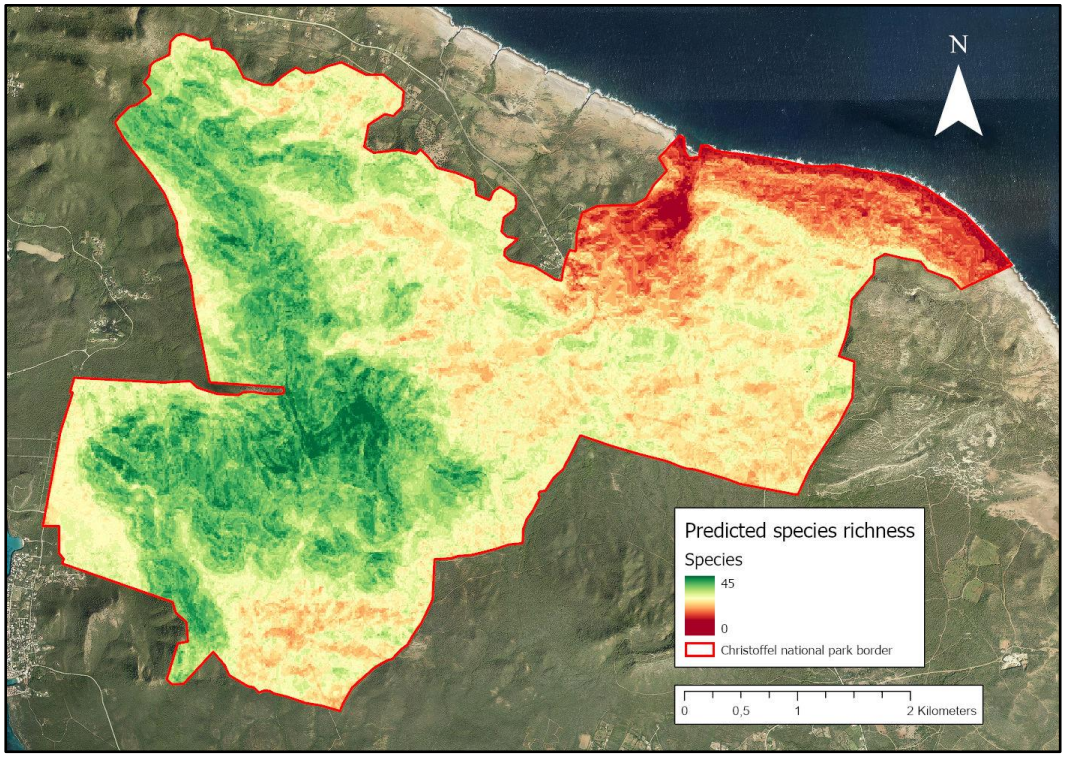
Geo-information Science and Remote Sensing

Thesis Report GIRS-2020-13

High-resolution prediction of plant species richness in the Christoffel national park

Robbert Ekkelenkamp



March 1, 2020

WAGENINGEN
UNIVERSITY & RESEARCH

CARMABI
CARIBBEAN RESEARCH & MANAGEMENT OF BIODIVERSITY

High-resolution prediction of plant species richness in the Christoffel national park

Robbert Ekkelenkamp

Registration number 96 07 11 219 260

Supervisors:

Dr. Harm Bartholomeus (WUR)
Erik Houtepen, MSc (CARMABI)

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

March 1, 2020

Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2020-13
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Abstract

Previous attempts at mapping the vegetation of the Christoffel national park on the island of Curaçao were done in times of intense grazing pressure and are likely not valid anymore after the removal of goats from the park because grazers have significant effect on the native vegetation of island ecosystems. In 2018, a 2-year fieldwork campaign was started to revisit the sampling points of Bokkestijn & Slijkhuis (1987) with the aim of remapping the vegetation communities and studying the change that occurred in the last decades. This thesis aims to assess the changes in vegetation distribution and use the newly acquired data to predict plant species richness across the entire national park at a high resolution using a macroecological modelling strategy. A trend of secondary vegetation succession has been found since 1985, with an increase in the coverage of trees, orchids and bromeliads and a decrease in grasses and herbs. The large-scale recovery of the native vegetation is found especially on the coast and midland of the park, while the Christoffel mountain and its surroundings have remained relatively stable. An aerial photograph interpretation of the vegetation communities found significant dependence of vegetation communities on elevation and slope aspects. High-resolution plant species richness prediction models were built and it was found that elevation and slope aspects have the most predictive weight. Little research has been done on high-resolution species richness prediction models; however, it is shown that these models can be utilized to characterize the variables influencing species distribution at high resolution and local scale, with comparable accuracy to coarser prediction models.

Acknowledgements

This thesis is the result of 6 months of work, of which 3 months were spend gathering field data on the island of Curaçao. The creation of this thesis would not be possible without the help of a couple of people, which I would like to thank.

First, I would like to thank Erik Houtepen, Cindy Eman and Luka Goudsblom for joining me on the wild adventures throughout the Christoffel national park. Getting dozens of cacti needles and sharp thorns stuck in your body and getting stung by approximately 20 bees are experiences best enjoyed as a group.

In particular, I would like to thank Erik for helping me make sense of the fieldwork data and ecological background in addition to supplying me with useful references and internal documents from CARMABI. I would like to thank Cindy for taking us shopping while the car was getting fixed and for organizing the barbeques. I would like to thank Luka for essentially being my private driver and for doing the plant determinations and database management so I could focus on the data analysis for my thesis.

I am also grateful for the skype meetings, suggestions and questions from my supervisor, Harm Bartholomeus, who gave me a lot of freedom to create this thesis in my vision. I would like to thank my family for the continuous support over the years and lastly, I would like to thank the Alberta Mennega stichting for the financial contribution to make this all possible.

Robbert Ekkelenkamp

Wageningen, March 2020

Table of contents

Table of figures

1. Introduction

1.1 Site description and short overview of previous research

Curaçao is a constituent country of the Kingdom of The Netherlands. The island is part of the lesser Antilles in the Southern Caribbean Sea and together with Aruba and Bonaire it forms the ABC-islands, situated just off the coast of Venezuela. The island has an area of approximately 444km^2 and the highest point is the summit of the Christoffel mountain. The climate is semi-arid and classified as BSh in the Köppen climate classification system. In the North of Curaçao lies the Christoffel national park, encompassing much of the area surrounding the Christoffel mountain and Savonet (figure 1). The Christoffel national park lies on top of 3 major geological formations: knip, diabase and limestone (Buisonjé, 1974). The Christoffel national park is the most biodiverse area of Curaçao, containing many endemic vegetation and animals.

In the 1980's, conservation and management activities were gradually getting initiated when conservation awareness began to grow (Putney, 1982; de Freitas & Wakkee, 1984). As part of these activities, research was done to map the vegetation communities in the Christoffel park and elsewhere on the island, the most notable are Bokkestijn & Slijkhuis (1987) and Beers et al. (1997). Another major research area was the effect of grazers on the Christoffel national park. Grazers tend to have a significant destructive effect on native island ecosystems and subsequent research in the Christoffel park has found that roughly the entire park has been affected by them. Later in the 1980's, the park authority started the removal of free roaming grazers in the park to let the vegetation recover.

Since then, not much research has been done in mapping the vegetation of the Christoffel national park. In 2018, a 2-year fieldwork campaign was started to revisit the sampling points of Bokkestijn & Slijkhuis (1987) with the aim of remapping the vegetation communities and studying the change that occurred in the last decades.

Figure 1: The Christoffel national park on Curaçao

1.2 Problem definition and objectives.

The previous attempts at mapping the vegetation in the Christoffel park are all based on data from the 1980's. Almost exactly after the data was acquired, the park authorities started removing herds of goats from the national park which have large effects on island ecosystems (Coblentz, 1978). As a result, the maps that are currently available are likely not valid anymore. In 2018 and 2019, the sampling points of Bokkestijn & Slijkhuis (1987) were revisited to explore the changes in floristic composition and to remap the vegetation communities in the Christoffel national park.

The mapping of vegetation communities in the Christoffel national park and on other Caribbean islands has traditionally been done using a landscape guided method developed at the International Institute for Aerospace Survey and Earth Sciences (ITC), which combines aerial photograph interpretation with stratified field sampling (Zonneveld, 1979; Bokkestijn & Slijkhuis, 1987; Beers et al., 1997; de Freitas et al., 2005, 2014, 2016). These aerial photograph interpretation maps are hard to validate because it is not based on underlying data, they can differ depending on the interpreter and it is not possible to see the undergrowth on an aerial photograph which makes it hard to differentiate certain vegetation communities. The lack of underlying data also prevents any analysis of factors that determine floristic composition or species richness.

Previous research has mostly looked at the floristic composition of the vegetation and not at the spatial variance in plant species richness, which is an important indicator of total species diversity and is relevant to develop conservation and management strategies. Species richness modelling is a research domain that aims to predict species richness based on climate and landscape variables. Current prediction models are often coarse resolution due to the lack of fine-scale global datasets. As a result, not much research has been done on the high-resolution predictions of species richness. In the context of species richness modelling, high resolution has been defined as being between 10 and 100m resolution (Nezer et al., 2017) and for the remainder of this thesis we will define it as 10m resolution. Coarse resolution species richness models are not useful in depicting local scale phenomena in relation to species richness. Since high-resolution species richness map can be useful for ecological management purposes in the Christoffel park, this thesis will explore the ability to predict species richness using high-resolution GIS data that is commonly available. The aim of this thesis can be divided into three main research questions:

1. *How and where has the floristic composition of vegetation changed in the national park since the last fieldwork in 1985?*

   To answer this question, vegetation data from 1985 is compared with the data from 2018 and 2019 to study the change in coverage of plant species. The decrease in grazing pressure since the 1980's has led to the hypothesis that large areas will have seen recovery in vegetation.

2. *What does the current spatial distribution of vegetation communities look like based on aerial photograph interpretation?*

   Grazers can destroy a lot of the vegetation in island ecosystems, so it is expected that the maps from the 1985 fieldwork are not valid anymore. New aerial photographs from 2018 can be used together with the new vegetation data to remap the vegetation communities

3. *Can a machine learning model accurately predict species richness in the national park based on high resolution geographic variables, and can this give insight in the features that influence species richness on a local scale?*

There has been a lack of research in high-resolution prediction of species richness due to the absence of fine scale global environmental datasets. To answer this question, multiple models are used to predict species richness based on available high-resolution GIS datasets.

2. Background

2.1 Climate and vegetation adaption

2.1.1 Contemporary climate

Curaçao is part of the leeward island of the Dutch Caribbean together with Bonaire and Aruba and lies in the Caribbean Sea off the coast of Venezuela. The average precipitation between 1980 and 2010 was 600.6mm with an average evaporation of 6.9mm per day (Meteorological Department Curaçao, 2016). The year-to-year variations are very large, as the standard deviation is larger than the mean. The precipitation divided by the evapotranspiration, also known as the aridity index, yields a fraction of 0.238 which makes Curaçao on average a semi-arid environment (Salem, 1989). Recent 1-km resolution Köppen-Geiger climate classification maps shows that most of the island is classified as semi-arid (BSh), while some of the Northern areas around the Christoffel national park are classified as tropical savannah (Aw; Beck et al., 2018). The dry climate is caused by the extension of the Azores high and an upwelling zone with colder sea surface temperatures along the Venezuelan coast (Lahey, 1958; Trewartha, 1981). Most of the precipitation is the result of very local convective events which makes it possible that two stations that are 10km apart can differ up to a factor of 2 for monthly rainfall (Martis et al., 2002). October to January is considered the rain season while February to May is considered the dry season. The other months are transitional with small rain events.

2.1.2 Future climate

Small islands such as Curaçao are vulnerable for climate change, and the latest IPCC report for small islands shows an average increase of 1.4°C in temperature and a decrease of 5-6% in average rainfall for the RCP4.5 scenario in the Caribbean region (Carabine & Dupar, 2014). Within the Caribbean, the Southern region in which Curaçao is situated is projected to become drier than the North, with extended dry seasons. The vegetation of Curaçao will therefore be subjected to harsher conditions in the future. The climate projections were used to derive future Köppen-Geiger climate classification maps in Beck et al. (2018). Comparing present and future maps shows a transition between a semi-arid and tropical savannah climate to a desert and semi-arid climate for Curaçao, which could affect the species richness on the island in the future.

2.1.3 Vegetation adaption

The dominant semi-arid climate in the region inhibits the growth of many species and in general you will only find vegetation that have adapted to the low rainfall throughout the year. Many plants on Curaçao have thorns, which characterize extreme seasonal types of vegetation. The plants on Curaçao adapt to the climate with root, leaf and propagation strategies (Stoffers, 1956). There are five categories of plant specializations when we look at how loss of water is limited through their leaves (Stoffers, 1956; de Freitas, 1991). First, there are plants that have thick and hard leaves which limit the exposure of the inner part of the leaf that contains the water. Many evergreen plants belong to this category, examples are *Jacquinia arborea* and *Coccoloba swartzii.* Second, we have deciduous plants with leaves that are not adapted to the climate and shed their leaves during the dry season, examples are *Bourreria succulenta* and the endemic *Bursera karsteniana.* Third, there are plants that opt for many small leaves instead of bigger leaves. The temperature of many small leaves increases slower under the influence of the sun as a result of the cooling effect of the wind (Stoffers, 1956; de Freitas, 1991). Examples of plants that use this strategy are *Vachellia tortuosa* and *Acacia glauca.* Fourth, there are plants that have hair growth on their leaves, which limits the influence of the sun

and the wind, an example is *Croton flavens*. The shape of a leaf can also affect the survivability of a species. Lastly, we have plants such as *Capparis indica* and *Phyllanthus bothryanthus* that have curled leaves that protect the stomatal surface. A special group of plants on the islands are the cacti species, as they do not have leaves to limit loss of water through transpiration. Cacti can store water in each part of its body and have shallow roots to benefit from the smallest amount of rainfall (de Freitas, 1991).

2.2 Geology

2.2.1 Geological formations in the Christoffel national park

The island of Curaçao consists of 4 geological formations (Buisonjé, 1974). The Christoffel park in the North of Curaçao can be divided in three main geological formations: The Curaçao lava formation, the Knip group and limestone (van Buurt, 2009; de Freitas, 1991; Buisonjé, 1974). The Curaçao lava formation is the largest formation on the island and underlies the other formations, it consists of pillow lava that is a few kilometers thick and was created under water (Klaver, 1987). The gentle rolling hills found in most of the island are the eroded surface of this formation (van Buurt, 2009). The second major geological formation is the Knip formation. This formation consists of finely grained marine sediments which were deposited in deep water and lifted tectonically afterwards. The highest mountain on the island, the Christoffelberg, and the surrounding mountains called "Zevenbergen" consist of the Knip formation. The youngest formation is the limestone found mostly around the edges of the island. When the island started to emerge from the sea, it reached the zone where corals and marine life could grow, which deposited marine sediments on top of the Curaçao lava formation, eventually these limestone ledges were uplifted above the sea level.

2.2.2 Effect of geology on vegetation distribution

The geological formations have differing water holding capacity which affects the ability of certain plants to grow on them. In general, the limestone and knip formations have a better capability to accumulate water in the rainy season and retains water for a longer time during the dry season. For this reason, knip and limestone have a higher variety of species growing on them. Diabase does not have this retention capacity, which means that deciduous species that lose their leaves in the dry season dominate on the formation.

2.3 The influence of goats on island ecosystems

The influence of goats on island ecosystems is a problem that has been described in detail in multiple papers, the following paragraphs are a summary of Baker & Reeser (1972), Coblentz (1978 & 1980), Meyboom (1994) and Coolen (2015).

One of the most serious ecological problems on oceanic island is the degradation caused by feral herbivorous animals. Many animals have the potential for a feral existence; however, goats are particularly well suited. The feral goats that exist on oceanic islands were mostly introduced by seafarers in the 17th and 18th century as a source of food. The biota of oceanic island evolved over long periods of isolation from mainland faunas and are not resistant to many kinds herbivores as a result. The introduction of goats resulted in high levels of grazing pressure that wiped out endemic plants or made it hard for the endemic plants to compete with better resistant exotic plants. Many plants

disappeared or were limited in range and number. Examples of islands where the native vegetation has suffered immensely from the introduction of goats and other feral herbivores are Saint Helena, Hawaii, the Galapagos and Santa Catalina. Not only do goats have effect on the types of vegetation communities, they also decrease the total cover of vegetation, which leads to extensive erosion along with an associated decrease in soil fertility and moisture retention. Quantitative studies on the ecological effects of goats have found that vegetation inside enclosers had a higher coverage and contained more species compared to non-enclosed areas that were reachable by goats (Baker & Reeser, 1972). Some plant species only exist as mature trees because their fruits, seedlings and shrubs are consumed, in such a situation the forest exist until the life expectancy of the mature tree is reached, afterwards the character of the habitat changes entirely. To control excessive grazing, measurements have been taken on many islands by eradicating or capturing the goats to help the regeneration of vegetation.

2.4 The influence of goats in the Christoffel national park

The effect of goats on the island of Curaçao has been significant. Similar to goat food habits elsewhere, the goats on Curaçao eat most species that are available, although their exact diet depends slightly on whether it is the dry or rain season. Notable exceptions are croton (*Croton flavens*), prickly pear cactus (*Opuntia* spp.) and the rubber vine (*Cryptostegia grandiflora*) which are often ignored by goats. As a result of the avoidance of these species, large areas are completely dominated by one or multiple of these species. *Vachellia tortuosa* is also present in affected areas even though this plant is often eaten by goats, but it has large, sharp thorns and a high regeneration level so it is able to survive. The goats have virtually altered the entire Christoffel national park, which is noticeable by the large areas of prickly pear. The goats eat all the other plants and transport the cactus to other places, which can easily regrow when it is dropped by the goats.

Debrot & Freitas (1993) studied the differences in vegetation on grazed and ungrazed patches of the Christoffel mountain. The most characteristic differences were the high coverage of orchids such as *Tillandsia fluxuosa* and *Brassavola nodosa* on the ungrazed rocks and the higher occurrence of *Vachellia tortuosa* and *Opuntia Caracassana* on the grazed rocks. In general, the vascular vegetation cover and the number of species found was a lot higher on ungrazed rocks. A dense growth of *bromeliads* is a typical feature of the herbaceous layer of deciduous seasonal forests in the Caribbean lowlands (Beard, 1944; Stoffers, 1956; Benzing, 1980; Garcia-Franco et al., 1991) and the introduction of grazers in the national park has affected this undergrowth. It can be hypothesized that the reduction of goats in the park can lead to the resurgence of bromeliads as the dominant undergrowth in grazed vegetation.

In the Christoffel national park, different stages related to grazing pressure of goats can be described based on literature and research results. Meyboom (1994) in particular has spent a lot of time on this subject. Heavily grazed vegetation consists of fields of prickly pear cactus, the next stage is a vegetation consisting of *Acacia glauca*, *Prosopis juliflora*, *Vachellia tortuosa*, *Caesalpinia coriaria*, *Croton flavens* and *Randia aculeata*. When the grazing pressure is low, the vegetation typically contains *Bourreria succulenta*, *Haematoxylon brasiletto*, *Cordia curassavica* and *Melochia tomentosa*. When goats have been absent for a long time, the vegetation can be rich in different tree species. The climax vegetation in the Christoffel national park can be divided into a seasonal formation and an evergreen formation. The seasonal climax vegetation consists of tree species such as *Tabebuia bilberghii, Ruprechtia coriacea, Trichilia trifolia, haematoxylon brassiletto*, *Capparis linearis* and will

include epiphytes such as *Brassavola nodosa, Schomburgkia humboldtii, Tillandsia flexuosa and Tillandsia recurvata.* The undergrowth will mostly consist of fields of *Bromelia humilis*. In the evergreen climax vegetation, the characteristic species is *Coccoloba swartzii* guided by *Portulaca venezulensis*, *Bernardia corensis* and *Brassavola nodosa*. Other possible species in this vegetation are *Haematoxylon brassiletto, Commelina elegans, Serjania curassavica*, *Schomburgkia humboldtii, Tillandsia flexuosa and Tillandsia recurvata*

2.5 Previous vegetation mapping attempts in the Christoffel national park

In the past, several vegetation mapping attempts have been conducted on Curaçao. The most important of these are Stoffers (1956), Bokkestijn & Slijkhuis (1987) and Beers et al. (1997). The fieldwork for Bokkestijn & Slijkhuis (1987) and Beers et al. (1997) were both done in the 1980's as conservation and management activities were gradually getting initiated when conservation awareness began to grow (Putney, 1982; de Freitas & Wakkee, 1984). Stoffers (1956) has studied the vegetation on Curaçao extensively. However, the small-scale vegetation map (1:110 000) only shows very generalized vegetation communities based on structure and it's clear that some areas are mapped inaccurately. For example, the large evergreen *Hippomane mancinella* forest near Savonet does not show up on the map, even though it contains some of the oldest and highest tree on the island. The vegetation map of Stoffers (1956) was unsuitable for habitat studies or decision making about environmental management. For this reason, Beers (1997) created a landscape ecological vegetation map of Curaçao that contains terrain characteristics, vegetation structure and species composition on a 1:50.000 scale. Bokkestijn & Slijkhuis (1987) specifically focused on the Christoffel national park and is therefore the highest resolution map of the different vegetation communities in the park. The vegetation maps in Bokkestijn & Slijkhuis (1987) and Beers et al. (1997) are both based on fieldwork data and aerial photographs of the mid 1980's. Large changes in vegetation are expected to have happened since the grazers were removed from the Christoffel national park in the late 1980's, so these maps are likely not valid anymore. A new fieldwork campaign was started in 2018 to remap the vegetation in the Christoffel national park and analyze the changes in species occurrence. The fieldwork campaign successfully finished in December 2019, and the acquired data will be used in this thesis.

2.6 Species richness modelling

2.6.1 Modelling strategies

There is often an absence of exhaustive data regarding the species distribution in a given area because the sampling of species distribution is costly. As a result, there has been great interest in statistical modelling methods that can predict the spatial distribution of biodiversity based on a limited amount of data, because this information can be important for conservation and scientific purposes (Ferrier, 2002; Stockwell & Peterson, 2002).

Statistical modelling is an often-used method to predict biological survey data based on environmental variables. Several different strategies exist to predict macroecological properties such as species richness, and they can be divided in 3 broad strategies (Ferrier & Guisan, 2006):

a. *Assemble first, predict later*
In this strategy, biological data from originally surveyed locations is first classified or aggregated based on ecological criteria or ordination analysis, after which community level attributes such as species richness are predicted as a function of environmental variables.

b. *Predict first, assemble later*
In this strategy, individual species are modelled first using single-species distribution models and are stacked afterwards to get a community prediction. The classification, aggregation or ordination analysis is done on the prediction of species occurrence, instead of the direct observations.

c. *Assemble and predict together*
Instead of modelling biological-environmental relationships as two separate steps, this strategy performs these tasks together by working with all the species data simultaneously within an integrated modelling process.

Figure 2 gives an overview of the three strategies.

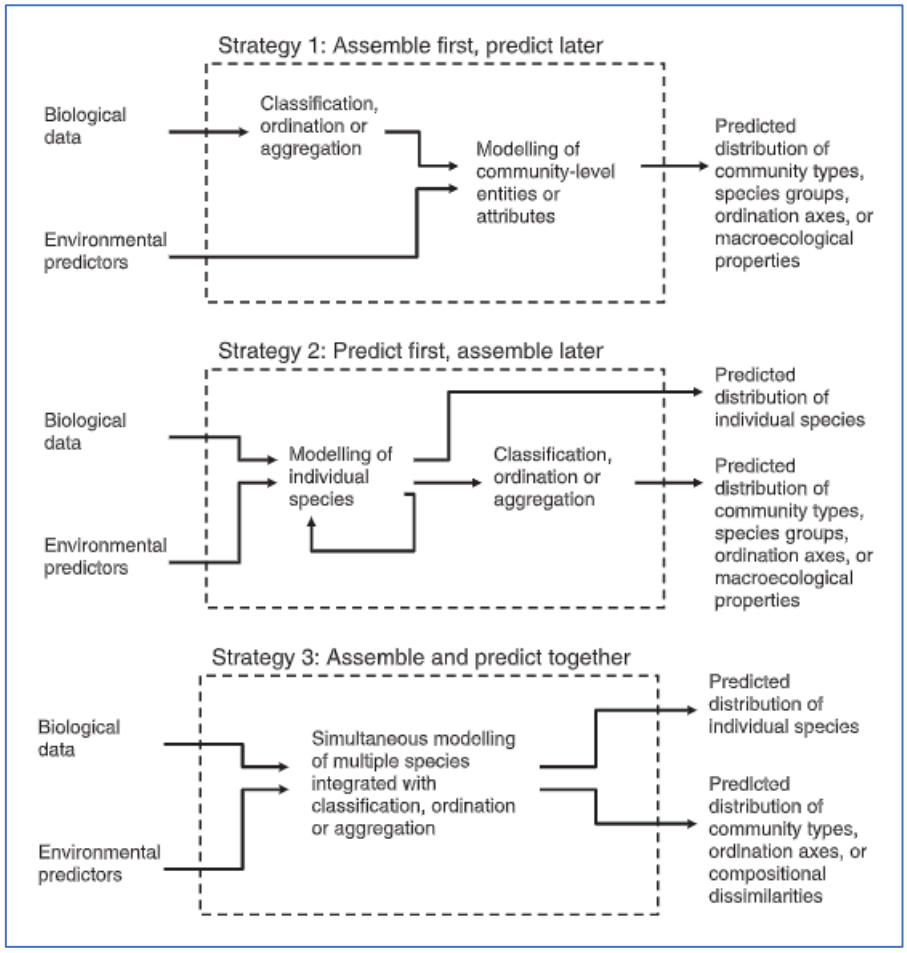


Figure 2: Modelling strategies for biodiversity. Source: Ferrier & Guisan (2006)

Environmental filters are important variables in deciding species distribution and biodiversity (Araújo & Rozenfeld, 2014). Curaçao is a semi-arid region projected to become more arid in the future, this relatively biologically stressful area increases the influence of environmental effects on species occurrence compared to more benign areas (Louthan et al., 2015).

Most modelling approaches use either a stacked species distribution model (SSDM) or a correlative macroecological model (MEM) to predict species richness. MEMs are a specific implementation of the assemble first, predict later strategy. These models propose a direct link between biodiversity metrics such as species richness and the environmental variables for their predictions (D'Amen et al., 2017). SSDMs are an implementation of the predict first, assemble later strategy. These models first try to predict the spatial occurrence of each species and then stacks them together to get the species richness.

2.6.2 Strategy comparison

Both MEMs and SSDMs have positive and negative attributes. A review of the different biodiversity modelling strategies found that a MEM following the "assemble first, predict later" strategy provided accurate and the most unbiased predictions overall (Dubuis et al., 2011; Zhang et al., 2019). A negative attribute of the MEM is that there will be no results for individual species, which SSDMs will provide (Zhang et al., 2019). SSDMs have three major drawbacks. Firstly, when stacking the output of multiple SDMs, the potential error that each SDM may have will be stacked and reduce the accuracy of the final prediction as well (Pineda & Lobo, 2009). Secondly, the SSDMs are species specific and will therefore require a minimum amount of occurrence information for the species, this can be troublesome when a rare plant species has a very small range (Wisz et al., 2008). The Christoffel national park contains some small ranging rare species so this could lead to larger errors. Lastly, SSDMs do not take into account the maximum carrying capacity of the soil, which leads to consistent overestimations of species richness. MEMs are able to predict the current and future patterns of species richness to similar levels of accuracy as the SSDMs but are less computationally expensive and only require the total species richness as response variable instead all the species-specific data (Harris et al., 2018; Biber et al., 2019).

2.6.3 High-resolution modelling of species richness

Little research has been done on the high-resolution modelling of species richness, usually because there is a lack of high-resolution environmental variables. Nezer et al. (2017) collected high resolution environmental variables and found that the high-resolution predictions (10m, 100m) gave the best predictions of species distribution and showed that high-resolution models could be utilized to characterize the variables influencing species distribution at high resolution and local scale, including anthropogenic effects and geomorphologic features. This suggests that high-resolution plant species richness modelling in the Christoffel national park should be possible because high-resolution geomorphological features such as elevation, slope and aspect are readily available through NASA's SRTM program. At the plot scale, plant species richness is largely dependent on factors such as geomorphology, soil, disturbance and competitive interactions (Tilman, 1982). Since the plant species richness during the fieldwork campaign was gathered on 10 by 10m plots, the species richness in the Christoffel national park should be predictable using these geomorphological variables.

3. Methodology

3.1 Data acquisition

3.1.1 Vegetation data

The vegetation data was acquired during 2 fieldwork campaigns in the months of September to December in 2018 and 2019. The sampling locations were the same as those in Bokkestijn & Slijkhuis (1987), so that the change in vegetation of the last decades can be measured. 19 extra vegetation plots were sampled compared to 1985 to increase the sampling coverage. The sampling locations were originally chosen in a stratified way across a range of vegetation communities and landscape units to get a good overview of the spatial distribution of vegetation and make aerial photograph interpretation possible. Figure 3 shows all locations that were sampled in 2018 and 2019.

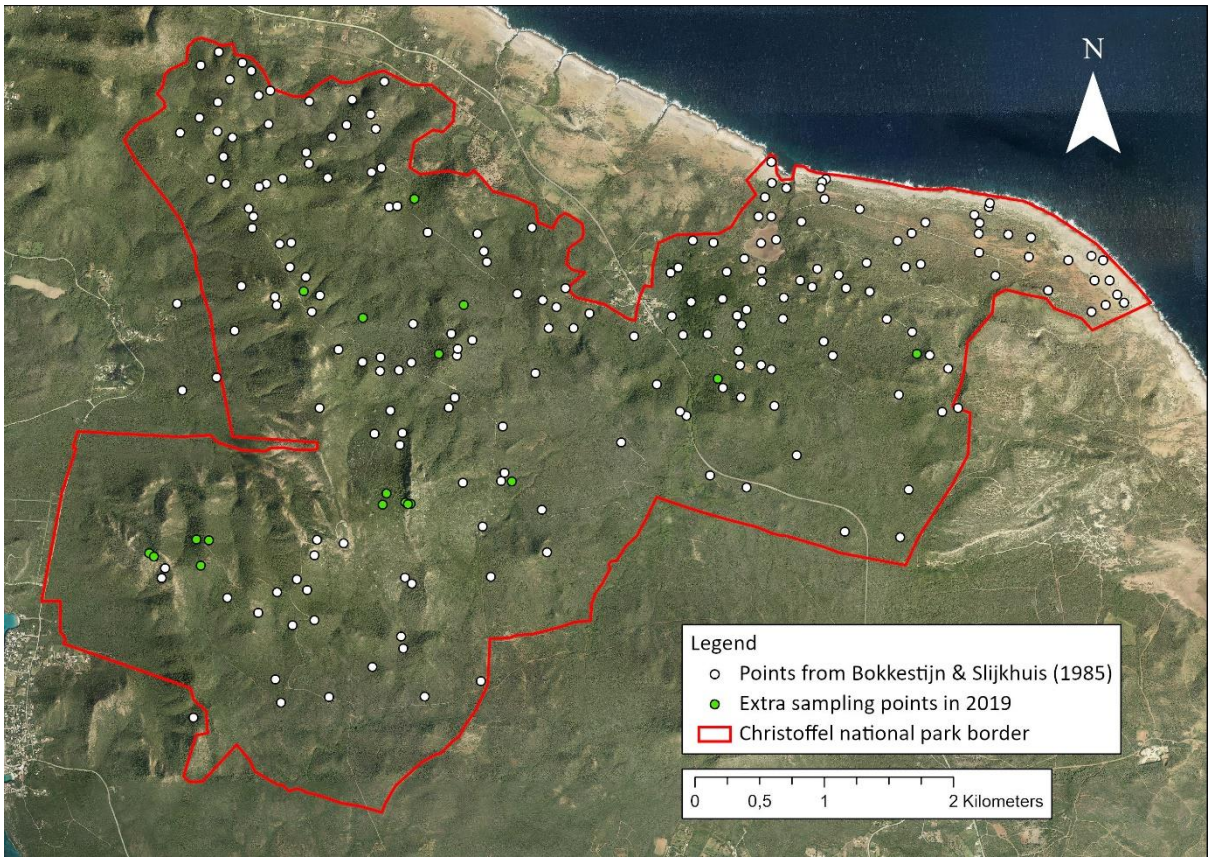


Figure 3: Sampling location of the fieldwork campaign in 2018 and 2019

For each sampling location of 10 by 10 meters, a relevé sheet was filled in with relevant information of the sampling points, consisting of:

Terrain characteristics: geological formation, relief type, slope steepness, exposure. And the percentage of surface stoniness.

Soil and water characteristics: the pH of the topsoil, an assessment of the soil colours using colour charts and the coverage of the soil with rocks and dead organic matter.

Grazer presence: Since grazers can have an influence on the floristic composition of the vegetation, any droppings that were found were noted down.

Vegetation structure and floristic composition: The total real cover as leaf area index, the cover of each vegetation stratum (herbs, shrubs, trees) and the occurrence and cover of each plant species. The identification of species was done using Arnoldo's zakflora (Arnoldo & Proosdij, 2012)

3.1.2 Satellite and aerial imagery

The aerial photograph is provided by CARMABI and has a resolution of 10cm. The aerial photograph was captured in November 2018, in the middle of the rain season. The satellite image that will be used as an input variable for the machine learning algorithm is a normalized difference vegetation index (NDVI) image in the dry season. NDVI quantifies the vegetation by measuring the difference between red light, which vegetation absorbs and infrared light, which vegetation reflects.

The NDVI image was created in Google Earth Engine and is based on a median composite of all the Sentinel 2 data captured in the dry season between April and June. The composite is made from dry season data because the NDVI image from this season will give a better contrast between the evergreen and deciduous dominated areas. In the rain season, the NDVI values are roughly the same for the entire national park which reduces the amount of information that can be extracted from it. Figure 4 shows the median composite NDVI image.

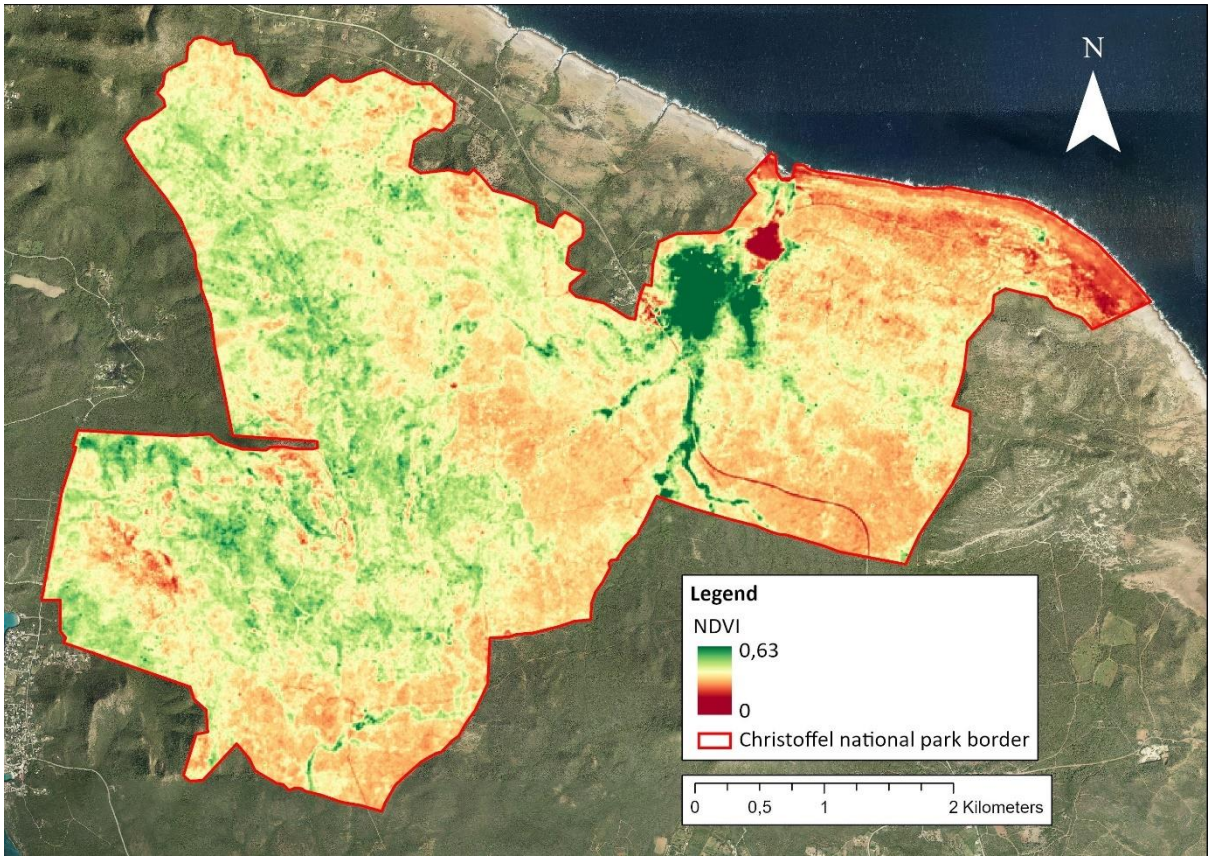


Figure 4: NDVI median composite of the dry season

The figure clearly shows the large evergreen *Hippomane mancinella* forest as bright green and the nearby salt flat and limestone coast as bright red. The mountainous region in the West of the national park is dominated by a greenish hue, indicating a higher amount of evergreen vegetation. The middle and eastern section has a yellow to orange hue, indicating a dominance of deciduous vegetation that

have lost most of their leaves. The NDVI composite will serve as one of the input variables for the species richness prediction models.

3.1.3 Topographic features

The elevation above sea level, the slope, the received solar radiation and the slope aspect are features used to predict species richness. The elevation feature is based on the digital elevation model (DEM) acquired through NASA's Shuttle Radar Topography Mission (SRTM). The original DEM has a 30m resolution and is resampled to the same resolution as the sentinel-2 image (10m). The DEM is used to calculate the slope, the aspect and the solar radiation for the entire national park in ArcGIS Pro. The geological formations that are present in the national park are also used as a feature in the model. The geological map is a digitized version of the map from Buisonjé (1974).

3.2 Software

The tabular data analysis is done using Python v3.7.6, while the spatial data analysis and visualization is done using ArcGIS Pro. Table 1 gives an overview of all the Python packages that are used and which steps they are used for.

Python package	**Description**
NumPy	NumPy gives support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy is necessary for the use of other packages such as Pandas and Rasterio.
Pandas	Pandas is a commonly used data analysis package that allows the creation of a data frame object, which is a two-dimensional tabular data structure that is also used in software like excel or the programming language R. The tabular data analysis for this thesis is mostly done using Pandas, and a Pandas data frame with independent variables serves as input for the machine learning model
Matplotlib + Seaborn	Matplotlib and seaborn are libraries that have a large variety of data visualization techniques. These packages will be used to visualize the results from the data analysis that is done with NumPy, Pandas and the machine learning libraries.
Scikit-learn	Scikit-learn is library that contains many machine learning models and functions. The linear regression model and the random forest model are used to predict species richness, as

	well as the randomized search function to find the ideal hyper parameters for the models.
XGBoost	The XGBoost package provides an implementation of a gradient boosting model which is one of the models used to predict species richness.
Eli5	The eli5 package is used to calculate the relative weight of the independent features for prediction species richness based on the Gini impurity index and can be used to find out the influence of each feature on species richness.
Rasterio	Rasterio can read GeoTIFF raster files as NumPy n-dimensional arrays. This package will be used to turn the model predictions into the final species richness prediction map.

Table 1: Python packages used for the data analysis and model creation

3.3 Methods

3.3.1 Comparative vegetation analytics

The sampling data from 2018 and 2019 was taken in the same locations and in the same season as the fieldwork of 1985 (Bokkestijn & Slijkhuis, 1987). During the fieldwork, every species was written down with its respective coverage. The total coverage of each species can be compared to calculate which species have seen significant growth and which species have decreased since 1985.

A second analysis is based on the change in each plot. Past research in the Christoffel national park and in national parks on Bonaire have found that the species composition is largely affected by the grazing pressure (Coblentz, 1978 & 1980; Debrot & Freitas ,1993; Meyboom, 1994; Coolen, 2015). As the grazing pressure decreases, the vegetation recovers in phases, each with different indicator species. Based on Meyboom (1994), the vegetation in each plot can be clustered and put into one of 4 recovery phases based on the coverage of the indicator species. A comparison of the grazing recovery phases between 1985 and 2018/2019 then shows whether a plot has recovered, deteriorated or has stayed the same. By plotting the changes on a map, it is possible to find out whether there is a spatial pattern in recovery and deterioration, or whether this is random.

3.3.2 Aerial photograph interpretation of vegetation communities

Aerial photograph interpretation is done by analyzing differences in photo-features such as tone, texture and spatial pattern of different classes. All the relevés containing the plant species occurrence and its coverage are classified in a hierarchical divisive way using TWINSPAN (Two-Way Indicator Species Analysis; Hill 1979). After each sampling point is classified as a certain vegetation community, it is labeled with a code and plotted onto the aerial photograph. The photo features of each plant community can be compared and serve as a basis to classify the unsampled areas (Beers et al., 1997).

3.3.3 Predictive modelling of plant species richness

3.3.3.1 Modelling strategy

Paragraph 2.6.2 gives a comparison of the common modelling strategies that are used for species richness predictions. The Christoffel national park has some rare species with small ranges, which would increase the error of the potential SSDM. Since the goal is to predict the spatial variance in total plant species richness and to assess the importance of different variables, the MEM strategy is chosen. With the MEM strategy, environmental variables are directly related to species richness for the prediction. The species richness is calculated for each surveyed plot by taking the sum of unique species occurrence.

3.3.3.2 Modelling techniques

Three modelling techniques were used to predict species richness: a generalized linear model (McCullagh & Nelder, 1989), random forests (Breiman, 2001) and XGBoost, an implementation of tree boosting (Chen & Guestrin, 2016). Linear models and random forests are well known and often used for the prediction of biodiversity, while XGBoost is a relatively new machine learning model. XGBoost is currently one of the most accurate predictive models in many machine learning competitions (Chen & Guestrin, 2016; Nielsen, 2016) and has also shown high accuracy in the prediction of species occurrence (Sandino et al., 2018; Aravindan & Jaisakthi, 2019; Herdter, 2019).

3.3.3.3 Dependent and independent variables

Table 2 gives an overview of the source and original resolution of all the variables used to predict the plant species richness in the Christoffel National park. ArcGIS pro was used to convert the polygon datasets into 10m resolution raster datasets. The coarser resolution raster datasets were resampled to 10m resolution using bilinear interpolation.

Variable	Source	Original resolution
Elevation	SRTM	30m
Slope	ArcGIS Pro (based on Elevation)	30m
Aspect	ArcGIS Pro (based on Elevation)	30m
Solar radiation	ArcGIS Pro (based on Elevation)	30m
Geology	Digitized from Buisonjé (1974)	Polygon
NDVI composite (dry season)	Sentinel 2; Google Earth Engine	10m
TWINSPAN vegetation map	Aerial photograph interpretation	Polygon
Vegetation species data	Fieldwork (2018 & 2019)	10m

Table 2: Variables used to train and test the plant species richness prediction model

The independent variables are used to predict the plant species richness in the Christoffel national park at 10m resolution. The species richness is defined as the sum of unique species per plot found in the vegetation species data mentioned in the table.

3.3.3.4 Model training and validation

The sampling points where species richness data was acquired are first plotted on top of the maps of the independent variables mentioned above. For each point, the underlying data is extracted in ArcGIS pro and the resulting table is exported to excel. The csv file can be loaded into a Pandas data frame and serve as training data for the models.

The models are validated using 5-folded cross validation: The training dataset is split in 5 parts and for each round of validation, 4 parts are used to train the model to measure accuracy on the 5th set. This is done 5 times so that each data point has been in the validation set once. The splitting of the data is done in a stratified way, as it was found that this gave a better prediction of out of sample accuracy (Kohavi, 1995).

The measures of accuracy estimation are the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE) and mean bias error (MBE). Finally, the Morran's I is calculated for the prediction residuals to check for spatial autocorrelation, which was found to be an issue in previous species richness modelling research and can lead to wrong conclusions regarding the feature importance (Kühn, 2007; Gaspard et al., 2019). The Gini impurity index is calculated for each feature using the eli5 package for python. The results show the relative importance of each feature and how different values for features relate to species richness.

3.3.3.5 Hyperparameter tuning

Random forest models and especially XGBoost models have many hyperparameters that need to be tuned, which can have significant impact on the accuracy of the prediction. To do this, the randomized search function from the scikit-learn package for Python is used to randomly select parameters and calculate the accuracy of the predictions using a stratified 5-folded cross validation for each set of parameters. Different combinations of independent variables will be used to search for the best combination, as too many independent variables can lead to overcomplication and overfitting of the data. The resulting parameters and combination of independent variables with the highest correlation coefficient will be used to train the final model and predict the species richness for every pixel in the Christoffel national park.

3.3.3.7 Creating a high-resolution species richness prediction map of the Christoffel national park

To predict the high-resolution species richness, every data source was transformed to the same resolution, which is 10m, through bilinear interpolation. Each 10 by 10-meter pixel in the Christoffel national park has a value for each of the independent features with which the species richness can be calculated using the tuned model.

The GeoTIFF files containing the spatial data were read as NumPy n-dimensional arrays using the rasterio package. The n-dimensional arrays were then transformed into 1D-arrays so that they could be turned into a data frame where each row is a pixel and the columns are the corresponding values for the independent features. The data frame with all the independent variables was put through the tuned and validated machine learning pipeline to predict the species richness for each pixel. The resulting 1D-array of predictions was transformed back into an n-dimensional array of the same width and height as the original data sets. Finally, the n-dimensional array of predictions was transformed into a GeoTIFF using the same coordinate system and transformation as the input rasters. Figure 5 is a flowchart that gives an overview of the creation of the final prediction map.

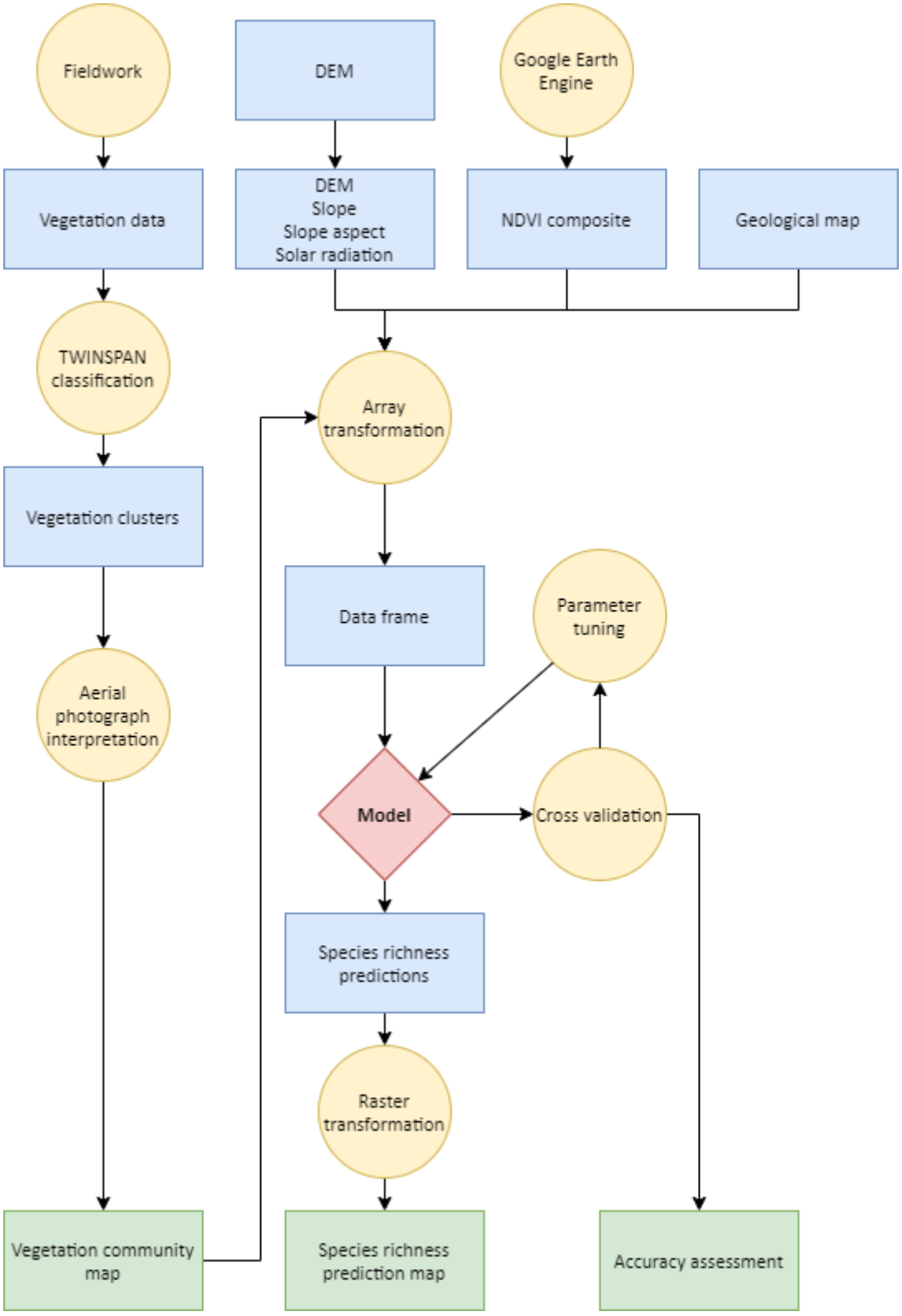


Figure 5: Methodology for creating the plant species richness prediction map

4. Results

4.1 Comparative vegetation analytics

4.1.1 Change in plant species coverage

Figure 6 shows the 15 species whose coverage has increased or decreased the most at the visited field sampling locations since 1985. *Croton flavens* is the species that shows the largest decrease in coverage and has been mentioned in the background chapter as being an indicator species for high grazing pressure. *Croton flavens* is one of the only plant species that goats would not eat, which resulted in a strong comparative advantage over other species and allowed it to grow unrestricted in many places. The culling of goats in the national park likely reduced its comparative advantage, causing a reduction in the coverage of the species as it had to compete with other species.

Figure 6: overview of the species with the largest change in coverage since 1985

Of the 15 species with the highest reduction in coverage, 10 are herbs or grasses, 3 are small shrubs, 1 is a vine and 1 is a tree. In contrast, of the 15 species with the most increase in coverage only 1 is an herb while the rest are tree species, high shrub species or bromeliads. *Bromelia humilis* and *Tillandsia flexuosa* are both bromeliads and show a strong increase in coverage, both are mentioned as being part of the undergrowth of climax vegetation (Meyboom, 1994). The increase in climax species together with the reduction of herb species is a clear indicator of secondary succession, which is the recovery of vegetation after a disturbance, in this case decades of grazing pressure.

By dividing all the individual species over their respective vegetation groups, a better overview of the changes can be seen. Figure 7 shows the changes in the major vegetation groups in the Christoffel national park. The strongest increase in coverage is found for trees and in the bromeliads and orchids group. A modest increase in cacti was found. Vines and shrubs show a modest decrease in coverage. The most significant change is the decrease in herbs and grasses. The increase in trees, bromeliads, orchids and cacti does not offset the decrease in vines, shrubs and especially herbs and grasses which means that a negative change in total coverage was found since 1985. An analysis of the individual plots shows that in 1985, a lot of coastal plots consisted of entire fields of grasses. in 2018 and 2019 these fields were found to be replaced by shrubs and tree species. Because the leaf area index of these species tends to be lower than the fields of grass, a negative total coverage was found. The findings of figure 7 are typical for vegetation succession. Figure 8 shows a violin plot that gives an overview of the distribution of species richness for both fieldwork campaigns. The species richness in 2018 and 2019 was found to be very close to a normal distribution, while the species richness distribution in 1985 has some local peaks. By comparing the quartiles of the species richness distribution, a small decrease in average species richness per plot can be found between 1985 and 2019. This can mostly be attributed to the decrease in herbs and grass species that were found in 2018 and 2019.

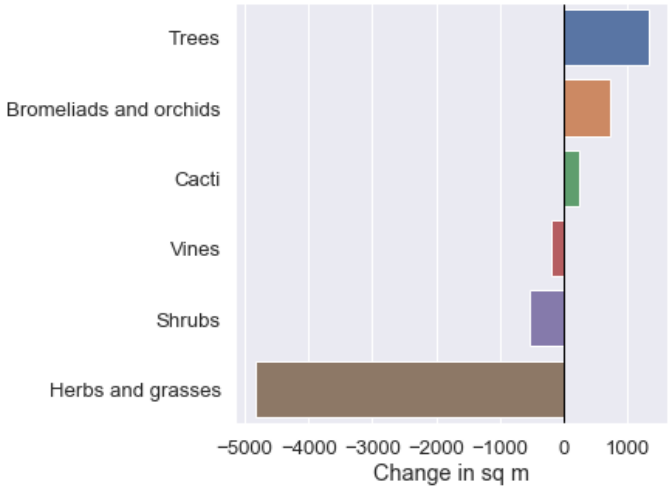


Figure 7: Change in coverage for each vegetation group

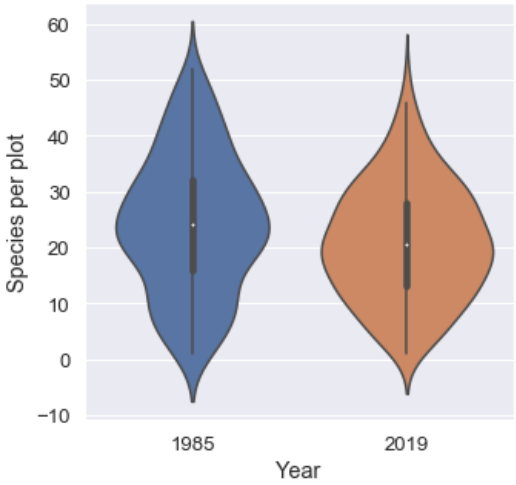


Figure 8: Distribution of species richness per year

4.1.2 Spatial recovery of vegetation

Figure 7 shows the spatial recovery and deterioration of vegetation based on the coverage of indicator species for different phases of vegetation recovery.

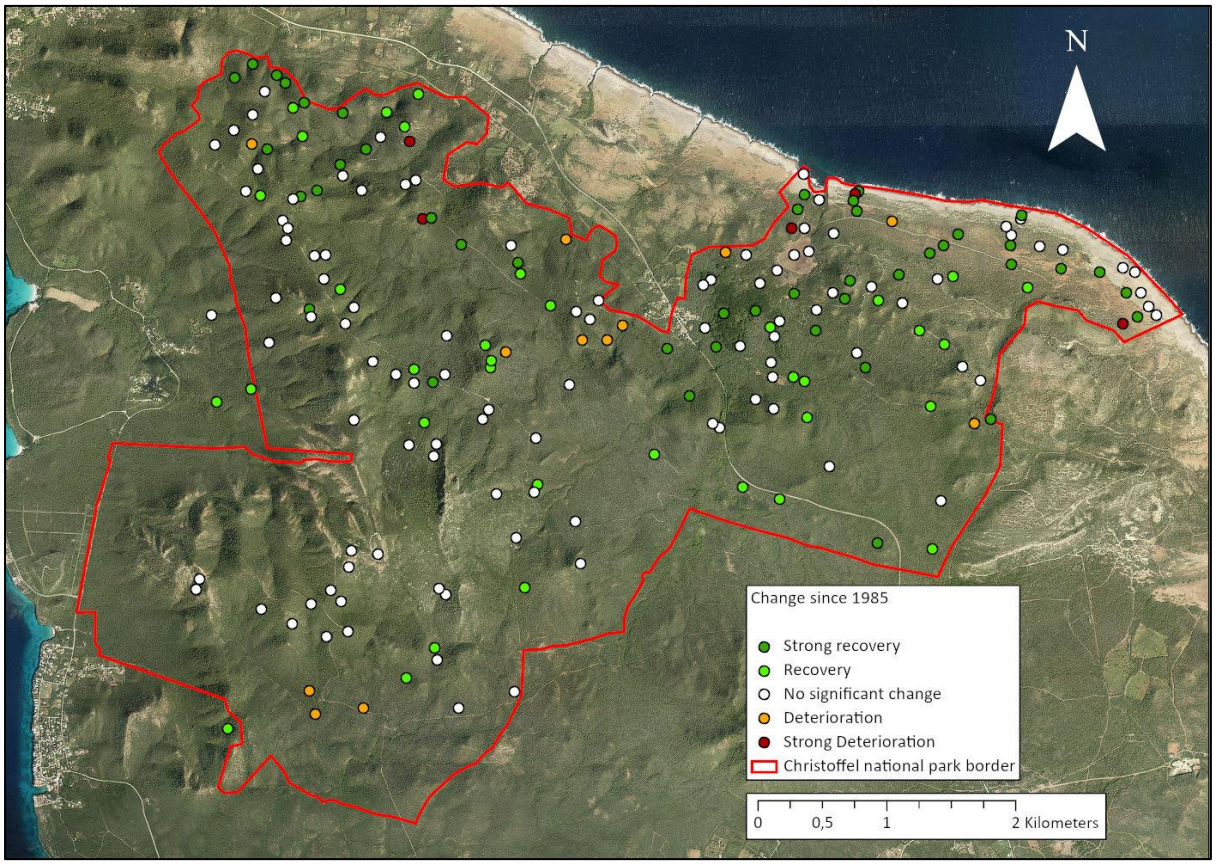


Figure 7: relative recovery of vegetation since 1985

Of the 200 plots that can be compared, 44 show strong recovery, 35 show recovery, 104 have not seen significant change, 12 have deteriorated and 5 have strongly deteriorated. Plots showing strong recovery or deterioration are those that are two phases higher or lower compared to 1985, while recovery and deterioration means that the plot is in one phase higher or lower compared to 1985.

On average, the sampled locations show a trend of recovery since 1985. The recovery and strong recovery can mostly be seen in the coastal and midland regions of the Christoffel national park, while the higher elevation areas surrounding the Christoffel mountain in the West do not show a significant amount of change.

The Western part of the national park consisting of the Christoffel mountain and the surrounding Zevenbergen area has not seen significant change, likely because most sampling locations already consisted of climax vegetation in 1985, so further recovery is not possible. In addition, this area has always had the lowest amount of grazing pressure as most grazers preferred to stay in the lowlands, which explains why there is no deterioration to be seen either. In contrast, the lowland areas show a large amount of recovery because the removal of grazers had the largest positive effect here. An interesting finding is that the climax vegetation that in 1985 mostly showed up in the Zevenbergen area was also found in some coastal locations, which suggests that If the vegetation is left alone long enough, the majority of the Christoffel national park might see a return to climax vegetation, which can give an idea of how the vegetation looked like before the arrival of grazers.

4.2 Aerial photograph interpretation

Figure 8 shows the 22 TWINSPAN clusters distributed over the national park based on aerial photograph interpretation.

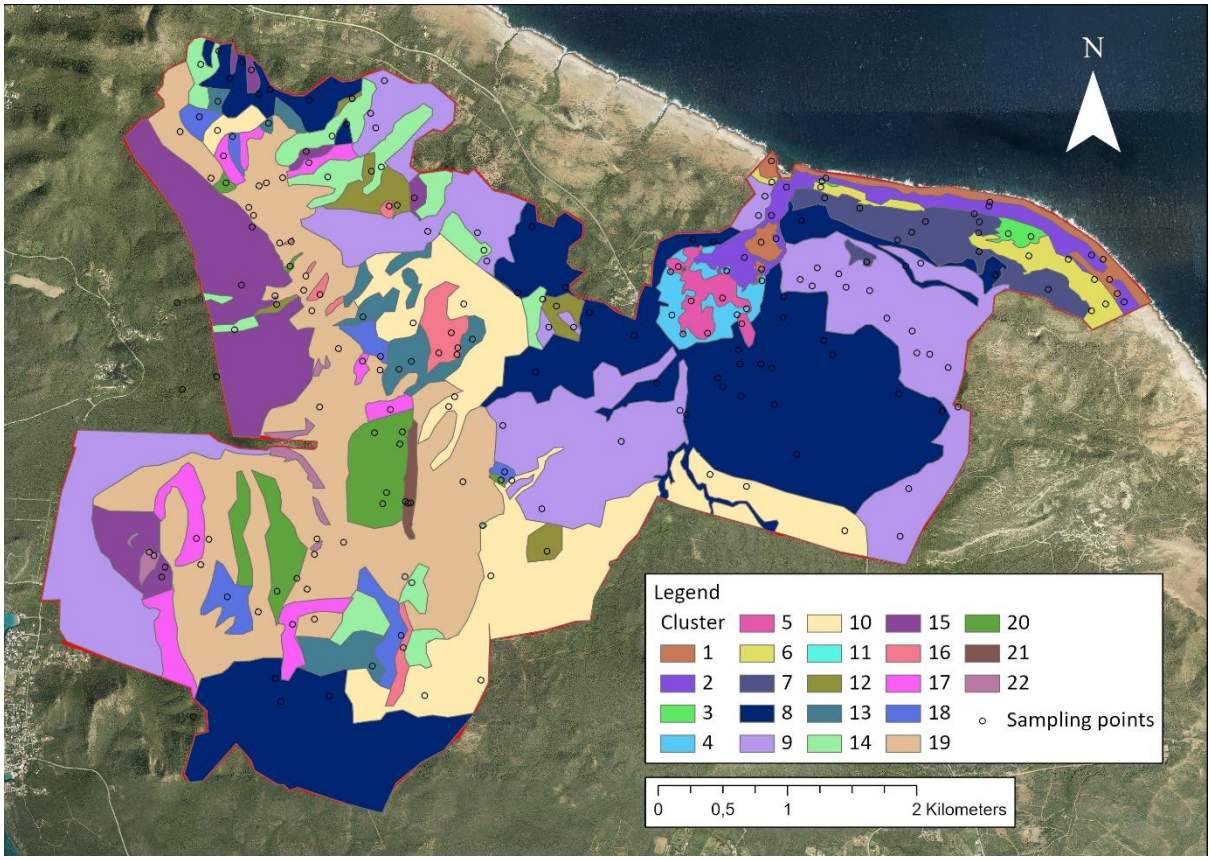


Figure 8: Vegetation community distribution

Based on the cluster distribution on the aerial photograph, there are some clear spatial patterns visible in the distribution of many vegetation clusters. The Christoffel national park can roughly be divided into 3 separate vegetation regions (figure 9):

1. Coastal vegetation
2. Midland vegetation
3. Mountain vegetation

The coastal vegetation consists of cluster 1 to 7, who show a rather unique composition. Cluster 1 is the vegetation of the very bare limestone and the “salina”. Only some saline resistant species can survive the harsh conditions found here. Cluster 2,3,6 and 7 are still affected by the salinity of the ocean but consists of a wider range of species and can be described as a thorny bushland. Figure 10 shows an example of a plot in cluster 7, which is dominated by thorny bushland. Cluster 4 and 5 are the *Hippomane mancinella* forest and its surroundings, including the invasive rubber vine that dominates in certain areas. In general, the coastal region contains more species of grasses than other regions and lacks in tree species, except for some mangrove species such as *Conocarpus erectus* and the *Hippomane mancinella* trees that thrive in brackish conditions.

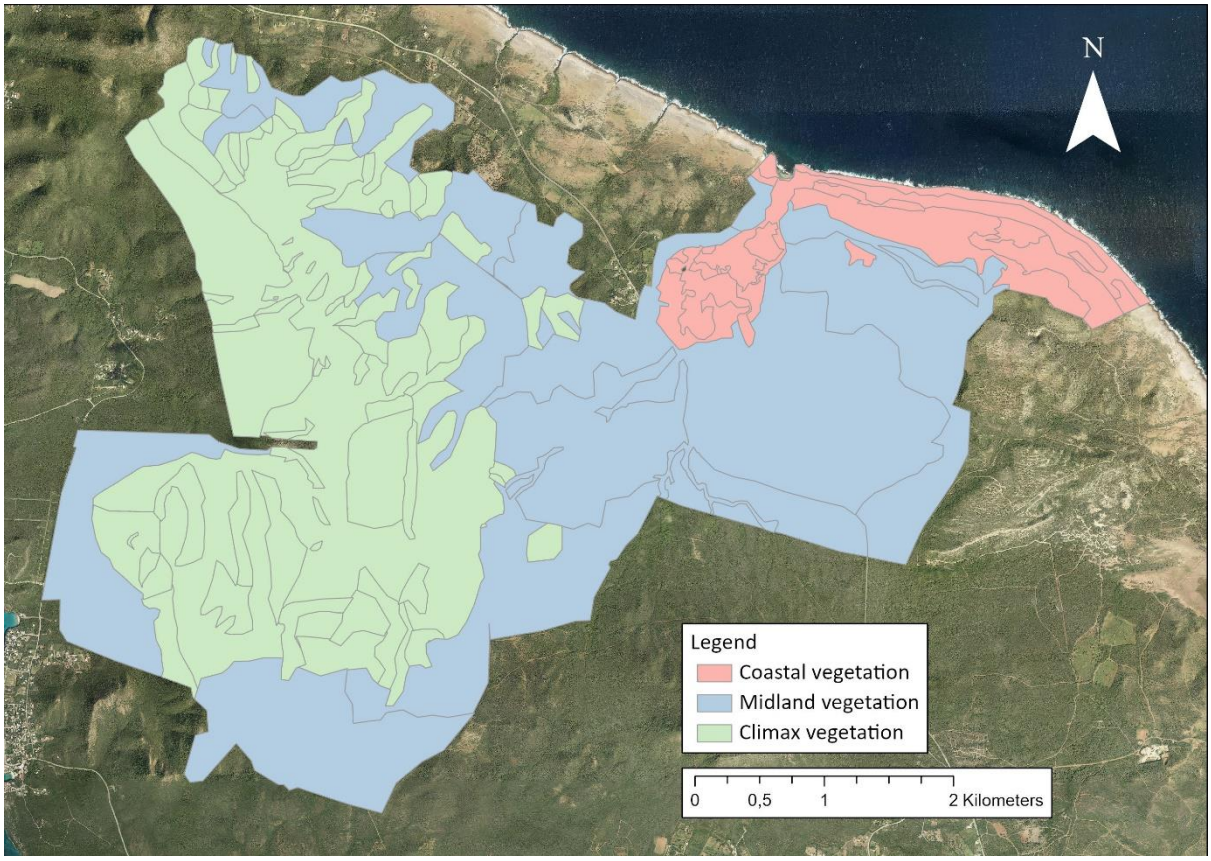


Figure 9: Vegetation clusters grouped by region

Figure 10: sample point 5, a thorny bushland typical of near-coastal vegetation.

Cluster 8,9 and 10 can be grouped as the midland region and have shown some of the largest rates of recovery as can be seen in figure 7. The midland region contains less species of grasses compared to the coastal region and contains a lot of common tree species such as *Bourreria succulenta* and *Haematoxylum brasiletto.* Some climax species will occur in this region, but it is not often dominated by them. Because this area has recovered significantly and some climax vegetation can be found, this might be an interesting area to monitor to research whether a return to full climax vegetation is possible in the future. Figure 11 shows a typical plot in the midland region.

Figure 11: sampling point 161, a typical midland vegetation.

Clusters 11 to 22 mostly consist of smaller niches of climax vegetation in combination with some of the more common tree species. Many of these clusters show dependence on landscape variables such as elevation and slope aspect. Most clusters only occur at higher elevation or on specific sides of the mountain, usually North. One noteworthy cluster in this section is cluster 22, which only occurs on the rock pavements at high elevations and contains the very rare *Sabal antillensis*, an endemic species to only Curaçao and Bonaire. Figure 12 shows a typical climax vegetation of the Zevenbergen area, with *Coccoloba swartzii* and a dense field of *bromelia humilis*, both being climax indicator species mentioned in Meyboom (1994).

Figure 12: sampling point 145 in vegetation cluster 20, a typical climax vegetation.

4.3 Biodiversity prediction

4.3.1 Validation

Table 3 shows the performance metrics for each model. When comparing the models, it becomes clear that XGBoost outperforms the other models in every metric except for the mean bias error. The random forest model scores better than the linear model in every metric except for the mean average error. Both the XGBoost model and the linear model have a small negative mean bias, suggesting a little more overestimation of species richness, while the mean bias of the random forest is minimal. Based purely on error metrics, the model performance can be ranked as follows:

1. XGBoost
2. Random forest
3. Linear model

Model	R^2	MAE	RMSE	MBE
XGBoost	0.72	4.19	5.51	-0.07
Random forest	0.69	4.44	5.78	0.01
Linear model	0.68	4.40	5.85	-0.14

Table 3: Error metrics for the predictive models

The random forest model and the linear model have the best performance while using all available features. In contrast, the XGBoost model shows the best performance with fewer features. Appendix A gives an overview of the combination of model parameters and features that give the highest accuracy for each model.

Figure 13 shows the scatterplot of the residuals for each model. Even though the mean bias error for the models do not show a clear bias, the scatterplot does show some areas where the models consistently over- or underestimate. The XGBoost model residuals show small overestimations at lower predictions of species richness and starting at around a species richness of 30 there is a clear case of underestimation. This trend is even more clear in the random forest model residuals, where almost all the plots with a species richness below 10 are overestimated and the plots with a species richness above 30 are consistently underestimated. The linear model residuals do not show clear trends in over or underestimation, but has more outliers in both direction, which decreases model performance.

XGBoost will be used for the creation of the predicted species richness map because it outperforms the other models in most measurements and the other models require the TWINSPAN map as input for the best accuracy, however the TWINSPAN map cannot be validated which increases the prediction uncertainty. To test whether there is any spatial autocorrelation in the residuals of the XGBoost model, a Moran's I test is run. Figure 14 shows the spatial distribution of the XGBoost prediction residuals in which positive values are underestimations and negative values are overestimations. Figure 15 shows the results of the Moran's I test. The Moran's index is 0.02 with a z-score of 0.72 and a p-value of 0.47, indicating a random distribution and no significant spatial autocorrelation in the residuals of the model.

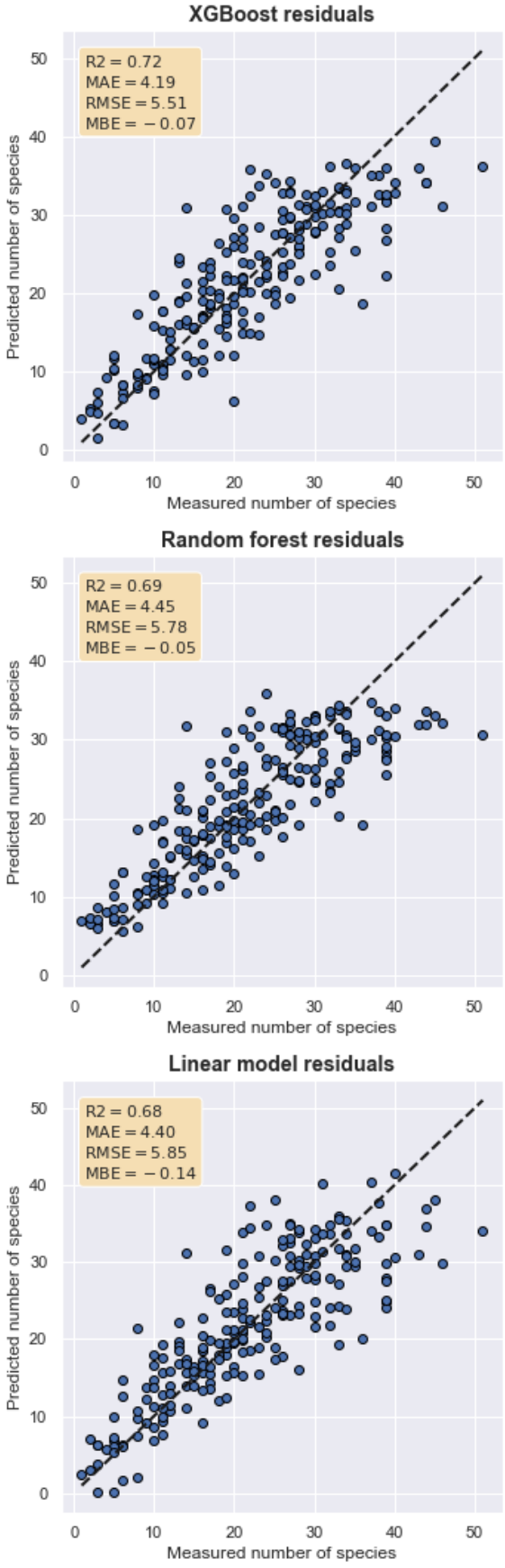


Figure 13: error residuals for each model

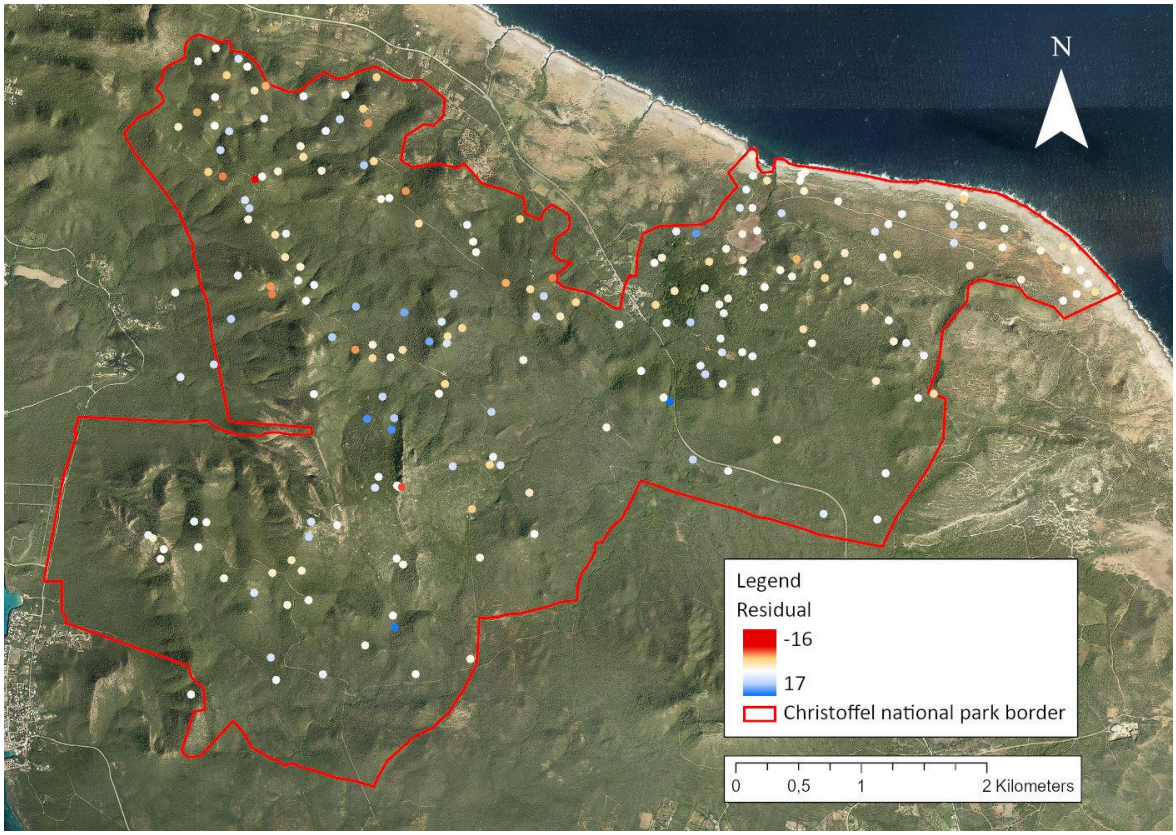


Figure 14: error residuals mapped over the Christoffel national park

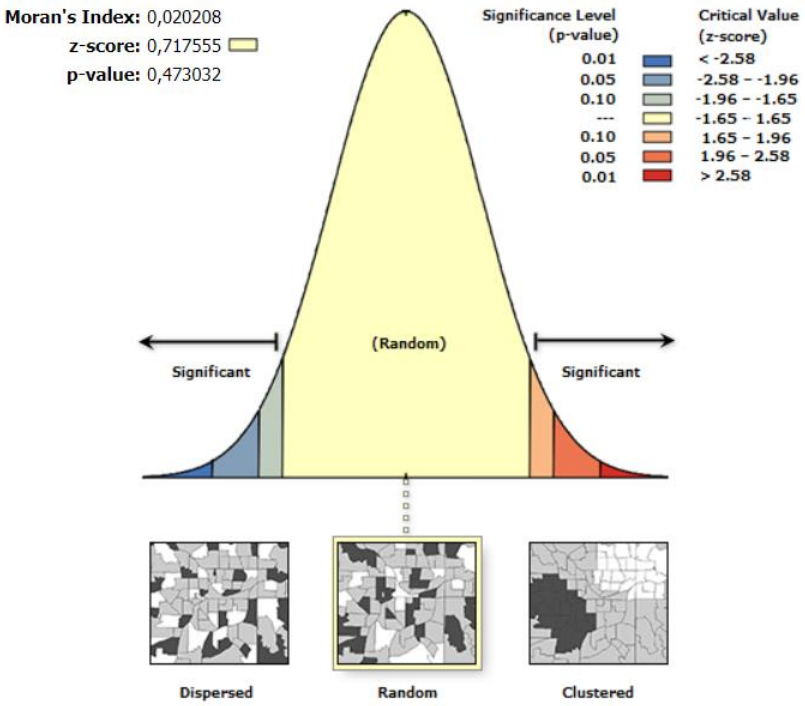


Figure 15: Moran's index showing no significant spatial autocorrelation

4.3.2 Feature importance

Table 4 shows the relative weight of each feature for predicting species richness using the XGBoost model based on the Gini impurity index. From all the possible input features, only 8 have any predictive weight. The weight distribution shows that elevation has by far the largest predictive weight, while a Northern facing slope is also a significant predictor for species richness. These factors were also noticeable in the aerial photograph interpretation, with several clusters only occurring at certain height or slope aspects. Together these features decide approximately 70% of the variance in predictions, with the other 6 features accounting for the other 30%. Some reasons for the relative importance of elevation in predicting species richness in the Christoffel national park will be discussed further in the next chapter.

Table 5 gives an example of how a species richness of 28.5 is calculated for a certain point and the positive or negative contribution of each feature. The bias is the average prediction on which extra species are added or subtracted depending on the feature. The elevation and the slope of this sample is rather high, which has a positive effect on the species richness. In contrast, the NDVI is low and the solar radiation is high which has a negative effect. For each prediction, only one slope aspect has a value of 1 while the rest are 0. The South aspect clearly has a negative impact on species richness, while the Northern aspect clearly has a positive impact as a value of 0 gives a negative contribution. An analysis of different species richness calculations gives the following insights:

a. Low values for slope, elevation and NDVI lead to a negative contribution, while higher values lead to a positive contribution to species richness

b. Low values for solar radiations give positive contributions while higher values give negative contributions to species richness

c. North and North-East aspects have positive effects on species richness while South and South-East aspects have a negative effect on species richness. Other slope aspects do not have predictive weight in this model.

Weight	Feature
0.5714	Elevation
0.1358	Northern aspect
0.0873	Dry season NDVI
0.0676	Slope in degrees
0.0411	Solar radiation
0.0360	South-East aspect
0.0353	South aspect
0.0254	North-East aspect

Table 4: Relative predictive weight of each feature

Contribution	Feature	Value
+20.928	<BIAS>	1.000
+8.855	Elevation	214.000
+2.044	Slope in degrees	18.028
+0.193	South-East aspect	0.000
-0.086	North-East aspect	0.000
-0.283	Solar radiation	105828
-0.316	North aspect	0.000
-0.997	NDVI	0.196
-1.828	South aspect	1.000

Table 5: Example of species richness calculation

4.3.3 Predicted species richness map

Figure 16 shows the high-resolution species richness map predicted by the XGBoost model. The feature importance's in table 4 can be clearly seen in the predictions, as areas with high elevations have high species richness and areas with low elevation has lower species richness. In addition, the differences in species richness for different slope aspects is also visible with Northern aspects showing higher species richness and Southern aspects showing lower species richness.

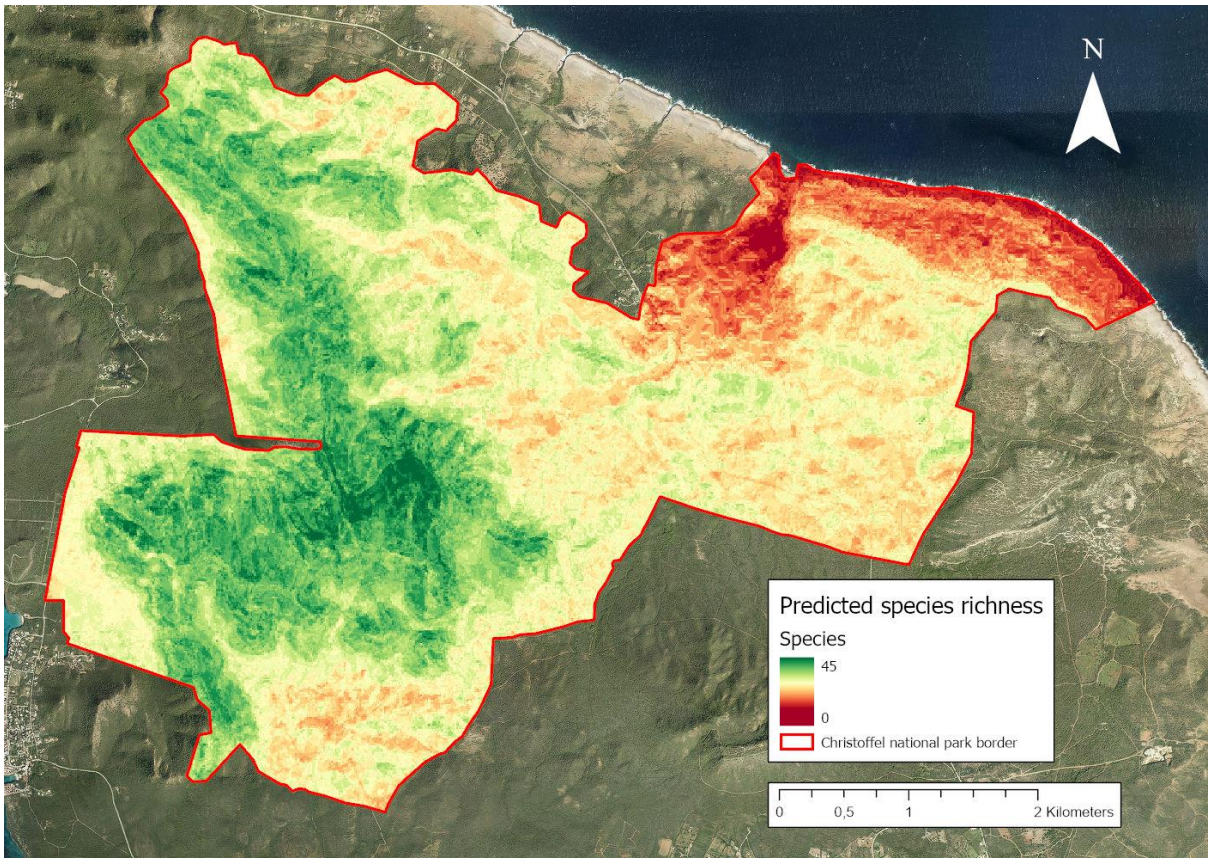


Figure 16: predicted species richness map

The residuals have shown that the highest species richness values are underestimated, and the lowest values are overestimated, this decreases the interpretability of the map when species richness is visualized in a stretched way. An alternative visualization is a relative classification from high to low species richness. Figure 17 show species richness as 5 classes ranging from high to low based on Jenks natural breaks. With this method, class breaks are identified that best group similar values and that maximize the differences between classes, this technique works well with normally distributed data such as the species richness predictions. In this visualization, the high species richness values that tend to be underestimated still fall in the highest class and vice versa for the lowest species richness. The relative species richness map is therefore better for interpretation purposes.

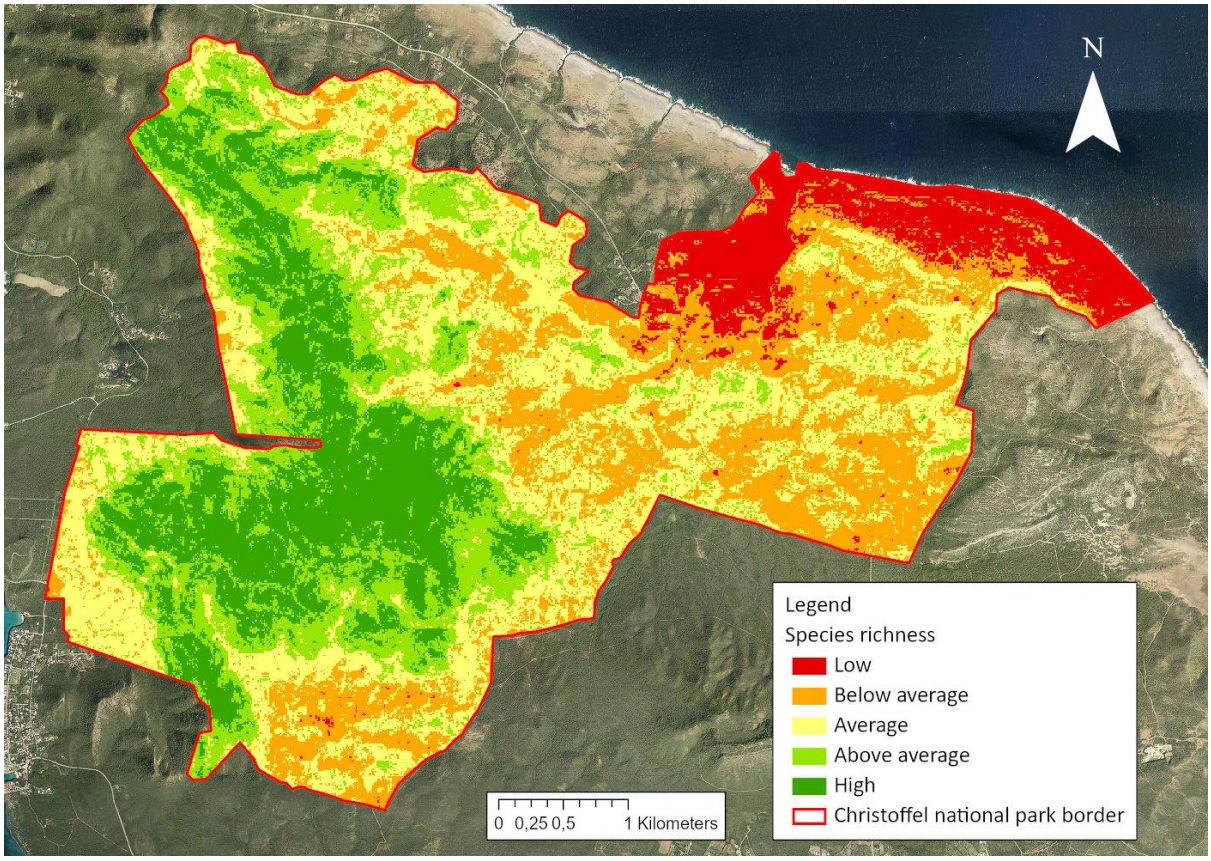


Figure 17: Relative species richness in the Christoffel national park

5. Discussion

5.1 Vegetation change

The analysis of vegetation change since 1985 shows large scale vegetation recovery and secondary succession in The Christoffel national park. These findings are proof of the immense threat that grazers are to island ecosystems and is consistent with other research showing strong recovery of native vegetation in island ecosystems after reducing grazing pressure (Scowcroft & Hobdy, 1987).

Much attention has been spent on researching the vegetation in the Christoffel national park in previous research, and in this thesis a trend of secondary vegetation succession has been found for the last decades. The effect of grazers on vegetation and the phases of recovery are decently clear, however the speed at which the vegetation recovers in the absence of grazers and how long it takes to reach each stage of recovery is unknown. Since 1994, CARMABI is also in charge of the Shete Boka national park just to the North of the Christoffel national park (figure 18). The geological background and location of the national park is nearly identical to the coastal areas of the Christoffel national park, but one of the biggest differences between the two is that herds of goats are still allowed to roam in the Sheta Boka park. Short expeditions to the park have shown that the area is very bare and that the vegetation is dominated by grasses, *Opuntia caracassana* and *Vachelia tortuosa* which are typical species that survive under high grazing pressure. Permanent plots could be set up in the Sheta Boka park to compare the recovery between ungrazed and grazed plots similar to Debrot & de Freitas (1993) to shed more light on the speed and mechanisms of recovery.

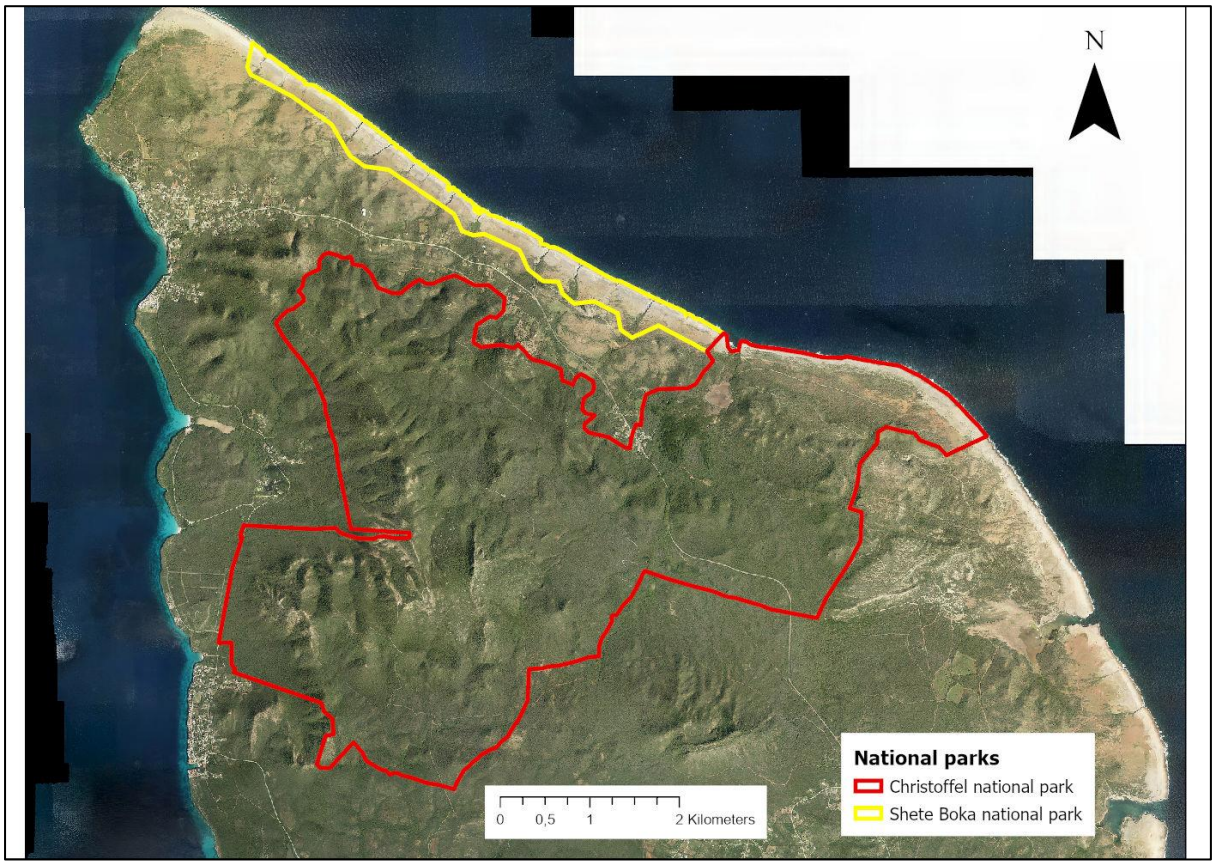


Figure 18: Shete Boka national park in relation to the Christoffel national park

5.2 TWINSPAN clusters map

5.2.1 Uncertainty

As has been mentioned in the problem definition, aerial photograph interpretations can be hard to validate. Some vegetation clusters were easily differentiable based on photo features, especially those along the coast. Inland it becomes a lot more difficult as the floristic composition becomes more diverse and some clusters will have similar photo features. It is especially hard when the differentiable vegetation between clusters are based on the undergrowth which cannot be seen from the aerial photograph. In some cases, such as with *Bromelia humilis*, the undergrowth is visible through sentinel-2 data, in other cases it is very hard to differentiate in which case location and distance to the nearest sampling point becomes the deciding factor. This is especially a problem in the broad midland clusters (8,9,10). For this reason, the exact borders of a few clusters are rather uncertain. Increasing the sampling points in underrepresented areas will give more information regarding the vegetation distribution and can drastically reduce the uncertainty. In the end, these results reinforce the shortcomings of aerial photograph interpretation for vegetation surveys compared to data-based modelling techniques.

5.2.2 Comparison with the species richness map

An analysis of the distribution of the TWINSPAN clusters on the aerial photograph interpretation map shows similarities to some of the findings in the species richness map. Many vegetation communities only occur on very specific location, such as high elevation or certain slope aspects. An inspection of the weight of predictor variables in the models also show that these features are important in deciding species richness. This points to a close relation between the type of vegetation community that is present at a location, and the species richness. The XGBoost model that was used to create the final predicted species richness map does not use the TWINSPAN aerial photograph interpretation map as input, but an analysis of the feature importance of the GLM and RF models does show that certain communities are linked to high species richness while other are linked to low species richness.

5.3 Species richness prediction

5.3.1 Model performance

There is a very low mean bias for all the models, which means that the main environmental and spatial trends were captured by the models. On the contrary, all models show an overestimation of lower values and an underestimation of higher values, with predictions centered around the mean. This is consistent with Guisan & Theurillat (2000), Thuiller et al. (2006), Algar et al. (2009), Newbold et al. (2009), Dubuis et al. (2011) and Biber et al. (2019) who also show that MEM predictions show bias at the extremes.

The XGBoost model yields a higher prediction accuracy than the other tested models. A comprehensive evaluation of predictive performance of models for vegetation data also showed XGBoost outperforming generalized linear models and random forests (Norberg et al., 2019). Nielsen (2016) studied the reason why tree boosting, and in particular XGBoost, excels at predictive tasks. Tree boosting is effective because it fits additive tree models, which have a good representational ability, using adaptively determined neighborhoods. XGBoost uses Newton boosting instead of the more common gradient boosting of other tree boosting algorithms, which allows it to learn tree structures better. Furthermore, XGBoost uses clever penalization of individual trees (Nielsen, 2016).

The XGBoost algorithm is still rather uncommon in academia, which might be due to its relative newness and the fact that tree boosting algorithms are harder to tune compared to random forests and other common algorithms. The consistent outperformance of XGBoost over other models should give it a higher priority in future predictive studies.

Aside from the higher prediction accuracy, the XGBoost model was chosen to be the final predictive model to create a species richness prediction map because it had the highest correlation coefficient and lowest error, without using the TWINSPAN clusters as independent variable. The predictive performance of the GLM and RF model does increase by using this variable, but it has been mentioned that the TWINSPAN vegetation cluster map cannot be easily validated, which would increase the uncertainty of the prediction map if it were included as an independent feature.

5.3.2 Model improvement

A lot of time has been spent on fine tuning the model, so it is unlikely that changing parameters will increase the prediction accuracy significantly. Model performance could be improved by adding more climatic data such as precipitation, evaporation and temperature. The currently available global climatic data sources are at a very coarse resolution and unsuitable for high resolution predictions, which means that local data must be gathered. There are already many small measurement stations across Curaçao to gather spatial precipitation data, additions around the national park could be used to create an interpolated map of precipitation and evaporation. The added benefit of using climatic variables is that the relation between climatic variables and species richness can be used to make future predictions regarding species richness distribution as a result of climate change (Biber et al., 2019). It is mentioned in the background chapter that the island of Curaçao is expected to become more arid as a result of climate change, which will change species distribution. Having predictive insight in the changes that might occur as a result of climate change can improve current management strategies.

Climatic variables are currently indirectly part of the model as higher elevations get more precipitation on Curaçao and the temperature is lower which makes for a more suitable habitat for many plant species. In addition, the Christoffel mountain and surroundings are dominated by the Knip formation which have a better water retention ability as has been mentioned in the background chapter. These factors partly explain why elevation has such a large predictive weight. Even with the addition of more data, there will always be a significant amount of unexplained variance because high resolution species richness is intrinsically hard to predict due to its reliance on very local factors.

5.3.3 Main contribution to the field of predictive plant species richness modelling

The significance and main contribution of this thesis to the wider field of plant species richness modelling is the finding that plant species richness can be predicted at a high resolution at similar accuracy levels as existing research that uses coarser resolution predictor variables (Thuiller et al., 2006; Dubuis et al., 2011). High-resolution species richness maps give better insights in more local factors compared to coarse resolution species richness maps and can be used to make better ecological management decisions on a local scale. In addition, this thesis shows that XGBoost should be considered more often in the academic field of predictive modelling, as it routinely outperforms other common regression models in academic studies (Norberg et al., 2019) as well as data science challenges (Nielsen, 2016).

6. Conclusion

Several conclusions can be drawn based on the initial research questions that were posed. Firstly, large scale changes were found in the coverage of species in the Christoffel national park and there was a clear spatial pattern of recovery. Since 1985, the coverage of grasses and herbs has drastically decreased while an increase in high shrub and tree species was found. This is a clear sign of secondary vegetation succession which is likely the result of a decrease in grazing pressure. The vegetation recovery was minimal on the Christoffel mountain and the surrounding areas, as this has traditionally been the area where most of the climax vegetation grew. The midland and coastal areas of the Christoffel national park have seen a large recovery of the endemic vegetation, with some plots reaching climax vegetation. These findings reiterate the threat that grazers pose to native island ecosystems, but also show that the vegetation can recover when the grazing pressure is reduced. More research in the Sheta Boka park can shed light on the speed and mechanisms of recovery as a result of grazing pressure reduction.

The 220 sampling points were divided in 22 vegetation clusters with TWINSPAN clustering. The resulting clusters show significant spatial dependence, with many vegetation communities only occurring on very specific elevations and slope aspects. The prediction model also showed that these factors were important in species richness predictions, suggesting that there might be a link between the vegetation communities and the species richness. The aerial photograph interpretation could be applied well in vegetation communities that had easily differentiable photo features, however some vegetation communities were hard to differentiate due to the occurrence of similar species and the lack of available information regarding the undergrowth from a top down view. This leads to some level of uncertainty in the exact borders between similar vegetation clusters and these borders would need to be validated with more sampling points in underrepresented areas.

The high-resolution species richness models predicted the species richness with similar levels of accuracy as coarser resolution models that are commonly used. XGBoost, an implementation of tree boosting, has the highest accuracy while using less features than the generalized linear model and the random forest model. Elevation and slope aspect have the highest predictive weight. Higher elevation, slope and NDVI correlated with higher species richness while higher solar radiation correlated to lower species richness. Northern facing slopes had a positive effect on species richness, while southern facing slopes had a negative effect. There have been few studies that have looked into high-resolution plant species richness modelling; however, this thesis shows that the predictions have similar accuracy as existing coarser resolution models found in literature, with the added benefit of getting more insight into the mechanisms of species richness at a local scale, which can be used for better ecological management.

7. References

Algar, A. C., Kharouba, H. M., Young, E. R., & Kerr, J. T. (2009). Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. Ecography, 32(1), 22-33.

Araújo, M. B., & Rozenfeld, A. (2014). The geographic scaling of biotic interactions. Ecography, 37(5), 406-415.

Aravindan, C., & Jaisakthi, S. M. (2019). Species Recommendation using Machine Learning-GeoLifeCLEF 2019.

Arnoldo, M., & van Proosdij, A. S. (2012). Arnoldo's zakflora: wat in het wild groeit en bloeit op Aruba, Bonaire en Curaçao. Walburg Pers.

Baker, J.K. & D. W. Reeser (1972) Goat management problems in Hawaii Volcanoes National Park. Natural Resources Report, No 2, National Park Service, U.S. Department of the Interior, Washington d.c.

Beard, J.S. (1944) Climax vegetation in tropical America. Ecology 25: 127-158

Beers, C. E., De Freitas, J., & Ketner, P. (1997). Landscape ecological vegetation map of the island of Curaçao, Netherlands Antilles. Natuurwetenschappelijke studiekring voor Suriname en de Nederlands Antillen nr. 138.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific data*, *5*, 180214.

Benzing, D.H. (1980) The biology of the bromeliads. Mad River press, Eureka, California

Biber, M. F., Voskamp, A., Niamir, A., Hickler, T., & Hof, C. (2019). A comparison of macroecological and stacked species distribution models to predict future global terrestrial vertebrate richness. Journal of Biogeography.

Bokkestijn, A., & Slijkhuis, J. (1987). Een vegetatiekundige detailkartering in het Christoffelpark, Curaçao. Landbouwuniversiteit, Vakgroep Vegetatiekunde, plantenoecologie en onkruidkunde.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Buisonjé, P. H. (1974). Neogene and Quaternary geology of Aruba, Curaçao and Bonaire. Natuurwetenschappelijke studiekring voor Suriname en de Nederlandse Antillen.

Carabine, E., & Dupar, M. (2014). The IPCC's fifth assessment report: what's in it for small island developing states.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Coblentz, B.E. (1978) The effects of feral goats (Capra hircus) on island ecosystems. Biol. Conserv. 13: p. 279-286

Coblentz, B.E. (1980) Goat problems in the national parks of the Netherlands Antilles. Carmabi internal document.

Coolen, Q. (2015) The impact of feral goat herbivory on the vegetation of Bonaire. MSc thesis resource ecology, Wageningen university & research.

D'Amen, M., Rahbek, C., Zimmermann, N. E., & Guisan, A. (2017). Spatial predictions at the community level: from current approaches to future frameworks. Biological Reviews, 92(1), 169-187.

Debrot, A.O. & de Freitas, J.A. (1993) A comparison of ungrazed and livestock-grazed rock vegetation in Curaçao. Biotropica 25:270-280

De Freitas, J. & Wakkee, P. (1984) Kwalitatieve natuur- en landschapsevaluatie van het eiland Curaçao ten behoeve van planningsdoeleinden op basis van twee case-studies. Carmabi & DROV, 34pp.

De Freitas, J. (1991) The Flora and Fauna of Curaçao - some aspects. Curaçao Tourism Development Foundation Awareness Programme 1991

De Freitas, J. D., Nijhof, B. S. J., Rojer, A. C., & Debrot, A. O. (2005). Landscape ecological vegetation map of the island of Bonaire (southern Caribbean). Royal Netherlands Academy of Arts and Science.

De Freitas, J. A., Rojer, A. C., Nijhof, B. S. J., & Debrot, A. O. (2014). Landscape ecological vegetation map of Sint Eustatius (Lesser Antilles). Koninklijke Nederlandse Akademie van Wetenschappen.

De Freitas, J., Rojer, A. C., Nijhof, B. S. J., & Debrot, A. O. (2016). A landscape ecological vegetation map of Saba (Lesser Antilles) (No. C195/15). IMARES.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J. P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. Diversity and Distributions, 17(6), 1122-1131.

Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: where to from here?. Systematic biology, 51(2), 331-363.

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. Journal of applied ecology, 43(3), 393-404.

Garcia-Franco, J.G., Rico-Gray, V., Zayas, O. (1991) Seed and seedling predation of Bromelia pinguin L. by the red land crab Gecarcinus lateralis Frem. In Veracruz, Mexico. Biotropica 23:96-97

Gaspard, G., Kim, D., & Chun, Y. (2019). Residual spatial autocorrelation in macroecological and biogeographical modeling: a review. Journal of Ecology and Environment, 43(1), 19.

Guisan, A., & Theurillat, J. P. (2000). Equilibrium modeling of alpine plant distribution: how far can we go?. Phytocoenologia, 30(3/4), 353-384.

Harris, D. J., Taylor, S. D., & White, E. P. (2018). Forecasting biodiversity in breeding birds using best practices. PeerJ, 6, e4278.

Herdter, E. (2019). Using extreme gradient boosting (XGBoost) to evaluate the importance of a suite of environmental variables and to predict recruitment of young-of-the-year spotted seatrout in Florida. bioRxiv, 543181.

Hill, M. O. (1979). A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. TWINSPAN.

Klaver, G. T. (1987). The Curaçao Lava Formation: an ophiolitic analogue of the anomalous thick layer 2B of the mid-Cretaceous oceanic plateaus in the western Pacific and central Caribbean.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

Kühn, I. (2007). Incorporating spatial autocorrelation may invert observed patterns. Diversity and Distributions, 13(1), 66-69.

Lahey, J. F. (1958). 'On the origin of the dry climate in the northern South American and the southern Caribbean', Scientific Report 10 University of Wisconsin.

Louthan, A. M., Doak, D. F., & Angert, A. L. (2015). Where and when do species interactions set range limits? Trends in Ecology & Evolution, 30(12), 780-792.

Martis, A., Oldenborgh, G.J., Burgers, G. (2002) Predicting rainfall in the Dutch Caribbean—more than El Niño? International Journal of Climatology. Volume 22, Issue 10

McCullagh, P. & Nelder, J.A. (1989) Generalized linear models, 2nd editionn. Chapman and Hall, London.

Meteorological Department Curaçao (2016) *Climatological Report 2016*. Meteorological Department Curaçao Seru Mahuma z/n. Curaçao, Dutch Caribbean.

Meyboom, M.J. (1994) De invloed van geiten op de natuurlijke vegetatie van het christoffelpark te Curaçao. Landbouwuniversiteit Wageningen, verslag Natuurbeheer Nr. 3144

Newbold, T., Gilbert, F., Zalat, S., El-Gabbas, A., & Reader, T. (2009). Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. Journal of Biogeography, 36(11), 2085-2095.

Nezer, O., Bar-David, S., Gueta, T., & Carmel, Y. (2017). High-resolution species-distribution model based on systematic sampling and indirect observations. Biodiversity and conservation, 26(2), 421-437.

Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win "every" machine learning competition? (Master's thesis, NTNU).

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., ... & Foster, S. D. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs, 89(3), e01370.

Pineda, E., & Lobo, J. M. (2009). Assessing the accuracy of species distribution models to predict amphibian species richness patterns. Journal of Animal Ecology, 78(1), 182-190.

Putney, A.D. (1982) Survey of conservation priorities in the Lesser Antilles : final report. Eastern Caribbean Natural area Management Program.

Salem, B.B. (1989) Arid Zone Forestry: A Guide for Field Technicians. Issue 20 of FAO conservation guide, Volume 20 of Fao Food and Nutrition Paper.

Sandino, J., Gonzalez, F., Mengersen, K., & Gaston, K. J. (2018). UAVs and machine learning revolutionising invasive grass and vegetation surveys in remote arid lands. Sensors, 18(2), 605.

Scowcroft, P. G., & Hobdy, R. (1987). Recovery of goat-damaged vegetation in an insular tropical montane forest. Biotropica, 208-215.

Stockwell, D. R., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. Ecological modelling, 148(1), 1-13.

Stoffers, A.L. (1956) Studies on the flora of Curaçao and other Caribbean islands. Natuurwetenschappelijke studiekring voor Suriname en de Nederlands Antillen, No. 15.

Thuiller, W., F. Midgley, G., Rougeti, M., & M. Cowling, R. (2006). Predicting patterns of plant species richness in megadiverse South Africa. Ecography, 29(5), 733-744.

Tilman, D. (1982). Resource competition and community structure. Princeton university press.

Trewartha, G. T. 1981. The Earth's Problem Climate University of Wisconsin.

Van Buurt, G. (2009). A short natural history of Curaçao. In Crossing Shifting Boundaries, Language and Changing Political status in Aruba, Bonaire and Curaçao. Proceedings of the ECICC Conference, Dominica (Vol. 1, pp. 229-256).

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. Diversity and distributions, 14(5), 763-773.

Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2019). How to predict biodiversity in space? An evaluation of modelling approaches in marine ecosystems. Diversity and Distributions, 25(11), 1697-1708.

Zonneveld, I. S. (1979). Use of aerial photographs in geography and geomorphology. International Training Centre for Aerial Survey ITC.

8. Appendix

A: Model parameters and features

Model	Parameters	Features
XGBoost	Learning rate: 0.23 Maximum depth: 1 Minimum child weight: 2.02 Number of estimators: 130 Alpha regularization: 0.28 Subsample: 0.90	- Altitude - Slope - Solar radiation - NDVI - Slope aspect
Random forest	Number of estimators: 400 Maximum depth: 181 Min_samples_split: 5 Min_samples_leaf: 1 Bootstrap: True	- Altitude - Slope - Solar radiation - NDVI - Aspect - Geology - TWINSPAN clusters
Linear model	-	- Altitude - Slope - Solar radiation - NDVI - Aspect - Geology - TWINSPAN cluster