted a small field visit in Amsterdam urban area. Thank Daniela Cianci offered suggestions for my proposal. Thank Jeroen Muller and Ries Visser for communicating about SGM point clouds.

I extend my sincere gratitude to Gerbert Roerink, Harm Bartholomeus, and Frans Rip for providing suggestions and help for satellite images, LAStools software and ideas for processing las data. I would like to thank our teacher Jan Verbesselt, and Arnold Bregt for giving comments for thesis proposal. Thank Ron van Lammeren, and Brice Mora for comments in midterm presentation. Thank Will Ten Haff for making suggestions for schedule. Thank Arun Pratihast for answering questions about point clouds.

I also would like to thank Latifah, Puja, Fahima, Kristin and other MGI students who gave me help and joined positive discussions during thesis life. Thank my graduated Chinese friends and corridor mates who encouraged me to conduct this thesis.

Finally I want to thank my dear family and dear friends Tianran and Keyang who supported me during this period in many ways.

Xiaxia Gao

At Wageningen, April 17, 2015

Table of Contents

List of Equations

List of Tables

List of Figures

Abbreviations

AHN2	Actueel Hoogtebestand Nederland/ Up-to-date Height Model of The Netherlands
CART	Classification And Regression Tree
GGD	The Public Health Service Bureau
GLM	Generalized Linear Models
IPM	Integrated Pest Management
KBKA10	Small Scale Base Map 1: 10,000 of Amsterdam
kNN	*k*-Nearest Neighbour
LiDAR	Light Detection and Ranging
m	Meter
ML	Maximum Likelihood
NA	no data
NDVI	Normalized Difference Vegetation Index
SD	Standard Deviation
SGM	Semi-Global Matching

1. Introduction

1.1 Background

1.1.1 Vector-borne diseases and vegetation structure

Vector-borne diseases such as Malaria, Lyme, Chikungunya, Dengue are a specific group of infections that present a re-emerging threat to Europe and need particular attention (ECDC 2012). These diseases that are carried by vectors like ticks and transmitted to humans are an increasing concern to public health bodies in many European countries (Kruijff et al. 2011). Tick, specifically *Ixodes ricinus*, is the main vector of Lyme borreliosis and tick-borne encephalitis in Europe (Swart et al. 2014; Gray et al. 2009). In the Netherlands, the Lyme disease has developed into an important disease in the past ten years (Hofhuis et al. 2006). The number of consultations of general practitioners for tick bites and Lyme disease has increased by three times since 1994 (Hofhuis et al. 2006; Gassner et al. 2011; Swart et al. 2014). This rise can be explained by an overall increase in abundance of questing infected ticks to some extent, and one factor contributing to this increase of questing infected ticks abundance is the expansion of tick-suitable habitats surface area including forest areas in the Netherlands (Sprong et al. 2012).

The distribution and abundance of ticks is affected by complex interplay of various abiotic and biotic factors such as the climatic conditions (temperature and humidity), habitat (vegetation) and vertebrate host community (Gassner and Hartemink 2013; Swart et al. 2014). Some studies have shown that spatial and temporal variation of tick occurrence is highly dependent on local microclimate conditions (Estrada-Peña 2001; Gassner et al. 2011; Greenfield 2011; Estrada-Peña et al. 2013), for example temperature, relative humidity, and saturation deficit. At temperatures < 7 °C, ticks remain inactive; ticks only venture out of the mat and up the vegetation to quest for a host when temperatures increase (Greenfield 2011; Tekenradar 2012). Humidity is a fundamental factor influencing tick survival primarily controlling the amount of time a tick can spend questing (Greenfield 2011).

Vegetation structure patterns and vegetation types have also been found strongly affecting tick abundance (Boyard et al. 2007; Gassner et al. 2011; Lindström and Jaenson 2003; Dobson, Taylor, and Randolph 2011). Vegetation provides different degrees of shelter for ticks (Swart et al. 2014). Most *Ixodes* ticks are inactive in the lowest layers of vegetation or in the leaf litter or soil before they begin to quest (Estrada-Peña et al. 2013). A thicker mat of leaf litter is able to retain more water and provides more niches for ticks to occupy thus improving the survival of the ticks (Greenfield 2011). At each life stage, ticks are present on different area of vegetation (adults on top and nymph and larva on lower part) (Guglielmone et al. 2006; Tack et al. 2013; Tack 2013). The dominant vegetation modulates the microclimate and also affects the host abundance (Estrada-Peña et al. 2013). Vegetation can be seen as an indicator of patterns of presence and abundance of ticks (Randolph and Storey 1999; Estrada-Peña et al. 2013).

Thus, knowledge about relationships between vegetation structure patterns and tick occurrence patterns can help to map potential tick distribution, and contribute to risk management of ticks and tick-borne diseases in the Netherlands.

The Public Health Service Bureau (GGD) of Amsterdam applies an Integrated Pest Management approach (IPM), which means using a variety of measures in combination to prevent the development of pests (van Adrichem et al. 2013). Tick is one of important pest species they keep focus on. GGD Amsterdam wants to have vegetation structure maps which can provide

necessary information for identifying tick suitable areas in Amsterdam, and expects to know more about the associations between environmental factors especially vegetation structure patterns and tick occurrence which can contribute to tick risk management in Amsterdam (J., Buijs & M. H., van Adrichem, personal communication, October 14, 2014).

1.1.2 Vegetation structure and tick occurrence in Amsterdam

From the above we understand that vegetation structure can affect the distribution and abundance of ticks. Different vegetation structure classes like grass, shrubs and trees can be expected to influence the habitat and life circle of ticks.

The occurrence of *Ixodes Ricinus* has been found to increase with an abundant tree layer and a high humidity on 61 grazed pastures in central France, by using a negative binomial model analysing tick abundance with environmental factors (Boyard et al. 2007). Another research shows thickness of the litter layer and moss cover is positively related to nymphal and adult tick densities from analysis of tick density and environmental conditions at 24 sites (mature forests, dune vegetation area, new forests, etc.) in the Netherlands (Gassner et al. 2011). In a study, ticks are found in all vegetation types including short grass close to car parks in parks in UK by sampling activity at 3-weekly intervals for 2 years; highest tick densities were consistently observed in plots with trees present (Dobson et al. 2011).

These studies have not been conducted in Amsterdam. Being an urbanized area, Amsterdam has different environmental and microclimatic conditions than natural areas. There is a need to study how local vegetation structure affects tick occurrence in Amsterdam. Besides, the dominant vegetation modulates the microclimate and also affects the tick hosts abundance (Estrada-Peña et al. 2013).

Thus it contributes to improve local risk management of ticks and provide better public health service that to research associations between environmental factors, particularly vegetation structure, and tick occurrence in Amsterdam.

1.2 Problem definition

1.2.1 Relationships between vegetation structure and tick occurrence

There are some papers concerning other regions of the Netherlands and statistical methods are commonly employed in these papers for assessing the associations of tick occurrence and environmental factors.

Gassner et al. (2011) undertook a longitudinal study at 24 forest area sites of the Netherlands on population dynamics of *Ixodes ricinus* and their infections with *Borrelia burgdorferi* using Generalized Linear Models (GLM). They found that habitat structure (tree cover) was an effective discriminant parameter in the determination of *Borrelia* infection risk (Gassner et al. 2011). Sprong et al. (2012) used historic data on land usage, temperature and wildlife populations to analyse with two longitudinal field studies data of questing ticks density by a negative binomial model and logistic model. They found circumstantial evidence for an increase in the risk of acquiring a bite of a tick infected with *B. burgdorferi* (Sprong et al. 2012). Wielinga et al. (2006) collected ticks in four habitats dunes, heather, forest, and a city park in the Netherlands. Authors analysed different kinds of infection rates of ticks applying reverse line blot analysis (Wielinga et al. 2006). Swart et al. (2014) used MODIS satellite image products (Enhanced Vegetation Index,

land surface temperature, Middle Infra-Red), roe deer population densities data, and soil moisture map with Fourier analysis, and quadratic discriminant analysis, aided by a Bayesian inclusion of expert opinion data. They got classified land-use types and tick suitability levels map in 1 km resolution for the Netherlands (Swart et al. 2014). They later applied a linear model to erythema migrans consultations incidence data by general practitioners in the Netherlands and the estimated probability of tick presence. Researchers stated a significant fraction of the tick bite consultations could be explained by the *I. ricinus* population outside the resident municipality (Swart et al. 2014).

There are no papers found which study tick occurrence and vegetation structure patterns relationships for Amsterdam urban region from literature, to the best of my knowledge. However, there are examples in the literature of studying relationships of tick abundance and environmental factors focusing on urban environments in other countries. A G-test and logistic regression were applied for analysing the relationships of tick occurrence and temperature, sward height, mat depth, etc. environmental parameters at 16 sites in a park of London (Greenfield 2011). Tick presence was found to be closely related to soil moisture, light levels and humidity throughout the park (Greenfield 2011). Using spatial statistics, Vatansever et al. (2008) investigated potential risk factors for ticks from reported tick bites in Istanbul and found *Ixodes* is highly reported in dense highly heterogeneous vegetation patches.

To summarize, tick occurrence has been found to be associated with vegetation structure but such relationships have not been studied for the Amsterdam urban region.

1.2.2 Need to map vegetation structure in Amsterdam

Currently there is no existing suitable vegetation structure map of Amsterdam area (at least containing height information) that can be used for studying relationships of vegetation structure and tick occurrence in Amsterdam, to the best of my knowledge.

Services of the Amsterdam municipality usually apply the Amsterdam Small Scale Base Map 1: 10,000 (Amsterdam KBKA10) (Municipality of Amsterdam 2014). It is augmented data made from the scale 1:10,000 Key Register Topography (TOP10NL) by adding information for the municipality of Amsterdam and a zone around it (Municipality of Amsterdam 2014). When the GGD of Amsterdam wants to use detailed vegetation structure information, they find the vegetation object classes of KBKA10 (city green, deciduous forest, etc.) are inappropriate. For example, the city green class is composed of a mixture of grass areas, shrubs and trees. Only part of the forest areas is included in the deciduous forest class. (J. , Buijs & M. H., van Adrichem, personal communication, November 13, 2014). The vegetation classes of KBKA10 map the green area of Amsterdam but they don't distinguish the required vegetation classes nor do they provide structure information of vegetation. In addition, the geographical department of Amsterdam municipality has mainly worked on the structure of paved areas (buildings, roads, etc.), while it did not yet consider vegetation structure (J. Muller, personal communication, January 29, 2015).

Besides, there are maps of some public parks and valuable old trees in Amsterdam (http://maps.amsterdam.nl/hoofdgroenstructuur) but these maps are not suitable for tick occurrence scientific research which requires higher accuracy and detailed information of vegetation.

So it is required to have vegetation structure map for monitoring and risk management of ticks in Amsterdam.

1.3 Research objective and research questions

Hence the main objective of this research is to map vegetation structure classes and to assess the relationship of these in combination with other environmental factors with tick occurrence patterns in Amsterdam.

This objective has then derived the following research questions:

RQ1. Which geo-data can be used to map the vegetation structure classes (grass, shrubs, and trees) which are deemed important for mapping pest species distribution in Amsterdam?

RQ2. Which methods can be used for mapping vegetation structure patterns?

RQ3. Which methods can be used to assess associations between vegetation structure classes maps and the occurrence data of ticks and to use the found relationships for prediction?

RQ4. What is the strength of correlations between occurrences of ticks and vegetation structure classes in Amsterdam?

RQ5. Where are predicted hot-spots with respect to occurrences of ticks in Amsterdam?

1.4 Outline of this thesis

This thesis is organized in five chapters.

Following the introduction in this chapter, Chapter 2 presents the materials and methods used in the thesis. It introduces the flow of whole research and explains the details of *k*-Nearest Neighbour method and Classification and Regression Tree method. Method of statistical relationship assessment is also illustrated. Accuracy assessment steps of main results are introduced.

Chapter 3 presents the result vegetation structure maps, their accuracy, the statistical relationship, validation of the logistic regression model, and predicted tick occurrence map.

Chapter 4 contains further interpretation, and discussion about the results.

Chapter 5 presents the conclusions and recommendations. This chapter contains an overview of the important results found in this thesis.

2. Materials and methodology

2.1 Study area

Figure 1 shows the study area (90 km^2). This research covers large part of the Amsterdam urban area. The study area covers the city centre, some residential areas, parks and more than 90% of the area where GGD Amsterdam sampled ticks. One delineated square in Figure 1 represents a 1 km^2 tile of the point cloud data containing height information.

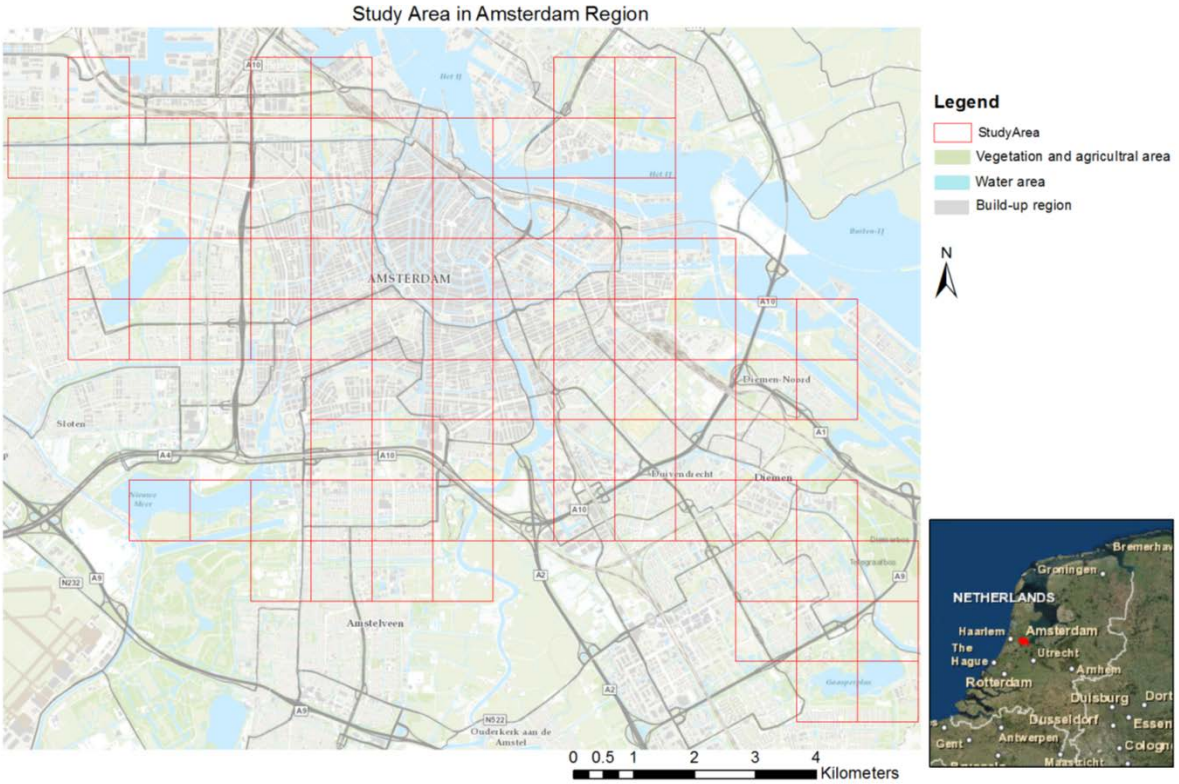


Figure 1. Study area in Amsterdam. A square corresponds to a 1 km^2 tile, in total there are 90 tiles. The base map is the Amsterdam area of World Topographic Map from OpenStreetMap contributors and the GIS User Community (ESRI 2015).

2.2 Data specifications

Point clouds (structural characteristics of vegetation)

Point cloud data supplied by the geographical department of the Civil Registry of Amsterdam were used to provide vegetation height data, to perform supervised classification and to create a vegetation structure map. The point clouds were acquired by a Semi-Global Matching (SGM) method (photogrammetry) from aerial images of March 2014, and the average point density is 16 points per square meter (Aerodata Surveys Nederland 2014). The estimated point cloud accuracy is between 0.4 and 0.8m vertically (Aerodata Surveys Nederland 2014). This SGM point cloud data is in LAS format. Unlike Light Detection and Ranging (LiDAR) data which also shows some inner structure of vegetation, this point clouds show mostly top outline of vegetation (see Figure 2 a) (Aerodata Surveys Nederland 2014).

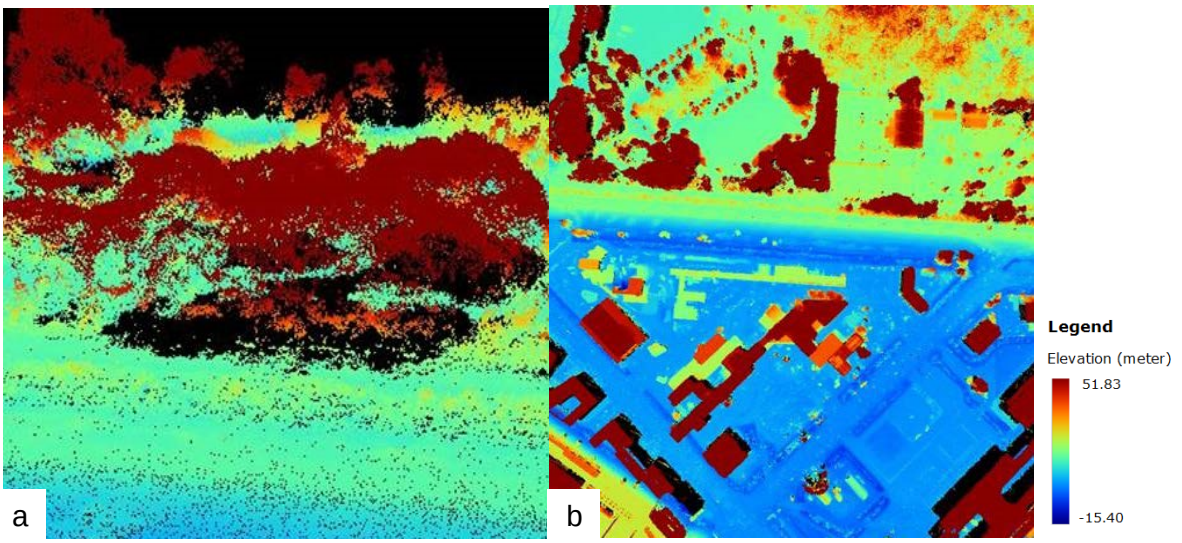


Figure 2. SGM point clouds. It shows top outline of vegetation (a, oblique view of trees); b is top view of some part of 1 tile (125000, 485000) point clouds.

AHN2 (terrain height, vegetation structure classes training data location selection)

The Actueel Hoogtebestand Nederland (AHN, Up-to-date Height Model of The Netherlands) file is detailed and precise elevation data created by LiDAR technique for the Netherlands (van der Zon 2013). The gathering of AHN2 data for Amsterdam area was carried out over a period of roughly 1st December 2009 to 1st April, 2010 (van der Zon 2013). In the western part of the data, the point density is possible slightly different due to the plane flying at a different speed (van der Zon 2013). In general the point density on average is between 6 and 10 points per square meter (van der Zon 2013).

In this study, the 0.5m resolution filled terrain raster was used for subtracting terrain height and for helping to select vegetation structure classes training data locations. The 0.5m resolution raw raster data were also utilized to help choosing training data sites. The 0.5m filled terrain raster is the ground level file where non-ground objects (trees, buildings, bridges and other objects) were removed from the point cloud and no data cells were filled (National Georegistry 2014). The 0.5m raw raster was resampled from the point cloud to a 0.5 meter grid, using both the ground and non-ground objects (National Georegistry 2014).

Aerial photographs (ground information, training data locations selection)

The geographical department of the Civil Registry of Amsterdam provided high resolution 0.11 m aerial photograph acquired in 2013 March for Amsterdam region. It contains RGB three bands and shows relatively clear ground information. Vegetation is not completely green yet in the images and there are dark shadows of buildings and trees.

Topography data (vegetation area, training data selection, other environmental factors)

The Amsterdam Small Scale Base Map 1: 10,000 (KBKA10) is augmented data made from the scale 1:10,000 Key Register Topography (TOP10NL) by adding information for Amsterdam and a zone around it (Municipality of Amsterdam 2014). The KBKA10 is updated annually on the basis of aerial photographs (Municipality of Amsterdam 2014). In this research, the KBKA10 of 2014 was used because it is detailed, comprehensive and available. It contains artificial object classes like buildings, local roads, railway, metro, etc.; green vegetation object classes like city green

area, deciduous forest, coniferous forest, poplars, etc.; water features like pool, basin, deep water, port, etc.; and not-used area classes. It should be noticed that KBKA10 contains vegetation area of Amsterdam but it doesn't contain very accurate distribution of separated vegetation classes like grass, trees nor vegetation structure information which we need. However, KBKA10 has been used for filtering point cloud, training data site selection and masking other land cover classes such as built-up region and water bodies.

Tick data

Ticks were collected by blanket dragging (blanket 1x1 m^2) at 282 sites in 23 parks in Amsterdam from 19 June to 17 July, 2013 by experts from GGD Amsterdam (M. H., van Adrichem, personal communication, November 4, 2014). Researchers recorded the coordinates, the absence and presence of ticks on the cloth, and counted the numbers. There were 31 presence records in a total of 282 records.

Auxiliary information (vegetation structure classes training data locations selection)

Google Earth and Google Street View provided some satellite images and photos which contributed to find locations of training data of vegetation structure classes (grass, shrubs, and trees).

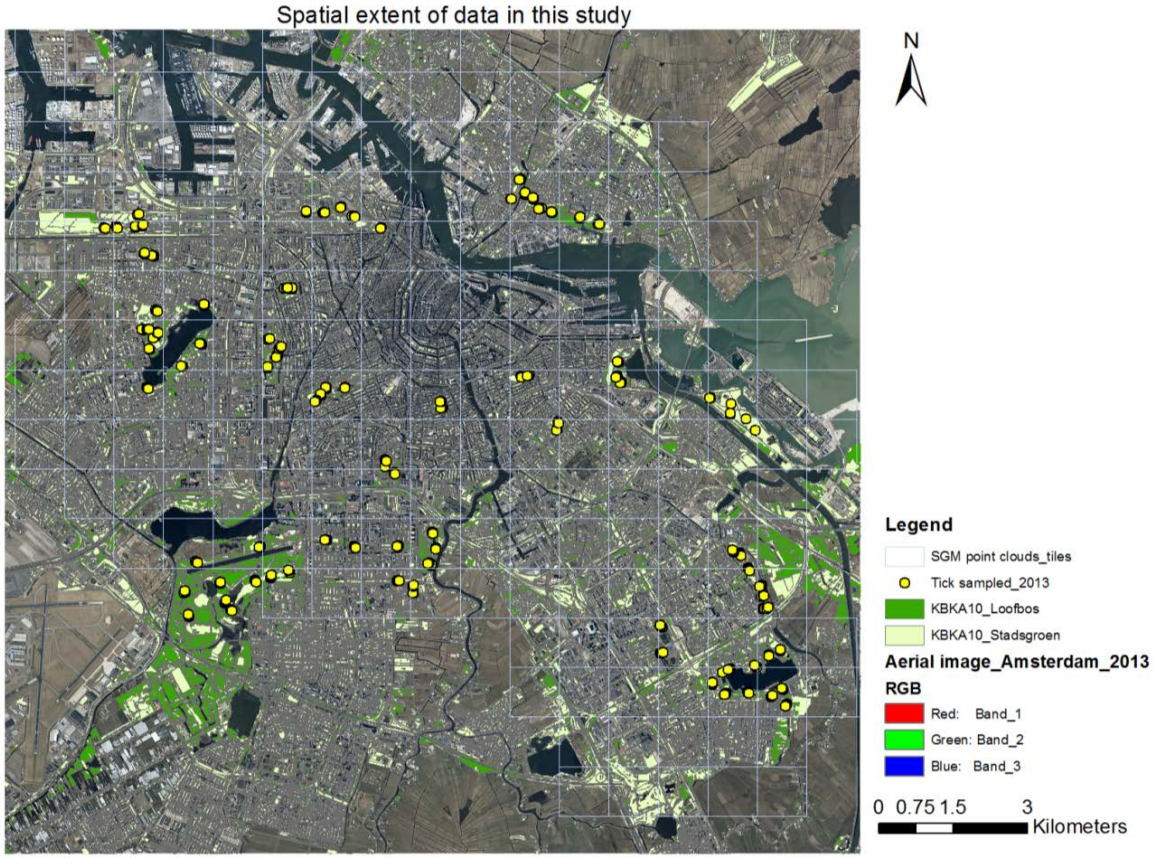


Figure 3. Display of tick data, some KBKA10 topo data, aerial image, and part of SGM point cloud tiles (boundaries)

Figure 3 presents the spatial extent of the tick data, two KBKA10 vegetation object classes city green and deciduous forest, the aerial image and some boundaries of SGM point cloud tiles.

2.3 General explanation on methodology

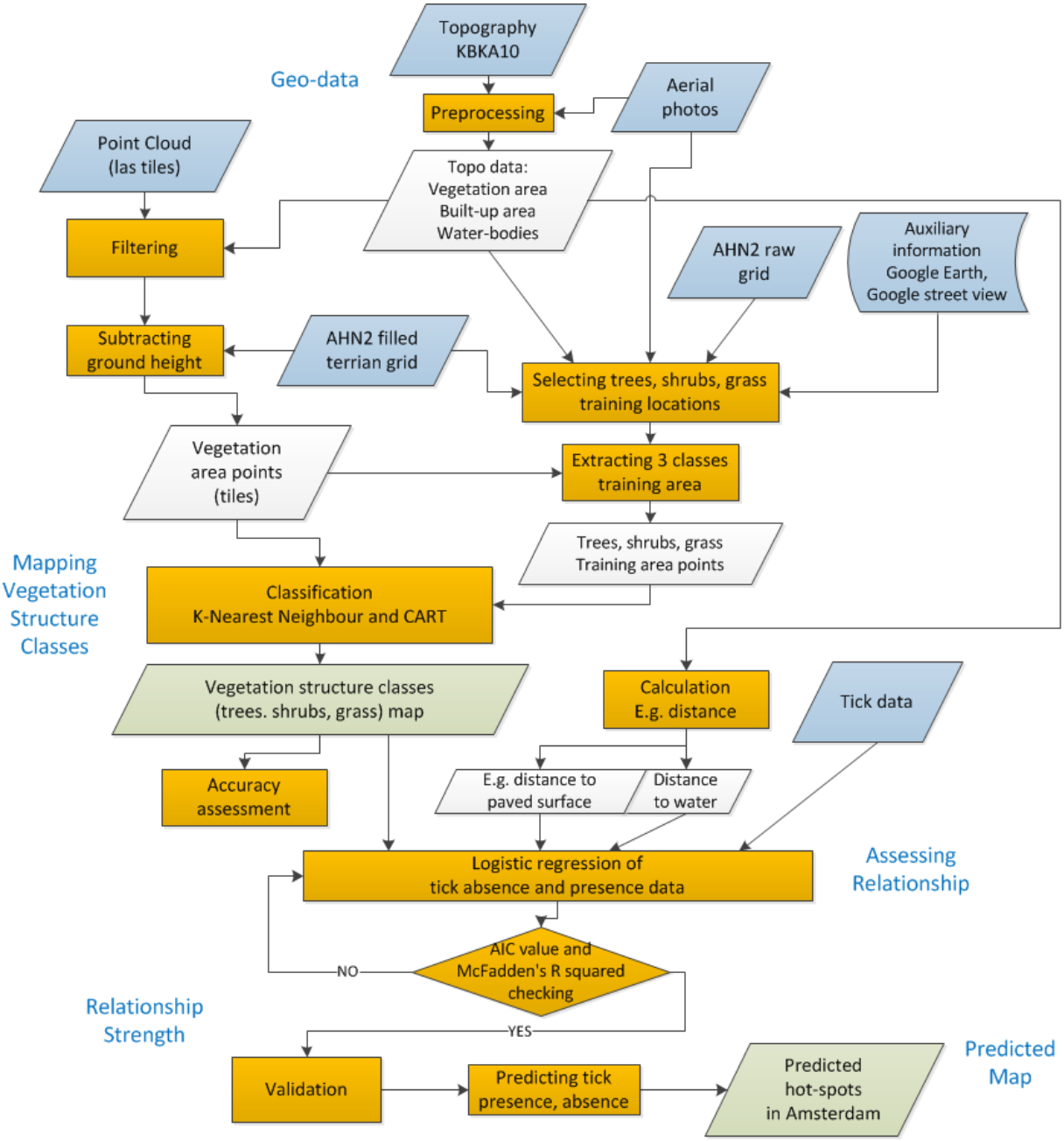


Figure 4. Methodology flow chart

A methodology was worked out to map vegetation structure classes and to assess the relationship of these in combination with other environmental factors with tick occurrence patterns in Amsterdam. Figure 4 illustrates the methodology of this research, which is divided into following sections: the vegetation area filtering, supervised classification of vegetation, logistic regression of tick occurrence and environmental factors, model assessment, and predicting tick occurrence in Amsterdam.

The whole process was implemented with ArcGIS for desktop 10.2 (ESRI 2014), LAStools 16 November 2014 version (rapidlasso 2014), R version 3.1.2 (R Core Team 2014) and RStudio Version 0.98.1102 (RStudio 2014).

2.4 Vegetation area filtering

Three steps were conducted in this section, i.e. pre-processing topography data KBKA10, vegetation area filtering, and subtracting terrain height from vegetation area point clouds.

Because SGM point clouds cover whole surface area and there are many other objects like buildings, cars, lamps have similar height with trees and shrubs, it was necessary to separate the vegetation objects with other items. In other studies mapping urban vegetation structure, the first step is also pre-classification to separate vegetation objects from other urban objects (Rutzinger et al. 2010; Pratihast 2010; Mathieu, Freeman, and Aryal 2007). In addition, water is a factor affecting tick habitat and there are water related object classes in KBKA10. So the first step was pre-processing topography data KBKA10 in ArcGIS model builder and making three groups: vegetation area, water area, and built-up area (details in the Figure 5).

Then the vegetation area was applied to filter out all non-vegetation areas from the point cloud dataset. This filtering process (details in Figure 6) was done in LAStools which is thought efficient for processing large amount of LAS format point clouds (rapidlasso 2014). The LAS file format is an industry-standard binary format for storing airborne LiDAR data and it is also a public file format for the interchange of 3-dimensional point cloud data between data users (Crosby 2011). Because LAS is a binary format, a reader of some kind is necessary to ingest the data (Crosby 2011). LAStools is recommend because it is fast in reading and writing LAS files, merging LAS files, converting format and so on (Crosby 2011).

Since the elevation of extracted vegetation area point clouds was the surface height, subtracting terrain height was also done in R for only keeping the vegetation objects height (Figure 6 right part). The R package ‘sp’ (Pebesma and Bivand 2005; Bivand et al. 2013), ‘rgdal’ (Bivand, Keitt and Rowlingson 2014), and ‘raster’ (Hijmans 2015) were used.

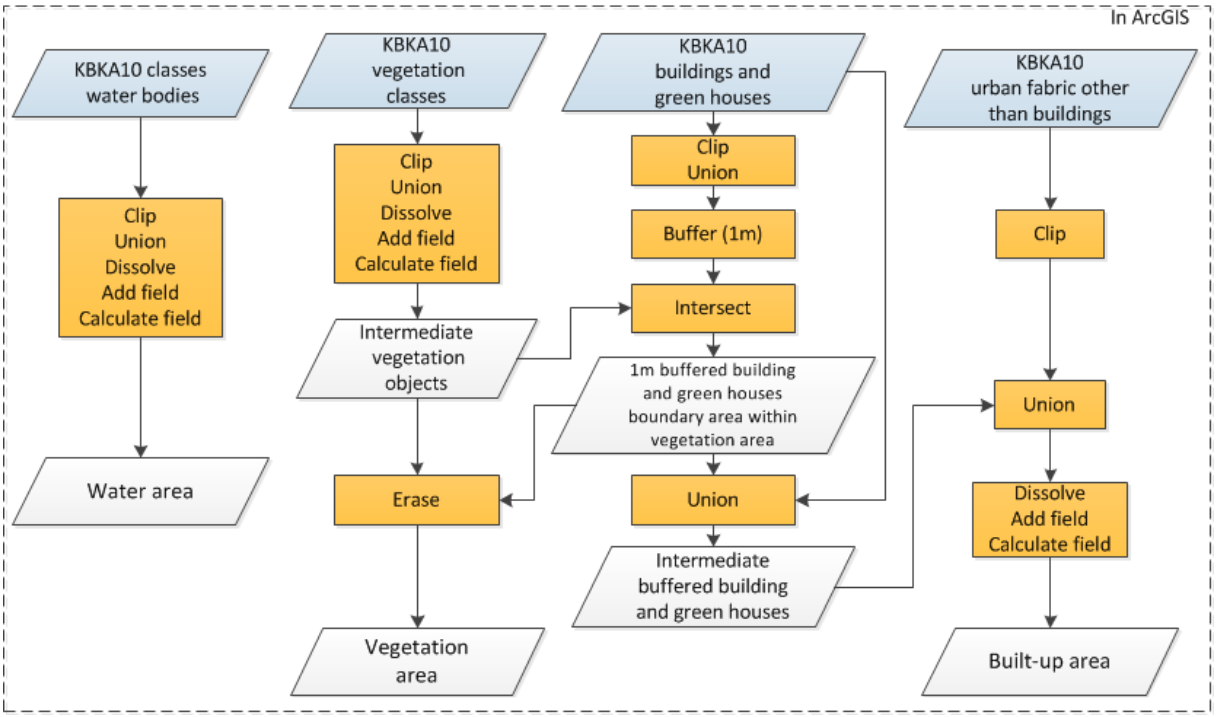


Figure 5. Pre-processing topography data KBKA10

Figure 5 explains the pre-processing of KBKA10 in ArcGIS. During a test of clipping vegetation area from point clouds, about 1m (estimation on average) boundary points of buildings and green

houses were found remaining in the vegetation area. So in order to get a more accurate vegetation area boundary, 1m buffering of building and green house classes, and erasing this 1m buffering area from vegetation objects were done.

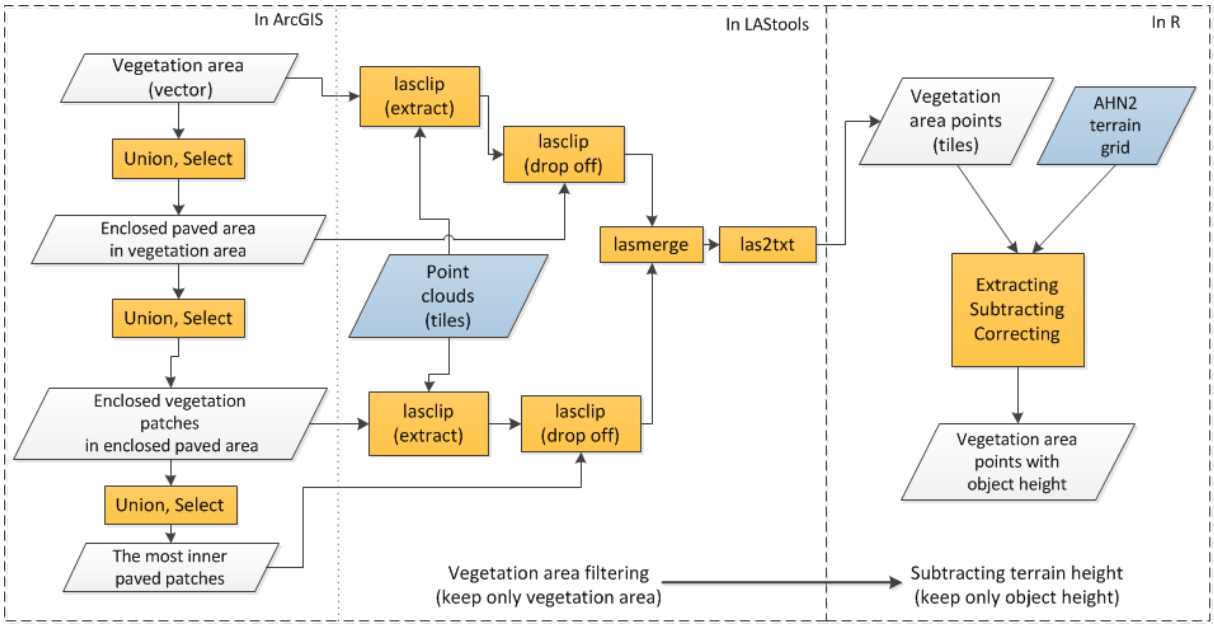


Figure 6. Vegetation area filtering (left ArcGIS and central LAStools part) and subtracting ground height (right: R processing part)

Figure 6 illustrates details of vegetation area filtering and subtracting ground height. After subtracting terrain height from vegetation area points, some negative values were found in the vegetation objects height data. So correcting these values, i.e. replacing negative values by 'zero' were conducted in R.

2.5 Overview of common vegetation mapping methods

Accurate representation of vegetation classification and structure information is a continuing challenge because of many factors. For instance, the range of climates, geomorphic substrates, natural disturbance and human encroachments has produced an incredible diversity of terrestrial vegetation (Rajaei Najafabadi 2014). Even with increasing availability of high resolution aerial images and satellite images and advancements in processing methods, it still remains a complex task for researchers to create detailed vegetation maps showing species-level vegetation and vegetation areas with heterogeneous structural composition (Harvey and Hill 2010).

The methods researchers used for vegetation structure mapping depend on applications and research purposes.

(1) Field measurement

The term vegetation structure has been used in some tick and mosquito studies to refer to horizontally and vertically distribution of vegetation. Researchers used ruler and tape measure in the field survey and recorded detailed information of local vegetation structure (types, species, height of vegetation, etc.) (Gardner et al. 2013; Greenfield 2011; Mejlon 2000). Field survey of vegetation species, height and canopy diameter (volume) with the help of aerial photos has been conducted to study impact of urbanization on vegetation of riparian areas (Hutmacher et al. 2013).

(2) Indirect mapping using existing maps

To investigate urban ecosystem services, Lehmann et al. (2014), predefined urban vegetation structure types based on species, dimension, height, etc. terms. Relationships of vegetation structure types, land use types, and urban biotope types were analysed and defined and vegetation structure was mapped based on land use map and urban biotope map with the help of aerial photos and field survey (Lehmann et al. 2014).

(3) Classification map based on satellite images

Object-oriented classification methods and very high-resolution multispectral satellite images have been used to map urban private gardens for ecological study (Mathieu, Freeman, and Aryal 2007). Study area was stratified into broad classes industrial area, residential area, dominated vegetation, and water in the beginning. Researchers firstly employed predefined criteria of vegetation structure classes and obtained training data for different garden types. Then they used a Nearest Neighbour method to classify residential garden areas (Mathieu, Freeman, and Aryal 2007). Wood et al. (2012) applied Spearman rank correlation to field measurements, sample-point pixel values and image texture measures from Infrared aerial photos and Landsat NDVI images. They found sample-point pixel values and texture measures from remote sensing images captured components of foliage-height diversity and horizontal vegetation structure in grassland, savanna, and woodland habitats (Wood et al. 2012).

(4) Detailed structural mapping based on multiple data sets

In recent years researchers have used hyperspectral data (Hyde et al. 2006; Mücher et al. 2013) and Light Detection and Ranging (LiDAR) data (Rutzinger et al. 2010; Kuilder 2012; Rajaei Najafabadi 2014; Hantson, Kooistra, and Slim 2012; Mücher et al. 2010) to classify and map vegetation structure at high resolution and broad scales. Accurate height measurements of shrubs and trees can be provided by LiDAR data (Mücher et al. 2010) and parameters of the outer shape 3D model of single urban trees have been extracted from mobile laser scanning data (Rutzinger et al. 2010). For mapping invasive woody species in dune ecosystems, maximum likelihood (ML) classification has been applied to multispectral aerial photos; ML classification combined with vegetation heights derived from LiDAR (ML+) and object-based classification were compared (Hantson, Kooistra, and Slim 2012). To map the vegetation structure classes along a floodplain (Kuilder 2012) and part of an island in the Netherlands (Rajaei Najafabadi 2014), spectral features, geometric features, and topographic features from multi spectral aerial images and AHN2 data were extracted and employed. The researchers acquired training data and applied object-based classification with a random forest classifier (Kuilder 2012; Rajaei Najafabadi 2014).

Among all available methods, on-site manually surveying is thought to be a highly precise but rather time-consuming method (Lehmann et al. 2014). In this study, such detailed level is not possible for the size of the Amsterdam area. Since the available data mainly are SGM point clouds and topographic data from KBKA10, we only focus on simple features of vegetation structure in Amsterdam, i.e. height and height variability in this research. This is also different from the term vegetation structure in LiDAR studies, which refers to many characteristics, like shape, volume, direction, etc.

2.6 Vegetation mapping methods used in this thesis

In this research, height is the first criterion of vegetation structure classes. We focus on height and variability in heights, which are considered the structural characteristics of vegetation in Amsterdam. The following major classes were considered:

Class 1: short vegetation with a typical height <= 0.3m, regarded as grass;

Class 2: medium height vegetation with heights between 0.3m to 6m, regarded as shrubs;

Class 3: tall vegetation, height >= 6m, regarded as trees.

It is noticeable that each vegetation class shows natural variation; therefore a set of simple height thresholds are not enough for classification of vegetation structure. We also need to know the variability of plant heights occurring within typical vegetation patches of each type. Therefore, 'true' vegetation structure class data (reference data) are to be used.

Among the classification methods for mapping, supervised classification is commonly used for producing classification maps from satellite images (Weih and Riggan 2008). Training data which thought as representing features of real ground objects are used in supervised classification as reference to map different classes. The *k*-Nearest Neighbour method is thought to be versatile, robust, and flexible with potential to combining different sources of information (Gao and Mas 2008; Samaniego and Schulz 2009; Franco-lopez, Ek, and Bauer 2001). Besides, Random Forest and Classification And Regression Tree (CART) are two popular machine learning methods used in the vegetation classification. The latter is regarded as effective, flexible, easy to implement, dividing complex problem into simpler sub-problems, and able to incorporate with different kinds of ancillary data (Bittencourt and Clarke 2003; Lawrence and Wright 2001; Qian et al. 2014). In this work, supervised classification using the *k*-Nearest Neighbour (kNN) and Classification And Regression Tree (CART) were applied for mapping vegetation structure classes.

2.6.1 Selecting and extracting training and validation data

Firstly, the locations of training area (certain locations of trees, shrubs, and grass) were selected using the aerial image, topography data, AHN2 raw grid, AHN2 terrain grid, Google street view and Google Earth. These training area sites were mapped by manually digitizing polygon areas in ArcGIS. I did my best to search and examine certain historical photos or high resolution remote sensing images to make sure the training area locations have real trees/shrub/grass growing and carefully draw the boundary of training area locations.

Each sample polygon was assigned to one of the vegetation structure classes defined in the previous section. The mean area of grass and tree training locations were about 45mx45m and the shrub training location area was about 24mx24m on average. The number of training area locations of each vegetation structure class is 60. Then training polygons data of three vegetation structure classes were extracted from the vegetation area points in R with package 'sp' (Pebesma and Bivand 2005; Bivand et al. 2013), 'rgdal' (Bivand, Keitt and Rowlingson 2014), and 'raster' (Hijmans 2015).

It appeared to be very difficult to obtain reliable training data for the shrub class. After several rounds of refinements it was decided to retain a subset of the original sites. The selected sites (30 shrub training sites) had histograms that match the above class definitions. These best 30 shrub training data were used in the training classifiers and also accuracy assessment step.

For 60 grass training data and 60 tree training data, half of them were selected by a random sampling approach for training the classifiers while the other half were put aside for accuracy assessment.

2.6.2 Training phase and classification

Next, *k*-Nearest Neighbour and CART were applied in classification and two classification result maps were compared in terms of accuracy and reliability.

2.6.2.1 *k*-Nearest Neighbour classifier and classification process

The 10%, 20%, ... 90% quantiles of height of each training location were calculated for representing the structure feature height and spread in heights. The 9 quantiles calculation was performed on each polygon of the training data.

The best 30 training locations of the shrub class together with randomly chosen 30 grass and 30 tree training locations were then used in the *k*-nearest neighbour computing step.

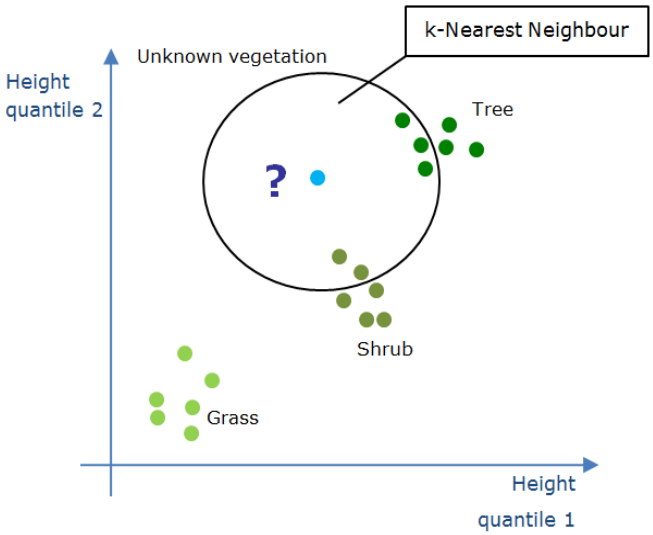


Figure 7. Example of k*-nearest (*k *= 5) neighbour classifier with two height quantiles. The two dimensions are used for illustration only; computations were based on 9 quantiles.*

K-Nearest Neighbour considered a 9-dimensional space, with each dimension being a quantile in the range (0.1, 0.2, ..., 0.9). It was calculated the distance (squared value of difference) between quantiles (10% to 90%) of the unknown vegetation and quantiles (10% to 90%) of the total 90 training areas. For every training area, the sum of squared difference was computed. Then it was ordered from smallest to largest the sum of squared difference of 90 training areas. The *k* nearest (minimum distance) training areas and their vegetation structure classes were recorded.

In this case, *k* was set to 5 considering the rather small size of the training data set (90 sites). Figure 7 illustrates the calculation using two height quantiles while in the actual process 9 quantiles were used. For the site with unknown vegetation, 5 nearest neighbour training data (in feature space) are found. The numbers of grass, shrubs, trees classes among total 5 nearest training locations were converted to probabilities of this unknown vegetation being grass/shrubs/trees class. In Figure 7 this case these probabilities amount to 0.4 for the shrubs class and 0.6 for the trees class.

For the classification process, the vegetation area point tiles were firstly rasterized to a 0.5m grid accounting for point clouds density 16pts/m^2 and software calculating capacity. After a field visit and some experiments, 5x5 focal window was chosen to apply on the vegetation height raster to computer quantiles and 5 nearest neighbour vegetation training areas. A smaller focal window size would be too small to compute the 9 quantiles and a larger window size would involve a large edge effect. When moving window operation was performed on the edge cells, because cells outside of the extent had no value, these edge cells would be assigned no data value. Lager window size would lead to have more no data edge cells (large 'edge effect'). If all cells of the 5x5 window were no data value, the returned value was set to 0 meaning there is no vegetation growing. After that the probability maps of grass/shrubs/trees class were produced.

The packages 'sp', 'raster' and 'rgdal' in R were utilized in this step. A function of calculating *k*-NN was programmed in R.

In the probability maps of grass/shrubs/trees, the pixels with more than 1 class probability represent mixture vegetation structure classes patches. In a patch with probability *p* for grass, the expected fraction of grass equals *p*. To present where the mixture vegetation structure patches are, Shannon entropy bits were computed on the probability maps of grass/shrubs/trees.

Shannon entropy is defined as (Shannon 1948):

Equation 1. Shannon entropy

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where X is a discrete random variable taking a finite number of possible values $x_1, x_2, ..., x_n$ with probabilities $P(x_1), P(x_2), ..., P(x_n)$. Here vegetation structure class is the variable, and the result entropy is the numerical measure of the uncertainty in labelling a pixel; it is the additional information--expressed in bits--needed by the classifier to have absolute certainty about class assignment. In the result entropy bits map, pixels of entropy value >0 indicates there are mixture vegetation structure classes.

2.6.2.2 CART classifier and classification process

CART (Breiman 1984) is a tree-like classifier and it is a non-parametric technique that can select from a large number of variables and assess their interactions that are most important in determining the outcome variable to be explained (Yohannes & Hoddinott 1999).

Here height values and standard deviation of height were used to show the structural features of vegetation. The training area points were first rasterized as 0.5m resolution grids and 5x5 focal window was applied on them to calculate the standard deviation (SD) of height value. The SD values were added to the training data set.

After that the sample (training areas) was analysed in CART, which divided all the training area data according to a set of "splitting rules" using a "goodness of split criterion"-Gini impurity (Therneau and Atkinson 2015). Gini impurity (generalized Gini index) measures the homogeneity of the target variable within the subsets. The Gini index calculates how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset (Therneau and Atkinson 2015). It is computed by summing the probability of each element being chosen times the probability of a mistake in classifying that item (Therneau and Atkinson 2015). Gini impurity is applied to each candidate subset and the resulting values are combined to provide a measure of the quality of the split. Minimization of Gini index is

the criteria used in CART to split a ‘’tree’’. Here splitting rules were built based on variables height value and standard deviation of height. The dependent variable was the categorical vegetation structure class-trees or shrubs or grass. A hypothetical example of a classification tree is shown in Figure 8.

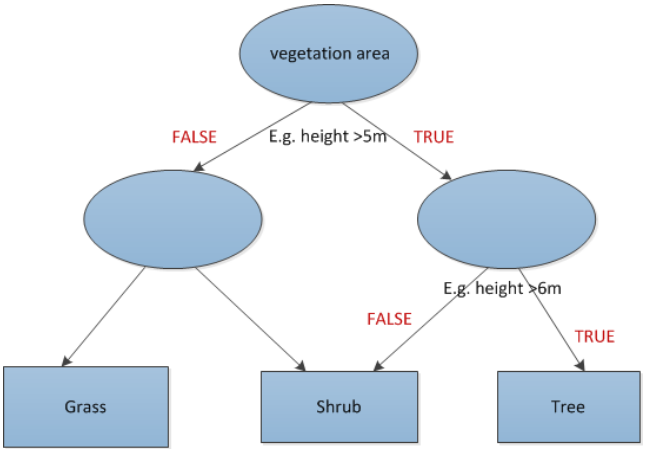


Figure 8. A hypothetical example of a classification tree

In the R ‘rpart’ package (Therneau, Atkinson, and Ripley 2014) of CART classifier, complexity parameter (cp) is a number between 0 and 1 (advisory parameter) which measures the 'cost' of adding another variable to the model and it reflects how complex/large the classification ‘tree’ is allowed to be (Therneau and Atkinson 2015). It is specified according to the formula (Therneau and Atkinson 2015):

Equation 2. The formula of complexity parameter and the 'cost' for the tree

$$R_{cp}(T) \equiv R(T) + cp * |T| * R(T_1)$$

where T_1 is the tree with no splits, $|T|$ is the number of splits for a tree, and R is the risk.

Cross-validation was used to choose a best value for complexity parameter (cp) and an estimate of the risk and its standard error are computed during the cross-validation: the relative error with the training data, cross validation error with one standard error values are computed with different complexity parameters (Therneau and Atkinson 2015). This step is in the process of mechanism called ‘pruning’ to decide the tree size. In actual practice, the ‘1-SE rule’ is used for selecting best cp value. A plot of different cp values versus cross validation error with the training data is obtained. The plot of cp versus the cross validation error often has an initial sharp drop followed by a relatively flat plateau and then a slow rise (Therneau and Atkinson 2015). It states that “*Any risk within one standard error of the achieved minimum is marked as being equivalent to the minimum (i.e. considered to be part of the at plateau)*” (Therneau and Atkinson 2015). The simplest model, among all those on the plateau, is selected. Additionally, it is noticeable that overfitting occurs when an over-complex tree describes random error or noise from the training data (Therneau and Atkinson 2015). These over-complex trees do not generalise well from the training data and can exaggerate minor fluctuations in the data. So it should be avoid overfitting.

In this study, a plot of different cp values versus cross validation error with the training data was obtained using ‘rpart’ package. After that the cp value near the plateau was decided to be selected as the suitable one.

The classification tree grown with the selected cp value was applied for classifying vegetation in the whole area. Vegetation area point tiles were firstly rasterized to 0.5m resolution grids and 5x5

focal window calculation was conducted on height grid to acquire SD grid. The selected classification tree was then applied on the height grid and SD of height grid. As a result, the CART classification vegetation structure maps were acquired.

2.6.3 Accuracy assessment of vegetation maps

A confusion matrix is a commonly used method for assessing classification results (Lillesand, Kiefer, and Chipman 2008). The confusion matrix is built on a category-by-category basis and stems from classifying sampled training data set pixels (Lillesand, Kiefer, and Chipman 2008).

The rows in the confusion matrix refer to the known 'true' categories (reference data, ground truth) and the columns are the actual classified result by the classifier (Clevers 2013). Overall accuracy is computed by sum of correctly classified pixels number of every class, the diagonal elements of confusion matrix (Clevers 2013).

Errors of omission (exclusion) correspond to the non-diagonal row elements in the matrix and commission errors (inclusion) are represented by non-diagonal column elements (Clevers 2013). The accuracy of individual category is computed by dividing correctly classified pixel number in one class by the number of training set pixels used for that class, the row total (Clevers 2013). This accuracy of individual class is often termed as 'producer's accuracy' and indicates how well training set pixels of the known ground truth class are classified (Lillesand, Kiefer, and Chipman 2008). The term reliability result from dividing correctly classified pixel number in one class by the total number of pixels that were classified in that category (the column total) (Clevers 2013). Reliability is also called 'user's accuracy'. It shows the commission error (equal to 1 minus commission error percentage) and indicates the probability that a pixel assigned to one known class really belong to that class on the ground (Lillesand, Kiefer, and Chipman 2008).

In this study the *k*-nearest neighbour classification was used to compute the probabilities a pixel belongs to each of the three classes. The probability value from [0, 1] was used in the calculation of confusion matrix for representing the degree of this pixel belonging to one class. Then values in the confusion matrix of *k*-nearest neighbour classification results represent summation of the probabilities of pixels belonging to one class (Freitas 2002). On the other hand, the values in the confusion matrix of CART classification results represent summed amount of pixels assigned to one class (pixels which had maximum probability of that class and were already assigned to that class).

Note that in the absence of a separate validation data set for the shrubs class, the same 30 shrub training areas were used both in training phase and accuracy assessment as reference data. Because the study area was adjusted during the study, 19 of the 30 shrub training areas were located inside the classified study area. These 19 shrub training areas were used as reference data in the calculation of two confusion matrices. Accordingly, the calculated overall accuracy, shrub class accuracy and reliability performance measures were partly over-estimated.

In terms of grass and tree reference data, 19 of the kept 30 grass validation areas and 19 of the kept 30 tree validation locations, which were not used in the previous classifier training phase, were selected by random sampling and applied in computing two confusion matrices as reference data.

2.7 Logistic regression of tick absence/presence data

Logistic regression is a type of probabilistic statistical classification model (Freedman 2009). Binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types (Hosmer and Lemeshow 2000). Logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors) (Hosmer and Lemeshow 2000). The model is given below (Cook et al. 2001):

Equation 3. Logistic regression model

$$E(Y_i \mid X_i) = \pi_i = \frac{e^{(\beta_0+\beta_1 X_i)}}{1+e^{(\beta_0+\beta_1 X_i)}}$$

Where X_i is the independent variable, Y_i is the dependent variable and π_i is the probability of $Y_i = 1$.

In the related research field of ticks and vector-borne diseases, logistic regression is a commonly used method (Greenfield 2011; Ageep et al. 2009).

In this research, the absence and presence data of ticks is a binary variable (the dependent variable). Vegetation structure classes (trees, shrubs, and grass), distance to water and distance to built-up class are the independent variables. Logistic regression was used as a tool to assess the relationship between vegetation structure classes, other two environmental factors and occurrence of ticks.

After accuracy assessment, the best vegetation structure map was used in the assessment of this relationship. Distance calculations of water and built-up class were performed in ArcGIS using the same 0.5m resolution with the vegetation structure map. Extraction of vegetation structure class at tick data locations and preparing training data for logistic model was done in R.

Logistic model fitting was done in R using the glm( ) function. It was used the binomial family option with logit link function (R Core Team 2014).

2.8 Model assessment

2.8.1 Internal measures

To choose a better fitted model, the internal quality of the fitted model (training data) was assessed using AIC and McFadden's R squared measure.

AIC is short for Akaike information criterion, a measure of the relative quality of a statistical model for a given set of data and AIC provides a means for model selection (Burnham and Anderson 2002). For any statistical model, the AIC value is

Equation 4. AIC value of statistic model

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the model, and L is the maximized value of the likelihood function for the model (Burnham and Anderson 2002). If there are a set of candidate models for the data, the preferred model is the one with the minimum AIC value (Burnham and

Anderson 2002). One can estimate, via AIC, how much more (or less) information is lost by one candidate model than by another candidate model (Burnham and Anderson 2002).

Besides, McFadden's R squared measure is one measure proposed for logistic regression inheriting the properties of the familiar R squared from linear regression (Bartlett 2014). In this study, McFadden's R squared measure was used to understand how much variability in the dependent variable (tick presence and absence) were explained by the model and how much improvement were got from the null model to fitted model. McFadden's R squared measure is defined as

Equation 5. McFadden's R square measure

$$R^2_{\text{McFadden}} = 1 - \frac{\log(L_C)}{\log(L_{null})}$$

where L_c refers the (maximized) likelihood value from the current fitted model, and L_{null} refers the corresponding value of the null model-the model with only an intercept and no covariates (Bartlett 2014). If McFadden's R squared measure is more close to 1, then the more variability of dependent variable are explained and the stronger predictive power of the logistic regression model is.

AIC value checking was conducted with the glm () function and stepwise regression in R to search for a better fitted model (selecting variables) among candidate models with independent variables (vegetation structure classes, distance to water and distance to built-up class). The model with smallest AIC value was chosen.

McFadden's R squared measure was computed with ‘pscl’ package (Jackman 2015) in R to measure the goodness-of-fit of the model, understand the model explained how much variability in tick absence and presence data and obtained how much improvement when compared to the null model.

2.8.2 Validation

After fitting and selecting the suitable model, the performance assessment step was conducted with calculating confusion matrix.

Owing to the small size of the data set, no independent validation data set (tick data) could be afforded. Therefore, the same tick dataset that was used for training was also used to compute the confusion matrix. Logistic regression model was applied to do prediction at sites where tick data located based on the data of significant predictor. Then the predicted tick occurrence and raw tick data (used again as reference data) were used to make confusion matrix. This step was performed to validate the selected logistic regression model and assess its performance.

As the predicted value of selected logistic regression model was probability between 0 and 1, a function in R was wrote to find the threshold which defines the possibility boundary separating predicted tick absence and tick presence. The function was designed to choose the threshold which maximizes the overall accuracy of confusion matrix of predicted tick occurrence and reference tick data.

After obtaining the threshold, predicted tick absence and presence were determined. The same tick data set was applied in accuracy assessment and consequently, the calculated accuracy is too optimistic.

To understand more about the relationship strength and performance of selected logistic regression model, except the confusion matrix, further study i.e. Goodman and Kruskal's tau-y, about the association of the significant predictor and tick occurrence was conducted.

Goodman and Kruskal's tau-y is a measure of association for nominal variables. It is defined as as follows:

Equation 6. Goodman and Kruskal's tau-y

$$tau_y = \frac{E_1 - E_2}{E_1}$$

where E_1 is the total expected cases of random assignment of cases to categories of the dependent variable, and E_2 is given the specific category of the independent variable, the total expected cases of mis-classification of cases into the dependent variable's categories (Lee 2002). Then E_1 and E_2 can be computed by following equations (Lee 2002):

Equation 7. Detailed formula in Goodman and Kruskal's tau-y

$$E_1 = \sum_{i=1}^{k} \left[\frac{N - f_i}{N} (f_i) \right]$$

where N is the total sample size, f_i is the frequency in the i th category of the dependent variable, and k is the number of categories of the dependent variable.

$$E_2 = \sum_{j=1}^{c} \sum_{i=1}^{k} \left[\frac{N_j - n_i}{N_j} (n_i) \right]$$

where n_i is the cell frequency in the i th category of the dependent variable within one of the c categories of the independent variable and N_j is the total (marginal) frequency for that category of the independent variable.

Here the dependent variable is the raw tick data absence and presence and based on this E_1 was calculated. The independent variable is the vegetation structure classes at tick data locations. Accordingly, E_2 was computed then. Goodman and Kruskal's tau-y of vegetation structure class and tick occurrence was obtained.

Codes in R were programmed to calculate confusion matrix and Goodman and Kruskal's tau-y.

2.8.3 Using the model for predicting tick occurrences

Finally, the fitted model with the significant predictors was used to make a map of the probability of tick occurrence using the packages 'raster', 'sp' and 'rgdal' in R. The predicted tick occurrence possibility map was again at 0.5m resolution. The threshold determined in the previous step was applied to the probabilities to label presence and absence as well. The resulting map was considered to represent the hot-spot map of tick presence.

3. Results

This chapter firstly introduces the vegetation structure classification results obtained by the *k*-nearest neighbour method and CART classification and the accuracy assessment results of both of them. Next, it presents the association analysis results between vegetation structure classes, other environmental factors and tick occurrence data in Amsterdam. Finally, the model performance evaluation and predicted map are presented.

3.1 Vegetation structure classification results

3.1.1 k-Nearest Neighbour results

The k-Nearest Neighbour classification results show the grass, shrub, tree probability distribution in the Amsterdam area. As an example, an area of one square kilometre (123000, 484000) was selected to illustrate the probabilities of grass, shrubs, and trees in Figures 9, 10 and 11, respectively. Figure 12 shows the aerial photo of this area. The 'hard classification' (label pixels according to their maximum probability) of three vegetation structure classes is shown in Figure 13.

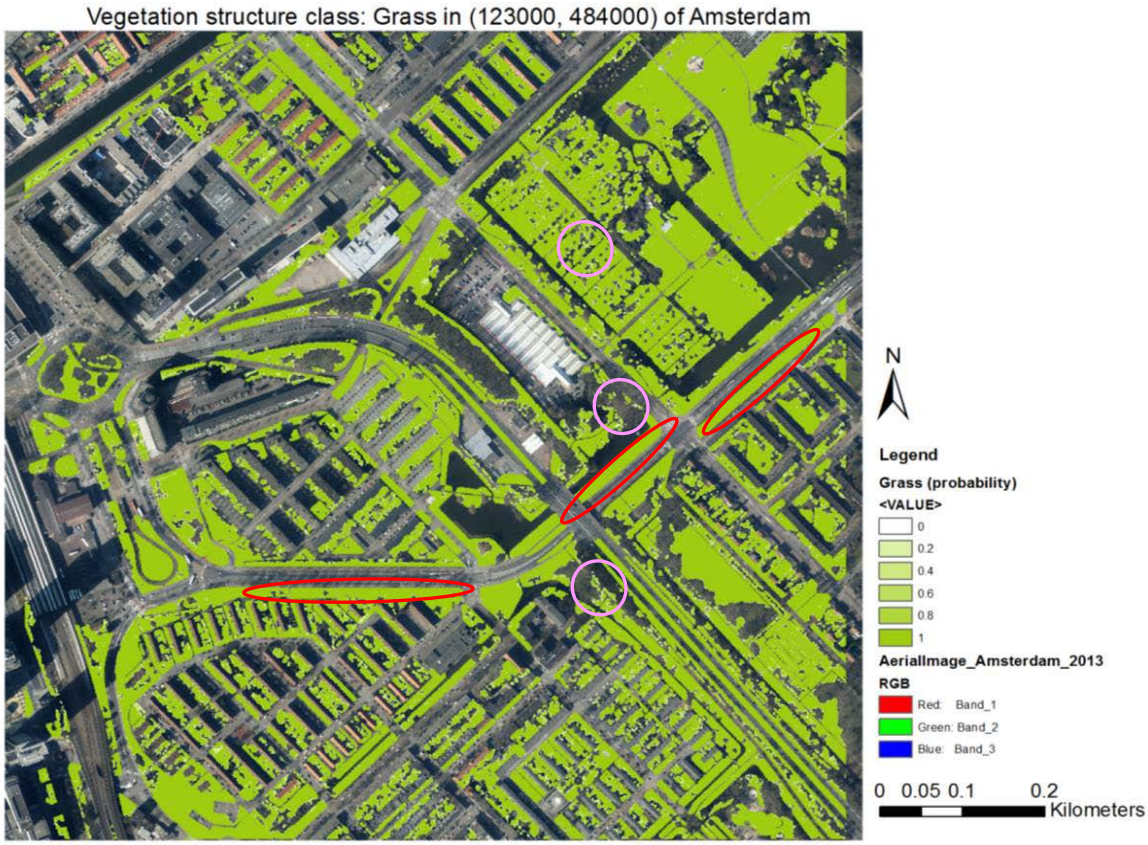


Figure 9. kNN result: Grass probability distribution in part of the Amsterdam area. When the probability=0, the colour is transparent; the pink circle areas indicate there are some obvious lower possibility grass patches; the red ellipse areas indicate some tree or shrub patches along the streets which were mistaken as grass.

Figure 9 denotes that there is large area of grass in this region, located around the buildings, or along the streets, or in the park, etc., which matching with mostly what we see from aerial photo (Figure 12). Lower possibility grass patches (probability <=0.4) are located close to where shrub/tree other vegetation grow, which can be observed from Figure 9 (pink circle areas indicate there are some obvious lower possibility grass patches). Besides, some patches of the trees or shrubs along the streets are mistaken as grass class (red ellipse areas) in Figure 9.

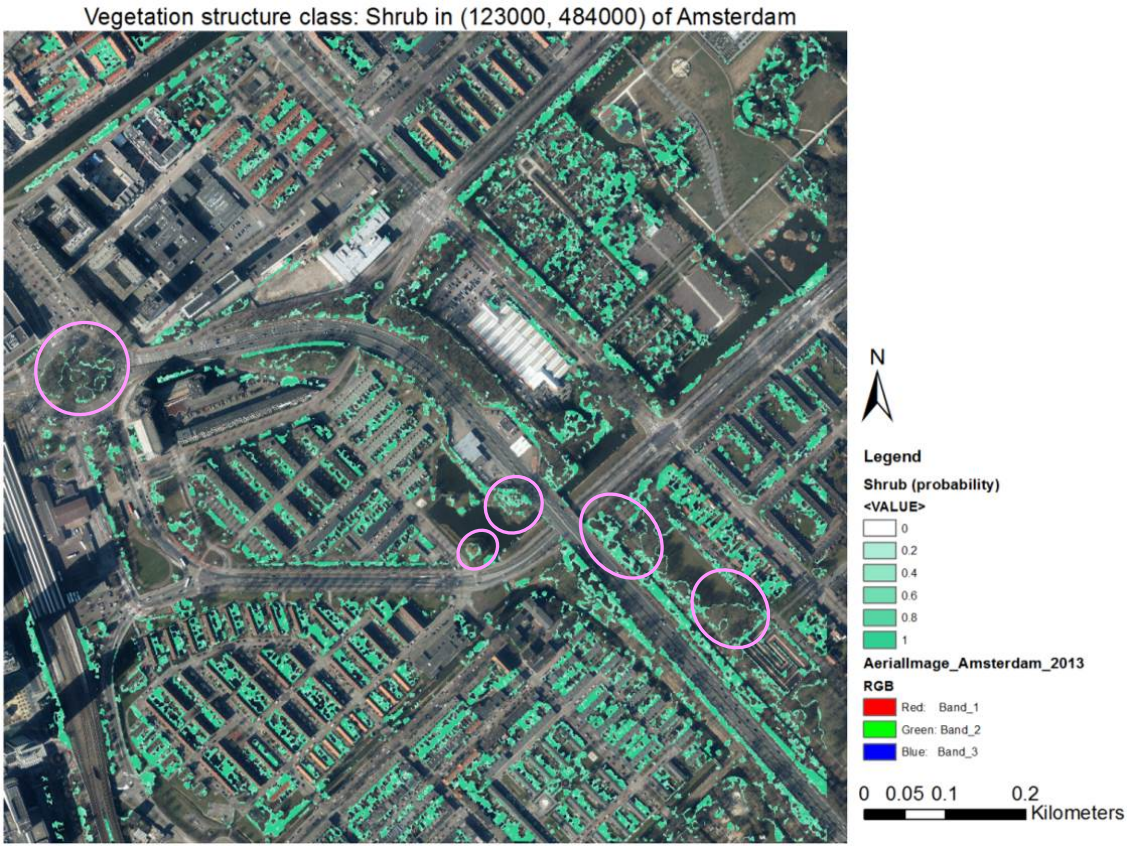


Figure 10. kNN result: Shrub probability distribution in part of the Amsterdam area. When the probability=0, the colour is transparent; the pink circle areas indicate there are some lower probability shrub pixels surrounding tree crowns or mixing with trees)

Figure 10 suggests there are many shrubs in this region. We can find some lower probability shrub patches (probability <=0.4) present surrounding the tree crowns, or mixing with trees (pink circle areas in Figure 10).

It is shown in Figure 11 that most tree patches with high probability (probability =1) present along the roads or in the park in this region. Some lower probability tree patches (probability <=0.4) can be clearly observed in the pink ellipse areas in Figure 11. Besides, in the red circle areas in Figure 11, from the background aerial photos we also can see tree crown shapes and the long shadow of the trees while these areas have only a small number of pixels with high probability for the tree class.

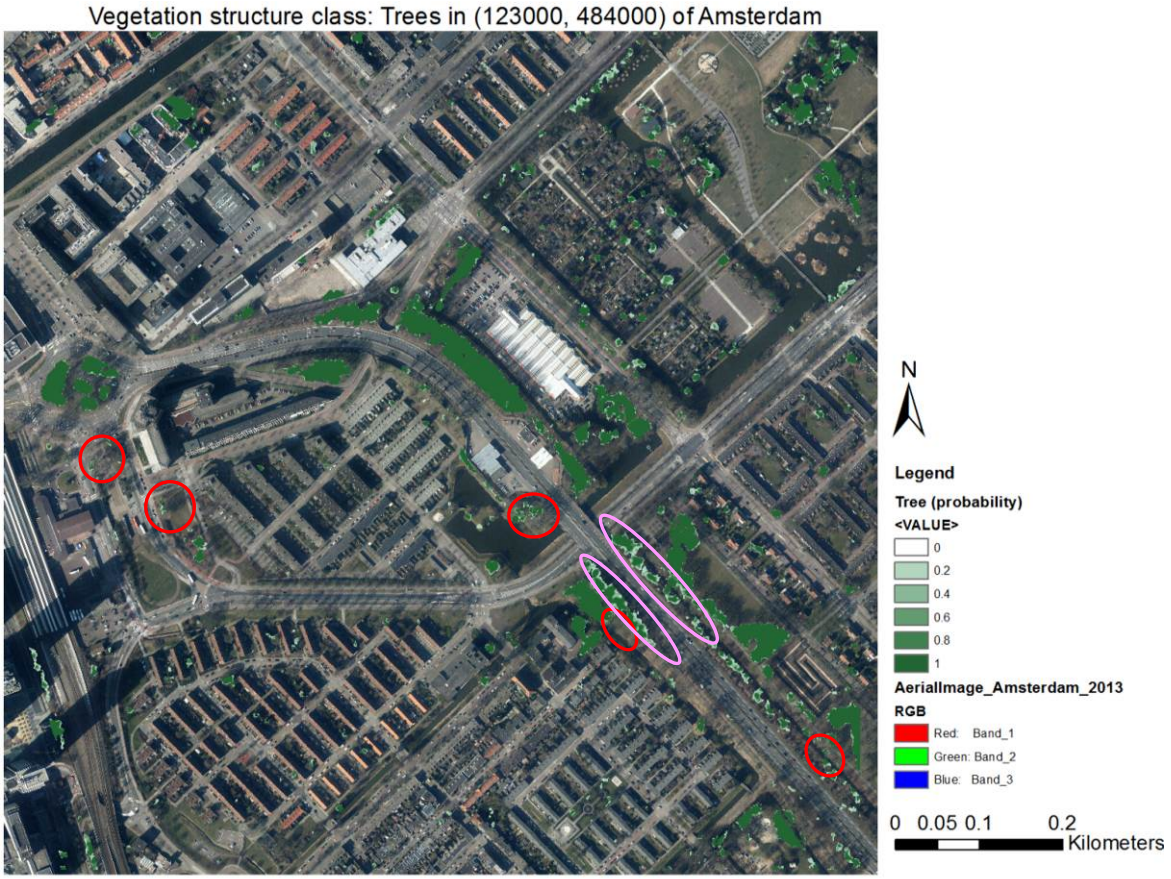


Figure 11. kNN result: Tree probability distribution in part of the Amsterdam area. When the probability=0, the colour is transparent; the pink ellipse areas indicate there are some lower probability tree patches; in the red circle areas, the map shows less tree class patches than expected on the basis of visual interpretation.

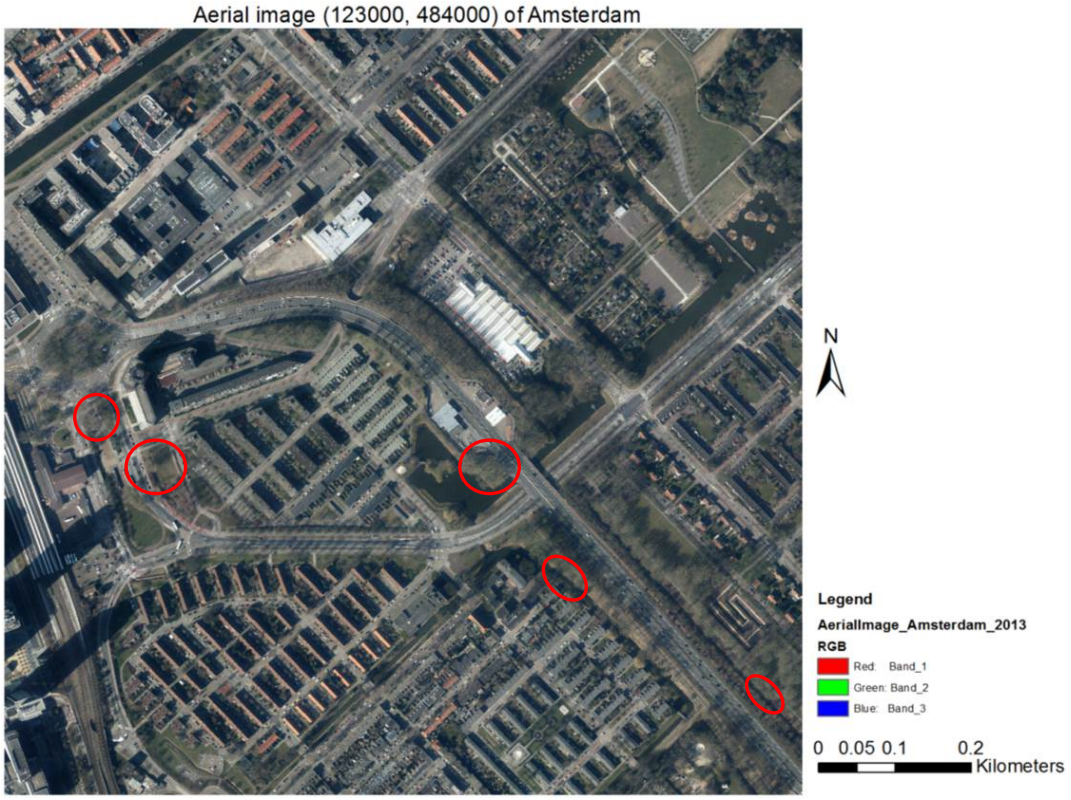


Figure 12. Aerial photo of this area

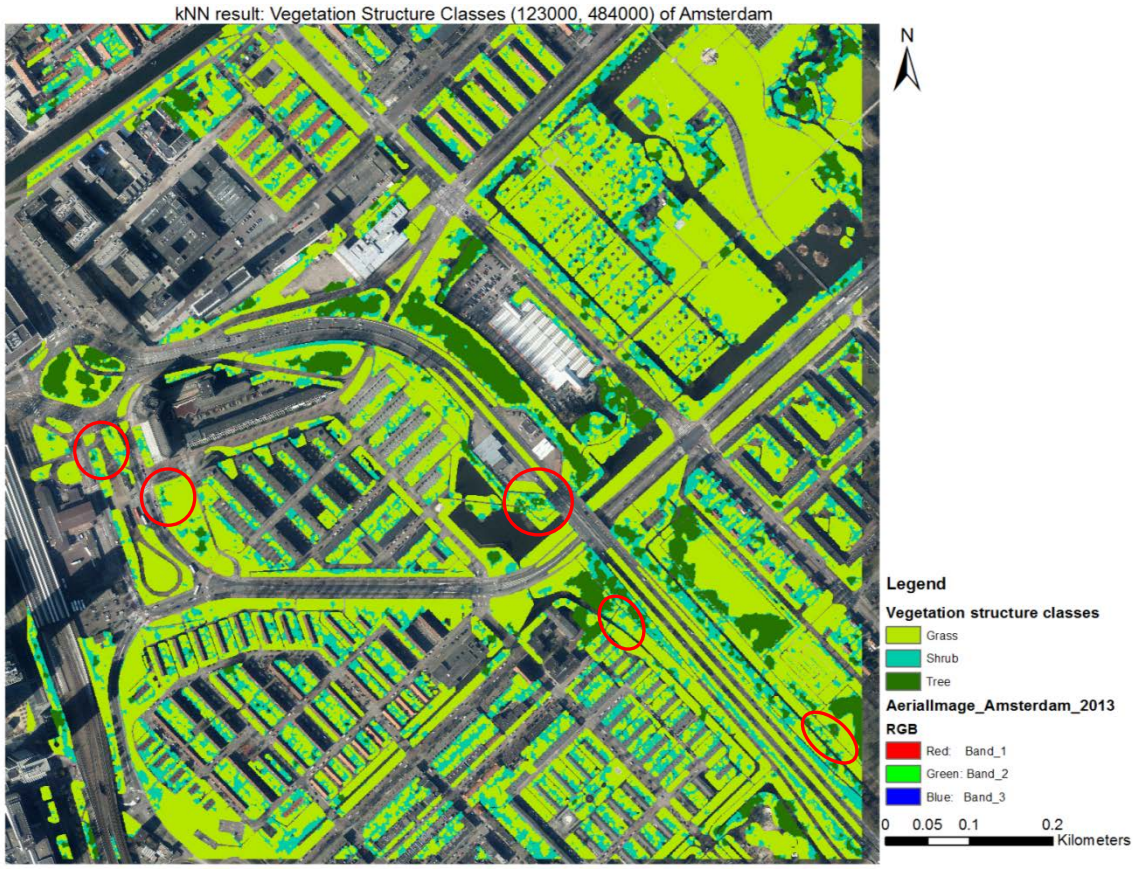


Figure 13. kNN result: vegetation structure classes distribution in part of the Amsterdam area ('hard' classification). Pixels were labelled according to the maximum probability. In the red circle areas, the

map shows less tree class patches than expected on the basis of visual interpretation of aerial photo and Figure 11.

The k-Nearest Neighbour classification result map shows not only the grass, shrub, tree possible distribution but also the possible mixture patches of vegetation class (Figure 14). The pixels with more than 1 class probability indicates mixture vegetation structure classes patches. In a patch with probability *p* for tree, the expected fraction of tree is *p*. Figure 14 shows the calculated entropy bits in part of Amsterdam representing mixture patches of vegetation structure classes.

In Figure 14, light yellow area (entropy value 0) means there are no vegetation or there is one certain kind of vegetation structure class (possibility = 1). Higher entropy value (>0) indicates there are mixture vegetation structure classes. Lager entropy value expresses greater degree of mixture vegetation structure classes. The pink circle areas in Figure 14 show some obvious high degree mixed vegetation classes patches.

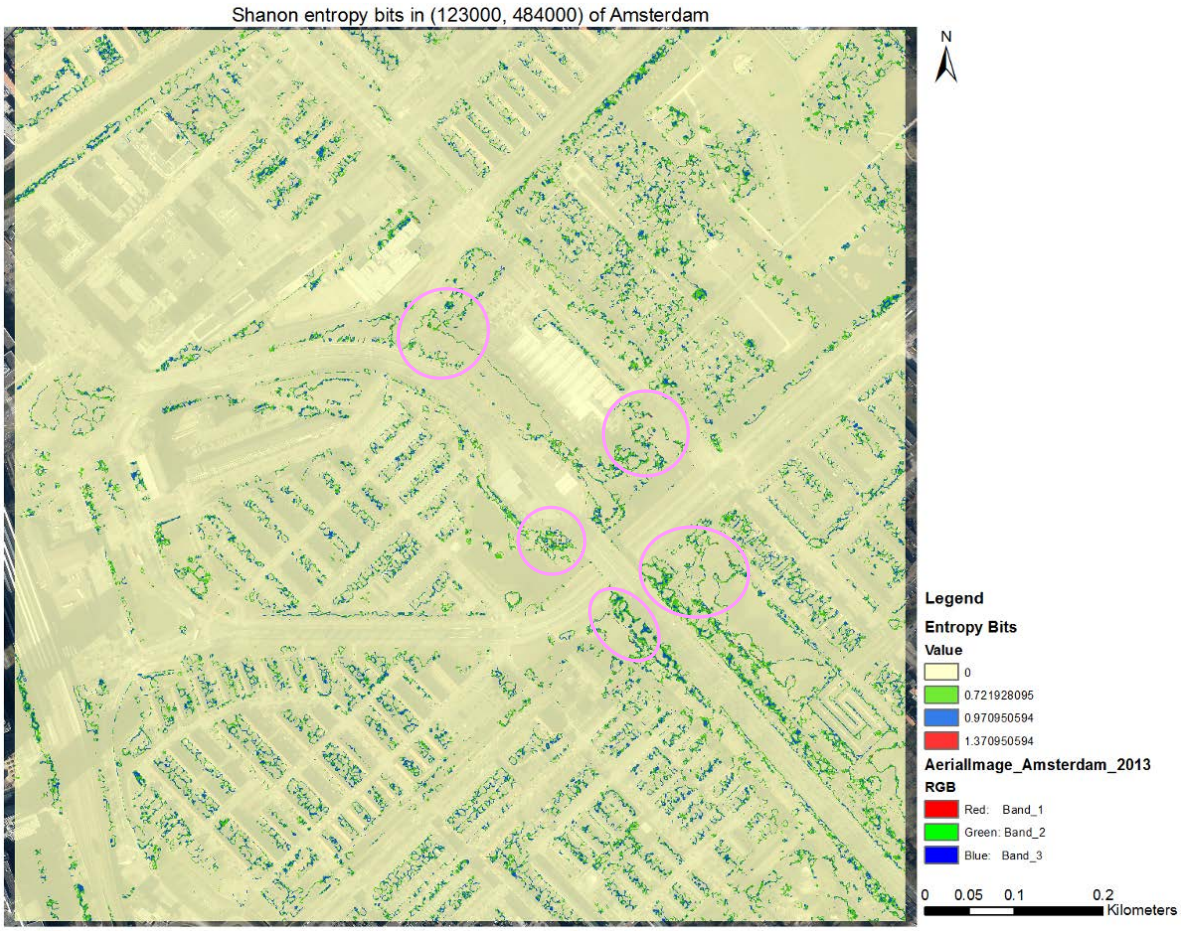
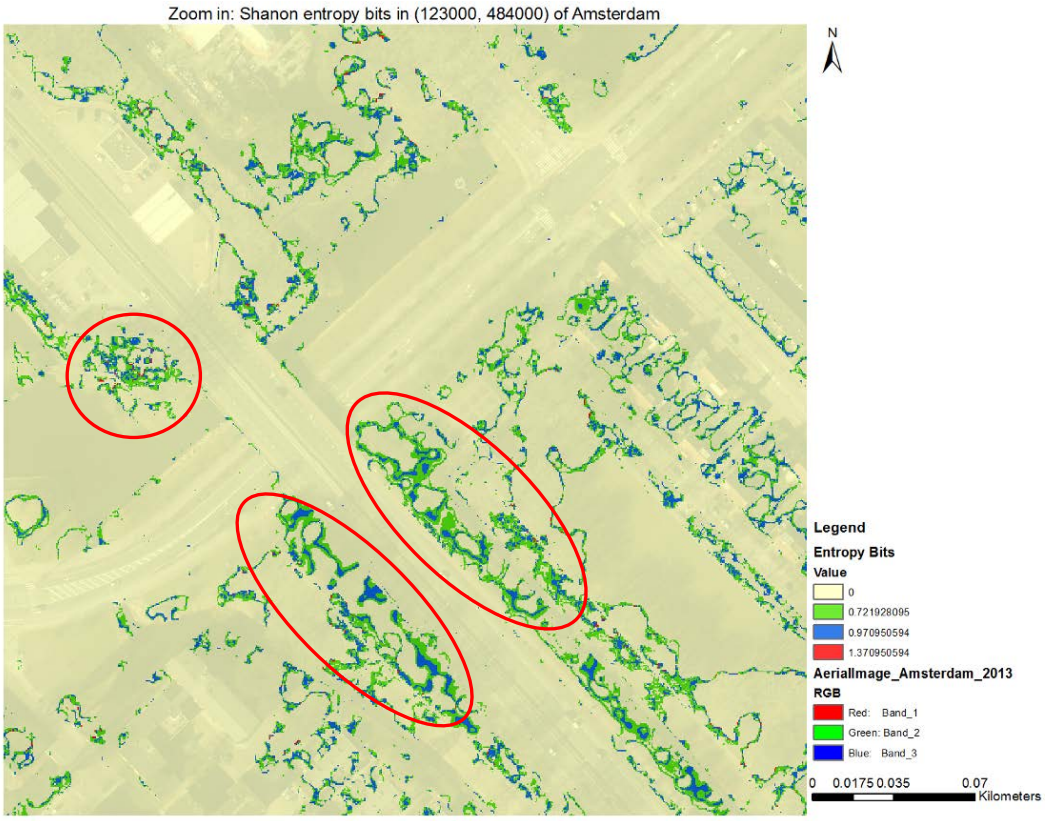


Figure 14. kNN result: Entropy bits (mixture classes patches) in part of the Amsterdam. Light yellow area (entropy value 0) refers there are no vegetation or there is one certain kind of vegetation structure class (possibility = 1). Higher entropy value (>0) indicates there are mixture vegetation structure classes. Lager entropy value expresses greater degree of mixture vegetation structure classes. The pink circle areas show some obvious high degree mixed vegetation classes patches.

If we zoom into one part showing obvious mixture pattern (Figure 15), it can be observed that the mixture patches are located between boundary area of trees and shrub near the roads.

(a)



(b)

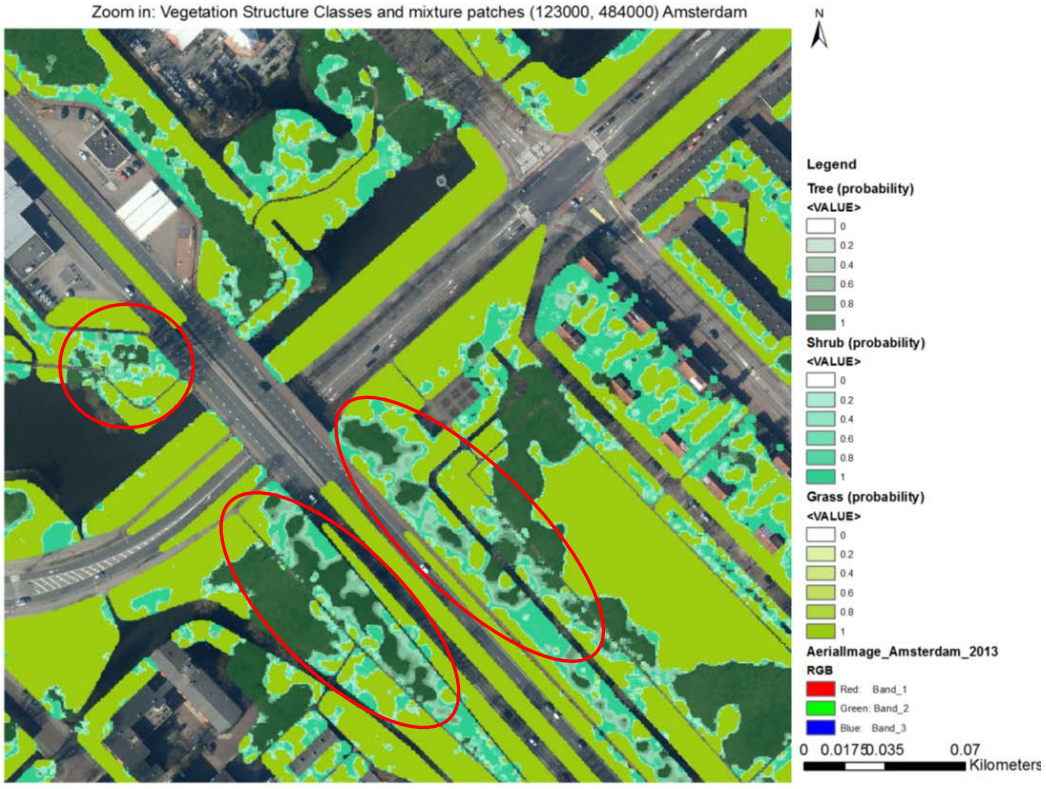


Figure 15. Zoom-in on an area showing patches with a mixture of classes. (a) is the entropy bits representing mixture patches and (b) is the three vegetation structure classes probability distribution. The red circle areas show mixed tree patches and shrub patches; since the grass layer is in the bottom, the low possibility grass patches can't be seen.

The confusion matrix of *k*-Nearest Neighbour classification result is shown in Table 1. Values in Table 1 represent summation of probabilities of pixels belonging to one class (accumulated fractional coverage).

Table 1. Confusion Matrix of k-*Nearest Neighbour classification map (probability map)*

		Prediction(classification result probability map)					
		Grass(band 1)	**Shrub(band 2)**	**Tree(band 3)**	**Zero(0)**	**NA**	Total
Reference classes	**Grass**	299618.8	420.8	0.4	2447	851	=303338
	Shrub	41736.2	28149.4	64.4	840	30	=70820
	Tree	77402.2	37934.4	219518.4	2679	156	=337690
	Total	=418757.2	=66504.6	=219583.2	=5966	=1037	=711848

The overall classification accuracy is 76.88%. In addition, Table 2 presents the accuracy (producer's accuracy) and reliability (user's accuracy) of each class.

Table 2. Accuracy and reliability of each class of k-*Nearest Neighbour classification map (probability map)*

	Accuracy (producer's accuracy)	**Omission Error**	**Commission Error**	**Reliability (user's accuracy)**
Grass	98.77%	1.23%	28.45%	71.55%
Shrub	39.75%	60.25%	57.67%	42.33%
Tree	65.01%	34.99%	0.03%	99.97%

The grass class shows very good accuracy (98.77%) and has 71.55% reliability. The omission error for grass is a mere 1.2%. The commission error indicates a pixel assigned to grass class has 28.45% probability belonging to shrub/tree class on the ground. Especially a pixel assigned to grass class has more than 18% possibility is actually tree class (column of grass class in Table 1).

The tree class has a high reliability approaching 100% and 65.01% accuracy. The predicted tree pixels almost all are real tree pixels (column of 'tree' class in Table 1). On the other hand, part of 'true' tree pixels were omitted (34.99%) and wrongly classified to the grass (22.92%) or shrub class (11.23%) (row of tree class in Table 1).

The shrubs class was mapped with lowest accuracy 39.75% and 42.33% reliability. It can be observed that many 'true' shrub pixels were omitted and misclassified having probability of grass (58.93%) (row of shrub class in Table 1). A pixel classified as having probability of shrub class has about 57% chance being actually tree class (column of shrub class in Table 1).

The NA (no data value) in Table 1 originate from the focal calculation edge effect (only at the edge of the raster) and this is discussed in the later discussion chapter.

As stated before in the method chapter, if all cells of the 5x5 window had no data value, then the returned value was a zero indicating that there is no vegetation growing. The zero value in Table 1 has two kinds of sources: original AHN2 terrain grid NA values and NA from wrongly drawing a part of a training area boundary inside water or built-up area. The AHN2 terrain data are claimed

to cover sites with buildings, trees and water but when the AHN2 terrain grid was used in this study, it was found that in some of these cases there was no data recorded in the data set.

Through examination with the ArcGIS measuring tool and checking in R, it was found that in the majority of cases, the zero vegetation values were due to lacking AHN2 terrain data.

3.1.2 CART results

Table 3 and Figure 16 show the complexity parameter (cp) values obtained during training of the classification tree.

Table 3. The complexity parameter table (different cp values and their corresponding error measure numbers)

	complexity parameter	number of splits	relative error	cross validation error	one standard deviation
	CP	nsplit	rel_error	xerror	xstd
1	7.352224E-01	0	1.0000000	1.0000000	0.0003803063
2	4.213878E-02	1	0.2647776	0.2647801	0.0002674022
3	8.542086E-03	3	0.1805000	0.1805161	0.0002265904
4	5.294082E-03	4	0.1719579	0.1719847	0.0002217362
5	1.230768E-03	6	0.1613697	0.1614035	0.0002154839
6	7.363487E-04	8	0.1589082	0.1589858	0.0002140172
7	1.933861E-04	9	0.1581719	0.1582494	0.0002135675
8	1.005545E-04	11	0.1577851	0.1578787	0.0002133406
9	7.439139E-05	14	0.1574834	0.1574916	0.0002131033
10	6.808704E-05	15	0.1574090	0.1574576	0.0002130825
11	5.000000E-05	16	0.1573409	0.1573674	0.0002130271

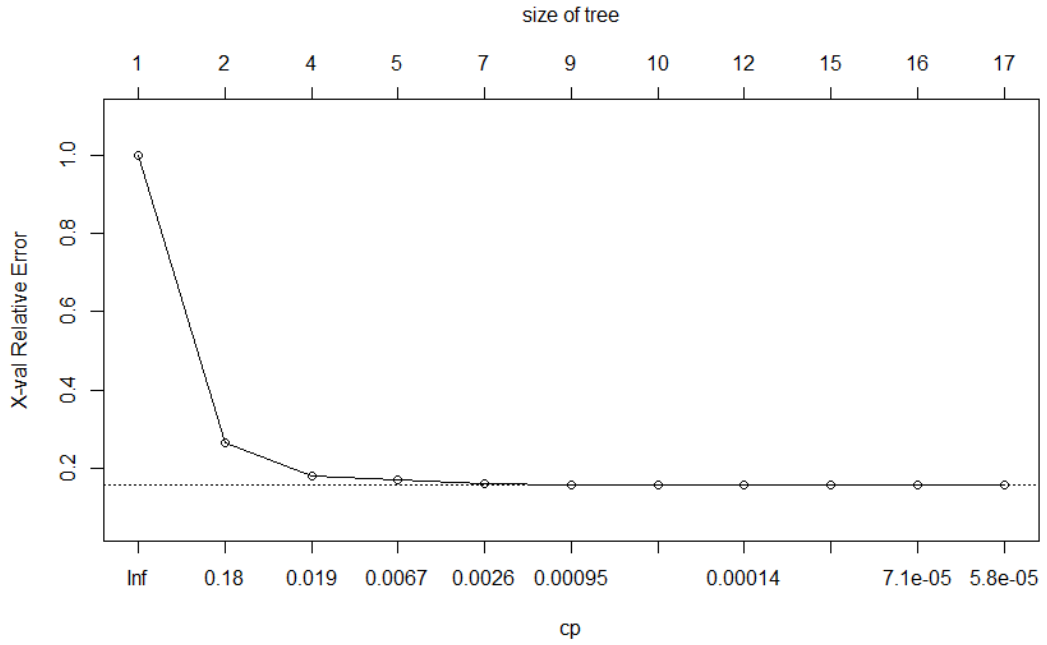


Figure 16. The plot of cp table. 0.0053 cp value corresponding cross validation error (x-val Relative Error) is almost on the flat plateau.

Checking the splitting rules of the classification trees, it was found that with a cp value smaller than 0.0053, some weird splitting rules were produced in the classification tree. In addition, in Figure 16, the cross validation error (x-val Relative Error) corresponding cp=0.0053 is almost on the flat plateau. Therefore, it was considered choosing 0.0053 as the suitable cp value. To further check the cp value suitability, the confusion matrices of classification trees of cp=0.01 and 0.005 were computed respectively based on the training data 'classes' and prediction by the two classification trees done to training data. The overall classification accuracy was 90.13% with cp=0.01 and 91.16% with cp=0.005. The 0.0053 cp value was thought a good choice.

The trained classification tree was acquired and is shown in Figure 17 (Appendix C. show a large figure of the same tree).

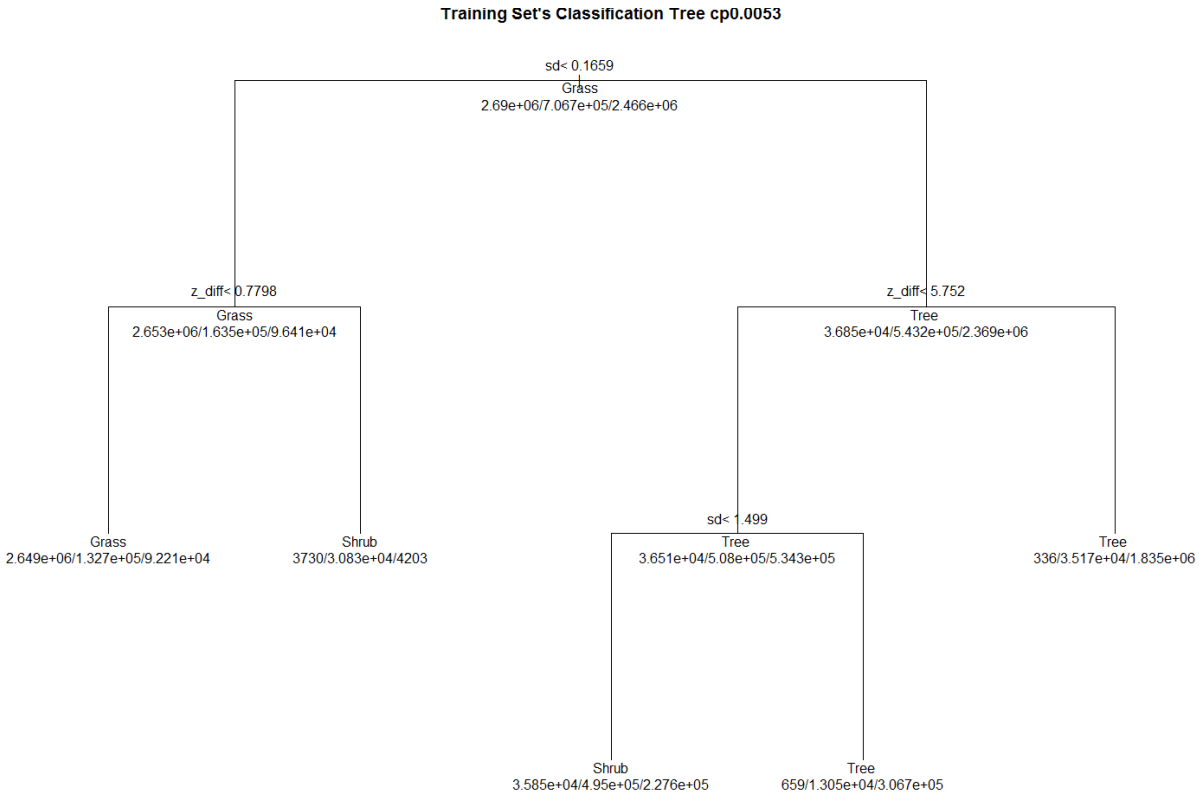


Figure 17. The trained classification tree for cp = 0.0053. Splitting rules are shown above the splits (lines). 'z_diff' refers to height value and 'sd' is the abbreviation for standard deviation of height. If data satisfies the rule, it goes to the right branch- meeting another splitting rule or becoming a final leaf node. At each leaf we obtain the predicted class for the members of that leaf using simple majority rule. Below the prediction class, the actual numbers of observations in the grass, shrubs, trees three categories are also displayed.

The selected trained classification tree was applied to classify whole vegetation area and then classification maps were acquired.

As an example, a square kilometre area of the CART results presenting grass, shrub, and tree distribution is displayed in Figure 18. Compared to Figures 11 and 13, the tree class in the CART result seems to cover a larger area than in the kNN result (explicitly in red circle areas). The tree class distribution in Figure 18 seems to match better with the aerial image shown in Figure 12.

The zoom-in of Figure 19 confirms more tree class area (in red circle areas) and more shrub class area (in red ellipse area) mapped by CART than by the *k*-nearest neighbour method (see Figure 15b).

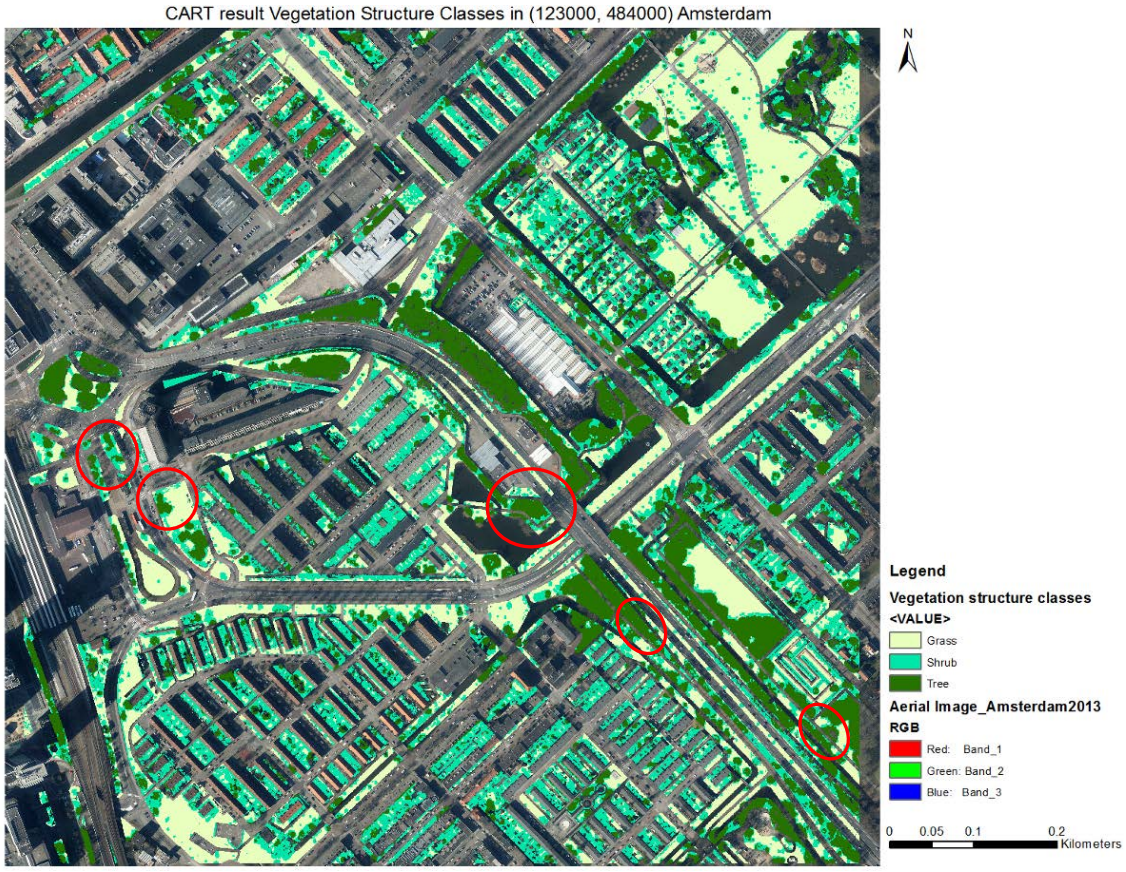


Figure 18. CART result vegetation structure classification map for part of Amsterdam. The red circles indicate locations where the tree class has a larger area than in the kNN result (Figure 11 and 13).

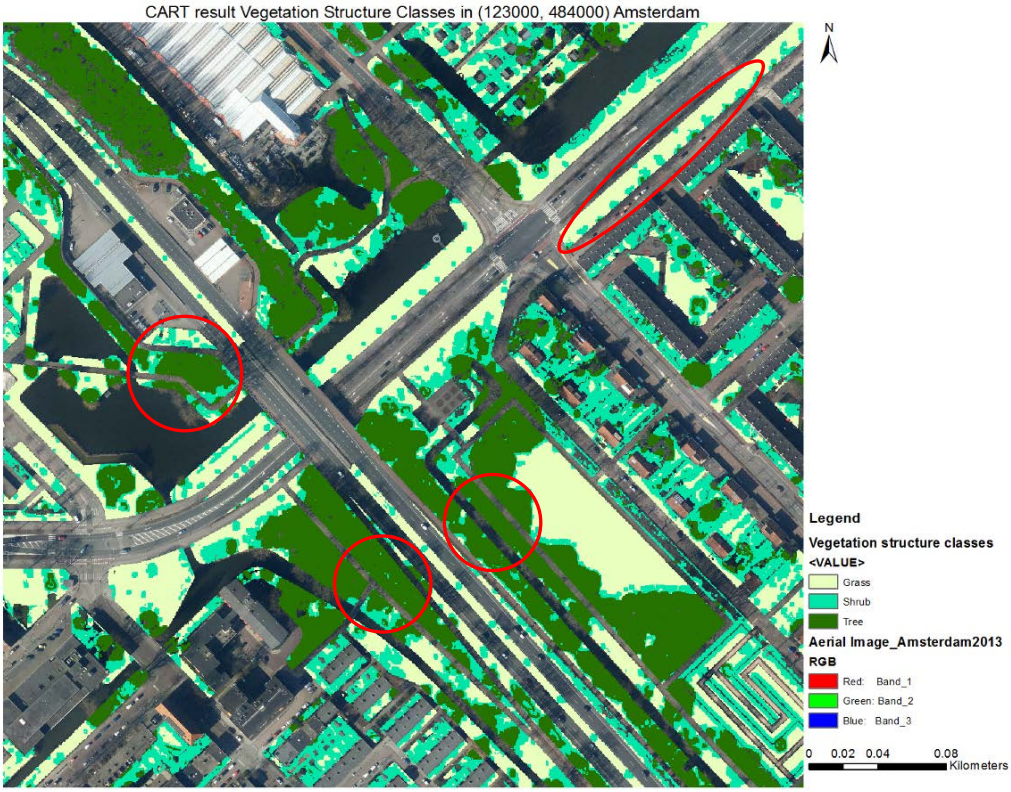


Figure 19. Zoom in CART result vegetation map for part of Amsterdam. There is more tree class area (in red circle areas) and shrub class area (in red ellipse area) mapped in the CART result than k-nearest neighbour result zoom-in figure (Figure 15b).

Table 4 shows the confusion matrix of CART result vegetation structure map and Table 5 presents the accuracy (producer's accuracy) and reliability (user's accuracy) of each class.

Table 4. Confusion matrix of CART result vegetation structure map (assigned class to pixels)

		Prediction(classification result map, sum of pixels)				
		Grass(1)	**Shrub(2)**	**Tree(3)**	**NA**	Total
Reference classes	**Grass**	346881	4445	923	6806	=359055
	Shrub	12523	44353	3442	5585	=65903
	Tree	13419	29882	314540	38737	=396578
	Total	=372823	=78680	=318905	=51128	=821536

Table 5. Accuracy and reliability of each class in CART classification map

	Accuracy (producer's accuracy)	**Omission Error**	**Commission Error**	**Reliability (user's accuracy)**
Grass	96.61%	3.39%	6.96%	93.04%
Shrub	67.30%	32.70%	43.63%	56.37%
Tree	79.31%	20.69%	1.37%	98.63%

The overall classification accuracy of CART result vegetation structure map is 85.91%.

The grass class has good accuracy 96.61% as well as a high reliability 93.04%. Few true grass pixels were omitted and wrongly predicted as belonging to the shrub class or tree class (row of grass class in Table 4).

The tree class has high reliability 98.63%. Tree category has an accuracy of 79.31%; 20.69% of the reference tree class pixels were wrongly assigned to the shrub or grass category (row of tree class in Table 4).

The shrub class has 67.30% accuracy and 56.37% reliability. A large part of reference shrub pixels were correctly mapped. Some true shrub pixels were wrongly classified to the grass or tree category (row of shrub class in Table 4). Predicted shrub pixels have some possibility being the actual tree class (37.98%) or grass class (5.65%) on the ground (column of shrub class in Table 4).

The NA (no data) values in the Table 4 have three sources: (1) the focal calculation edge effect, (2) NA values in the AHN2 terrain grid and (3) mistaken drawing training area boundary into water area/built-up area. The focal calculation edge effect can be solved in a later study as discussed in the discussion chapter.

By checking with the ArcGIS measuring tool and R, it was found that in most cases, NA values were obtained because of lacking terrain elevation values in the AHN2 terrain grid.

3.2 Relationship between vegetation classes and tick occurrence in Amsterdam

Compared with the *k*-Nearest Neighbour classification maps, the CART result vegetation structure class maps had better accuracy and reliability. Thus the CART result maps were used in the logistic regression.

The logistic model fitting was first conducted with vegetation structure classes, distance to water, distance to built-up area, and tick occurrence data. From the summary output of the model (details in Appendix A) in R, we got the information in following Table 6.

Table 6. The coefficients and AIC of original logistic regression model using environmental factors and tick occurrence data

Predictor Variables	Coefficients			
	Estimate	Standard Error	z value Wald z-statistic	The associated p-values
(Intercept)	-4.103219	0.809621	-5.068	4.02e-07
VegeClass2: shrub	1.806852	0.798196	2.264	0.02359
VegeClass3: tree	2.043076	0.685515	2.980	0.00288
Distance to water	0.004909	0.007982	0.615	0.53858
Distance to Built-up area	0.002856	0.008688	0.329	0.74235
AIC of the model	125.18			

From Table 6, we know the two environmental factors distance to water area and distance to built-up area are not significant (p-value > 0.05). The statistically significant predictors of tick occurrence are the terms for vegetation structure classes (p-value < 0.05). Also stepwise regression was conducted to choose a model with lowest AIC value.

After stepwise regression, the final model with smallest AIC value 121.67 was selected. Information of this model is shown in Table 7 (details in Appendix B).

Table 7. The information of final selected logistic regression model with smallest AIC

Predictor Variables	Coefficients			
	Estimate	Standard Error	z value Wald z-statistic	The associated p-values
(Intercept)	-3.7297	0.5842	-6.384	1.73e-10
VegeClass2: shrub	1.7487	0.7911	2.211	0.02707
VegeClass3: tree	1.9532	0.6690	2.919	0.00351
AIC of the model	121.67			

The selected model with smallest AIC value contains only vegetation structure classes (grass, shrub, and tree) this variable. The shrub and tree classes are sufficient. The shrub, tree and grass class are mutually exclusive (that is why grass class is not shown in the Table 7). The three classes together made up of the predictor vegetation structure classes. From Table 7, distance to water and distance to built-up area were not chosen in the lowest AIC model because they are not statistically significant.

McFadden's R squared measure of the final selected model is 0.092. It indicates that only a small part of the tick absence and presence was explained by the fitted model. The selected model with significant predictor vegetation structure classes had small part improvement from the null model.

A threshold probability which defines the boundary between predicted tick absence and tick presence was computed from the reference data. This cut-off maximizing the overall accuracy of tick data was established at 0.098. The confusion matrix of tick data was obtained after applying this threshold.

Table 8 is the computed confusion matrix with predicted tick occurrence and reference tick data. Table 9 shows the accuracy and reliability of predicted tick absence and presence. The calculated accuracy and reliability numbers are actually little optimistic due to the tick data which applied in the logistic regression model training was used again as reference data in validation (making confusion matrix).

The overall accuracy is 59.07% based on the confusion matrix of Table 8.

Table 8. The confusion matrix of predicted tick occurrence and reference tick occurrence

		Prediction(tick occurrence)		
		Absence(0)	Presence(1)	Total
Reference tick occurrence	Absence(0)	125	94	=219
	Presence(1)	3	15	=18
	Total	=128	=109	=237

Table 9. The accuracy and reliability of predicted tick absence and presence

	Accuracy (producer's accuracy)	Omission Error	Commission Error	Reliability (user's accuracy)
Absence(0)	57.08%	42.92%	2.34%	97.66%
Presence(1)	83.33%	16.67%	86.24%	13.76%

The predicted tick absence has 57.08% accuracy with reliability 97.66%. Many reference absences were left out (row of absence in Table 8) while predicted tick absences are reliable (column of absence in Table 8).

The tick presence has high accuracy 83.33% with low reliability 13.76%. Most reference tick presences are predicted correctly. Among all the predicted presence, a large number of presence predictions corresponded with absence on the ground (column of presence in Table 8).

Goodman and Kruskal's tau-y was calculated to amount to 0.047. It shows that the selected fitted model with the predictor vegetation structure classes has only marginally improved the prediction power when compared to randomly assigning the tick absence and presence classes (in the right proportion). So there is a weak association between vegetation structure classes and tick occurrence in Amsterdam. On the other hand, distance to water and distance to built-up area were not selected in the final logistic regression model. Distance to water and distance to built-up area were not found to be significantly related to tick absence and presence in the Amsterdam area.

Figure 20. Predicted tick occurrence map in part of the Amsterdam area

Using the final selected model with the predictor vegetation structure classes, and the probability threshold 0.098, the predicted tick occurrence maps were acquired and Figure 20 presents part of the predicted potential tick occurrence map. We can observe that the predicted tick presences are located where shrubs or trees grow while predicted tick absences are at sites with grass.

4. Discussion

The vegetation structure classification results of *k*-Nearest Neighbour and CART showed that supervised classification with the two methods can be applied to map vegetation structure classes, and can have good accuracy and reliability especially for grass and tree class. The shrub class need more attention no matter in selecting training data step or doing classification step. Many factors can influence the accuracy and reliability of the classification result, especially for shrubs class, for example reference data quality, support of calculating variables which represent vegetation structural features, variability in features of vegetation structure, and so on. Details are discussed in the following sections. Suggestions for improving classification accuracy and reliability in later study are provided.

The logistic regression analysis result indicated weak association between vegetation structure classes and tick occurrence in Amsterdam. Interpretation of accuracy of predicted tick occurrence is presented in the following sections as well suggestions for future study.

4.1 Accuracy of reference data

In this study, different datasets and their production time influenced the accuracy of selected grass, shrubs and trees training sites. Particularly the shrubs class was affected.

The used aerial image is from March 2013. Grey and dark colour of the vegetation objects influenced interpreting where the shrubs and trees are growing. Google Earth and Google street view have some historical satellite images and pictures while these photos are in different seasons in 2013 or 2014. The pictures were usually taken in the summer days for grass or tree areas. Some areas of Amsterdam have no detailed pictures showing what kinds of vegetation growing at that site in 2013. During selecting the training area of shrubs, it happened that from Google street view there were pictures showing the site has tall shrubs growing; while after extracting the site polygon from the SGM point data, height histogram of points in this site polygon showed there are only very short grass growing.

The SGM point clouds were created from March 2014 aerial images and it is in the spring time in which height of shrubs are not stable. This increased the difficulty to select the ‘true’ training areas of shrubs.

During the site visit of Amsterdam in January 2015 (section 2.6.2), several ‘true’ shrub sites were visited. When these sites were examined using topography data (KBKA10) and the SGM point cloud data, some sites were found to be partly not included in the KBKA10 topographic dataset, which means they are not known as ‘vegetation area’ and 1 site was found to have lacking terrain elevation values because of many NA values at that site in AHN2 terrain raster.

Besides, AHN2 terrain grids created many NA and zero values in vegetation height data.

AHN2 data was produced around winter of 2009 and spring time in 2010. The western part of Amsterdam and eastern part of Amsterdam of AHN2 data has different point density (van der Zon 2013). It was found that many locations where trees and shrubs growing have NA values in the filled AHN2 terrain grids and no consistency in the size of these NA ‘holes’ (patches). This lead NA values created in the vegetation area height data and influenced the height value distribution of some vegetation sites which only have small area of dominant vegetation structure class.

It was also found some negative values were produced by extracting AHN2 terrain height from SGM points representing surface elevation. This was because the two datasets have different produce time as well as different resolution and accuracy. These negative values were replaced by ‘zero’ in this study.

In other vegetation classes mapping researches, terrain model creation from AHN2 raw point clouds was conducted (Hantson, Kooistra, and Slim 2012); digital elevation model and interpolation were done in (Mücher et al. 2010; Kuilder 2012). In later study, it may help produce more solid vegetation object height data involving the step of interpolation on AHN2 terrain grid or creating terrain model from AHN2 point clouds. Further investigation is needed to find out whether SGM point cloud can provide data for producing vegetation height model of Amsterdam or not, since the SGM point clouds mostly show top outline of objects.

In this study both the training data and the validation data sets were not perfect. Probably better results could be obtained with the same methods but using better data sets. It may increase training data quality (select ‘true’ training data locations) if there were aerial images of the same time period with the SGM point cloud data.

4.2 k-Nearest Neighbour Result

In this study, the spatial support of computing quantiles in the training phase and classification phase didn’t agree with each other. The spatial support of the objects to be mapped, and that of the training data to represent the objects need to be treated with caution (Wood et al. 2012). In this research, the vegetation structure classes grass, shrub, and trees are not strictly defined in the training data polygon area. At first the polygon area of training data was selected with same area size (about 45mx45m) for grass, shrubs and trees (section2.6.1). During refinement of selecting the ‘true’ shrub training locations, shrub training data location area (size) were found have to be smaller to have the ‘true’ shrub patches. Then the area size of shrub training data locations gradually decreased in many trails.

At last the average area of 1 shrub training area was about 24mx24m and mean area of grass and tree training locations were about 45mx45m (section2.6.1). The 9 quantiles of height values used in the training phase were computed from such large area. At the predicting stage, a moving window of 5x5 was applied to compute quantiles of unknown vegetation patches and hence the spatial support is 2.5mx2.5m (section2.6.1). The differences in the area to calculate quantiles mean different numbers of height values and more important: different amount of variability in height within the support unit. This may cause in Table 1 and 2 reference pixels wrongly assigned into another class. The assumption of more than 24mx24m training data and 2.5mx2.5m unknown vegetation patches have same (similar) height variation did not work for all grass, shrubs and trees classes. From Table 1 and 2 we know this assumption for grass class may have worked because height variance of grass is small. As for shrubs and trees, it turns out that the height variances are large.

In future study, the focal window operation can be applied first in computing quantiles of shrubs, trees, grass training data (polygon area) and then same area size focal window operation should be employed on calculating quantiles of unknown vegetation patches at the prediction stage. So more height variance and same resolution height variance can be generalised in the training data and used in classifying vegetation.

It is not surprising that the shrub class was mapped with lowest accuracy (39.75%) and 42.33% reliability. Due to use the same shrub training data in accuracy assessment, the shrub accuracy and reliability were partly overestimated.

One main reason is probable the different spatial support mentioned above. Another possible explanation is that shrubs this class is an in-between class while trees and grass are clear extremes. Shrubs on the ground do have more variability in height, shape or other features. The tree, shrub and grass, they have different features in the patch size and spreads in height. Grass patches and tree patches are relatively homogeneous than shrub patches in height variance. Shrubs area certainly contained low vegetation patches. So part of reference shrub pixels were assigned have possibility of grass in Table 1. Similarly, low height part of reference tree pixels were found closest to shrub class based on 2.5mx2.5m window calculated quantiles.

Additionally, there were difficulties in selecting shrub training locations (section 4.1) and accordingly shrub training data quality was not very good.

Furthermore, in this study only best 30 quality shrub training areas were used. This number is small and the study area is large. It was found that number of training locations influences the classification accuracy especially using k-Nearest Neighbour method (Qian et al. 2014; Gao and Mas 2008). 30 shrub training locations maybe did not reflect the height spread feature of various real shrub patches. Similarly 30 tree training locations were also not enough.

In future research, more training locations need to be obtained for shrubs, trees and grass classes for generalising more variance of vegetation structure to improve classification accuracy.

From Table 1 and Table 2, we also know the grass class of k-Nearest Neighbour classification result shows very good accuracy (98.77%) and also has 71.55% reliability.

The good accuracy of grass class was acquired mainly because (1) variability of height and height spreads of grass is small (height from 0 to about 0.3m) as well as (2) the quantiles of all grass training data reflected well with the height variability of real grass patches. The quantiles of grass training locations didn't have strong variety no matter in 45mx45m area or 2.5mx2.5m area.

Besides, one probable reason for 28.45% grass commission error is in the training data of shrub and tree class, there were small patches of low height vegetation (<0.6m). The low height was reflected in the quantiles and since the focal window is 5x5 (2.5mx2.5m), when doing prediction, the low height vegetation patches was found having nearest neighbour 'grass'. It is also common in real life where grass patches occur between trees and shrubs.

The tree class has a high reliability (99.97%) and 65.01% accuracy (Table 2). Because the defined height of tree class is larger or equal to 6m, once the calculated quantiles of unknown vegetation have such large value, the nearest neighbour would only be tree class. Accordingly high reliability was obtained. On the other hand, the lower height values existed in the quantiles of tree class training data. When doing prediction, the lower height patches in tree training data was computed having nearest distance to 'grass' or 'shrub' class since 5x5 focal window size.

In other researches mapping vegetation structure (Kuilder 2012; Rajaei Najafabadi 2014), geometric feature of vegetation structure, and spectral features of vegetation or predefined various local vegetation structure classes increased classification accuracy. In this study, 9 quantiles were employed to represent height spreads one feature of vegetation structure. In later study, it can improve classification accuracy that adding more features showing more variability of

vegetation structure, like spectral feature or geometric feature. Predefining more than three local vegetation structure classes is also possible to reflect more variance of vegetation structure.

The NA values in Table 1 were from focal calculation edge effect. When moving window passed the edge area of raster, no data cells outside the raster were included in the moving window calculation and then NA values were left in the central cell of moving window result on the edge of result raster. The focal operation edge effect can be solved in later study by overlapping tiles of raster.

The zero values in Table 1 mostly came from NA 'holes' in the AHN2 terrain grid. As it mentioned in previous section 4.1, it can fill the NA values and may help produce more solid vegetation object height data that conducting interpolation on AHN2 terrain grid or creating terrain model from AHN2 point clouds. Some zero values were from wrongly drawing training area boundary into water area/built-up area and this can be solved by checking more carefully with water area/built-up area in topographic data when drawing training data boundaries or applying a (larger) inward buffer on selected reference sites.

4.3 CART results

CART classification performed well in general. However, since the same shrub training data were used to assess accuracy, the shrub class accuracy and reliability were overestimated.

From Tables 4 and 5, the grass class shows high accuracy 96.6% as well as high reliability 93%. The splitting rules based on height and SD of height for grass were distinct and worked well for grass. The tree class has high reliability 98.6% and it indicates the prediction based on 'large height value' and 'large SD value' well distinguished tree pixels with other class. About 21% reference tree class pixels were omitted and it was probably because the lower height parts of trees and very short vegetation pixels in the tree reference data have similar height value and SD value with the reference grass/shrub pixels. Likely, in the shrub reference data, some pixels have similar height value and SD value with grass/tree reference data. That is the mainly reason why the shrub class has only 67.30% accuracy and 56.37% reliability.

One way to reduce the omission and commission error is to improve the quality of three classes training data, making sure they only have pure area of real grass/shrub/tree class.

However in reality, trees and shrubs also have lower height part and may grow with grass patches together. Good quality training data cannot reduce all the omission and commission error. It would be better adding more features such as spectral features, or geometric features instead of only height feature of vegetation structure into this CART method, like researchers did in (Kuilder 2012; Rajaei Najafabadi 2014). More features can generalise more variability of vegetation structure classes. These features can be used like height and height spreads in this study.

In this study, weird splitting rule produced in classification tree pruning (section 3.1.2) probably was the result of overfitting. There were low height pixels in some reference shrub and tree data. These patches were also 'learned' by the complex classification tree and weird splitting rules were produced then. It was wisely decided selecting cp value as 0.0053 which had almost smallest cross validation error (Figure 16) and at the same time did not produce weird splitting rules (Figure 17). Except improving purity of training data and adding more features (discussed

above), in the later study, it is suggested that pruning and carefully selection of cp value should be always conducted when using CART to avoid overfitting. It maybe need consider using 'minsplit' and 'minbucket' two control parameters in rpart package (Therneau and Atkinson 2015) to control the number of observations in a tree node or other pruning techniques (Kohavi and Quinlan 2002) to get better classification tree for new training data in vegetation structure classification.

CART performed better than kNN method in this study (Tables 4 and 5). Probably one reason was that kNN used different 'spatial support' of computing quantiles of training data and focal window size in classification. On the other hand, CART applied same operation unit in training and classification. Height data was rasterized as 0.5m resolution grid and 5x5 focal window operation was conducted on height grid to acquire SD grid in training as well as classification process. It can bring better result if in kNN, the focal window operation can be applied firstly in training phase and then same area size focal window operation can be employed on calculating quantiles in classification process.

In Table 4, there are similar observation with kNN about NA values applying to CART. The NA values have three sources: focal operation edge effect, AHN2 terrain grid NA values and mistaken drawing training area boundary into water area/built-up area. The solutions have been discussed in the previous section.

In this research only height feature and spreads in height are used in CART to map vegetation structure classes. It already showed that the results have reasonable accuracy. For grass and tree classes, the accuracy and reliability can be concluded as 'high'. If more features from multi spectral bands data and other kinds of data involving into this kind of researches, it is very promising to obtain even higher accuracy and reliability result. As mapping vegetation structure classes can also bring other benefits to the local society, for instance, helping biodiversity studies (Wood et al. 2012), or urban ecology research (Nagase and Dunnett 2012; Inkiläinen et al. 2013), it is significant to conduct more studies in this direction.

4.4 Tick data, environmental factors and relationships

In the result of tick occurrence prediction, predicted tick absence is highly reliable (reliability more than 97%, Table 8 and Table 9). On the other hand, the predicted presence is not reliable (a large number of presence predictions corresponded with 'actual absence' on the ground, column of presence class in Table 8).

One possible reason is that the small number of presences in the tick data affected the result. The raw tick data contains 31 presence records in a total 282 records (section 2.2). Only 264 records including 24 presence records are within the area which has height data of vegetation. After accounting for NA values in the vegetation structure class map only 18 presence in total of 237 records remained (Table 8). The number of tick presence data is small and on the contrary the number of tick absence is 219. It is hard to predict presence of a sparse class with small commission error. The low reliability of predicted presence is reasonable.

McFadden's R squared measure (0.092) also indicated that tick occurrence in this case were not strongly related to the vegetation structure classes. Bartlett (2014) stated that one should really not be surprised if, from a fitted logistic regression McFadden's R^2 is not particularly large. Bartlett

(2014) has compared different predictors and only found the variable X which changed possibility of dependent variable Y from 0.01 and 0.99 can get the high McFadden's R^2 value 0.93. It means extremely strong predictors are required for McFadden's R squared to get close to 1. In this study, McFadden's R^2 value 0.092 has just confirmed vegetation structure classes affect tick occurrence to small extent. These influence are not strong. Goodman and Kruskal's tau-y of the logistic regression model is 0.047 (section 3.2) and it also illustrates weak relationship between vegetation structure classes and tick occurrence in Amsterdam. Accordingly using vegetation structure classes as the only predictor to predict tick occurrence did not perform well.

It may bring better prediction if we use other potential strong predictors, for instance temperature, humility, soil moisture, etc. (Estrada-Peña et al. 2013; Greenfield 2011).

In Figure 20 predicted tick occurrence map, predicted tick presences are located at where shrubs or trees grow and predicted tick absences are in the grass area. It is not meaningful for risk management of tick from practical view because there are large area of shrubs and trees in Amsterdam. It has just provided some idea in grass area tick seems has more chance to be absent.

It is not a surprise that ticks are predicted at where shrubs and trees grow. In other research (Boyard et al. 2007; Vatansever et al. 2008; Dobson et al. 2011) high tick densities were found in patches with tree present and dense heterogeneous vegetation patches. The prediction result in this study confirmed their founds. The researchers indicated trees or heterogeneous vegetation influence the humidity which is the fundamental factor affecting tick survival (Estrada-Peña et al. 2013; Greenfield 2011; Boyard et al. 2007).

However it is meaningful that the whole logistic regression in this study confirmed the relationship between vegetation structure classes and tick occurrence in Amsterdam and found it is weak association. Distance to water and distance to built-up area were found not statistically significant associated with tick occurrence in Amsterdam. It is a surprise that distance to water is not significant related to tick occurrence as many literature stated the humidity is the fundamental factor. However in this study the distance to water can be described as a 'stable' factor. It is stable over seasons and actually the humidity influences tick life described in the literature is more about micro-climate and in terms of short time condition. The researchers sampled ticks and measured humidity in a short period of summer (Greenfield 2011; Boyard et al. 2007). It is possible that distance to water the factor in this study did not reflect the variance of 'humidity' at a local scale and in a short time interval.

In some tick studies about vegetation structure, environmental factors and tick occurrence patterns, researchers stressed that the vegetation is an 'indicator' which affects local climate and indirectly affects the tick life circle (Randolph and Storey 1999; Estrada-Peña et al. 2013). In (Dobson et al. 2011) which studied tick occurrence in urban parks for two years, they found ticks in all vegetation types sampled, including short grass close to car parks. Highest tick densities were consistently found in plots with trees present (Dobson et al. 2011).

Based on these papers and the result (weak association in section 3.2) in this study, vegetation may be a proxy for other predictors of tick occurrence in Amsterdam urban area. Vegetation is the 'relatively stable predictor'. The strong predictors can be the local micro-climate factors (more variance in short time). The dominant vegetation modulates the microclimate (temperature, humidity, etc.) and affects the abundance of hosts (Estrada-Peña et al. 2013). The actual driver of the processes is the microclimate; local vegetation affects local microclimate (Estrada-Peña et al.

2013). Anyway, vegetation provide necessary shelter for ticks and without vegetation there are no ticks.

For investigating other strong predictors of tick occurrence in Amsterdam region, further studies need to be conducted about the relationships between potential strong factors such as humidity (Greenfield 2011; Boyard et al. 2007; Estrada-Peña et al. 2013), temperature (Estrada-Peña et al. 2013; Greenfield 2011; Vatansever et al. 2008), soil moisture (Greenfield 2011; Lindström and Jaenson 2003) and tick occurrence in Amsterdam region. It is suggested to use other variables instead of distance to water bodies to represent variance of 'humility' in the environment. The data of these factors can be transferred into same resolution spatial data together with vegetation structure map and tick occurrence data. Statistical analysis like logistic regression or Poisson regression can be used for analysing the relationships of these factors and tick occurrence in Amsterdam.

5. Conclusions and recommendations

5.1 Conclusions

This study concerned mapping vegetation structure classes and assessing the relationship of these in combination with other environmental factors with tick occurrence patterns in Amsterdam. To be specific, supervised classification with *k*-Nearest Neighbour and CART were applied in mapping vegetation structure classes and logistic regression was conducted with CART result vegetation maps, tick data and other two environmental factors distance to water area and distance to built-up area. The following research questions have been answered:

RQ1. Which geo-data can be used to map the vegetation structure classes (grass, shrubs, and trees) which are deemed important for mapping pest species distribution in Amsterdam?

- Semi Global Matching point cloud data were applied in this research successfully providing height values and spreads in height this structural feature of vegetation. The topography dataset KBKA10 provided relatively detailed locations of vegetation for filtering the point cloud data. AHN2 terrain grid offered terrain height values for extracting objects height of vegetation area. The aerial image of Amsterdam was used for providing ground information and selecting training locations of the three classes. Google Earth and Google street view provided auxiliary information for helping selecting training areas of grass, shrub, and tree classes.

RQ2. Which methods can be used for mapping vegetation structure patterns?

- Supervised classification method using *k*-Nearest Neighbour and CART were conducted in this study for mapping vegetation structure classes. CART was most successful. The overall classification accuracy of CART result vegetation structure map was 85.91%. The grass class shows high accuracy 96.61% and high reliability 93.04%. The tree class acquires high reliability 98.63%. The *k*-Nearest Neighbour result overall classification accuracy is 76.88%. The grass class in the result shows 98.77% accuracy. The tree class obtains a high reliability 99.97%.

RQ3. Which methods can be used to assess associations between vegetation structure classes maps and the occurrence data of ticks and to use the found relationships for prediction?

- Logistic regression was used to assess relationships between the vegetation structure class map, distance to water, distance to built-up area and tick absence and presence data in Amsterdam. It was found that distance to water and distance to built-up area were not statistically significant related to tick absence and presence. Vegetation structure classes were found the significant predictor (p-value < 0.05) for tick absence and presence in Amsterdam.

RQ4. What is the strength of correlations between occurrences of ticks and vegetation structure classes in Amsterdam?

- There is a weak association between vegetation structure classes and tick presence and absence in Amsterdam. Calculated Goodman and Kruskal's tau-y is 0.047. McFadden's R squared measure of the logistic model only containing vegetation structure classes the predictor is 0.092. Only a small part of the tick absence and presence was explained by the fitted model with vegetation structure classes this variable.

RQ5. Where are predicted hot-spots with respect to occurrences of ticks in Amsterdam?

- Tick presence is predicted where shrubs or trees grow.

5.2 Recommendations

Better results could be obtained with the same methods using better training and validation data sets. It is suggested to have aerial images of the same time period with the SGM point cloud data, which may increase training data quality (easy to select 'true' training data locations). In later study, it may help produce more solid vegetation object height data that conducting the step of interpolation on AHN2 terrain grid or creating terrain model from AHN2 point clouds.

It is suggested that when applying kNN method or CART method in the future study, use the same spatial support in training and classification. To be specific, for kNN method, the focal window operation can be applied in computing quantiles of shrubs, trees, grass training data (polygon area) and then same area size focal window operation should be employed on calculating quantiles of unknown vegetation patches at the classification stage.

In addition, more training samples for every vegetation structure class, and more 'pure' training data may produce higher classification accuracy in kNN and CART methods. Overlapping tiles of raster in focal calculation and checking with attention about water area/built-up area boundary can reduce producing NA values in the classification process.

Besides, to improve classification accuracy, it is recommended that adding more features (not only height and height spreads) into kNN method showing more variability of vegetation structure, like spectral feature or geometric feature. Predefining more than three local vegetation structure classes is also possible to reflect more variance of vegetation structure to obtain higher classification accuracy.

It is also recommended that adding more features such as spectral features, or geometric features instead of only height feature of vegetation structure into this CART method. In this study, the Near-Infrared (NIR) band data was delivered at a too late stage, so I couldn't use the data anymore. If more features from multi spectral bands data and other kinds of data involving into this CART method, it is very promising to obtain even higher accuracy and reliability result than this study result.

Additionally, mapping vegetation structure classes can also bring other benefits to the local society, like helping biodiversity studies or urban ecology research. So it is significant to conduct more studies in this direction.

Besides, it is suggested to use a more balanced tick distribution dataset (more presence data) to get better result in the relationship assessment. Further studies need to be conducted about the associations between potential strong factors such as humidity, temperature, soil moisture, etc. and tick absence and presence in Amsterdam urban region.

References

Aerodata Surveys Nederland. 2014. ‘’Project proposal: Amsterdam dense matching proposal’’. QN14-0104. Amsterdam, April 18, 2014.

Ageep, Tellal B, Jonathan Cox, M’oawia M Hassan, Bart G J Knols, Mark Q Benedict, Colin a Malcolm, Ahmed Babiker, and Badria B El Sayed. 2009. “Spatial and Temporal Distribution of the Malaria Mosquito Anopheles Arabiensis in Northern Sudan: Influence of Environmental Factors and Implications for Vector Control.” *Malaria Journal* 8 (January): 123. doi:10.1186/1475-2875-8-123.

Bartlett, J. 2014. "R squared in logistic regression". Logistic regression. November 30, 2014. http://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/

Bittencourt, H. R., and Clarke, R. T. 2004. ‘’Feature selection by using classification and regression trees (CART)’’. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Bivand, R.S., Pebesma, E., Gomez-Rubio. V., 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

Bivand, R., Keitt T.,and Rowlingson B. 2014. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.9-1. http://CRAN.R-project.org/package=rgdal

Boyard, C, J Barnouin, P Gasqui, and G Vourc’h. 2007. “Local Environmental Factors Characterizing Ixodes Ricinus Nymph Abundance in Grazed Permanent Pastures for Cattle.” *Parasitology* 134 (Pt 7): 987–94. doi:10.1017/S0031182007002351.

Breiman, L., Friedman, J., Olshen, R., Stone, C., Steinberg, D., and Colla, P. 1984. *CART: Classification and regression trees*. *Wadsworth: Belmont, CA*, *156*.

Burnham, K. P., and Anderson, D. R. 2002. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer.

Clevers, J. 2013. Classification of remote sensing. Course material. Master of Geoinformation Science. Wageningen University.

Cook, D., Dixon, P., Duckworth, W. M. , Kaiser, M. S., Koehler, K., Meeker W. Q., & Stephenson, W.R. 2001. *Binary Response and Logistic Regression Analysis. Part of the Iowa State University NSF/ILI project Beyond Traditional Statistical Methods*. Retrieved from http://www.public.iastate.edu/~stat415/stephenson/stat415_chapter3.pdf

Crosby Chris. 2011. “LAS Conversion Tools: LAStools and LASUtility”. National Science Foundation Open Topography Facility. April 2, 2015. Retrieved from http://www.opentopography.org/index.php/blog/detail/las_conversion_tools_lastools_and_lasutility

Dobson, Andrew D M, Jennifer L Taylor, and Sarah E Randolph. 2011. “Tick (Ixodes Ricinus) Abundance and Seasonality at Recreational Sites in the UK: Hazards in Relation to Fine-Scale Habitat Types Revealed by Complementary Sampling Methods.” *Ticks and Tick-Borne Diseases* 2 (2): 67–74. doi:10.1016/j.ttbdis.2011.03.002.

ECDC, European Centre for Disease Prevention and Control. 2012. Technical report: "Guidelines for the surveillance of invasive mosquitoes in Europe". Stockholm: 2012. doi 10.2900/61134

ESRI, Environmental Systems Research Institute. 2014. ArcGIS for desktop. Retrieved from http://desktop.arcgis.com/en/

ESRI, Environmental Systems Research Institute. March 1, 2015. ArcGIS – World Topographic Map. Retrieved from http://www.arcgis.com/home/item.html?id=30e5fe3149c34df1ba922e6f5bbf808f

Estrada-Peña, A. 2001. "Distribution, abundance, and habitat preferences of Ixodes ricinus (Acari: Ixodidae) in northern Spain". *Journal of Medical Entomology* 38(3): 361-370.

Estrada-Peña, Agustín, Jeremy S Gray, Olaf Kahl, Robert S Lane, and Ard M Nijhof. 2013. “Research on the Ecology of Ticks and Tick-Borne Pathogens--Methodological Principles and Caveats.” *Frontiers in Cellular and Infection Microbiology* 3 (August): 29. doi:10.3389/fcimb.2013.00029.

Franco-lopez, Hector, Alan R Ek, and Marvin E Bauer. 2001. “Estimation and Mapping of Forest Stand Density , Volume , and Cover Type Using the K -Nearest Neighbors Method” 77: 251–74.

Freedman, D. 2009. *Statistical models: theory and practice*. Cambridge University Press. 128-135.

Freitas, A. A. 2002. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science & Business Media: 228-230

Gao, Y., and Mas, J. F. 2008. "A comparison of the performance of pixel-based and object-based classifications over images with various spatial resolutions". *Online journal of earth sciences*, *2*(1): 27-35.

Gardner, Allison M, Tavis K Anderson, Gabriel L Hamer, Dana E Johnson, Kate E Varela, Edward D Walker, and Marilyn O Ruiz. 2013. “Terrestrial Vegetation and Aquatic Chemistry Influence Larval Mosquito Abundance in Catch Basins, Chicago, USA.” *Parasites & Vectors* 6 (1). Parasites & Vectors: 9. doi:10.1186/1756-3305-6-9.

Gassner, Fedor, and Nienke Hartemink. 2013. “7 . Tick – Borrelia Interactions : Burden or Benefit ?”, 141–54.

Gassner, Fedor, Arnold J H van Vliet, Saskia L G E Burgers, Frans Jacobs, Patrick Verbaarschot, Emiel K E Hovius, Sara Mulder, Niels O Verhulst, Leo S van Overbeek, and Willem Takken. 2011. “Geographic and Temporal Variations in Population Dynamics of Ixodes Ricinus and Associated Borrelia Infections in The Netherlands.” *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)* 11 (5): 523–32. doi:10.1089/vbz.2010.0026.

Gray, J. S., H. Dautel, a. Estrada-Peña, O. Kahl, and E. Lindgren. 2009. “Effects of Climate Change on Ticks and Tick-Borne Diseases in Europe.” *Interdisciplinary Perspectives on Infectious Diseases* 2009: 1–12. doi:10.1155/2009/593232.

Greenfield, B. P. J. 2011. "Environmental parameters affecting tick (Ixodes ricinus) distribution during the summer season in Richmond Park, London". *Bioscience Horizons*, *4*(2): 140-148.

Guglielmone, a a, L Beati, D M Barros-Battesti, M B Labruna, S Nava, J M Venzal, a J Mangold, et al. 2006. “Ticks (Ixodidae) on Humans in South America.” *Experimental & Applied Acarology* 40 (2): 83–100. doi:10.1007/s10493-006-9027-0.

Hantson, Wouter, Lammert Kooistra, and Pieter a. Slim. 2012. “Mapping Invasive Woody Species in Coastal Dunes in the Netherlands: A Remote Sensing Approach Using LIDAR and High-Resolution Aerial Photographs.” Edited by Geoffrey Henebry. *Applied Vegetation Science* 15 (4): 536–47. doi:10.1111/j.1654-109X.2012.01194.x.

Harvey, K. R., and G. J. E. Hill. 2010. “Vegetation Mapping of a Tropical Freshwater Swamp in the Northern Territory, Australia: A Comparison of Aerial Photography, Landsat TM and SPOT Satellite Imagery.” *International Journal of Remote Sensing* 22 (15): 2911–25. doi:10.1080/01431160119174.

Hijmans, R. J. 2015. "raster: Geographic data analysis and modeling". R package version 2.3-24. http://CRAN.R-project.org/package=raster

Hofhuis, A., Van der Giessen, J. W., Borgsteede, F. H., Wielinga, P. R., Notermans, D. W., and Van Pelt, W. 2006. Lyme borreliosis in the Netherlands: strong increase in GP consultations and hospital admissions in past 10 years. *Euro Surveill,* 11(6), E060622.

Hosmer, D. W., and S. Lemeshow. 2000. *Applied logistic regression*. Wiley. 276-290.

Hutmacher, Amy M., George N. Zaimes, Jonathan Martin, and Douglas M. Green. 2013. “Vegetation Structure along Urban Ephemeral Streams in Southeastern Arizona.” *Urban Ecosystems* 17 (1): 349–68. doi:10.1007/s11252-013-0293-4.

Hyde, Peter, Ralph Dubayah, Wayne Walker, J. Bryan Blair, Michelle Hofton, and Carolyn Hunsaker. 2006. “Mapping Forest Structure for Wildlife Habitat Analysis Using Multi-Sensor (LiDAR, SAR/InSAR, ETM+, Quickbird) Synergy.” *Remote Sensing of Environment* 102 (1-2): 63–73. doi:10.1016/j.rse.2006.01.021.

Inkiläinen, Elina N.M., Melissa R. McHale, Gary B. Blank, April L. James, and Eero Nikinmaa. 2013. “The Role of the Residential Urban Forest in Regulating Throughfall: A Case Study in Raleigh, North Carolina, USA.” *Landscape and Urban Planning* 119 (November). Elsevier B.V. 91–103. doi:10.1016/j.landurbplan.2013.07.002.

Jackman S. 2015. "pscl: Classes and Methods for R Developed in the Political Science". Computational Laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. R package version 1.4.8. URL http://pscl.stanford.edu/

Lawrence, R. L., and Wright, A. 2001. "Rule-based classification systems using classification and regression tree (CART) analysis". *Photogrammetric engineering and remote sensing,* 67(10): 1137-1142.

Lee, K. M. 2002. "Statistical Tests for Categorical Data, Research methods". Lecture Notes from division of social science, City University of Hong Kong. March 1, 2015. http://www.cityu.edu.hk/dss/adpam/rm/main1b.htm

Lehmann, Iris, Juliane Mathey, Stefanie Rößler, Anne Bräuer, and Valeri Goldberg. 2014. “Urban Vegetation Structure Types as a Methodological Approach for Identifying Ecosystem Services – Application to the Analysis of Micro-Climatic Effects.” *Ecological Indicators* 42 (July). Elsevier Ltd: 58–72. doi:10.1016/j.ecolind.2014.02.036.

Lillesand, T. M., Kiefer, R. W., and Chipman, J. W. 2008. *Remote sensing and image interpretation* (No. Ed. 6). John Wiley & Sons Ltd. 585-590.

Lindström, Anders, and Thomas G T Jaenson. 2003. “Distribution of the Common Tick , Ixodes Ricinus (Acari : Ixodidae), in Different Vegetation Types in Southern Sweden Distribution of the Common Tick , Ixodes Ricinus (Acari : Ixodidae), in Different Vegetation Types in Southern Sweden” 40 (4): 375–78.

Kohavi, Ronny, and J. Ross Quinlan. 2002. "Data mining tasks and methods: Classification: decision-tree discovery." *Handbook of data mining and knowledge discovery.* Oxford University Press, Inc., 267-276.

Kruijff, M., Hendrickx, G., Wint, W., and Ginati, A. 2011. "Mapping habitats for vectors of infectious disease: VECMAP". October 16, 2014. Retrieved from http://artes-apps.esa.int/sites/default/files/KruijffIAC-11-B5%201%2010%20-%20VECMAP%20-%20v2.2.pdf

Kuilder, E. 2012. "Mapping river flood plain vegetation structure with a consistent mapping scale by color-infrared aerial images and LiDAR data in object-based random forest classification". *Master Thesis.* Retrieved from Wageningen Theses Online record. (WTO/ 2037139)

Mathieu, Renaud, Claire Freeman, and Jagannath Aryal. 2007. “Mapping Private Gardens in Urban Areas Using Object-Oriented Techniques and Very High-Resolution Satellite Imagery.” *Landscape and Urban Planning* 81 (3): 179–92. doi:10.1016/j.landurbplan.2006.11.009.

Mejlon, Hans. 2000. *Host-Seeking Activity of Ixodes Ricinus in Relation to the Epidemiology of Lyme Borreliosis in Sweden*. Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 577. 42 pp. Uppsala.

Mücher, C A, L Roupioz, H Kramer, and B Bunce. 2010. “Use of LiDAR to Map and Monitor Habitats”, no. November: 1–10.

Mücher, Caspar A., Lammert Kooistra, Marleen Vermeulen, Jeroen Vanden Borre, Birgen Haest, and Rense Haveman. 2013. “Quantifying Structure of Natura 2000 Heathland Habitats Using Spectral Mixture

Analysis and Segmentation Techniques on Hyperspectral Imagery." *Ecological Indicators* 33 (October). Elsevier Ltd: 71–81. doi:10.1016/j.ecolind.2012.09.013.
Municipality of Amsterdam, November 14, 2014. "Registration BRT in Amsterdam - City of Amsterdam". Retrieved from http://www.amsterdam.nl/gemeente/organisaties/organisaties/dbi/stelselpedia/brt-index/registratieproces/
National Georegistry. December 12, 2014. "Metadata service, AHN2 WCS". Retrieved from http://www.nationaalgeoregister.nl/geonetwork/srv/dut/search#|fff9d7cf-9929-4dde-98b8-06ceda7e5610
Nagase, Ayako, and Nigel Dunnett. 2012. "Amount of Water Runoff from Different Vegetation Types on Extensive Green Roofs: Effects of Plant Species, Diversity and Plant Structure." *Landscape and Urban Planning* 104 (3-4). Elsevier B.V. 356–63. doi:10.1016/j.landurbplan.2011.11.001.
Pebesma, E.J., and R.S. Bivand. 2005. "Classes and methods for spatial data in R". R News 5 (2), http://cran.r-project.org/doc/Rnews/.
Pratihast, Arun Kumar. 2010. "3D Tree Modelling Using Mobile Laser Scanning Data 3D Tree Modelling Using Mobile Laser Scanning Data."
Qian, Yuguo, Weiqi Zhou, Jingli Yan, Weifeng Li, and Lijian Han. 2014. "Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery." *Remote Sensing* 7 (1): 153–68. doi:10.3390/rs70100153.
R Core Team. 2014. "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
Rajaei Najafabadi, M. 2014. "Mapping vegetation structure on Ameland using machine based learning technique to support risk management of vector-borne diseases". Master Thesis. Retrieved from Wageningen Theses Online record. (WTO/2082308)
Randolph, Sarah E, and Katie Storey. 1999. "Impact of Microclimate on Immature Tick-Rodent Host Interactions (Acari: Ixodidae): Implications for Parasite Transmission." *Journal of Medical Entomology* 36 (6): 741–48. http://jme.oxfordjournals.org/content/36/6/741.abstract.
rapidlasso GmbH. 2014. rapidlasso GmbH, LAStools. December 1, 2014. Retrieved from http://rapidlasso.com/lastools/
RStudio. 2014. RStudio, desktop. Retrieved from http://www.rstudio.com/products/rstudio/#Desk
Rutzinger, M., Pratihast, A. K., Oude Elberink, S., and Vosselman, G. 2010. "Detection and modelling of 3D trees from mobile laser scanning data". *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 38: 520-525.
Samaniego, Luis, and Karsten Schulz. 2009. "Supervised Classification of Agricultural Land Cover Using a Modified K-NN Technique (MNN) and Landsat Remote Sensing Imagery." *Remote Sensing* 1 (4): 875–95. doi:10.3390/rs1040875.
Shannon, C. 1948. "A mathematical theory of communication". *Bell System Technical Journal* 27 (3): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x
Sprong, Hein, Agnetha Hofhuis, Fedor Gassner, Willem Takken, Frans Jacobs, Arnold J H van Vliet, Marijn van Ballegooijen, Joke van der Giessen, and Katsuhisa Takumi. 2012. "Circumstantial Evidence for an Increase in the Total Number and Activity of Borrelia-Infected Ixodes Ricinus in the Netherlands." *Parasites & Vectors* 5 (1). Parasites & Vectors: 294. doi:10.1186/1756-3305-5-294.
Swart, Arno, Adolfo Ibañez-Justicia, Jan Buijs, Sip E van Wieren, Tim R Hofmeester, Hein Sprong, and Katsuhisa Takumi. 2014. "Predicting Tick Presence by Environmental Risk Mapping." *Frontiers in Public Health* 2 (November): 238. doi:10.3389/fpubh.2014.00238.
Tack, Wesley. 2013. "Impact of Forest Conversion on the Abundance of Ixodes Ricinus Ticks". Ghent, Belgium: Ghent University. Faculty of Bioscience Engineering.
Tack, Wesley, Maxime Madder, Lander Baeten, Margot Vanhellemont, and Kris Verheyen. 2013. "Shrub Clearing Adversely Affects the Abundance of Ixodes Ricinus Ticks." *Experimental & Applied Acarology* 60 (3): 411–20. doi:10.1007/s10493-013-9655-0.
Tekenradar. 2012. "Where ticks are active". October 16, 2014. Retrieved from https://www.tekenradar.nl/teken/teken/wanneer-zijn-teken-actief
Therneau, Terry M, and Elizabeth J Atkinson. 2015. "An Introduction to Recursive Partitioning Using the RPART Routines", 1–62.
Therneau,T., Atkinson E.J. and Ripley Brian. 2014. "rpart: Recursive Partitioning and Regression Trees". R package version 4.1-8. http://CRAN.R-project.org/package=rpart
van Adrichem, M. H., Buijs, J. A., Goedhart, P. W., and Verboom, J. 2013. "Factors influencing the density of the brown rat (Rattus norvegicus) in and around houses in Amsterdam". *Lutra,* 56(2): 77-91.
van der Zon, N. 2013. "Quality Document AHN2, Kwaliteitsdocument AHN2". *The AHN, AHN Netherlands*. March 15, 2015 Retrieved from http://www.ahn.nl/pagina/het-ahn/het-ahn.html
Vatansever, Z, a Gargili, N S Aysul, G Sengoz, and a Estrada-Peña. 2008. "Ticks Biting Humans in the Urban Area of Istanbul." *Parasitology Research* 102 (3): 551–53. doi:10.1007/s00436-007-0809-z.
Weih, Robert C, and Norman D Riggan. 2008. "OBJECT-BASED CLASSIFICATION VS . PIXEL-BASED CLASSIFICATION : COMPARITIVE IMPORTANCE OF MULTI-RESOLUTION IMAGERY" XXXVIII.
Wielinga, Peter R, Cor Gaasenbeek, Manoj Fonville, Albert de Boer, Ankje de Vries, Wim Dimmers, Gerard Akkerhuis Op Jagers, Leo M Schouls, Fred Borgsteede, and Joke W B van der Giessen. 2006. "Longitudinal Analysis of Tick Densities and Borrelia, Anaplasma, and Ehrlichia Infections of Ixodes Ricinus Ticks in Different Habitat Areas in The Netherlands." *Applied and Environmental Microbiology* 72 (12): 7594–7601. doi:10.1128/AEM.01851-06.

Wood, Eric M., Anna M. Pidgeon, Volker C. Radeloff, and Nicholas S. Keuler. 2012. “Image Texture as a Remotely Sensed Measure of Vegetation Structure.” *Remote Sensing of Environment* 121 (June). Elsevier Inc. 516–26. doi:10.1016/j.rse.2012.01.003.

Yohannes, Y., and Hoddinott, J. 1999. "Classification and regression trees: an introduction". *International Food Policy Research Institute*, *2033*.

Appendices

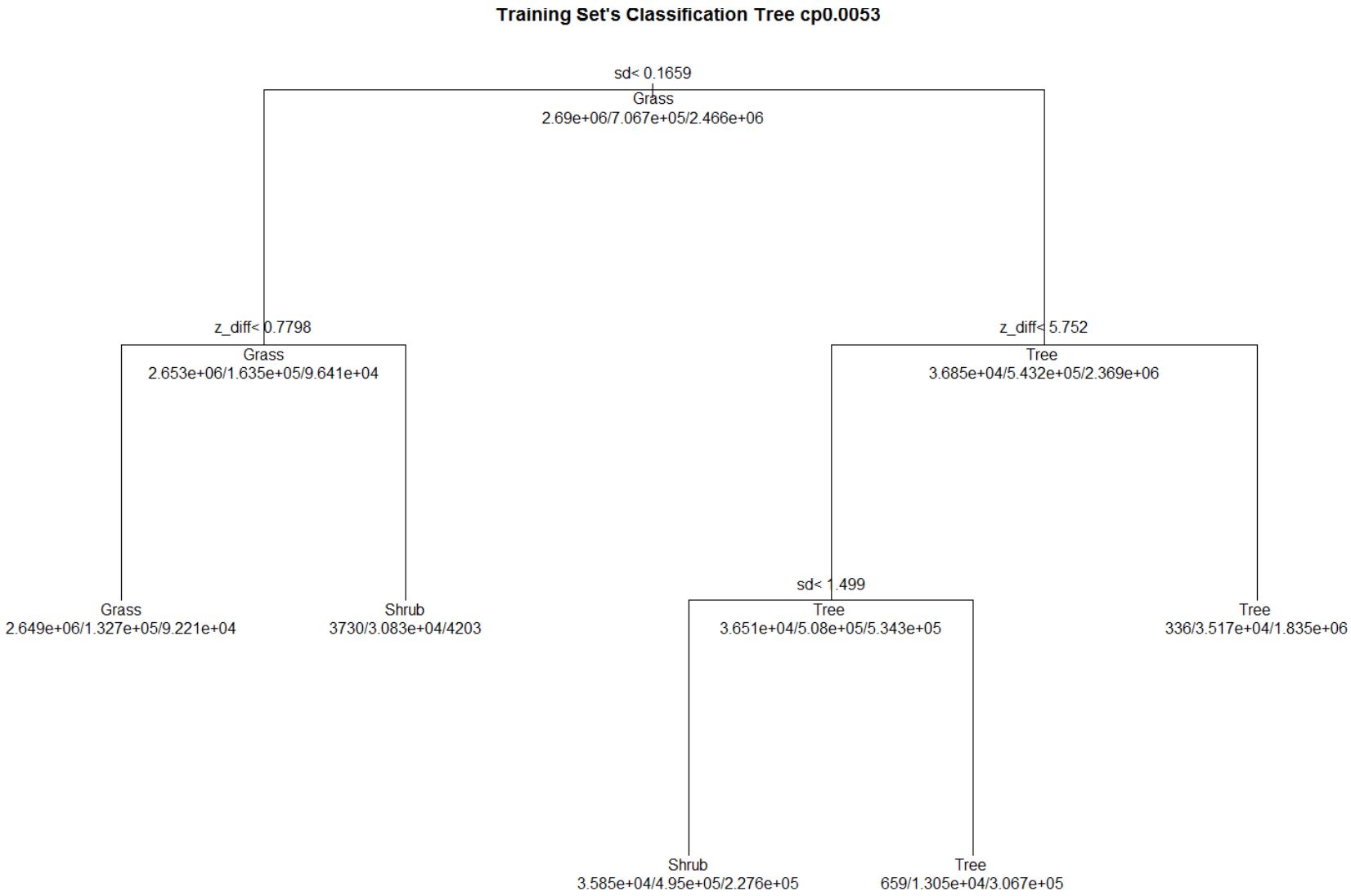
Appendix A. The summary output of original logistic regression model using environmental factors and tick occurrence data in R

```
Call:
glm(formula = Presence ~ VegeClass + DistWater + DistBldup, family = binomial,
    data = TickTrainValid)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7234  -0.5264  -0.2316  -0.2035   2.8436

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.103219   0.809621  -5.068 4.02e-07 ***
VegeClass2   1.806852   0.798196   2.264  0.02359 *
VegeClass3   2.043076   0.685515   2.980  0.00288 **
DistWater    0.004909   0.007982   0.615  0.53858
DistBldup    0.002856   0.008688   0.329  0.74235
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 127.39  on 236  degrees of freedom
Residual deviance: 115.18  on 232  degrees of freedom
AIC: 125.18

Number of Fisher Scoring iterations: 6
```

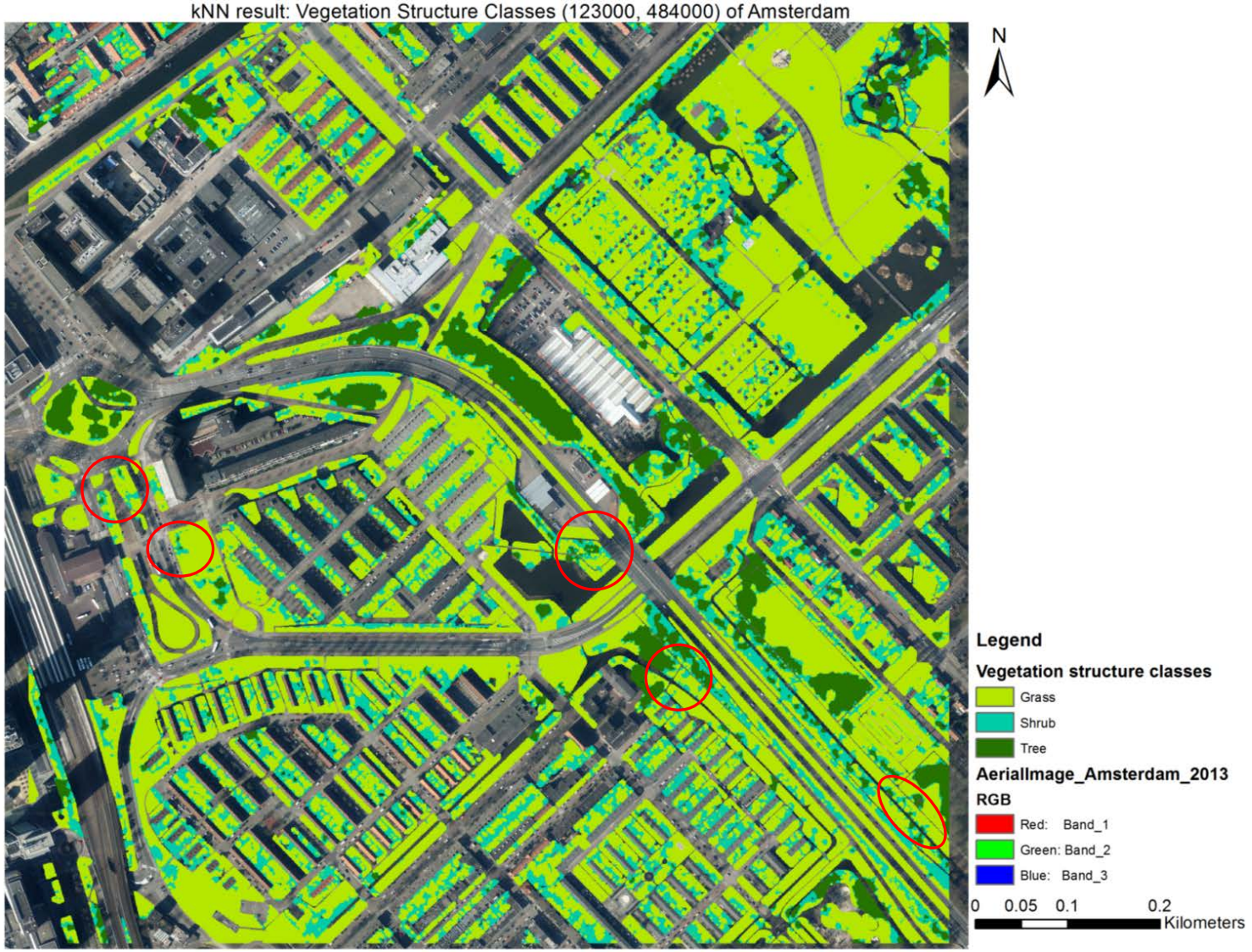
Appendix B. The summary output of final selected logistic regression model with smallest AIC in R

```
Call:
glm(formula = Presence ~ VegeClass, family = binomial, data = TickTrainValid)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5592  -0.5592  -0.2178  -0.2178   2.7399

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7297     0.5842  -6.384 1.73e-10 ***
VegeClass2    1.7487     0.7911   2.211  0.02707 *
VegeClass3    1.9532     0.6690   2.919  0.00351 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 127.39  on 236  degrees of freedom
Residual deviance: 115.67  on 234  degrees of freedom
AIC: 121.67

Number of Fisher Scoring iterations: 6
```

Appendix C. The trained classification tree for cp = 0.0053 of CART method (same with Figure 17, just for visualisation)



Appendix D. The kNN result: vegetation structure classes distribution in part of the Amsterdam area (‘hard’ classification). (same with Figure 13, just for visualisation)



Appendix E. CART result vegetation structure classification map for part of Amsterdam. The red circles indicate locations where the tree class has a larger area than in the kNN result (Figure 11,13, Appendix D). (same with Figure 18, just for visualisation)