Jasmine Planitzer, BSc

Satellite Remote Sensing for Environmental Modeling of Aedes species in Kinshasa District, DRC

Master's Thesis

to achieve the university degree of

Master of Science

Master's degree programme: Geospatial Technologies

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dr.rer.nat. Dipl.-Forstwirt Mathias Schardt

Institute of Geodesy
Head: Univ.-Prof. Dr.rer.nat. Dipl.-Forstwirt Mathias Schardt

Graz, January 2018

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date	Signature

Acknowledgement

I would first like to express my sincere gratitude to my advisor Univ.-Prof. Dr.rer.nat. Dipl.-Forstwirt Mathias Schardt of the Institute of Geodesy at the Graz, University of Technology, for the continuous support of my master thesis. His patience, motivation, and knowledge helped me all the time of my research and writing of the thesis. The inspiring discussions kept me continuously motivated to conduction my study. I would not have imagined having a better advisor.

Beside my advisor, I would also like to thank Univ.Ass. Dr.techn. MSc Mariette Vreugdenhil, Dr.rer.nat. MSc Matthias Forkel and Univ. Prof. Dipl.-Ing. Dr.techn. Wolfgang Wagner of the Department of Geodesy and Geoinformation at the Vienna, University of Technology, for their insightful comments and encouragement. During the summer of 2017, I had the tremendous opportunity to join their team as Women4Geo intern. Without their support, it would not have been possible to conduct this research. I would also like to acknowledge Ass.Prof. Dr. Stefan Lang of the Interfaculty Department of Geoinformatics – ZGIS at the University of Salzburg, who inspired and motivated me to write the present thesis. Further, I would like to acknowledge Edith Rogenhofer and Idriss Ait-Bouziad from Médecins sans Frontières (Doctors without Borders) for providing me medical advice and test samples.

Finally, I must express my very profound gratitude to my parents, family, and friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

15th January 2018 Jasmine Planitzer

Abstract

The major motivation for assessing environmental indicators for epidemiological studies is to provide information about diseases conditions arising from the physical environment to policy-makers and health authorities. Communicable diseases such as malaria, dengue fever, west nile virus or yellow fever are transmitted by different mosquito species (vector-borne diseases). Thus, the evaluation of potential breeding habitats contributes to diseases conditions and dynamics. Yellow fever, occurring in the tropics and subtropics, is transmitted by the Aedes mosquito species. The aim, therefore, is to identify and define the most important environmental indicators detected by remote sensing imagery that imply to the distribution of the vector and consequently the potential abundance of yellow fever. The risk zone mapping and the assessment of environmental indicators are based on the random forest classifier, a state-of-the-art ensemble machine learning approach. To evaluate the use of multiple remote sensing data active and passive remote sensing products are employed on variable importance models. The model solely based on active microwave remote sensing data shows better results than the other implemented models, indicating that microwave remote sensing products are more suitable for predicting the abundance of yellow fever in endemic countries.

Zusammenfassung

Die Evaluierung von Umweltindikatoren für epidemiologische Untersuchungen liefert nützliche Informationen für Entscheidungsträger und Gesundheitsministerien über Krankheiten, welche durch eine sich veränderte Umwelt auftreten. Übertragbare Krankheiten wie Malaria, Dengue Fieber, das Zika- oder das West Nil Virus sowie Gelbfieber werden mittels unterschiedlicher Stechmückenspezies übertragen (durch Moskitos übertragene Krankheiten). Resultierenden daraus sollten potenzielle Brutstätten und Habitate dieser Stechmückenspezies identifiziert werden. Gelbfieber, welches ausschließlich in subtropischen sowie tropischen Regionen vorkommt, wird durch die Stechmückenspezies Aedes übertragen. Das Ziel dieser Arbeit ist es daher, die wichtigsten und ausschlaggebendsten Umweltindikatoren, welche sich durch Fernerkundungsdaten ableiten lassen, zu identifizieren und definieren und welche auf das Vorkommen von Stechmücken und daraus möglicherweise auf das Auftreten von Gelbfieber schließen lassen. Die Kartierung von Risikozonen sowie Beurteilung von Umweltindikatoren basiert auf der Verwendung des Random Forest Algorithmus, welcher aus dem Bereich des maschinellen Lernens stammt. Aktive sowie passive Fernerkundungsdaten werden verwendet und anhand von Modellen evaluiert, welche den Einfluss jeder Variable bestimmt. Jenes Modell, welches nur auf Mikrowellendaten aus aktiven Fernerkundungssensoren beruht, liefert das beste Klassifikationsergebnis. Resultierend daraus kann abgeleitet werden, dass Indikatoren basierend auf Mikrowellensensoren besser für die Vorhersage des Auftretens von Gelbfieber in endemischen Regionen geeignet sind als jene von passiven Sensoren.

Contents

List of Figures

List of Tables

1 List of Acronyms

ANN Artificial Neural Network
AOT Aerosol Optical Thickness
AVHRR Advanced Very High Resolution Radiometer
BOA Bottom of Atmosphere
CCI Climate Change Initiative (ESA)
DF Dengue Fever
DT Decision Tree Classifier
DRC Democratic Republic of Congo
DWSI Disease Water Stress Index
EO Earth Observation
ERTS Earth Resource Technology Satellite (LANDSAT-1)
EVI Enhanced Vegetation Index
FAO Food and Agriculture Organization of the United Nations
FPAR Fraction of photosynthetically active radiation
GMES Global Monitoring for Environment and Security
HR High Resolution
LAI Leaf Area Index
LST Land Surface Temperature
LULC Land Use Land Cover
MLC Maximum Likelihood Classification
MODIS Moderate-resolution Imaging Spectroradiometer
MSAVI Modified Soil Adjusted Vegetation Index
MSF Médecins Sans Frontières (Doctors without Borders)
MSI Multi Spectral Instrument
NDWI Normalized Difference Water Index
NDVI Normalized Difference Vegetation Index
NOAA National Oceanic and Atmospheric Administration
OLI Operational Land Imager (LANDSAT-8 OLI)
RF Random Forest

1 List of Acronyms

RS Remote Sensing
RVF Rift Valley Fever
SAR Synthetic Aperture Radar
SAVI Soil Adjusted Vegetation Index
SVM Support Vector Machine
TOA Top of Atmosphere
VBD Vector Borne Disease
VI Vegetation Index
VHR Very High Resolution
WHO World Health Organization
WNV West Nile Virus
WV Water Vapor
YF Yellow Fever
YFV Yellow Fever Virus

2 Introduction to Vector Borne Diseases and Habitat Analysis

2.1 Background

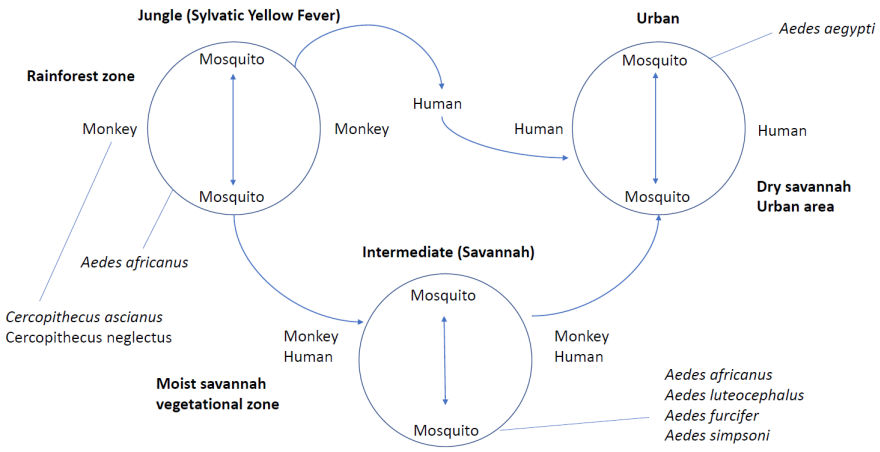
Current evidence suggests that human health is dependent on the dynamics of the Earth's climate system. Global change including environmental, climatic, societal and demographic change, has a significant impact on humans as well as natural systems. Several direct and indirect effects of global change have been observed, including the increase of extreme weather events such as droughts or floods leading to insufficient food production and freshwater supplies or the re-emerge of vector-borne diseases (VDBs) (IPCC, 2014; Hagenlocher, 2014).

According to the World Health Organization (WHO), vector-borne diseases such as malaria, dengue fever or yellow fever account more than 17% of all infectious diseases, causing more than 1 million deaths annually. Vectors are living organisms such as mosquitoes or ticks that transmit contagious diseases between humans or from animals, mostly vertebrates to humans. Many of these vectors are blood-soaking insects, which get infected by the blood meal from infected hosts (humans, animals) and later inject it to new hosts (WHO, 2014). This thesis focuses on one mosquito-borne disease which was selected due to recent outbreaks in Africa and impacts on vulnerable populations and national and international health systems: yellow fever.

Yellow fever (YF) virus is an arthropod-borne virus (arbovirus) from the Flavivirus genus, occurring in tropical and subtropical areas, which strikes an estimated 84 000 – 170 000 persons worldwide each year and causes an estimated 29 000 – 60 000 deaths. It is transmitted by the *Aedes* species in

Africa and *Haemagogus* species in South America (Jentes et al., 2011; WHO, 2016a). The "yellow" refers to jaundice that appears during the disease progression. Symptoms include high fever, headache, muscle pain, nausea, vomiting, and fatigue. Epidemics occur when infected humans introduce the virus into populated areas with high vector density and where most people lack vaccination, because of poor maintenance of vaccination campaigns or an increasing number of people borne since the last epidemics. Under these circumstances, infected vectors (mosquitoes) transmit the virus from person to person. The transmission is divided into three cycles: the sylvatic yellow fever cycle (jungle yellow fever cycle), the intermediate yellow fever cycle (savannah yellow fever cycle) and the urban yellow fever cycle. Depending on the cycle different vectors and hosts are responsible for the spread of the disease (WHO, 2016a).

Figure 2.1: Yellow Fever Transmission Cycle



Redrawn from WHO 2016

The transmission stages are best understood if ecological zones and indicators such as rainfall patterns and vegetation are considered. The transmission cycle is visualized in figure 2.1. Within the jungle cycle, characterized by tropical rainforest, the transmission is predominantly monkey-monkey via the mosquito species *Aedes africanus*, and human infection is sporadic.

The cycle zone is characterized by year-round enzootic transmission. However, the virus activity is generally at a low level and local outbreaks rarely occur. The second cycle occurs within the moist savannah zone. During the rainy season this zone is responsible to repeated emergence of YF activity, because of a high vector abundance including the vector *Ae. Africanus, Ae. Luteocerphalus, Ae. Furcifer* and *Ae. Simpsoni*. The urban area is characterized by dry savannah, precipitation tends to be lower than in the moist savannah and the rainy season is abbreviated. The virus is introduced to the urban area by the *Ae. Aegypti* vector, which occur predominately in urban, populated areas, where humans provide artificial breeding containers. If the virus is introduced in urban areas, because of migrating humans, that travel between different ecological zones, explosive outbreaks of *Ae. Aegypti* – borne yellow fever may induce (Vainio and Cutts, 1998).

Despite the availability of a vaccine and mass vaccination campaigns, yellow fever (YF) is considered as a reemerging disease caused by an increase in incidences in the past 25 years (Barrett, Higgs, 2007). The virus originates in Africa and was imported to America during the 16^{th} century and claimed thousands of lives during the 18^{th} and 19^{th} centuries. In 1937, Max Theiler invented a yellow fever vaccine that is still in use today and provides lifetime immunity in up to 99%. However, outbreaks continue to occur especially in Western Africa and Southern America. Beginning in December 2015, a large yellow fever outbreak has been reported in Angola and the Democratic Republic of Congo (DRC), with 7293 suspected cases and 137 deaths. The outbreak, which was first detected in Angola, spread to the entire country and to the neighboring country DRC, where local transmission was first detected in March 2016 (Kraemer et al., 2017; Paules and Fauci, 2017; WHO, 2017).

The physical environment is characterized by many aspects including land cover, air and climate characteristics. Remote sensing and earth observation is applied to health studies when the physical environmental and physical indicators need to be assessed. One main scope of remote sensing in health studies is the characterization of environmental indicators for vector-borne diseases. Remote sensing imagery offer several advantages compared to traditional health studies. They enable surveys on large or inaccessible areas on different spatial and temporal resolutions and a growing rate of images

are free available. And the availability of long archives provides the possibility of performing long term studies to investigate the influence of global environmental change on human health (Kotchi et al., 2016).

2.2 Objectives and Research Questions

The overall aim of this thesis is to investigate the relations between environmental characteristics derived from satellite remote sensing imagery and the risk of yellow fever occurrence in Africa. The central aspect is to assess the distribution of yellow fever vectors (mosquitos) using environmental and spatial indicators. The major advantage will be that the proposed method, which is based on data level machine learning, avoids cost-extensive in-situ surveys and thus easy access to information about the risk of the disease distribution.

The research is based on active and passive remote sensing products including an ESA CCI Land Cover Map and imagery from SENTINEL-2A, LANDSAT-8, MODIS Terra and Aqua, and ASCAT. Concerning the study objective, several aspects will be considered and analyzed. SENTINEL-2A image data will be preprocessed using the recommended Sen2Cor preprocessor and further classified using automated image classification. LANDSAT-8 scenes will be preprocessed and used for extracting pixel values and calculation of several data-level metrics and indices. MODIS Terra and Aqua data will be applied to the study to assess the importance of land surface temperature, vegetation indices, leaf area index and the fraction of photosynthetically radiation. ASCAT soil moisture data, which is provided by the Vienna University of Technology, will be employed to investigate the importance of soil moisture on the occurrence of the yellow fever vector. A land cover map provided by the Climate Change Initiative of the European Space Agency will be utilized to evaluate the use of different remotely sensed land cover types. With regard to all data sources, the impact of each remote sensing product for epidemiological studies will be assessed.

Thus, this research is located at the interface of ecological and environmental aspects, earth observation science (remote sensing) and public health.

The following research questions (RQs) are addressed to fulfill the objectives:

RQ 1: What are possible environmental indicators influencing the occurrence of the yellow fever virus (YFV) in Africa?

RQ 2: What kind of environmental indicators can be derived from remote sensing data and be applied for habitat analysis of hosts and vectors?

RQ 3: Is it possible to infer a yellow fever virus risk hazard index?

3 State of the Art in Epidemiological Research

The following chapter is a summary of the impact of Remote Sensing (RS) and Earth Observation (EO) in health science, including the modeling of VBDs and habitat analysis using remotely sensed data, and environmental and ecological indicators that affect the abundance of Ae. Species and distribution of Yellow Fever. It is based on a systematic review of the available literature. The research was realized using Internet search engines and keywords associated to VBDs, RS, and epidemiology (e.g., RS and VBDs or RS and yellow fever, malaria, dengue fever), as well as indices of relevant scientific journals. Articles referring to similar research studies were clustered together in groups.

3.1 Mapping of Habitats regarding Vector-Borne Diseases

The accessibility of multi-temporal satellite imagery for 45 years now, expedites the use of RS in epidemiology to gain a better understanding of disease dynamics. Since the launch of the satellite LANDSAT-1 in 1972 (initially named Earth Resource Technology Satellite ERTS-1), remotely sensed data had been used to acquire information about objects or phenomena on the Earth's surface, without direct physical contact. However, only a small number of researchers have applied RS data for detecting environmental indicators that might have impacts on VBDs (Beck et al., 2000; Herbreteau et al., 2006; Zhang et al., 2013).

One of the first research in RS and epidemiology was conducted by Cline in 1970, who highlighted the use of aerial photography and remote sensing techniques for epidemiological study. During the HAO (High Altitude Observatory) initiative by NASA scientists in 1971 color and color-infrared aerial photography was employed to map vegetation assemblages associated with the abundance of Aedes species, which are intermediate hosts for several infectious diseases such as Yellow Fever, Rift Valley Fever or Dengue Fever (Hay et al., 1997). Eight years later Wagner et al. (1979) conducted research also using color-infrared aerial photography to map open wetlands, marshes and residential areas (mosquito habitats) for mosquito control in Michigan (Wagner et al., 1979). Several studies in the following years highlighted the fact that the usage of aerial photography is more cost-effective than conventional ground surveys of mosquito breeding sites. The known flight range of the mosquitoes in combination with the distance to residential areas and classified habitats were used to identify control priorities (Hay et al., 1997). Rejmankova et al. (1988) investigated that the NDVI (Normalized Difference Vegetation Index) is considered as an essential variable for the abundance of mosquitoes. They demonstrated that rice crop cover is positively correlated with mosquito larval production. Higher NDVI values were associated with high mosquito producing rice fields. A discriminant analysis was conducted to distinguish between high and low mosquito producing fields resulting in an overall accuracy of about 75 % (Rejmankova et al. 1988). Wood and Muir (1988) proposed to use remote sensing imagery to derive environmental indicators such as land cover, precipitation or temperature for understanding vector abundances and dynamics. They highlighted to focus on monitoring the environmental consequences of deforestation and resources to detect potential breeding habitats (Wood and Moor, 1988).

During the 1990 various researchers have investigated the hypothesis that the spatial and temporal aspects of mosquito population dynamics can be predicted to environmental indicators observed by remotely sensed imagery. In 1990 the second phase of the HAO initiative started (Hay et al., 1997).

"*The second phase of the project investigated the population dynamics of Anopheles albimanus and Anopheles pseudopunctipennis in the tropical wetlands of Chiapas, Mexico, where malaria is endemic (Roberts et al. 1991). Pope et al. (1994) used two*

Landsat-TM scenes of the area, one from the dry and one from the wet season, to provide an unsupervised classification of the region. The resulting clusters were assigned to land-cover types on the basis of color-infrared aerial photographs and field inspection of 30 test sites. These sites were independently sampled for mosquito density and information was collected on environmental variables affecting water and vegetation characteristics (Rejmankova et al., 1991). These sites were then grouped into 16 habitat types using cluster analysis, and correlations were performed between the habitat types and land-cover units (Rejmankova et al. 1992). The cover-units were subsequently ranked as having high, medium or low mosquito production potential on the basis of these correlations. Incorporating this information into a GIS, sites of high mosquito production [...] were found to occupy only 9 percent of the designated control area, allowing the potential for substantial streamlining of control campaign effort and resources." (Hay et al., 1997)

Dry and wet season LANDSAT-TM scenes were again used by Beck et al. (1994) for stepwise discriminant analysis and linear regression to assess the relationship between the abundance of vectors and landscape (Hay et al., 1997).

"*Rejmakova et al. (1995) have also shown that the density of Anopheles albimanus mosquitoes around villages in Belize could be reliably predicted using multi-spectral Satellite Pour l'Observation de la Terre (SPOT) High-Resolution Visible (HRV) data. Productive larval habitats were first identified as marshes containing relatively few emergent aquatic plants and high coverage of cyanobacterial mats. An unsupervised Bayesian maximum likelihood classification was then applied to a single SPOT scene [...]. The classes generated were subsequently assigned to individual 'landscape elements' based on field observations. Human settlements were identified with ancillary map data and located more precisely on subsequent field visits with a global positioning system (GPS).*" (Hay et al., 1997)

At about the same time two researchers were conducted using coarse resolution satellite imagery such as the AVHRR (Advanced Very High-Resolution Radiometer) from the US National Environment Satellite, Data and Information Service (NOAA) to predict Rift Valley Fever (RVF) in Kenya. Satellite-derived green vegetation associated with RVF resulting in the assumption that satellite data may become a forecasting tool for RVF (Linthicum et al., 1987). Regional scale NDVI values were derived to identify

shallow wetlands *(dambos)*, which are highly suitable for mosquito breeding. Based on that work RVF epidemics were correctly predicted in 1989 in Kenya following exceptional NDVI values (Bailey and Linthicum, 1989).

Besides, the importance of soil moisture for prediction malaria transmission factors was investigated. Patz et al. (1998) conducted a study to get insights how modeled soil moisture influences malaria in Kenya using a water balance model. The authors concluded that modeled soil moisture improves the prediction of biting rates compared to usually used environmental indicator precipitation.

Further, additional airborne Synthetic Aperture Radar (SAR) data was investigated to detect the flooding status of wetland areas. In a study conducted by Pope et al. (1992) passive (LANDSAT-TM) and active (SAR) satellite imagery was combined to examine the feasibility of detecting RVF habitats by analyzing soil and land cover. Thomson et al. (1995) studied AVHRR data together with the Cloud Coverage Data (CCD) from the Food and Agricultural Organization (FAO) to predict malaria in the Gambia. They also noted the connection between malaria and environmental indicators including precipitation and land cover derived from satellite data is complex since sociological factors may play an essential role in the epidemiology of the disease. Because in areas where the biting rates were high due to a higher abundance of malaria mosquitoes people protect themselves using bed nets and insect repellents (Hay et al., 1997). Hay et al. (1997) conclude that there might be a possibility to initiate mosquito efforts before disease outbreaks using RS imagery and that the cost-effectiveness of RS data compared to ground surveys is a decisive factor to investigate further the use of remotely sensed imagery in vector control and epidemiology.

Hay et al. 1997 concluded: *"Our understanding of the spatial and temporal distribution of invertebrate intermediate hosts and the disease they transmit has been enhanced by understanding the statistical associations between the ecological variables and processes observed remotely [...]."* (Hay et al., 1997)

Masuoka et al. (2003) employed LANDSAT 7 ETM+ and IKONOS data, which is a commercial very high-resolution satellite operated by DigitalGlobe, to estimate the area covered by the malaria mosquito larval habitats

in the Democratic Republic of Korea (Satellite Imaging Corporation, 2017). Larval samples were collected to validate the classification results. Several potential larval habitats were identified: (1) rice fields, (2) streamside ponds, (3) irrigation ponds, (4) irrigation ditches, (5) drainage ditches, (6) swamps, and (7) rivers. They implemented multiple classification algorithms on both LANDSAT 7 and IKONOS scenes, a minimum distance, a maximum likelihood and a parallel-epiped classifier. The parallel-epiped classifier with a maximum likelihood as a tiebreak was the most accurate on both images. Though, small habitats such as small ponds were not classified on the LANDSAT scene due to the moderate resolution of the product. They concluded rough estimates of habitats of the malaria vector could be performed by LANDSAT 7 ETM+ images on local level; however, a very high-resolution product from satellites such as IKONOS allows better distinctions of vector habitats.

A similar study on malaria at the same year was conducted by Vasconcelos and Novo (2005). They investigated the influence of precipitation, deforestation and human operations on malaria incidences using precipitation and water level information as well as RADARSAT-1 and LANDSAT 5 Thematic Mapper (TM) remote sensing data. RADARSAT-1 was launched by the Canadian Government in 1995 to provide SAR images. It acquired images of the Earth day and night, in all weather and through clouds and haze. It was designated to provide information on risk assessment, interferometry, agriculture and other earth observing fields of application (Canadian Space Agency, 2014). They found out that malaria incidence locations are spatially related to deforestation areas and area with human operations such as road construction areas.

The use of IKONOS and LANDSAT 7 ETM+ data was again explored in 2006 identifying *Culex quinquefasciatus* which are responsible for transmitting West Nile Virus (WNV) and Rift Valley Fever (RVF) in Kenya (Jacobs et al., 2006). A LULC classification was conducted using a maximum likelihood classification to identify potential breeding habitats. Results show that LULC rice-cultivation fields are associated with high *Culex* abundance. Concluding the study LULC change especially the increase in rice-cultivation fields contribute the abundance and transmission of WNV and RVF.

Zeng et al. (2006) highlighted the fact that environmental conditions may be important determinants of VBDs outbreak and transmission. They defined several indicators that may influence the abundance and transmission vectors. First, climate as essential indicators including temperature, precipitation, pressure, humidity and short-term climatic variations such as El Nino events or the minimum/maximum of the mean monthly temperature. Second, land use land cover (LULC) including vegetation type and density. Third, hydrological indicators such as precipitation, runoff and water content may influence the abundance of VBDs. And fourth, terrain indicators are associated with VBDs because of the fact the different terrain features support different land cover types, thus, provide specific habitats for vectors. Further three types of epidemiological risk assessment are defined: (1) Identifying breeding habitats of vectors, (2) linking between environmental indicators and disease outbreaks and transmission, and (3) risk assessment based on landscape elements. Rotela et al. (2007) went further and investigated the use of RS and GIS for vector surveillance of dengue fever classifying indicators in several spatial scales ranging from micro to macro scale. The spatial pattern of dengue is a cooperative result of multiple indicators that can be clustered in environmental, demographic, entomologic and epidemiologic indicators. These indicators may be classified into micro-scale (e.g., small-scale vector breeding habitats), medium-scale (e.g., houses), and macro-scale (e.g., blocks of houses, roads, agricultural fields). The macro-scale includes some environmental indicators such as vegetation types derived from LULC classification, wetness or temperature, which may be observed by remote sensing images. The authors created an incidence map of Tartagal, a city in Northern Argentina, showing the hot spot areas during the outbreak of dengue fever in 2004. A maximum likelihood classification was implemented to classify rivers, roads and vegetation coverage and vegetation indices were calculated the examine the vegetation's greenness and wetness using LANDSAT 5 TM data. The authors concluded river nearness and humidity are the most relevant indicators are probably linked to the abundance of the dengue fever vector (Rotela et al., 2007).

Another research on malaria was conducted comparing three satellite sensors at three different spatial scales to predict larval presence in the US. Beside a larval survey, images from the LANDSAT-5 TM satellite, ASTER satellite as well as Hyperion hyperspectral satellite were acquired (Brown

et al., 2008). The Advanced Spaceborne Thermal Emission and Reflectance Radiometer (ASTER) is on board of the Terra satellite operated by NASA providing data for a wide range of science research including land surface climatology, vegetation and ecosystem dynamics, volcano monitoring, hazard monitoring, hydrology, geology and land surface and land cover change. Aster delivers high spatial resolution data in 14 bands, ranging from the visible to the thermal infrared wavelengths (NASA, 2004). The Hyperion sensors as part of the Earth Observing-1 satellite launched by NASA provide hyperspectral images consisting of 220 spectral bands with a 30-meter resolution. Hyperspectral imaging has wide-ranging applications in agriculture, environmental studies forestry (USGS, 2011). Within the study, Brown et al. (2008) extracted several vegetation indices such as the NDVI or the Disease Water Stress Index (DWSI), a measure of leaf water content from all acquired images of the three sensors. A stepwise regression model was implemented to identify the abundance of *Anopheles* larvae at three different scales: pixel, wetland perimeter, and wetland area. The best models were developed using ASTER data on wetland area scale. Thus, models based on higher spatial and spectral resolution (ASTER and Hyperion) are resulting in more accurate products than LANDSAT 5 TM.

Tran et al. (2008) employed an object-based classification approach on LANDSAT-7 ETM+ images to determine and quantify the spatial distribution of *Anopheles* mosquitoes in France. Referring to changes in climate and landscape, the evaluation of the risk of emergence of re-emergence of such major diseases is of great importance in Europe. They investigated the relationship between vector breeding sites and land cover, and their results highlighted a strong link between land cover and presence or absence of Anopheles larvae. Irrigated cropland, rice-cultivation fields are positively correlated with the abundance of the vector.

Besides malaria, several other VDBs were investigated at the beginning of the new century. Nakhapakorn and Tripathi (2005) explored the relationship between the climate factors rainfall, temperature and humidity and the occurrence of dengue fever (DF). The environmental indicators were derived from LULC classification implementing a maximum likelihood classifier on LANDSAT-TM imagery. Medical data was applied to validate the results.

Another approach is using time-series analysis of remotely sensed data for modeling the abundance and transmission of VDBs. Ugorji et al. (2009) proposed to process a time series with LANDSAT-5 TM+ and LANDSAT 7 ETM+ performing a supervised image classification of the land use and land cover change in Nigeria to infer to climate and landscape as determinates of the prevalence of malaria. They found out that although climate has the most significant impact on the *Anopheles* mosquito, the detection of land use change may be used to limit the spread of arthropod vectors.

In the last few years, several studies focused on the global and regional distribution of Ae. Species, which are responsible for the transmission of several VDBs including dengue fever, yellow fever or the chikungunya virus. The underlying assumption behind the modeling of global Ae. Species distribution is that diseases can only persist where mosquito vectors are present. Understanding the distribution may help to overcome VDBs. Therefore, several approaches were tested including time series analysis, analysis of environmental indicators, which favor the abundance of mosquito species and analysis including socio-economic factors such as demographic factors or migration. A study in 2015 was conducted based on literature research combined with the environmental indicators temperature, vegetation indices, precipitation and land cover to a regression tree machine learning model. Resulting in a global distribution map of Ae. Species (Kreamer et al., 2015). Another study on the worldwide distribution of the dengue fever and yellow fever vectors Ae. Aegypti and Ae. Albopictus using temperature and precipitation time series data was conducted by Santos and Meneses (2017). The authors applied a Maxtent Correlative model, which rely on the association between the species distribution and a set of environmental indicators, to predict the global distribution of the vectors. The model output revealed areas with a high probable abundance of Ae. Aegypti and Ae. Albopictus. However, they also highlighted the fact that environmental suitability for virus transmission in an area does not necessarily mean that the virus will arrive and become established at suitable territories.

Besides studies on global scale, also studies on regional level were performed to correlate remote sensing data with the abundances of Ae. Species. A study in Central Mexico was undertaken by Moreno-Madrinan et al. (2014) using field survey data, remotely sensed data from the MODIS, TRMM and

SRTM satellites. Spearmen correlation models were applied using spectral bands including the visible and near-infrared range from the TRMM satellite, land surface temperature derived from MODIS and elevation from the SRTM mission, resulting in the fact that land surface temperature was positively and significantly correlated with the abundance of Ae. Species. Elevation estimated through SRTM also showed a correlation with the presence of Ae. Species.

Another modeling approach to assess the temporal and spatial patterns of Ae. Species is to employ hotspot and cluster analysis socio-economic and environmental indicators. A study conducted by Espinosa et al. (2016) using hotspot and cluster analysis examined the abundance of Ae. Species in Argentina. The authors concluded that land cover has the most significant impact on the distribution of the vector within the study site.

A more complex model introduced by Buczak et al. (2012) was used for the epidemiological prediction of dengue fever using local and remote sensing data in Peru. Temperature, vegetation indices, the southern oscillation index, sea surface temperature, socio-economic and demographic data as well as elevation data were employed on a fuzzy data mining model using artificial neural networks, resulting in a reliable prediction of dengue fever for up to seven weeks. Thus, environmental indicators may have the potential to predict the abundance of Ae. Species and model the probability of VBDs outbreaks at different spatial scales.

LULC indicators derived from classification may be used to describe the habitats of Aedes species. A study conducted by Landau and Leeuwen (2012) on fine-scale urban land cover indicators and factors assessed to potential use of land cover classes for the distribution modeling of Aedes species in Tucson, Arizona. The authors performed an in-situ mosquito collection and realized a LULC classification using 1 m spatial resolution multi-spectral aerial imagery and LiDAR elevation data. The LULC classification was composed of eleven classes including bare soil, pavement, structure, pool, water, shadow, herbaceous, shrub, low height tree, medium height tree, and high height tree. The LULC classes structure and medium height tree were positively correlated with the abundance of Aedes mosquitoes, whereas bare soil had a negative correlation. They concluded that land cover classes

may influence the distribution of the Aedes mosquito within an urban environment and should be therefore further investigated.

Vanwambeke et al. (2011) performed a study on the abundance of Aedes species on the island of Oahu, Hawaii. Land cover data including 13 land cover classes were obtained from the National Oceanic and Atmospheric Administration Coastal Services Centre. Mosquito abundance data was acquired from the Department of Health, which regularly traps different mosquito species. Regarding LULC highest abundances were associated with areas with mixed vegetation and built-up areas.

Vegetation coverage, in general, is associated with the abundance of Aedes species. A study conducted by Vezzani et al. (2005) demonstrated the importance of vegetation in the context of habitat suitability for the Aedes species. The study assessed microhabitats of Aedes species in Argentina using in-situ mosquito collection and several indicators including the height of vegetation. Areas with greater vegetation heights were associated with the abundance of the mosquito, whereas areas with little or no vegetation coverage were not associated with the abundance. The authors concluded that vegetation might have favorable effects on the abundance and development of the species. However, the authors do not provide information about the data source for obtaining vegetation coverage and height.

A similar study was conducted by Hayden et al. (2010) who assessed microclimate and human indicators that influence habitats of Aedes species in Tucson, Arizona. They defined several microclimatic variables including temperature, relative humidity, and vegetation coverage and performed in-situ mosquito collections. Climate indicators such as hourly average, minimum and maximum temperature and relative humidity were acquired from weather stations. Further human factors such as the presence/absence of running water in households or mosquito nets were investigated. A visual assessment of vegetation coverage at the test site was performed categorized into three classes of low, medium and high vegetation coverage. Additionally, to the visual assessment vegetation coverage was examined using IKONOS 1m remote sensing satellite imagery. The authors found out that more significant vegetation coverage is favorable for Aedes species habitats. Further moisture was identified as an important indicator of the

abundance of the mosquito. Thus, they concluded the land cover type vegetation and the climatic variable moisture play a vital role in the distribution of Aedes habitats.

Rey et al. (2006) investigated the importance of LULC for defining Aedes mosquito habitats in South Florida. Favorable habitats were identified by using areal imagery for each test site, visual image interpretation for obtaining LULC classes and mosquito sampling. The study revealed that only built-up areas were positively correlated to the abundance of Aedes species, whereas bare soil and open areas had a significant negative influence on the habitats. Hence, urbanization and the increasing expansion of built-up areas may favor mosquito habitats.

Further research was carried out by Lockaby et al. (2016) who tried to associate climatic, ecological and socioeconomic indicators with the abundance of the West Nile Virus in Atlanta, Georgia. They investigated the use of the land cover classes forest, water, impervious surfaces and non-impervious surfaces using mosquito collection and object-based image classification on high-resolution areal images. Besides, precipitation, temperature, evapotranspiration, and moisture was acquired from the US National Weather Service. Linear regression and correlation analysis were implemented to assess the importance of LULC. Lockaby et al. (2016) concluded that the class impervious (built-up areas) is having the most significant impact on the abundance of the West Nile Fever vector. Consequently, Aedes habitats more likely occur in regions with anthropogenic activities.

Consequently, LULC classes especially vegetation and built-up areas are having a significant influence on Aedes mosquito habitats and therefore on the distribution. Thus, vegetation coverage, as well as impervious surfaces, should be considered as indicators. To analyze the seasonal variation and density of vegetation in addition to the LULC vegetation class, vegetation indices may be employed. The Enhanced Vegetation Index (EVI) and the Leaf Area Index (LAI), which are derived from time-series remote sensing imagery, are the most widely used indices for extracting information about phenology (Wang et al., 2017). Messina et al. (2016) demonstrated the usage of the EVI for mapping the global environmental suitability for Zika virus, which is transmitted via Aedes species. *The mean EVI product is indicative of*

amount of photosynthesis taking place in the environment over the course of a year, which is positively correlated with the density of vegetation, and is thus a proxy for the level of moisture available given the relationship between precipitation and vegetation growth."

3.2 Mapping of Habitats using Machine Learning Techniques

Machine learning approaches such as Support Vector Machine (SVM), Artificial Neural Networks (ANN) or Random Forest Ensemble Learning (RF) are more and more used in a wide variety of applications including habitat suitability mapping since the past decades. A detailed definition of machine learning and the above-quoted methods will be given in chapter 3.

Several authors propose to use random forest classifiers for mapping avian species habitats via absence and presence modeling (Martinuzzi et al., 2009; Wilsey et al., 2012). Another application field for random forest classifiers is mapping land cover classes for suitable modeling habitats for large mammals (Bleyhl et al., 2017). Further, random forest is employed to link the abundance and distribution of mosquitoes which are responsible for several VBDs (Ibañez-Justicia and Cianci, 2015). Regarding oceanographic applications, machine learning algorithms such as linear and generalized additive models (GAM), bagging, RF, boosted regression trees and SVM may be employed to map reef fish habitats and benthic habitats (Zhang, 2015; Knudby et al., 2010). Beside animal habitats, vegetation habitats may also be modeled using machine learning. Several studies examined the usage of random forest in mapping natural and semi-natural habitats of agricultural landscape using airborne and spaceborne remote sensing imagery (O'Connell et al., 2015; Duro et al., 2014). Mangrove and grassland habitats may be detected by using support vector machine (SVM). Moreover, SVM is used for mapping habitats for bogs. Further, decision tree rules may be employed to model invasive insects which cause damage to forests (Williams et al., 2017).

Forest habitats may be detected using multi-task learning based machine learning algorithms or ensemble learning (Li et al., 2017; Healey et al., 2018).

3.3 Environmental Indicators

Indicators in general quantify and simplify information about complex phenomena (e.g., ascending barometric pressure may introduce excellent weather conditions). Thus, indicators may reveal trends or events, which are not straightly detectable (Hammond et al., 1995). One definition by the US Environmental Protection Agency quotes: *"an environmental indicator is a numerical value that helps provide insight into the state of the environment or human health. Indicators are developed based on quantitative measurements or statistics of environmental conditions that are tracked over time. Environmental indicators can be developed and used at a wide variety of geographic scales, from local to regional to national levels."* (EPA, 2017)

Remote sensing imagery provides an efficient way to, directly and indirectly, measure biodiversity and species distribution. Environmental indicators are applied to make predictions about the earth's biodiversity. Indicators may be derived from measurements of the physical environment. Temperature and moisture, for example, are well-known environmental drivers of biodiversity, which can be linked to vegetation productivity. Other indirect remotely sensed indicators such as land use and land cover (LULC), terrain, climate or the fraction of photosynthetically active radiance are employed for habitat suitability and distribution modeling of different species (Michaud et al., 2014).

According to the ESA Centre for Earth Observation (ESRIN, 2016) remotely sensed environmental indicators have three primary purposes including the support of policy-makers regarding environmental problems, to identify critical factors that cause pressure on the environment and to monitor responses of policy-makers.

The World Health Organization published a working document on risk assessment on yellow fever virus circulation in endemic countries in 2014, with the goal to prevent yellow fever (YF) epidemics. As part of the Initiative, a strategy to control and reduce the risk of YF outbreaks was implemented. The working document was implemented in two African YF endemic countries, Cameroon and the Central African Republic. The report includes risk assessment methodologies, assessment of ecological and environmental indicators influencing the distribution of YF, data handling and analysis methods as well as ethical considerations regarding the work with health data. For the present thesis, the attention is directed to ecological and environmental factors. According to the WHO ecological and environmental indicators may provide complementary information for risk assessment of YF in endemic countries. Indicators may support the interpretation analysis of yellow fever virus activity and offer potential relevant details. Table 3.1 shows environmental, ecological and general indicators that potentially impact yellow fever virus activities according to the WHO and the London School of Hygiene and Tropical Medicine and if the defined indicators are detectable using remote sensing imagery (RS):

Table 3.1: Environmental indicators

	Indicator	Relevance	Detectable
Climate	Temperature: min., max., average	Mosquito distribution, transmission	Yes
	Precipitation: min., max., average, spatial distribution	Mosquito distribution, transmission	Yes
	Humidity	Mosquito distribution, transmission	Yes
Environment	Elevation	max. altitude for habitats is 2300 m	Yes
	Vegetation coverage	Mosquito distribution	Yes
Human	Land use	Understanding socio-demographic factors including demographic change	Yes
	Population movement	Human migration	No
	Human behavior	Creating artificial breeding containers, hunting monkeys (YF hosts), political unrest, forestry practice	Partly
Medical	Virus	Amount of virus, virulence	No
Vector and host	Ae. Species	Abundance, number of blood meals, length of incubation of YF virus in the vector, vector competence	Yes: suitable habitats
	Non-human primates	Abundance, immunity rate, susceptibility	Yes: suitable habitats

Environmental indicators include temperature, precipitation, vegetation coverage and land use. Regarding the indicator temperature, several statistics should be considered such as the average temperature over a specified period (daily, monthly) or the range between minimum and maximum temperature. Temperature may increase the understanding of the relationship between climate and YF transmission and outbreaks caused by Ae. Species. Precipitation should be analyzed as it is critical for the development of the vector. The best period of conducting research on yellow fever is according to the WHO at the end of the rainy season. Therefore, parameters such as minimum or maximum rainfall and average rainfall should be involved in the study. Areas above 2300 m altitude are considered as non-endemic for the transmission of YF via Ae. Species. Thus, elevation should be regarded as in mountainous territories. Land use may give insights including socio-demographic changes such as urbanization because increasing urbanization will burden future outbreaks (see studies above). Beyond those indicators, human behavior may be considered in studies. Population movement and migration should be considered especially when conducting cross-border studies as human migration increases the spread of YF between countries and regions. Further, the creating of artificial breeding containers by humans have a significant influence on the abundance of the vector. When applying studies cycle overlapping scales, primarily if the jungle cycle is investigated, the hunting behavior of monkeys and forestry practices should be incorporated. Regarding the medical aspect of YF the virus and the amount of virus takes a central role in the transmission and spread of the disease. Finally, depending on the study and expected results the vector and host should be modeled to deduce convincing information about the probability of transmission and spread of YF (WHO, 2014; Vainio and Cutts, 1998).

Regarding the abundance of Ae. Species several indicators may have significant influence. One crucial environmental indicator is average temperature. Although Ae. Species survive at low temperatures near 0 °C and very high temperatures above 30 °C the optimal temperature for breeding is between 18 °C and 26 °C. Another significant variable is the precipitation. According to several studies, the peak of the Aedes distribution is at the beginning of the rainy season, following a dry season with low precipitation. Aedes species, especially Aedes aegypti species prefer breeding in artificial breeding containers that collect rainfall water. Consequently, the Aedes mosquito is counted to the domestic mosquitoes living in urban areas and therefore near humans. Further, vegetation indices such as the Normalized Difference Vegetation Index (NDVI) were positively correlated with the abundance of the vector. Greater vegetation coverage is favorable for the mosquito habitats. Several land cover types including bare soil, urban areas or vegetated areas were also positively correlated with the abundance of Ae. Species. Considering the optical remote sensing data, information from spectral bands may be used to identify mosquito habitats. The LANDSAT 7 ETM+ blue, green, NIR and thermal Infrared spectral bands were successfully associated with the occurrence of Ae. Species. Table 3.2 gives a summary of all indicators, which can be detected with remotely sensed data and therefore employed for the study:

Table 3.2: Remote sensing environmental indicators

Indicator	Specification	Source
Temperature	Optimum between 18 C and 26 C	Nakhapakorn and Tripathi (2005);
Precipitation	Peak at the beginning of the rainy season	Regis et al. (2005) ; Diallo et al. (2011); Wilke et al. (2017)
Habitat	Artificial breeding containers that collect rainwater	Gubler and Clark (1995)
Vegetation Indices (NDVI)	High NDVI values were associated with Ae. Species abundance	Hay et al. (1998); Estallo et al. (2016)
Land Cover	Bare soil, urban areas, water bodies	Espinosa et al. (2016)
	Vegetation cover (medium height trees), urban areas	Landau and van Leeuwen (2012)
	Dry river beds and shorelines	Samson et al. (2015)
Land surface temperature (day and night)	positive correlation with abundance	Moreno-Madriñán et al. (2014); Estallo et al. (2016)
LANDSAT ETM 7+ spectral bands	B1, B2, B4, B6	Arboleda et al. (2012)
Soil Moisture	Increase in biting rates	Patz et al. (1998)

4 Machine Learning in Habitat Modelling

Automated image classification in remote sensing is the process of separating different types of patterns within a digital image. Thus, image classification is strongly related to computer science and pattern recognition and therefore also connected to machine learning. Each pixel in remote sensing data is a pattern describing a specific phenomenon. Depending on the decision rule of the classifier, spectral or spatial pattern recognition can be distinguished. Spectral classification is based on the spectral reflectance or radiance of the digital image band. Whereas, spatial classification solely relies on the size or shape of objects. Classifications are usually conducted in features spaces, which containing all measurements for pixels for all wavelengths. The classification as a complex process requires the consideration of several factors including the selection of suitable remotely sensed imagery, the choice of training data, the data pre-processing, the feature extraction and, the selection of an appropriate classifier (Lu and Weng Q, 2007; Richards and Jia X., 1999).

Table 4.1 gives a summary of classification categories, examples and specifications (Lu and Weng, 2007):

Table 4.1: Remote sensing classification approaches

Criteria	Categories	Examples of Classifier	Specification
Use of training data	Unsupervised classification	MLC, DT, ANN, Min. Distance	Cluster-based
	Supervised classification	K-means, ISODATA, DT, ANN, SVM	Reference and training data
Use of parameters (e.g., covariance, mean vector)	Parametric classifier	MLC, Linear Discriminant Analysis	Gaussian distribution of data is assumed
	Non-parametric classifier	ANN, DT (RF), SVM	No assumption about the data is required
Use of pixel information	Per-pixel classifier	MLC, DT, ANN, Min. Distance	Development of a signature
	Sub-pixel classifier	Fuzzy-set Classifiers, Sub-pixel Classifiers	Spectral value of each pixel is assumed to be linear
	Object-based classifiers	Object-based Classifier	Image segmentation merges pixels into objects
Output of classification	Hard classifier	Most of the Classifier	Final land cover classification
	Soft classifier	Fuzzy-set Classifier, Sub-pixel Classifier	Similarities of pixels
Use of spatial or spectral information	Spectral classifier	MLC, ANN	Spectral values
	Spatial classifier	Contextual Classifier	Spatially neighboring pixel information

Non-parametric Classifiers and Machine Learning

The following chapter gives an overview of non-parametric classifiers and state-of-the-art machine learning approaches in remote sensing and the application for the present thesis.

Machine learning algorithms have the goal to optimize performances iteratively via learning from data and are powerful for regression and classification of nonlinear systems. Thus, a machine learning algorithm enables the computer to determine from existing data using inference strategies. Over the past two decades, machine learning has been an active research field in several disciplines for exploring patterns and relationships in data. Applications of machine learning are often connected to pattern recognition, data mining and artificial intelligence. In opposite to parametric models, which assume that data underlies the theory of Gaussian distribution, non-parametric machine learning models are mostly flexible enough to discover complex nonlinear relationships. Machine learning algorithms are applied where the theoretical knowledge of the studied phenomenon is insufficient but for which enough observations exist. Machine learning is a powerful tool for a wide variety of remote sensing based applications including land, ocean and atmosphere studies. The most implemented algorithms in remote sensing are artificial neural network (ANN), support vector machine (SVM), decision trees (DT), ensemble methods such as random forest (RF), etc. Consequently, statistical parameters are not calculated for separating classes within the classification procedure. Therefore, non-parametric classifiers are suitable for spectral and non-spectral input data. (Zhao et al., 2011; Benediktsson, 2015).

The **artificial neural network** (ANN) machine learning approach, used as non-parametric classifier was developed in the 1990s to model high nonlinear functions. A neural network is an interconnection of neurons in a network, similar to the neurons in the human brain. As all non-parametric classifiers, ANN does not rely on statistical relationships. Thus, continuous functions are estimated from data without specifying mathematically how outputs depend on inputs. The network learns the function during an automatic training process. The goal of the process is to find a suitable set of

parameters which describe the input-output relation. The architecture of an ANN is composed of processing elements or nodes (the neurons), the network topology and the weighted connections between the processing elements (the graphs)(Gopal and Woodcock, 1996; Gardner and Dorling, 1998; Mas and Flores, 2008).

"A neuron has a set of inputs x1, . . . , xm. Each connection from the input to the processing unit is affected by different strengths called synaptic weights. A signal xj at the input of synapse or connection j, connected to neuron k, is multiplied by synaptic weight wkj. An adder sums all inputs forming a linear combination of them. An activation function is used for limiting the output of the neuron." (Mas and Flores, 2008)

Depending on the input, single layer perceptron models or multi-layer perceptron models are distinguished. The single layer perceptron model is the simplest from an ANN, using one layer as input including all nodes or source elements, connected to an output layer. The multi-layer perceptron model, a so-called feedforward ANN model is based on a set of input nodes and a set of output nodes, connected on a layer-to-layer basis (Mas and Flores, 2008)

Support vector machine (SVM), a non-parametric supervised statistical learning technique locates optimal boundaries between classes. The algorithm aims to find a hyperplane that separates the dataset into a predefined number of classes. The simplest form of an SVM classifier is a linear binary classifier that assigns a test sample a class from one or two possible labels. One central aspect is that not all available training samples are used for the definition of the hyperplane. In remote sensing SVM is used primarily with small training data, resulting in a higher classification accuracy than traditional parametric methods such as the maximum-likelihood classifier. The fundamental principle behind the SVM is the structural risk minimization (SRM). SVM minimize misclassifications without prior knowledge of the distribution of data (Mountrakis et al., 2011; Huang, 2002).

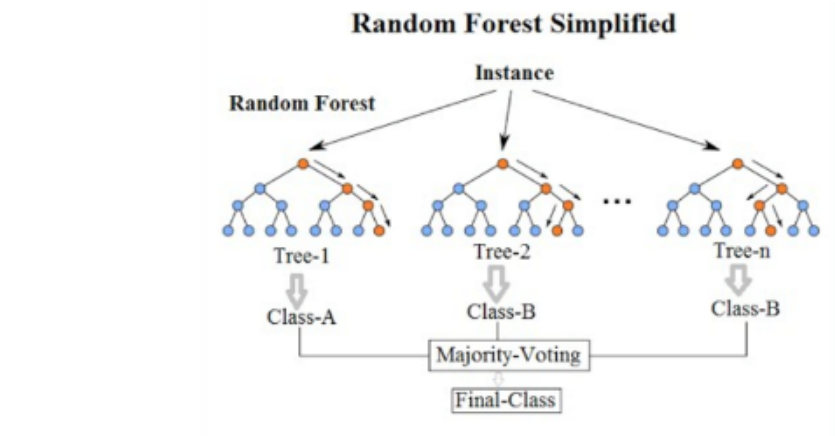
Decision tree classifiers, such as Random Forest, are multi-stage hierar-

chical classifiers meaning that a series of decisions are taken to determine the correct label (class) of a pixel. The simplest approach in decision tree classification is to apply a binary tree in which each component classifier performs a classification into only one or two possible classes. Within decision tree algorithms different data sources and different sets of data may be used. The design of decision tree classifiers is composed of three principle tasks: the finding of the optimum structure of trees, the selection of the optimum variables and the selection of the tree to use at each node (Richards and Jia, 1999).

The **random forest** (RF) machine learning algorithm is part of the so-called ensemble learning algorithms, which iteratively produce multiple classifications. Either one base classifier or a combination of various base classifiers is applied to the classification process. The underlying assumption behind the RF classifier is that several classifiers perform better than an individual classifier. Ensemble learning methods are based on two principle concepts boosting and bagging. Bagging trains many classifiers on bootstrapped samples from the training data and boosting uses iterative re-training. (Rodriguez-Galiano et al., 2012). Random forest uses decision trees for the classification process. Decision trees are part of non-parametric supervised learning methods and are applied to create a model that predicts values of a target variable by learning decision rules derived from input data. The process within the RF classifier is based on the growth of several trees (classifiers) and the combination of their results through a voting process. Trees are created by drawing a subset of training samples trough replacement, meaning that each sample can be selected multiple times, while others may not be selected at all during the classification process. About one-third of the samples are not used to train the classification trees. Those samples are the so-called out of bag (oob) data. The algorithm needs two input parameters, the number of trees that are grown during the process and the number of predicted variables. The number of predicted variables is constant during the whole classification process (Breiman and Cutler, no date; Rodriguez-Galiano et al., 2012; Gislason et al., 2004).

Figure 4.1 shows three decision trees and a classification obtained from each

Figure 4.1: Random Forest Trees



Source: Venkata, 2017

of the trees. The final prediction is based on a majority voting, whereby each tree votes for one class.

In remote sensing random forest is increasingly applied to land cover classifications using multi- and hyperspectral data, radar – and LiDAR data or a combination of different types of data. In contrast to other ensemble learning algorithms, RF is computationally lighter, resulting in the fact that RF is efficient also on large data sets, which means it can estimate the importance of hundreds of input variables. Thus, multisource remote sensing data may be used to extract as much information as possible for the area being classified. The random forest classifier overcomes the problem of the Hughes phenomenon, which can occur when the number of variables is larger than the number of training samples. Besides, high dimensional remote sensing data may be used as the classifier computes the importance for each variable within the classification tree (Rodriguez-Galiano et al., 2012; Gislason et al., 2006; Belgiu and Drăguţ, 2016).

Within this thesis, a non-parametric random forest approach was implemented as the density function of the available data is unknown, which is the case in most classification applications. No additional assumptions can be assumed and therefore only available information is the training

data from which the classification rules must be determined. Moreover, the number of variables is much more significant than the available training samples. 38 training samples are employed to evaluate the importance of 274 indicators. Thus, to overcome the occurring Hughes phenomenon, machine learning is applied to the present study (Cortijo et al. 1996).

Besides, the classification approach is mainly based on data-level. In contrast to semantic approaches, were image patches extracted from multiple remote sensing images that contain specific semantic classes (e.g., built-up areas, agriculture, and vegetation) are used, data-level approaches employ statistics on image bands (Cheng et al., 2017). The data-level method was chosen due to the limited available optical remote sensing data from SENTINEL-2 and LANDSAT-8 for the study area. However, an ESA CCI land cover map was used to assess the importance of semantic information for the present research. Beyond that, the primary objective of the thesis is to investigate the significance of multiple remote sensing environmental indicators. Apart from semantic land cover information, additional indicators including temperature, vegetation indices, and soil moisture are employed assessing the distribution of the yellow fever vector and the potential abundance of the disease (chapter 7).

5 Data Source

5.1 SENTINEL-2A

5.1.1 SENTINEL-2A System Overview

For this thesis, SENTINEl-2A images were obtained from the Copernicus Open Access Scientific Data Hub. The SENTINEL-2 mission is part of the Copernicus previously known as the Global Monitoring for Environment and Security (GMES) joint initiative of the European Commission (EC) and the European Space Agency (ESA) providing continuous operational monitoring information for various fields of application. ESA's role is the allocation of space- and ground related system elements. SENTINEL-2 offers wide-swath, multi-spectral, high-resolution optical data over a global terrestrial surface. SENTINEL-2A has been successfully launched on 23 June 2015, and roughly 20 months later SENTINEL-2B has been equally successfully sent to the orbit on 3 March 2015. The system design is based on the simultaneous operations of two identical satellites (SENTINEL-2A and 2B) flying on a single sun-synchronous orbit but phased at 180° at an altitude of 186 km, each hosting a Multi-Spectral Instrument (MSI) covering a broad spectral range from the visible to the shortwave range and providing high spatial and temporal resolution imagery.

The twin satellites of SENTINEL-2 provide image data on the behalf of Copernicus services and applications including risk management (floods and forest fires, subsidence and landslides), European land use/land cover states and change mapping, forestry (forest monitoring), agriculture (food security and early warning), water management and soil production, urban

mapping, natural hazard mapping and terrestrial mapping for humanitarian aid and development. The SENTINEL-2 mission addresses several new necessities in remote sensing such as the need for high temporal coverage and high revisit frequency, the need for large-scale coverage for global land mass acquisition and the need for data in a broad spectral range (Drusch et al., 2012; ESA, 2017; Gatti A., Bertolini A., 2015; Suhet, 2013, Gatti A., Naud C., 2017).

5.1.2 Specifications

"The SENTINEL-2 orbit is sun-synchronous. Sun-synchronous orbits are used to ensure the angle of sunlight upon the Earth's surface is consistently maintained. Apart from small seasonal variations, anchoring of the satellite's orbit to the angle of the sun minimizes the potential impact of shadows and levels of illumination on the ground. This ensures consistency over time and is critical in assessing time-series data." (Suhet, 2013)

Both satellites operate at a mean altitude of 786km with an orbit inclination of 98.62°. The Mean Local Solar Time (MLST) during the descending operation is 10:30 a.m. and ensures a suitable level of solar illumination. Regarding the geographical coverage, SENTINEL-2 acquires global data over land and coastal areas from 56° South to 83° North.

The temporal resolution, the revisit frequency of the satellite to a location is 5 days in the two-satellite constellation configuration and 10 days in a single configuration. The radiometric resolution defines the capacity of the instrument to distinguish differences in intensity or reflectance. The MSI has a 12-bit radiometric resolution. Thus, intensity values ranging from 0 to 4095 can be acquired by the system. SENTINEL-2A is designed to measure the earth reflected radiance trough the atmosphere in 13 spectral bands (spectral resolution) ranging from Visible (VIS) and Near-Infra-Red (NIR) to the Short Wave Infra-Red (SWIR) in different high spatial resolutions (Figure 5.1). Depending on the spectral three different spatial resolutions

are available. Four bands at 10m spatial resolution (Table 5.1) including blue (B2), green (b3), red (B4) and NIR (B8) ensure the comparability with SPOT4 and SPOT5 image data (Drusch et al., 2012; Gatti and Bertolini, 2015; Suhet, 2013).

Table 5.1: Sentinel-2 10m Spectral Bands

	SENTINEL-2A		SENTINEL-2B	
Band Number	Central Wavelength (nm)	Band-width (nm)	Central Wavelength (nm)	Band-width (nm)
2	496.6	98	492.1	98
3	560	45	559	46
4	664.5	38	665	39
8	835.1	145	864	32

Six bands at 20m spatial resolution (Table 5.2) including three vegetation red edge bands (B5, B6, B7) a narrow NIR band (B8a) and two SWIR bands (B11, B12) are designed for applications including vegetation, snow, and ice detection as well as cloud discrimination:

Table 5.2: Sentinel-2 20m Spectral Bands

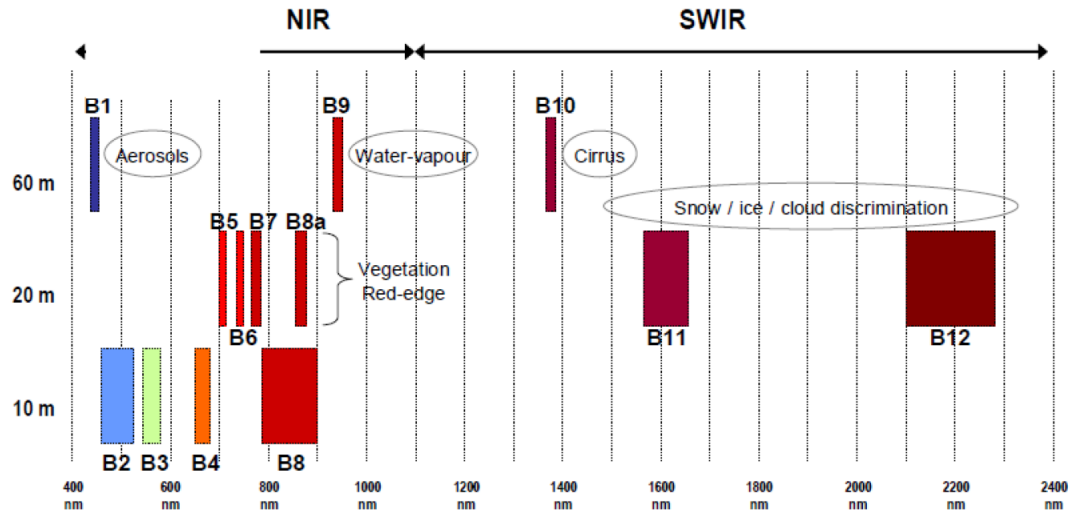
	SENTINEL-2A		SENTINEL-2B	
Band Number	Central Wavelength (nm)	Band-width (nm)	Central Wavelength (nm)	Band-width (nm)
5	703.9	28	703.8	20
6	740.2	18	739.1	18
7	782.5	28	779.7	28
8a	864.8	33	864	32
11	1613.7	143	1610.4	141
12	2202.4	242	2185.7	238

Further, three bands at 60m spatial resolution (table Table 5.3) allow cloud screening and atmospheric corrections of the acquired data. B1 in the domain of 443nm enables an aerosol correction. The design of the water vapor band (B9) allows the correction of atmospheric water vapor and B10 facilitates cirrus correction:

Table 5.3: Sentinel-2 60m Spectral Bands

	SENTINEL-2A		SENTINEL-2B	
Band Number	Central Wavelength (nm)	Band-width (nm)	Central Wavelength (nm)	Band-width (nm)
1	443.9	27	442.3	45
9	945	26	943.2	27
10	1375.5	75	1376.9	76

Figure 5.1: Sentinel-2 Spectral Bands



Source: Gatti and Bertolini, 2015

5.1.3 Product Type Overview

Different product types are available to the user including Level-0, Level-1, and Level-2. The processing of Level-0 is performed in real-time during the data-reception operations. The processing of Level-0 consists of packaging the MSI and satellite ancillary raw-data with annotations and metadata to ensure further processing. The Level-0 User products contain raw image data and all the information required to generate Level-1 (and upper) product levels organized per granules (Gatti and Naud, 2017).

The Level-1A User products are obtained by decompressing image data and implementing a geometric model to locate any pixel in the image. Each image pixel value of Level-1 is encoded on 12 bits, and image data are provided as separated files in GML/JPEG2000 file format. Following Level-1A, Level-1B products are radio-corrected and geo-refined by performing radiometric corrections on Level-1A and refining the geometric model. Level-1C products are resampled to 10m, 20m, and 60m. The projection of the product is UTM (WGS 84). Level-1C products are delivered in tiles, and each tile is defined by its projection code, ground coordinates of the upper-left pixel of the tile, pixel size in line and column, and tile size. The image data are provided in separated raster files in GML/JPEG200 file format (Gatti and Naud, 2017).

5.2 LANDSAT-8 OLI

LANDSAT-8 data was obtained from the USGS Earth Explorer, which allows the user to view and download different earth observing product including satellite images, aerial photography, and cartographic products. The LANDSAT-8 mission is part of the National Aeronautics and Space Administration's (NASA's) Science Directorate (SMD) LANDSAT program, which is a component of the U.S. Global Change Research Program (USGCRP). The mission provides long-term studies of the Earth's global environment. LANDSAT-8 is the latest satellite in a series of 7 operational satellites (LANDSAT 6 was lost immediately after the launch in 1993) starting in 1972 with the launch of the first satellite. Thus, LANDSAT data have been

used in a variety of disciplines including agriculture, pollution monitoring, global change research, forestry, geology, land cover mapping, resource management, water, and coastal studies (USGS, 2013; Zanter, 2016).

5.2.1 LANDSAT-8 System Overview

LANDSAT-8 has been successfully launched on 11 February 2013 from the Vandenberg Air Force Base California (U.S.A) providing ongoing moderate-resolution multispectral data from the Earth's surface. The LANDSAT-8 satellite carries two sensors on board, the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) (Figure 5.2). The satellites orbits in 705 km height, repeating in a 16 days cycle. The OLI collects image data from nine spectral bands ranging from Aerosol to SWIR with a 30m spatial resolution in the multi-spectral (MS) bands and 15m resolution in the panchromatic (PAN) band. The thermal infrared sensor (TIRS) measures the energy emitted by the Earth's surface in two bands at a spatial resolution of 100 m (USGS, 2013; Zanter, 2016).

5.2.2 Specifications

LANDSAT operates in a sun-synchronous near-polar orbit at a mean altitude of 705 km with a 16-days repeat cycle. The satellite makes a complete orbit every 99 minutes and completes about 14 orbits per day. Two sensors on board of the satellite, the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS), simultaneously image every scene (USGS, 2013).

The OLI sensor collects data from nine spectral bands (Table 5.4) with a 14-bit radiometric resolution and over a 190km swath. However, only 12-bits are recorded to the Solid-State Recorder (SSR) and transmitted to the ground by the satellite. The spatial resolution of all bands is 30m, except

the 15m panchromatic band (B8). The Coastal band (B1) is mainly used for coastal studies and shallow water observation as well as aerosol and smoke detection. The Blue band (B2) helps scientist in bathymetric mapping for understanding the aquatic system geomorphology. Further, it is used in soil and vegetation discrimination studies and forest type mapping. Peak vegetation on the earth's surface can be assessed using the Green band (B3). The Red and NIR bands (B4, B5) are used primarily in vegetation detection analysis. For vegetation moisture or drought analysis and mapping of burned areas, the two SWIR bands (B6, B7) on board of LANDSAT-8 could be used. The panchromatic band (B8) is used for sharpening multispectral imagery to higher resolution. And the cirrus band (B9) at 30m resolution facilitates the detection of clouds.

Table 5.4: Landsat-8 OLI Spectral Bands

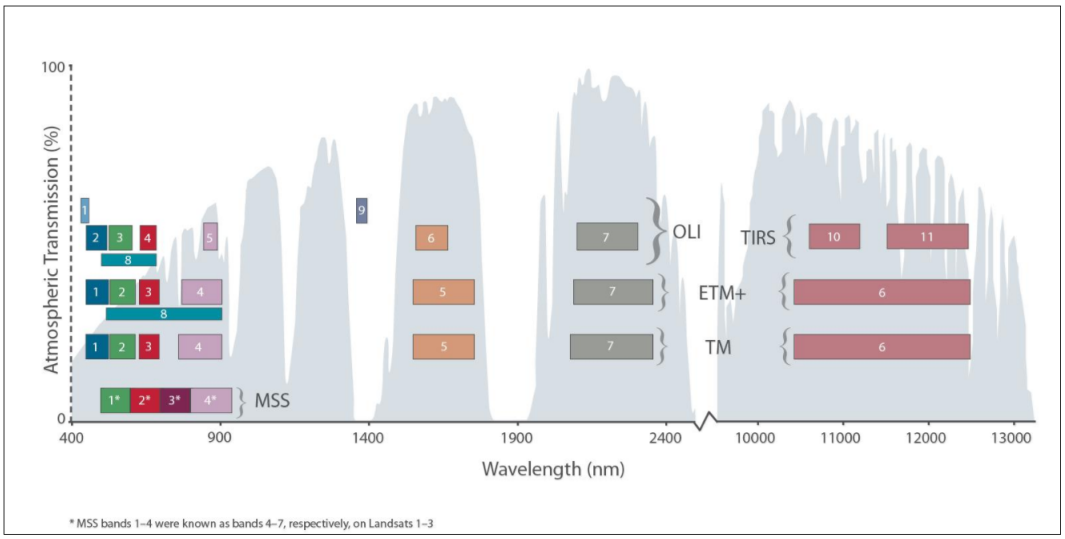
Band Number	Wavelength Range (nm)	Spatial Resolution and Name
1	435 – 451	30 m Coastal/Aerosol
2	452 – 512	30 m Blue
3	533 – 590	30 m Green
4	636 – 673	30 m Red
5	851 – 879	30 m NIR
6	1566 – 1651	30 m SWIR 1
7	2107 – 2294	30 m SWIR 2
8	503 – 676	15 m Panchromatic
9	1363 – 1384	30 m Cirrus

The TIRS sensor uses Quantum Well Infrared Photodetectors (QWIPs) to measure Thermal Infrared energy emitted by the earth's surface. Image data from two spectral bands (Table 5.5) over a swath of 190km and with a 100m spatial resolution are obtained (USGS, 2016; Zanter, 2016; USGS, 2013)

Table 5.5: Landsat-8 TIRS Spectral Bands

Band Number	Wavelength Range (nm)	Spatial Resolution and Name
10	10600 – 11190	100 m TIRS 1
11	11500 – 12510	100 m TIRS 2

Figure 5.2: Landsat Sensors Spectral Bands



Source: USGS, 2017

5.2.3 Product Type Overview

Level-1 products are available to the user and are produced by the Level 1 Product Generation System (LPGS). The Level-1 processing algorithm is composed of several essential processing steps and two refinement processing tasks. The basic processing includes the ancillary data processing, the geometric model generation, the correction grid generation and the

terrain-corrected image resampling. After the basic processing, two refinement tasks are applied to the image data, the geometric model precision using GCPs and the refinement of the terrain-corrected image resampling task. The L1 product consists of 13 files in GeoTIFF format, including the 11 image bands (9 bands from the multispectral sensor OLI and two bands from the thermal infrared sensor TIRS), a product specification metadata file and a quality assessment image (could mask, fill data) (Zanter, 2016).

5.3 MODIS Terra and Aqua

5.3.1 MODIS System Overview

The Moderate Resolution Imaging Spectroradiometer (MODIS) instruments onboard of the Terra and Aqua satellites are part of the NASA Earth Observing System (EOS). The two satellites are operating in a sun-synchronous, near-polar orbit at a mean altitude of 705 km above the earth's surface. The Terra satellite (originally known as EOS AM-1) collects data from the earth's bio-geochemical and energy system using five sensors the Advanced Spaceborne Thermal Emission and Reflectance sensor (ASTER), the Clouds and Earth's Radiant Energy System sensor (CERES), the Multi-Angle Imaging Spectroradiometer (MISR), the Moderate Resolution Imaging Spectroradiometer sensor (MODIS) and the Measurements of Pollution in the Troposphere sensor (MOPITT). The Terra satellite was launched on December 18, 1999 (NASA, 2017).

The second satellite Aqua (originally known as EOS PM-1) collects data about the earth's water cycle, including evaporation from the oceans, water vapor in the atmosphere, clouds, precipitation, soil moisture, sea ice, land ice, and snow. Aqua was launched on May 4, 2002, and acquires data via six sensors including the Atmospheric Infrared Sounder (AIRS), an Advanced Microwave Sounding Unit (AMSU), CERERS, MODIS, reduced quality data from an Advanced Microwave Scanning Radiometer for EOS (AMSR-E).

The sixth sensor, the Humidity Sounder for Brazil (HSB) failed after a nine months operation. The viewing swath of both satellites is 2330km; thus the revisit frequency at a location is about one to two days. The ground track repeat cycle is every 16 days. The MODIS sensor acquires image data in 36 spectral bands with a spatial resolution of 1000m, 500m and 250m. The spectral wavelength range of the 36 bands ranges from 400 to 14400 nm. The radiometric resolution of the image data is 12-bits (NASA, 2017).

5.3.2 Land Surface Temperature

MODIS Land Surface Temperature (LST) products are available with a temporal resolution of one day, an eight-day, and a monthly composite. MODIS LST data is achieved in Hierarchical Data Format (HDF) files containing metadata information and scientific data sets (SDS). Several LST datasets are available for the user. The MOD11_L2 is an LST product at a 1km spatial resolution for a swath. The MOD_11A1/2 products are tiles of LST with a temporal resolution of one day/eight days. The MOD11_B1/2/3 products are tiles including information on the LST and emissivities. The MOD11_C1/2/3 products include information on the LST, but in contrary to the MOD_B and MOD_A products in a geographic projection.

For this thesis, the MOD11A2 product was obtained via Reverb, a NASA Earth Science Discovery Tool. The MOD11A2 is an eight-day composite calculate with a simple average method from MOD11A1. The MOD11A2 product includes 12 SDS, Day Land Surface Temperature (+ Quality Indicators), Night Land Surface Temperature (+ Quality Indicators), local time and view zenith of day and night observations, band 31 emissivity, band 32 emissivity and clear day coverage during day and night. The image data is projected in in Sinusoidal projection (Wan, 2015).

5.3.3 Vegetation Indices

The MODIS 13 series provide consistent information about global vegetation conditions via vegetation indices (VI) calculated at each pixel. Two indices, the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) are used. Six products are available within the MODIS Vegetation Indices series. Three daily level 2 products, MOD13Q1 (16-days composite at 250m spatial resolution), MOD13A1 (16-days composite at 500m spatial resolution) and MOD13A2 (16-days composite at 1km spatial resolution) are available. Further, three products are available at a finer resolution (MOD13A3, MOD13C1, MOD13C2). The MOD13A2 product, which was used for this thesis, provides 12 SDS layers including 16-days composite EVI and NDVI average values, the bands Red, NIR, Blue and MIR, Quality indicators and the view and sun zenith angle (Didan et al., 2015).

Vegetation indices such as the NDVI or EVI are based on spectral reflectance signatures of leaves. Within the visible spectrum, the reflectance is very low due to high absorption of energy caused by photosynthesis of plants. Little absorption and therefore high reflectance can be detected within the near-infrared wavelength range of the electromagnetic spectrum. Consequently, comparing red and NIR spectral bands may indicate the vegetation amount, with a maximum red-NIR difference over a full canopy and minimum over areas with little or no vegetation (Didan et al., 2015).

The NDVI is a normalized transform of the NIR to the red reflectance ratio, calculated using standard equation:

$$NDVI = NIR–Red/NIR + RED \tag{5.1}$$

The second index, which can be acquired from the MOD11A2 product is the Enhanced Vegetation Index (EVI) and implemented using the equation:

$$EVI = G * (NIR–Red/NIR + C1 * Red–C2 * Blue + L) \tag{5.2}$$

Where NIR, Red and Blue are the reflectance values of the MODIS bands, L the canopy background adjustment, C1 and C2 the coefficients of the aerosol resistance term, and G the scaling factor (2.5).

Using rations for assessing vegetation cover has several advantages such as the fact that certain types of band-correlated noise may be reduced or influences caused by variations in could cover, sun and view angle, topography or atmospheric attenuation may be minimized. One of the main drawbacks in using ratio-based is their non-linearity exhibiting asymptotic behaviors. Thus, over certain land coverage ratios tend to be insensitive to vegetation variation (Didan et al., 2015).

5.3.4 LAI and FPAR

The MODIS 15 (older Version 5) series provide information about the Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation (FPAR). The LAI is defined as the total one-sided leaf area per unit ground area in broadleaf canopies and as one-half to total needle surface area per unit ground area in coniferous canopies. Thus, LAI is one of the most significant variables to model canopy photosynthesis and evapotranspiration. The second, the FPAR is a relative measure of the vegetation absorbed radiation in the 400 – 700 nm spectral region, and indicate the energy used during the process of photosynthesis (Weiss, 2004; Zhu, 2013).

Four different standard products are available within the MODIS 15 series. Two identical products MOD15A2 and MYD15A2 from platform Terra and Aqua with a spatial resolution of 1000 m and a temporal resolution of eight days. Further, a combined Terra and Aqua platform product (MCD15A2) is available with a spatial resolution of 1000 m and a temporal resolution of eight days. The fourth product MOD15A3 is also a combined product from both satellites, but with a finer temporal resolution of four days. MYD15A2 products were acquired including the eight-days composites of the FPAR and LAI at 1km spatial resolution (Myeni, 2012).

5.4 Metop ASCAT

5.4.1 Metop ASCAT System Overview

MetOp ASCAT soil moisture data were obtained from the Vienna University of Technology (TU Wien). MetOp is a series of three polar orbiting satellites from monitoring weather and climate from space operated by the European Organization of Meteorological Satellites (EUMETSAT) as part of the Copernicus Earth Observation Programme. MetOp-A (launched on October 19, 2006) and MetOp-B (launched on September 17, 2012) are operating at a mean altitude of 817km proving observation data of the atmosphere, oceans, and continents. MetOp carries several sounding and imaging instruments including the real aperture radar Advanced Scatterometer (ASCAT). ASCAT is operating at 5.255 GHz using C-Band and vertically polarized antennas, with morning and evening overpasses at approximately 09:30 a.m. and 09:30 p.m. local time. The spatial resolution of ASCAT is 25km. ASCAT provide useful information for several applications including soil and vegetation analysis, sea ice and ice extent monitoring and monitoring of permafrost or desertification. Two sets of three antennas measure the electromagnetic backscatter from the earth's surface in 500km swaths (EUMETSAT, 2017).

5.4.2 ASCAT Soil Moisture Retrieval Algorithm

The TU Wien soil moisture retrieval algorithm is based on a change detection approach and underlies the following assumptions:

"*One basic assumption for the change detection method is that backscatter, expressed in decibels, is linearly related to surface soil moisture content. Furthermore, backscatter is strongly dependent on incidence angle, and the slope and curvature of the relationship between backscatter and incidence angle are assumed affected only by vegetation density but not by changes in soil moisture. Subsequently, the slope and curvature obtained from the backscatter observations under different incidence angles are used to parameterize the vegetation for every day of the year and for every grid point. To account for noise in the backscatter measurements and to ensure that an entire range of incidence angles is covered, 366 slope and curvature pairs (i.e.,*

one for each day of the year) are determined by averaging backscatter observations of several years."

5.5 ESA CCI Land Cover Maps

A global CCI Land Cover (LC) map, which provides information about land cover classes, was obtained from the TU Wien from the year 2010. CCI Land Cover maps are part of the ESA Climate Change Initiative (CCI), Global Monitoring of Essential Climate Variables. Three LC maps were developed, representing the epoch 1998-2002 (LC maps 2000), the epoch 2003-2007 (LC map 2005) and 2008-2012 (LC map 2010) on a spatial scale of 300m. Multiple suitable data were combined using a multi-year and multi-sensor strategy for producing consistent global Land Cover maps including sensors such as the Moderate Resolution Imaging Spectrometer (MERIS), SPOT or the Advanced Very-High Resolution Radiometer (AVHRR). The World Geodetic System 84 (WGS-84) was used as a reference system. Land cover types were defined using the Food and Agriculture Organization (FAO) Land Cover Classification System (LCCS) (Defourny et al., 2017).

The study area is covered by 18 land cover classes (Table 5.6) including agriculture, vegetation, urban areas and water bodies:

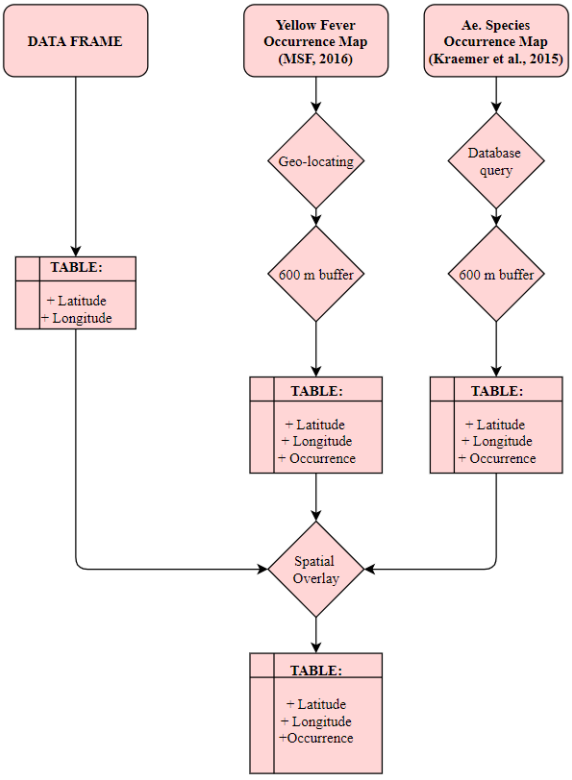
Table 5.6: ESA CCI LULC classes

ID	Land cover type
10	Cropland rainfed
11	Cropland rainfed - Herbaceous cover
30	Mosaic cropland (>50%) / natural vegetation (tree/shrub/herbaceous cover) (<50%)
40	Mosaic natural vegetation (tree/shrub/herbaceous cover) (>50%) / cropland (<50%)
50	Tree cover broadleaved evergreen closed to open (>15%)
60	Tree cover broadleaved deciduous closed to open (>15%)
61	Tree cover broadleaved deciduous closed (>40%)
62	Tree cover broadleaved deciduous open (15-40%)
100	Mosaic tree and shrub (>50%) / herbaceous cover (<50%)
110	Mosaic herbaceous cover (>50%) / tree and shrub (<50%)
120	Shrubland
122	Shrubland deciduous
130	Grassland
160	Tree cover flooded fresh or brackish water
170	Tree cover flooded saline water
180	Shrub or herbaceous cover flooded fresh/saline/brackish water
190	Urban areas
210	Water bodies

5.6 Training Data Preparation

To train the random forest algorithm a response vector including training areas needs to be created. The training data preparation consisted of combining Ae. Species occurrence and YF occurrence (during the outbreak 2016) by a spatial overlay and empirically defining a buffer around each located occurrence point (Figure 5.3).

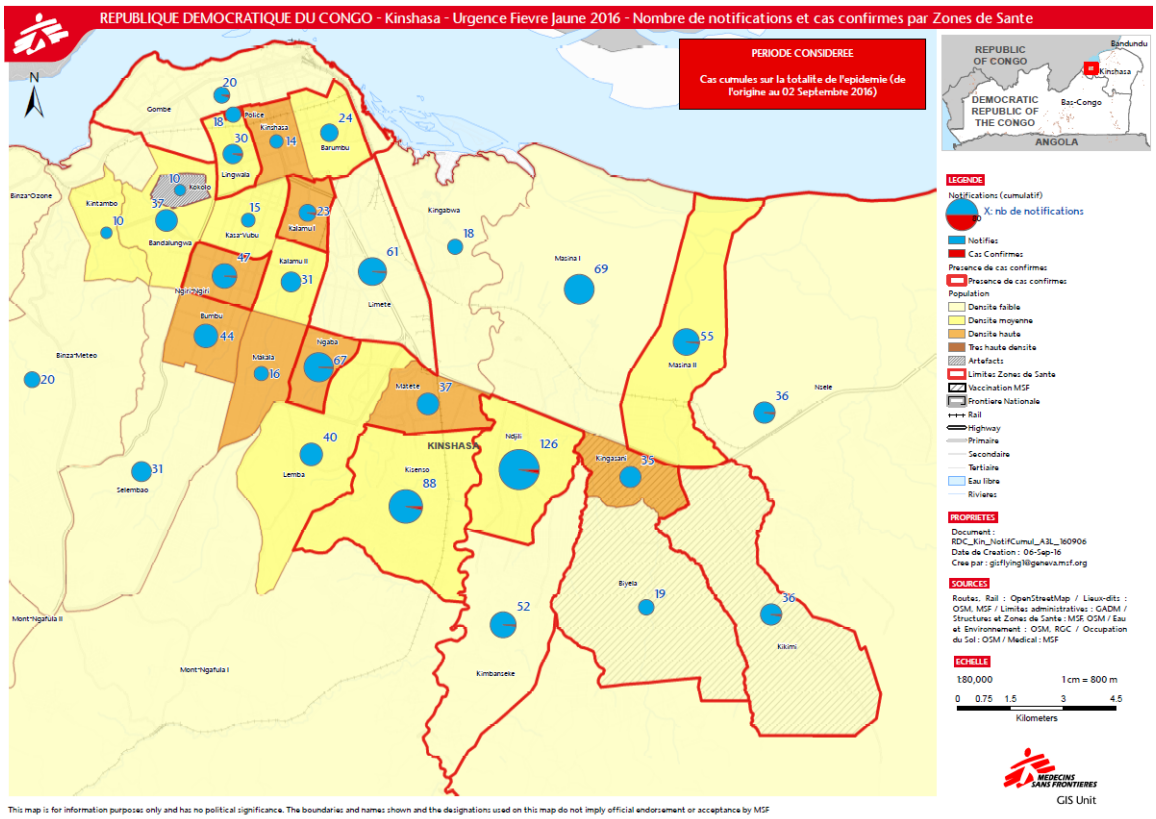
Figure 5.3: Training Data Preparation



Because medical data is difficult to access especially in developing countries Doctors without Borders (MSF) provided a map including all YF cases during the outbreak 2016. Unfortunately, no medical data was supplied due to restriction from MSF. MSF requests a full research protocol, an approval from the Institutional Ethics Committee, and a plan to obtain approval from

the National Ethics Committee (the country where the data subjects are coming from). Thus, the YF map was taken as first training data source (fig. msf).

Figure 5.4: Yellow Fever Cases during the Outbreak 2016 in the DRC



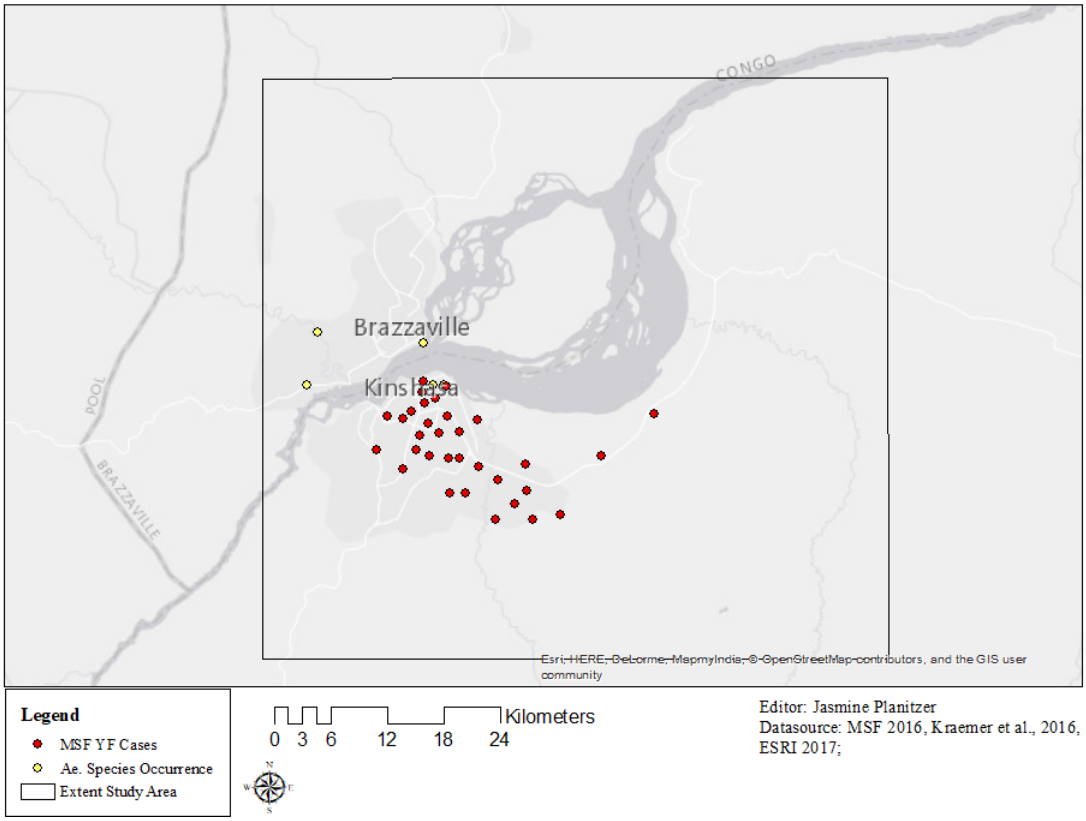
Source: MSF, 2017

The map (Figure 5.4) shows suspected and confirmed YF cases during the outbreak in 2016 on medical zones level. The map was produced by the Manson Unit, which is a team of experts that aims to improve the quality of MSF's medical programmes worldwide. The team consists of medical doctors, laboratory specialists, epidemiologists, a GIS officer, and a medical editor (MSF; no date).

Further, a global *Ae. Aegypti* and *Ae. Albopictus* database, which was compiled by including several sources such as published literature or primary occurrence data from national and international surveys, was used to analyze the Ae. Occurrence for the study area (Kraemer et al., 2015)

The first step for preparing the training data was to locate the suspected and confirmed YF cases by using the ArcGIS editor. The editing was conducted by identifying the locations of the cases (blue and red circle on the map, Figure 5.4) and manually placing a point within the center of each circle, resulting in 32 points including detected YF cases during the outbreak. Following the first step, a database query was conducted on the global Ae. Species point database created by Kraemer et al., 2015). The query detected six occurrence points within the study area (Figure 5.5). Thus, 38 points including YF cases and Ae. Species occurrences were edited. A combination of both was necessary to create enough training points for the random forest. As YF appears solely jointly with the abundance of Ae. Species both data sources may be combined and taken as training data. Subsequently, an empirically defined buffer of 1000 m (approx. flight range of the Ae. Species is 600 m) was created around all points. The assumption behind the use of buffers was that the risk of infection is not limited to one location defined by a point, but to a larger area (polygon). By editing features in ArcGIS attribute tables are created. In this case, attribute tables including coordinates of the edited points were generated. Additionally to the coordinates, a column including information about occurrence (0 = no occurrence, 1 = occurrence), was added to the table. After that, both point data sets were converted from vector into raster files for conducting a spatial overlay.

Figure 5.5: Yellow Fever Cases and Ae. Occurrence



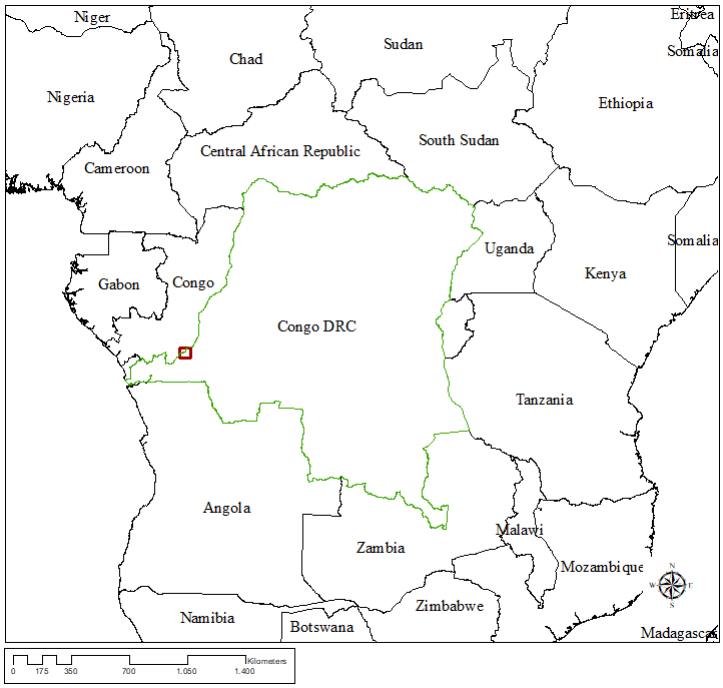
Finally, the data frame table including the training data information was exported and imported again as a data frame in R to apply the random forest. Regarding the reliability of the training data, it can be assumed that MSF is providing reliable medical data as the NGO is working in cooperation with national health authorities in the Democratic Republic of Congo. However, the uncertainty of using an empirically defined buffer for estimating infection zones remains. Besides, the number of training areas need to be discussed. In order to gain convincing classification results, comprehensive training data is required. In this case, since the training data preparation was solely based on the YF cases map (Figure 5.4) and the Aedes occurrence data, a limited amount of training areas were available.

Further, to obtain fully reliable data sets field surveys, need to be conducted, or medical data from national health authorities need to be requested. Though working with medical data remains critical concerning several aspects including the protection of humans privacy.

6 Study Area

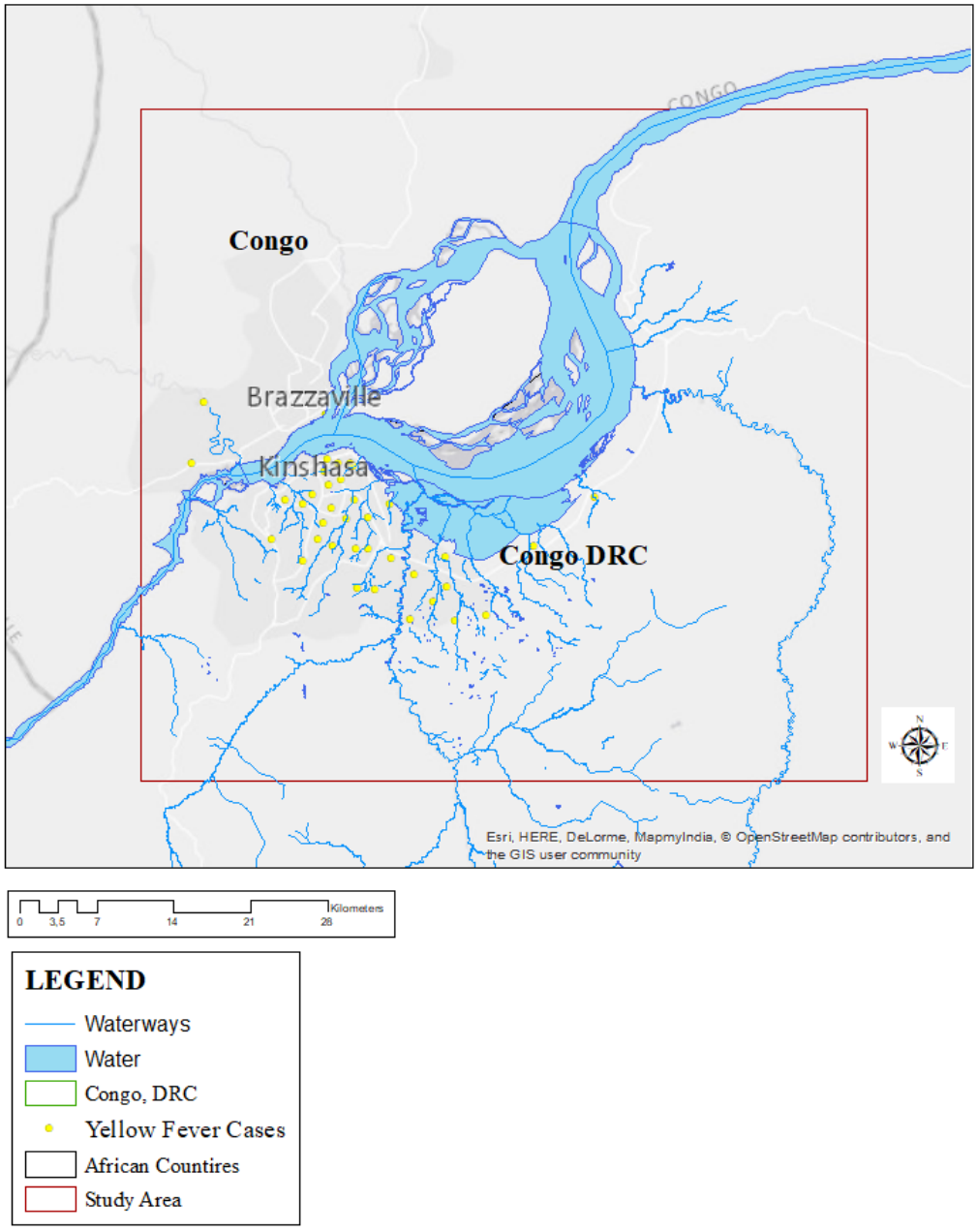
To test the machine learning classifier using environmental indicators a study site within the Democratic Republic of Congo was chosen (Figure 6.1). The study area extends over the capital of the Democratic Republic of Congo, Kinshasa and covers about 4.550 square kilometers (Figure 6.2).

Figure 6.1: Democratic Republic of Congo



Source: Open Street Map, 2017

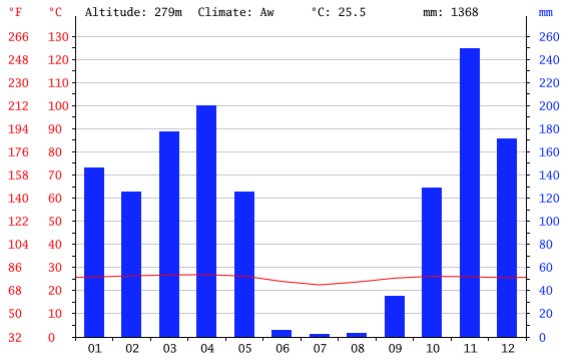
Figure 6.2: Study Area



Source: Open Street Map, 2017

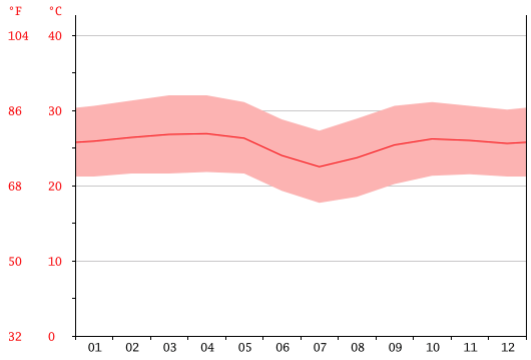
According to the Food and Agriculture Organization of the United Nations (FAO) global ecological zoning for the global forest resources assessment 2000 the study site is located within the tropical rainforest ecological zone. The tropical rain forest zone covers the central part of Africa, on both sides of the Equator and is characterized by high precipitation, ranging from 1000 to over 2000 mm/year. The dry season with lower precipitation exceeds not more than 3 to 4 months. Temperature is high in the low latitude areas, generally more than 20 °C, except in mountainous areas (FAO, 2017). According to the Köppen and Geiger climate classification, the study site is characterized by the Aw zone. The equatorial (A) savannah zone is characterized by dry winter months (w) with mean min precipitation lower than 60 mm (Kottek et al., 2006).

Figure 6.3: Monthly Mean Temperature and Precipitation Kinshasa



Source: Climate-data.org, 2017

Figure 6.4: Monthly Mean Precipitation Kinshasa



Source: Climate-data.org, 2017

Figure 6.3 and Figure 6.4 show the mean precipitation and temperature from Kinshasa between 1982 and 2012. According to the figure June, July and August are the driest months with average precipitation lower than 5 mm. The rainy season start in October, following a peak in November with a mean precipitation of about 250 mm and ends in May. The mean annual precipitation averages at 1368 mm. The mean temperature hardly varies within a yearly cycle, ranging from an average of 22.5 °C in July to 26.9 °C in April.

According to the Climate Change Initiative (CCI) from the European Space Agency, the study site is characterized by several land use and land cover classes:

Table 6.1: CCI LULC in Square Kilometres and Percent

Land Use Land Cover Class	Area in km^2	Area in %
Cropland, rainfed	190	4.83
Herbaceous cover	305	7.7
Mosaic cropland (>50%) / natural vegetation (<50%)	409	10.41
Mosaic natural vegetation (>50%) / cropland (<50%)	58	1.48
Tree cover, broadleaved, evergreen, closed to open (>15%)	203	5.18
Tree cover, broadleaved, deciduous, closed to open (>15%)	144	3.67
Tree cover, broadleaved, deciduous, closed (>40%)	1.4	0.04
Tree cover, broadleaved, deciduous, open (15-40%)	865	22.02
Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	366	9.34
Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	67	1.71
Shrubland	81	2.08
Shrubland deciduous	1.9	0.05
Grassland	78	2.01
Tree cover, flooded, fresh or brackish water	25	0.66
Tree cover, flooded, saline water	61	1.55
Shrub or herbaceous cover, flooded, fresh/saline/brackish water	176	4.48
Urban areas	483	12.32
Consolidated bare areas	0.9	
Water bodies	410	10.44

7 Workflow for Risk Zone Mapping and Modeling

The risk zone mapping (Figure 7.1) consists of the preprocessing of the different data sources, the pixel value extraction and calculation of metrics and indices and the random forest modeling approach. First, a SENTINEL-2A scene was preprocessed using the recommended Sen2Cor preprocessor. Further, the scene was resampled, cropped and re-projected. Water bodies from the SENTINEL-2A scene were extracted using a maximum likelihood classifier to calculate the mean distances to water bodies (Figure 7.4). Then an accuracy assessment was undertaken to assess the classified water bodies using training data. LANDSAT-8 scenes were acquired and resampled, cropped and re-projected. Data from the MODIS Terra and Aqua sensors and the ASCAT sensor were converted into TIFF format and preprocessed using the same processing steps as with LANDSAT-8 images. Beyond, a CCI Land cover map was acquired and preprocessed by resampling and cropping the map to conduct a plausibility check.

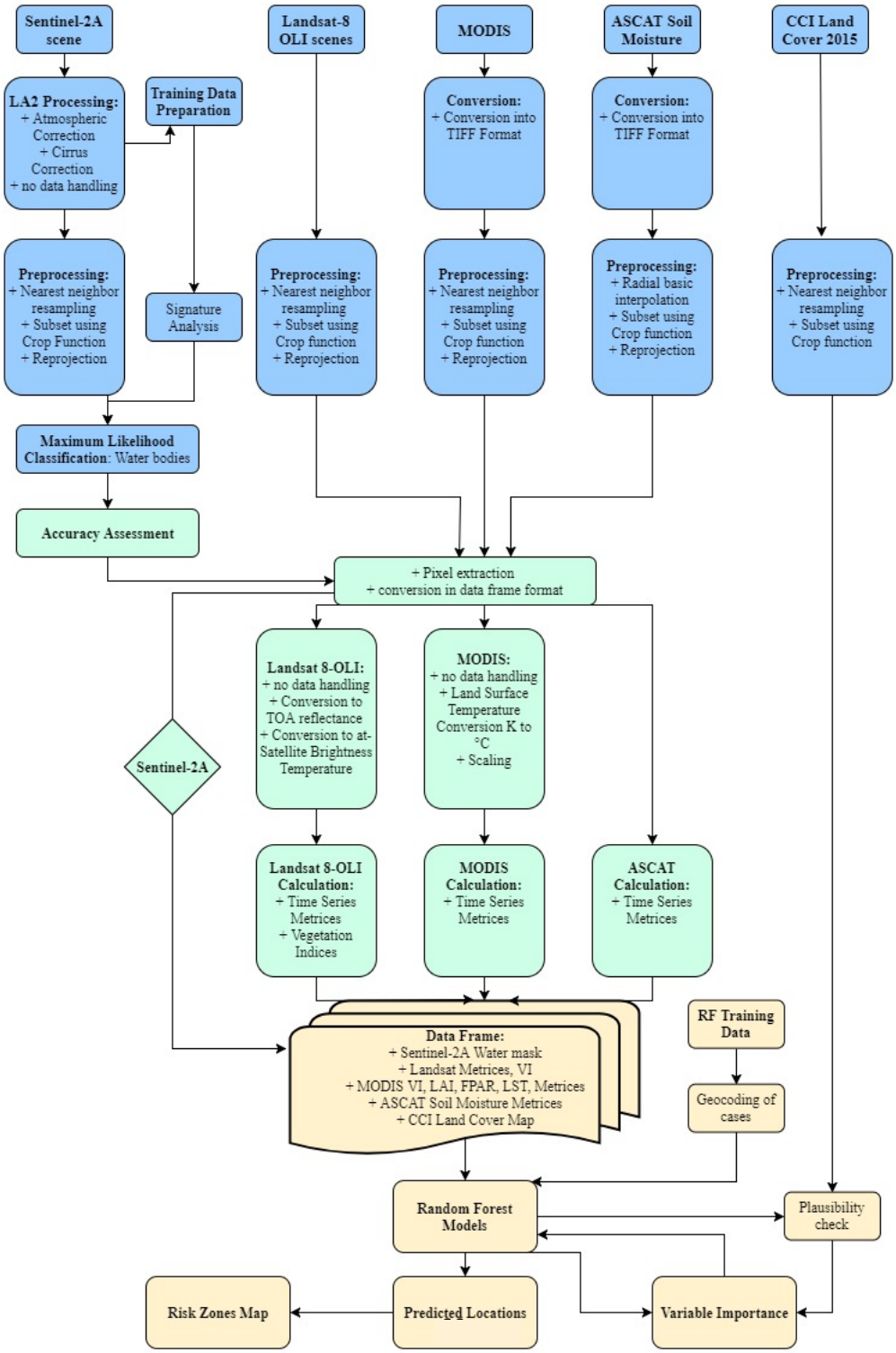
After the preprocessing steps, pixel values were extracted and converted into data frame format to process the data further using R. The LANDSAT-8 pixel values were converted into top-of-atmosphere (TOA) reflectance values and at-satellite brightness temperature using the proposed NASA equations. Following the conversion of the LANDSAT-8 data, time series metrics from all bands and vegetation indices were calculated. Moreover, MODIS data scaled and land surface temperature values were converted from Kelvin to Degree Celsius. Time series metrics were computed from both MODIS and ASCAT data. Although several SENTINEL-2 images were acquired from the ESA Science Hub, only one scene was used due to the risk zone mapping due to the high cloud coverage especially covering the metropolitan area.

Consequently, the limited availability and utility of SENTINEL-2 images for the study area lead to the use of LANDSAT-8 data. Thus, instead of calculating indices and metrics from SENTINEL-2 data, indices and metrics from nine LANDSAT-8 scenes were computed. However, as mentioned above, the SENTINEL-2 image was employed to determine water bodies.

Following the computation of metrics and indices a data frame including values from all available data sources was implemented (Table 7.8). Training data was prepared using geocoding of recorded YF cases in the study area to train the random forest models. Several random forest models were implemented, and the importance of each indicator was assessed. Finally, a risk zone map for the study area was created from the predicted locations.

7 Workflow for Risk Zone Mapping and Modeling

Figure 7.1: Workflow for Risk Zone Mapping and Modelling



7.1 Preprocessing

SENTINEL-2A Preprocessing

Atmospheric Correction with Sen2Cor

For SENTINEL-2 data products the SENTINEL-2 Level-2A Prototype Processor, which is labeled "Sen2Cor for Sentinel 2 (atmospheric) Correction", was used. The processor corrects the effects of the atmosphere from SENTINEL-2 Level-1C products. Further, it performs the tasks of terrain and cirrus correction. The implementation of the Sen2Cor Processor, which can be operated either as a purely command line application or from the Sentinel-2 Toolbox (SNAP), combines several state-of-the-art techniques for performing Atmospheric Correction (AC) and a Scene Classification (SC) (Mueller-Wilm, 2016). In this thesis, one SENTINEL-2 Level 1C product was acquired from the ESA Science Hub and processed to Level 2A. Level 1C products provide geo-coded TOA reflectance values. The first processing step using the Sen2Cor command line application performs an atmospheric correction. The atmospheric correction is based on the Atmospheric/Topographic Correction for Satellite Imagery algorithms, which performs a libRadtran radiative transfer model. The atmospheric correction algorithm is divided into subtasks including Aerosol Optical Thickness (AOT), Water Vapor (WV) and terrain retrieval. The atmospheric correction generates three different user outputs: AOT and WV tables on pixel level as well as the Bottom-of-Atmosphere (BOA) corrected reflectance images for all available bands. The AOT retrieval method is based on the Kaufmann Y.J. et al. (1997) algorithm, which mainly can be divided into the following three steps: (1) detection of dark pixels in the infrared wavelength range, (2) reflectance estimations at 0.47 µm and 0.66 µm and (3) exploration of optical thickness and mass concentration from the identified radiance. However, some small alterations were applied to the algorithm including reduction of negative reflectance values and fixed rural/continental aerosol types. The AOT can be deducted from the images themselves above reference areas, which are characterized by known reflectance behavior. Those reference areas are either dark, dense vegetation (DDV) targets or water bodies. The DDV uses the short-wave infrared (SWIR) and correlates its reflectance with the red and blue band.

Water vapor is a crucial driver for weather and absorption of solar radiation in the atmosphere. WV is performed using the Atmosphere Pre-Corrected Differential Absorption algorithm, which is applied to band B8a (reference channel in the atmospheric window) and B9 (measurement channel in the absorption region). Water vapor together with aerosol retrieval defines the accuracy of the surface reflectance product. (Mueller-Wilm, 2016; Makarau et al., 2017; Drusch et al., 2012; Kaufmann, 1997).

"*The absorption depth is evaluated by calculating the radiance for an atmosphere with no WV, assuming that the surface reflectance for the measurement channel is the same as for the reference channel.*" (Mueller-Wilm U., 2016)

A Cirrus Correction is required to minimize the cirrus could affect in the visible, near- and short-wave infrared spectral bands. Cirrus clouds are partially transparent and therefore difficult to detect, especially over inhomogeneous land cover. To overcome these affects the cirrus band reflectance was employed to correlate with the visible- and near-infrared (VNIR) and SWIR using the Cirrus Correction Algorithm (Mueller-Wilm, 2016). The cirrus correction was applied on the acquired SENTINEL-2 scene, resulting in a cirrus corrected image.

Figure 7.2: Cirrus Correction of the SENTINEL-2 scene

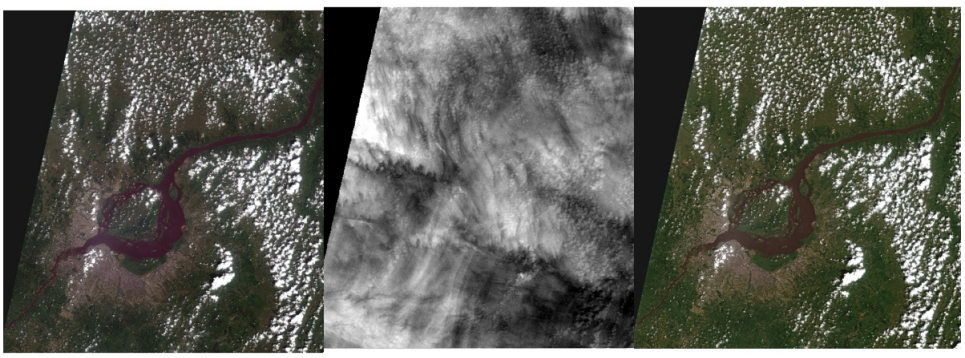
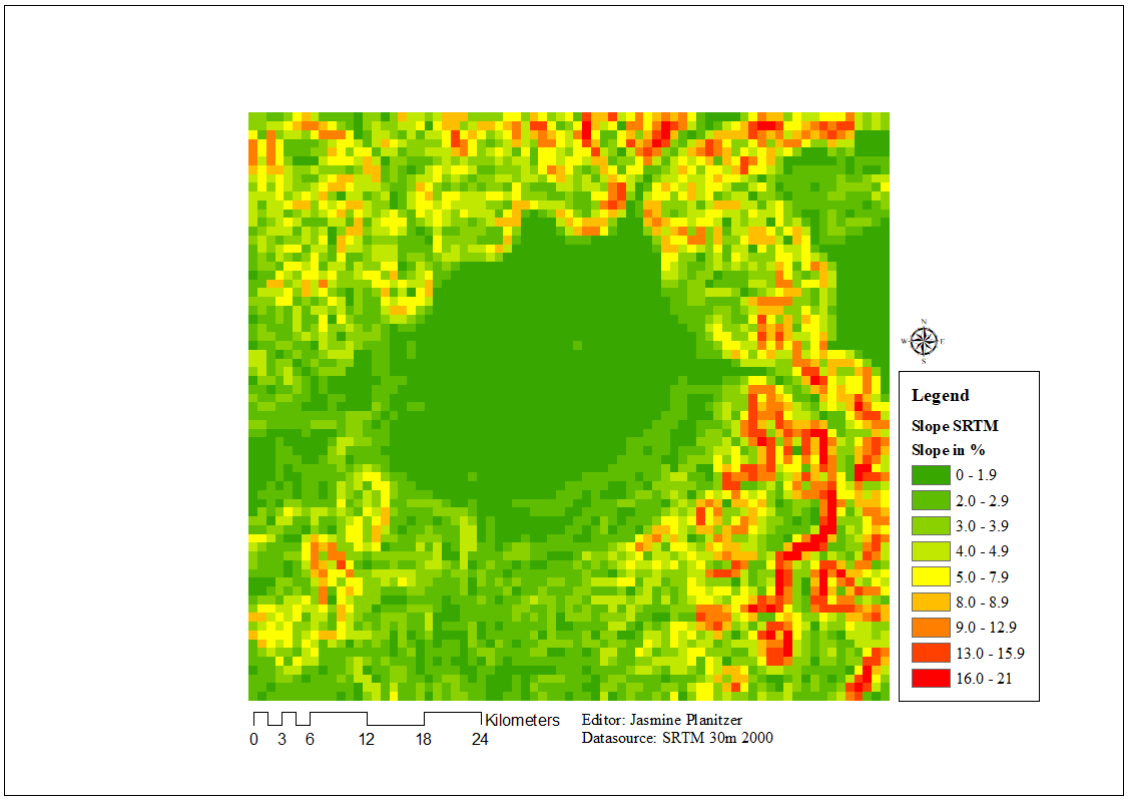


Figure 7.2 shows the Level 1-C product, the cirrus band, and the Level 2-A product. As can be seen from the visualization, clouds at lower levels of the

atmosphere remain after the cirrus correction.

Further, during the surface reflectance retrieval performance a Digital Elevation Model (DEM) may be employed to enhance the scene classification and to improve the terrain correction for rugged terrain. To assess if terrain correction needs to be performed, slopes from the SRTM DEM were calculated using the ArcGIS Spatial Analyst (Figure 7.3).

Figure 7.3: Terrain Analysis



Further statistics on the basis of the calculated slope were computed (Table 7.1). The slope was classified into nine classes based on the maximum

rate of change per pixel (in %). Then raster statistics were computed, to analyze the percentage of each class.

Table 7.1: Slopes

Slope in %	% of class
0 – 1.9	43.82
2.0 – 2.9	12.21
3.0 – 3.9	10.41
4.0 – 4.9	8.42
5 – 7.9	12.85
8.0 – 8.9	3.02
9 – 12.9	7.32
13.0 – 15.9	1.2
16.0 - 21	0.74

A terrain correction is recommended if more than 5 % of the pixels have slopes greater than 8°. In this case, more than 98 % of the pixels are having slopes smaller than 12.9% (= 7.407°). Thus, assuming the SRTM DEM is reliable, the terrain can be considered as flat, and no terrain correction is applied (Richter et al., 2012).

"*The atmospheric model of SEN2COR (L2A_AtmCorr) is dependent on the calculation of radiative transfer functions for different sensor and solar geometries, ground elevations, and atmospheric parameters [...]. The processor reads the parameter in the form of Look Up Tables (LUTs) pertaining to this parameter space and interpolates if required. The LUTs have been generated via libRadtran, a library for the calculation of solar and thermal radiation in the Earth's atmosphere. The LUT selection is configurable via the user configuration file*". (Mueller-Wilm, 2016).

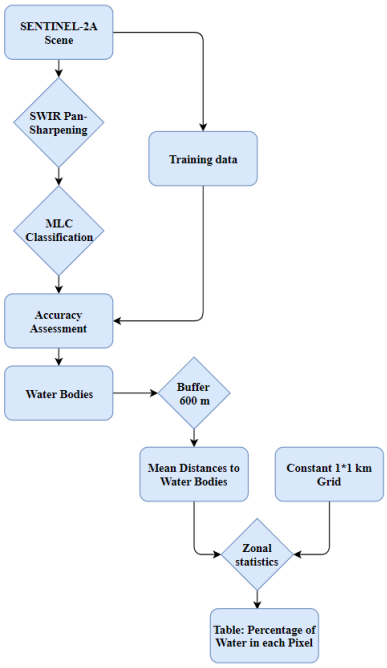
Three different parameters for calculating LUTs were set within the configuration file: the aerosol type, the, and the ozone content. The aerosol type can either be set to Auto, which enables the processor to automatically determine the aerosol type, Maritime or Rural. The mid-latitude can either be set to Auto or Summer/Winter. Further, six different types of ozone concentrations may be set depending on the mid-latitude configuration.

For the processing of the downloaded L1C scene the following settings were selected: the aerosol type was chosen to be rural, the mid-latitude was selected to summer as the scene was acquired on March 2017, which during the summer period on the southern hemisphere, and the ozone content which is measured in Dobson Units (DU) was selected based on the mid-latitude summer atmosphere to be 331 DU (standard ozone concentration) (Mueller-Wilm U., 2016).

Water Body Extraction

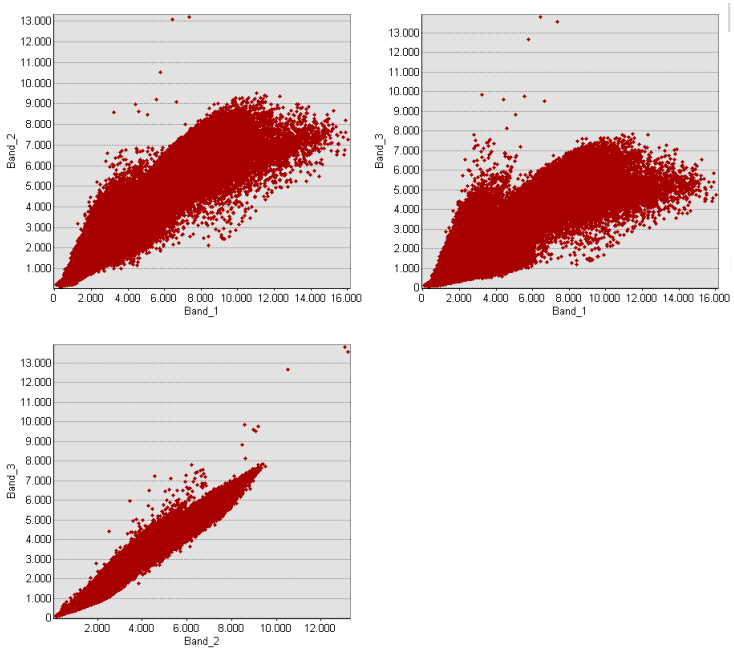
One crucial indicator for the risk zone mapping is the mean distance to water bodies. Therefore, after employing the SENTINEL-2 Prototype Processor to generate a level 2A product, water bodies were extracted and mean distances to water bodies were calculated using a buffer and zonal statistics (Figure 7.4).

Figure 7.4: Water Body Extraction Workflow



The first step was to pan-sharpen the both SENTINEL-2 20-m SWIR bands. To perform a supervised classification on images with different spatial resolution either bands with higher resolution need to be upscaled or those with lower resolution downscaled. Concerning downscaling, the two most popular approaches are spatial interpolation and image fusion. Pan-sharpening which is one image fusion technique is applied to coarser multispectral bands and a high-resolution panchromatic band. Pan-sharpening requires a correlation between the bands; therefore the both SWIR and the NIR band were stacked together and correlated (Figure 7.5), resulting in a positive correlation of both the SWIR 1 (Band_3) and SWIR 2 (Band_2) with the NIR (Band_1) band. However, the fact the the bands do not fully correlate (100%), residual errors remain and further classification errors might occur. To produce a high-resolution stack the spatial resolution of both SWIR bands were enhanced by treating the NIR band as a PAN-like band. Thus, the 10-m NIR band was used during SWIR pan-sharpening process. The pan-sharpening was performed using the ERDAS Imagine wavelet resolution merge, which allows multispectral images to be sharpened using panchromatic images (Kaplan and Avdan, 2017; Du et al., 2016; Intergraph, 2013).

Figure 7.5: Correlation of NIR and SWIR SENTINEL-2 bands



The next step was comprised of a supervised Maximum Likelihood Classification (MLC) to extract water bodies using the 10-m spatial resolution bands Blue, Green, Red and NIR and the 10-m spatial resolution pan-sharpened bands SWIR1 and SWIR2. The classification process was implemented on the classes water, non-water, clouds and cloud shadows. Despite correcting cirrus clouds within the level-2 preprocessing clouds in lower levels of the atmosphere remained; thus, clouds were classified within the Maximum Likelihood Classification. The MLC assumes a Gaussian distribution of data and is based on an estimated probability function for each of the reference classes (chapter 4.1.). Within the MLC pixels are allocated to their most likely class of membership (Foody M.G. et al., 1992).

The classification was conducted using the ERDAS Imagine MLC which is based on the following equation:

$$D = ln(ac) - [0.5ln(Covc)] - [0.5(X - Mc)T(Covc - 1)(X - Mc)] \quad (7.1)$$

"Where D is the weighted likelihood, c a class, X the measurement vector of the candidate pixel, MC the mean vector of the sample of class c, ac the percent probability that any candidate pixel is a member of class c, Covc the covariance matrix of the pixels in the sample of class c and T the transposition function."(Intergraph, 2013)

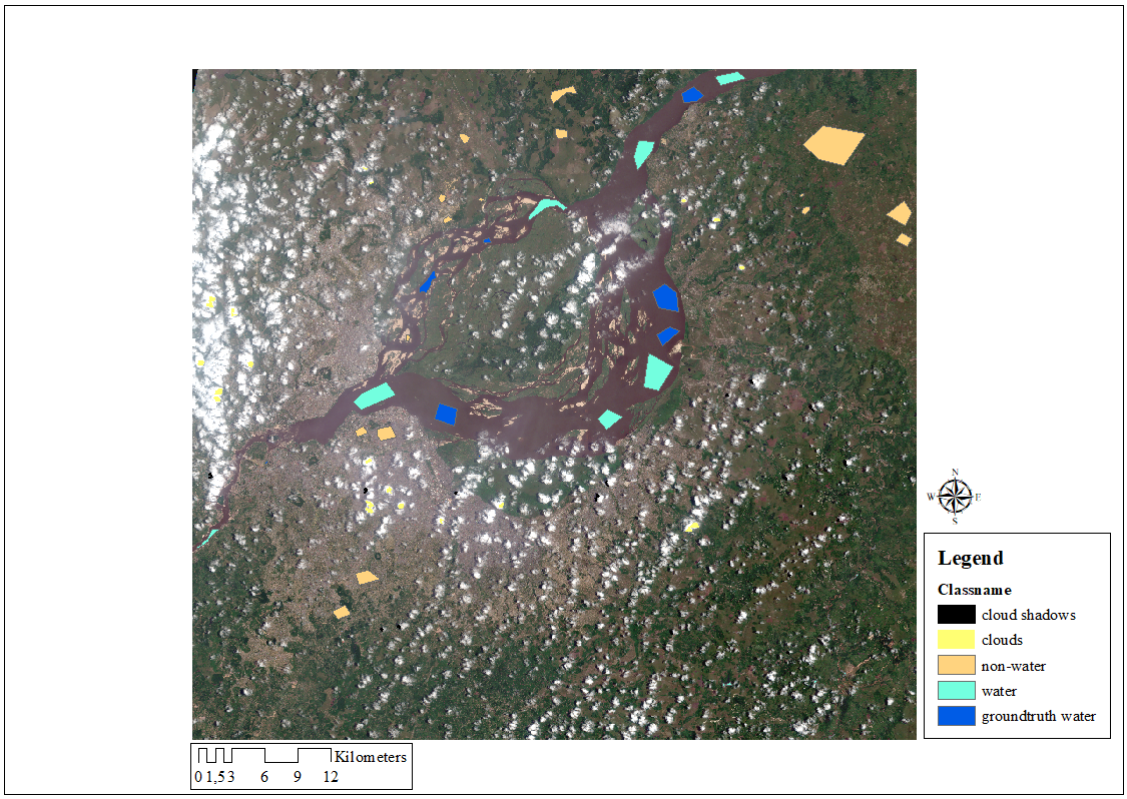
The Maximum Likelihood Classifier was applied for the classification process, since spectral unmixing was not required to apply for this study.

To conduct an accuracy assessment training data was selected on the basis of the SENTINEL-2 scene. Therefore, solely training areas were selected, which can be detected by the use of the SENTINEL-2 spatial resolution. The training data selection was performed using the ArcGIS training sample manager. 18 well over the scene distributed training polygons were created for each class including water, non-water, clouds and cloud shadows (Table 7.2, Figure 7.6).

Table 7.2: MLC Training Data

Class Name	Pixels Count
Clouds	37889
Water	208832
Non-water	259904
Cloud shadows	7762

Figure 7.6: MLC Training Polygons and Water Groundtruth Polygons



A significant drawback in solely using the SENTINEL-2 scene for the selection of training polygons is the resolution of the data, meaning small ponds or pools cannot be detected. Thus, those small aquatic habitats may not be identified during the classification process. One possible solution for this problem would be using VHR remote sensing imagery and comparing the SENTINEL-2 scene and the VHR data. However, an accurate assessment of the water bodies was performed using R and the package caret, which enables a simple accuracy assessment. Since only water bodies were needed for the calculation of the mean distance to water bodies, only the class water was evaluated during the accuracy assessment. The function confusion matrix calculates statistics about observed and predicted classes. The input arguments for the function are the predicted classes and reference

classes. Several statistics are calculated including the sensitivity, specificity, positive predictive value and negative predictive value (Kuhn M., 2017). To validate the classification results validation data is needed. Because of the fact, that no validation was available for the test site, observation data was created by comparing the Sentinel-2 image stack with Google Maps images from the test site. Consequently, the accuracy assessment might be unrepresentative, since the the training data, which was derived from HR SENTINEL-2 imagery was validated with VHR Google Maps data. One approach, which might overcome this drawback would be in using VHR data (e.g. Google Maps imagery) to create training and validation data. However, 18 polygons including 121586 pixels for water bodies were compiled using the ArcGIS training sample manager (Figure 7.6). Then, the polygons were converted into a raster file and imported in R. Within the R script both the classification and the validation data were clipped using the crop function. All classes besides water were removed and assigned to 0 as only the water bodies were validated. And finally, the confusion matrix was calculated via the confusionMatrix function (Table 7.3, Table 7.4)

Table 7.3: Confusion Matrix

	Reference	
Prediction	Non-water	Water
Non-Water	40.001.743	3.571
Water	0	283.054

Table 7.4: Accuracy Assessment Water Body Extraction

Accuracy	0.9999
95% CI	(0.9999, 0.9999)
Kappa	0.9937
Mcnemar's Test P-Value	<2.2e-16
Sensitivity	1.0000
Specificity	0.9875
Detection Rate	0.9929
Balanced Accuracy	0.9938

The confusion matrix (Table 7.3) shows that 3.571 non-water pixels were

classified as water pixels. This could be affiliated with the fact that some could shadow pixels were falsely classified as water body pixels. The overall accuracy with 0.999 and the Kappa parameter (0.9937) give an indication that the classification process was successfully applied on water bodies.

Figure 7.7: Results of the Water Body Extraction

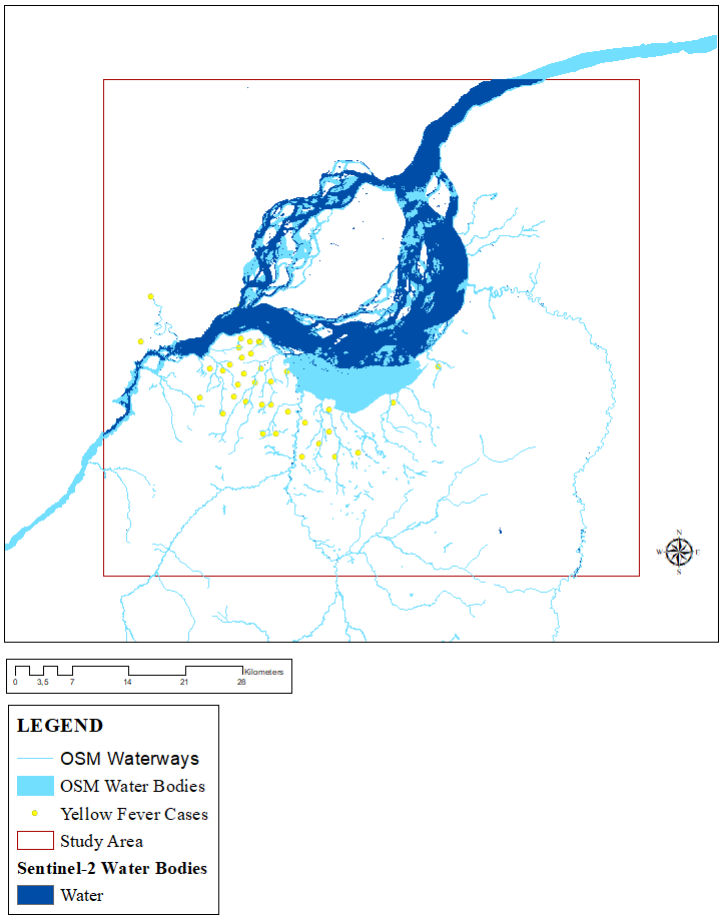
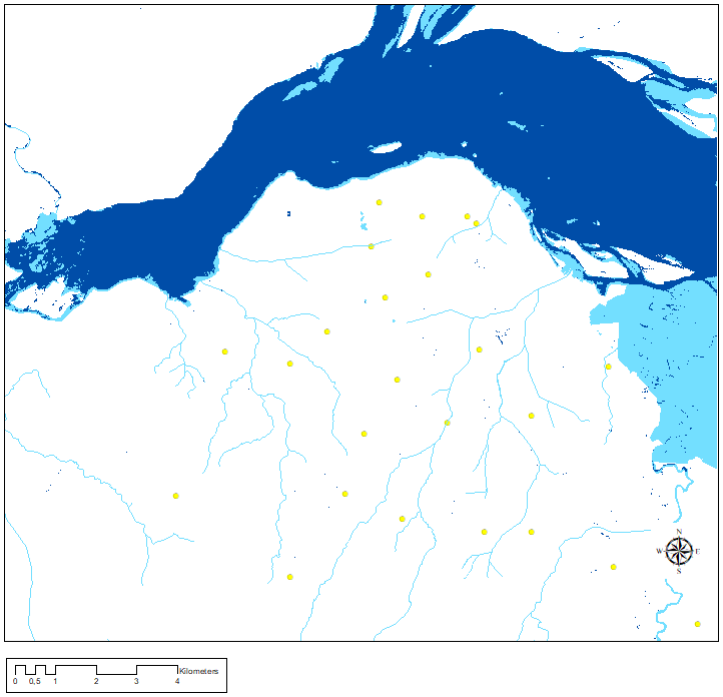


Figure 7.8: Results of the Water Body Extraction - Details



The water bodies were extracted over the whole Sentinel-2 MWR scene (Figure 7.7, Figure 7.8) The extraction was most accurate for the Congo river and its branches. Further, by visually interpreting the classification result most of the small rivers and ponds were classified successfully. However, classification problems arose during the classification of cloud shadows as the signatures of cloud shadows differ only slightly from the water body signatures (see Appendix 2). Moreover, it is not possible to accurately define the minimum size of successfully classified water bodies. It would have been meaningful to validate a subset area, which includes smaller ponds using additional ground truth polygons derived from VHR imagery. Nonetheless, it is assumed that small ponds were successfully classified.

Calculation of mean Distances to Water Bodies

Following the water body extraction, mean distances to water bodies were calculated. First, a buffer of 600 m (estimated flight range of the Ae. Species)

was calculated using the QGIS buffer function representing the mean distances to water bodies (including the water areas itself). A better strategy might have been calculating mean distances to water bodies and then excluding areas covered by water since the risk of infection remains low on the surface of the water. However, a constant grid of 1*1 km was computed to reach the same resolution as the other data sources. Further, zonal statistics for the 1*1 km were calculated using the zonal statistics function in QGIS. Resulting in a 1*1 km grid including the percentage of each pixel covered by water including the distances to water bodies. Consequently, a table was created including a column consisting of the percentage of water plus the surrounding water buffer. To prepare the grid for further processing, a cropping function using R was used to create a subset of the water grid.

Finally, the grid (including the table with the information about the percentage of water and buffer) was converted to a data frame object using the R base function *as.data.frame* which is defined as follows: *as.data.frame(x)* where x is the R object that should be converted.

Landsat-8 Preprocessing

Six LANDSAT-8 OLI scenes were acquired from the USGS Earth Explorer download portal in GeoTIFF format including eleven spectral bands and a metadata file for each scene. The metadata file includes information about image attributes, radiometric rescaling factors, and projection parameters. To analyze a time series over 2016 scenes from five months with minimal cloud coverage were acquired:

Table 7.5: Acquired Landsat-8 Scenes

Landsat Product Identifier	Acquisition date
LC08_L1TP_182063_20161211_20170316_01_T1	December 11, 2016
LC08_L1TP_182063_20161125_20170317_01_T1	November 25, 2016
LC08_L1TP_182063_20161109_20170318_01_T1	November 09, 2016
LC08_L1TP_182063_20160906_20170321_01_T1	September 06, 2016
LC08_L1TP_182063_20160618_20170323_01_T1	June 18, 2016
LC08_L1TP_182063_20160211_20170330_01_T1	February 11, 2016

Conversion to TOA radiance and reflectance

For LANDSAT-8 products several applications and algorithms are available to process level 1 products to Top-of-Atmosphere (TOA) corrected level 2 products. The LANDSAT-8 satellite is equipped with two different sensors, the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). Preprocessing has the goal to minimize sensor, solar, atmosphere, and topographic effects, which can emerge during the image acquisition process. LANDSAT level 1 products are delivered as digital numbers (DNs), which need to be converted in absolute reflectance and radiance values (TOA products). One major processing step is the conversion from the DNs to TOA reflectance and radiance values via solar correction. The following equation is applied to convert DNs to TOA radiance values for all 11 bands (Young et al., 2017):

$$L_\lambda = M_L * Q_cal + A_L \tag{7.2}$$

where L is the TOA spectral radiance in Watts/(m2 * srad * µm), ML the band-specific multiplicative rescaling factor, Qcal the DN values and AL the additive rescaling factor. Both the radiance multiplicative and additive rescaling factors can be obtained from the delivered metadata file, RADIANCE MULT BAND x and RADIANCE ADD BAND x, where x is the band number (Zanter, K., 2016; USGS, 2017).

The acquired Landsat-8 data was processed using R and the following

function. The radiance was calculated via a function based on the equation above.

Furthermore, a conversion from DNs to TOA reflectance values is necessary. Therefore, a conversion using the following equation on band 1 – 9 is implemented:

$$p_\lambda = M_P * Q_cal + A_p \tag{7.3}$$

where p is the TOA spectral reflectance value, Mp the reflectance multiplicative scaling factor for the band, Qcal the DN pixel values and Ap the reflectance additive scaling factor. Both the reflectance multiplicative and additive scaling factor can be obtained from the delivered Metadata file specified as REFLECTANCEW MULT BAND N and REFLECTANCE ADD BAND N, where N is the band number (Zanter, K., 2016; USGS, 2017).

The reflectance conversion was done by using the R package *'Landsat 8'*. The function which is the following: reflconv(x, Mp, Ap) uses three input parameters. X is the scene to be converted either in matrix, data frame or SpatialGridDataFrame type. Mp is as can be obtained from the equation the reflectance multiplicative scaling factor and Ap the additive scaling factor.

```
data(band2)
band2.dn<- as(band2, 'matrix')
band2.refl<-reflconv(band2.dn,2.0000E-05,-0.100000)
```

The first line reads the data, in this case the blue (band 2) Landsat-8 OLI band. The second line takes the input and writes it in a matrix. And the third line converts the DN value to TOA reflectance values using the function reflconv with the band 2 DNs and the multiplicative and additive scaling factors from the metadata file (Santos A., 2017).

Conversion to At-Satellite Brightness Temperature

Finally, the thermal infrared bands (TIRS) 10 and 11 need to be converted

to At-Satellite Brightness Temperature to get temperature values in Kelvin. The following equation was implemented in R to get temperature values from the TIRS bands:

$$T = \frac{K2}{ln(\frac{K1}{L\lambda}) + 1} \tag{7.4}$$

where T is the temperature in Kelvin, K1 and K2 are the band-specific thermal conversion constants, and L the TOA radiance values calculated in the previous step. The thermal constants can be obtained from the metadata file specified as K1 CONSTANT BAND N and K2 CONSTANT BAND N, where n is the number of the band which is processed (Zanter, K., 2016; USGS, 2017).

Preparation for Processing

After the calculation of TOA and At-Satellite Brightness Temperature values, further pre-processing including image resampling and cropping was conducted. The resampling of the LANDSAT-8 scenes was performed using the ERDAS Imagine 2015 resampling software tool, to fit the LANDSATt-8 scenes to the MODIS scenes. The nearest neighbor resampling was applied to all LANDSAT-8 scenes to obtain images with the same spatial resolution as the MODIS data (1*1 km). A significant drawback of this resampling approach is that the nearest neighbor algorithm assigns output pixel values using the values of the nearest neighbors (Pouncey et al., 1999). Thus, this resampling method might be a source of error in assigning pixels to neighbors that do not have the same or similar pixel values. However, one novel approach would be the calculation of the mean of all pixels within an area of 1*1 km.

All images were cropped to the study area extent to guarantee the overlapping of all input data using the crop function implemented in the R package raster. Finally, the image data from all acquired and processed scenes was converted to a data frame object using the R base function as.data.frame.

MODIS Terra/Aqua Preprocessing

Several MODIS products from 2016 were downloaded from the Reverb Portal including global Surface Temperature and Emissivity 8-day data (MOD11A2), Vegetation Indices 16-day data (MOD13A) and global Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation (FPAR) 8-day data (MOD15A2). Table 7.6 specifies the used MODIS products, the temporal granularity, the pixel size in meters, the coordinate system and the scaling factors for each product:

Table 7.6: My caption

Product	Product Type	Temporal Granularity	Pixel size (m)	Scaling Factors
MOD 11 A2	LST Day and Night	8-Day	1000	0.02
MOD 13 A2	EVI and NDVI	16-Day	1000	0.0001
MOD 15 A2	LAI and FPAR	8-Day	1000	0.1 / 0.01

A spatial resolution of 1000m was chosen for the risk zone mapping because of the further processing of the random forest. A finer resolution with all the used variables would result in extensive computation time. However, the coarse spatial resolution may also have some limitations concerning the microhabitats of the investigated mosquito species.

Product type conversion and re-projection

After the acquisition, the MODIS products were re-projected to the WGS-84 reference frame and converted from HDF to GeoTIFF using the MODIS Reprojection Tool (MRT).

"*The MODIS Reprojection Tool (MRT) was developed to support higher level*

MODIS Land products which are distributed as Hierarchical Data Format-Earth Observing System (HDF-EOS)2 files projected to a tile-based Sinusoidal grid. MRT software facilitates the use of MODIS Land tiles by providing map projection, format conversion, and spectral and spatial subsetting options and is compiled for use on multiple operating systems."

The tool can either be run from a Graphical User Interface or the command line. For this thesis, the MODIS data was processed from the Graphical User Interface using a parameter file, which contained all needed information for re-projection and format conversion.

Preparation for Processing

Both the Land Surface Temperature (LST) and Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation (FPAR) were aggregation to 16-days composites as the vegetation index products are delivered only as 16-days composites. The aggregation was conducted by conflating two images each time and calculating pixel value averages resulting in 23 scenes for all products for 2016. The aggregation process was conducted using the overlay function from the raster package in R. The overlay function creates a new raster object based on two or more input raster objects (Hijmans et al., 2017).

```
sel <- subset(sel2016, 1:2)
a1 <- overlay(sel, fun=mean)
```

The example code above shows how the aggregation works. A subset was calculated from the first and second scene in 2016, and the first combined layer a1 was calculated via the overlay function using the computation function "mean", resulting in one scene including a 16-day composite of the input values.

After the aggregation process, all products were rescaled using the rescaling factors from the MODIS product specification table. Several functions were

implemented for rescaling the products:

```
sfactor1 <- 0.0001 # rescaling factor EVI and NDVI
rasterRescale1<-function(r){
  (r*sfactor1)
}
```

Where sfactor 1 is the rescaling factor for the vegetation index products and r the MODIS input.

The Land Surface Temperature products were converted from Kelvin to Degree Celsius to ensure easier readability of temperature values. The temperature conversion was performed using the convert temperature function from the *'weather metrics'* packages implemented in R (Anderson et al., 2016).

```
convert_to_C<-function(t){
  convert_temperature(LSTD_rescale, old_metric = "kel-
vin", new_metric = "celsius")
}
```

The *convert_temperature* function needs the input, the old metric and new metric as arguments. In this example, the rescaled Day Land Surface Temperature (LSTD rescale) was taken as input, the old metric was defined as Kelvin and the new desired metric was set to Degree Celsius.

After the rescaling process all products were clipped to the study area extent using the crop function and converted into data frame format. Resulting in three data frames, one for each product, with all pixel values.

ESA CCI Land Cover

The CCI Land Cover Map 2015 at a spatial resolution of 300 * 300 m was provided by TU Vienna. The map is delivered in TIF format in the WGS-84 reference frame (EPSG: 4326). The Land Cover Map was preprocessed using the ERDAS Imagine nearest neighbor resampling approach. The Map was resampled from 300 * 300 m to 1 * 1 km. Further, it was clipped to the study area extent. The following table shows the appearing LULC types, the total area of each type in square kilometers and approximated percentage. The area statistics were calculated using the QGIS GRASS r.stats algorithm (Shapiro, 2017).

Table 7.7: CCI LULC before and after resampling

Land Cover Type	Area in %	Area in % (resampled)	Area in km (resampled)
Cropland, rainfed	4.83 %	54.14%	2106.92
Cropland irregated and post flooded	7.77%	0.18%	6.87
Herbaceous cover	4.47 %	0.15%	6.01
Mosaic cropland (>50%) / natural vegetation (<50%)	10.41%	4.28%	166.66
Mosaic natural vegetation (>50%) / cropland (<50%)	1.48%	6.58%	256.01
Tree cover, broadleaved, evergreen, closed to open (>15%)	5.18%	0.24%	9.45
Tree cover, broadleaved, deciduous, closed to open (>15%)	3.67%	0.77%	30.06
Tree cover, broadleaved, deciduous, closed (>40%)	0.04%	0.09%	3.44
Tree cover, broadleaved, deciduous, open (15-40%)	22.02%	3.82%	148.64
Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	9.34%	4.28%	166.68
Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	1.71%	0.86%	33.51
Shrubland	2.08%	1.99%	77.32
Shrubland deciduous	0.05%	0.26%	10.31
Grassland	2.01%	3.97%	154.64
Tree cover, flooded, saline water	1.55%	0.07%	2.58
Shrub or herbaceous cover, flooded fresh/saline/brackish water	4.48%	0.60%	23.19
Urban areas	12.32%	8.06%	313.51
Consolidated bare areas	0.01 %	0.04%	1.72
Water bodies	10.04 %	10.04%	390.87

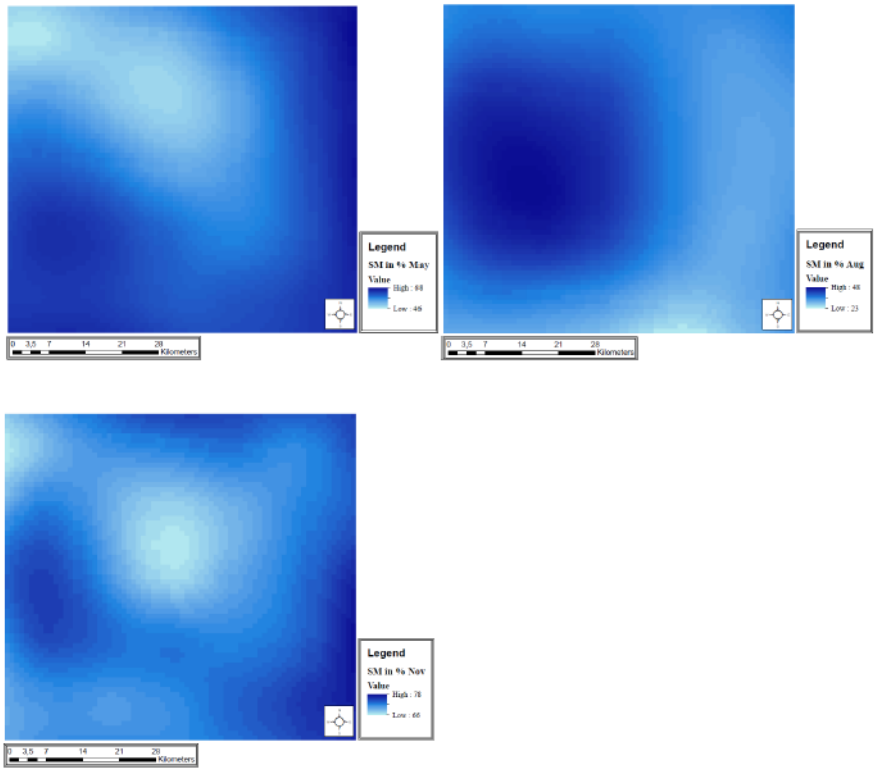
As can be seen from Table 7.7 the most essential LULC classes are rainfed cropland with more than 2106 square kilometers, covering more than 54 percent of the study area, water bodies which cover about 390 square kilometers and roughly 10 percent, and urban or built-up areas which cover about 313 square kilometers and more than 8 percent of the total study area.

ASCAT Soil Moisture

Spatial Interpolation and Preparation

ASCAT soil moisture data is produced on 12.5 * 12.5 km. Thus, the soil moisture data needed to be interpolated to 1 * 1 km. The interpolation was performed and provided by TU Vienna using SciPy interpolation algorithm. SciPy is a collection of mathematical algorithms built on the Numpy extension of Python. The interpolation sub-package includes several approaches for interpolating data: 1) 1-D interpolation, 2) Multivariate data interpolation for multidimensional data, 3) Spline interpolation in one- or two dimensions and procedural and object-oriented, and 4) Radial basis functions for smoothing and interpolation (SciPy, 2017). The interpolation was conducted using a radial basis function.

Figure 7.9: Soil Moisture Content Kinshasa



The three Figure 7.9 show the soil moisture measured within the study after the wet season in May 2016, the before wet season in August 2016 and after the wet season in November 2016. The post-wet season is characterized by soil moisture contents between 46 and 68 %. The pre-wet season between August and October, which is after the dry season during the winter months June, July, and August, is characterized by lower soil moisture contents between 29 and 48 %. The wettest season, which is during the southern hemisphere summer in November, December and January, is characterized by overall high soil moisture contents with low variability over space. The soil moisture content is between 68 and 78 %.

After interpolating the soil moisture scenes, the projection was set to WGS 84 reference frame. Then subsets of scenes were created and rescaled by

dividing the soil moisture values by two. Finally, the soil moisture values were converted to a data frame.

7.2 Calculation of Risk Zone Variables

After the processing, the input data including resampling to 1*1 km spatial resolution, cropping and calculation of TOA reflectance values and At-Satellite Brightness Temperature values from the available LANDSAT-8 scenes, several indices, and metrics were computed. Table 7.8 shows beside the already compiled mean distances to water bodies and the land cover classes all calculated indices (from LANDSAT-8) and metrics. Further, all SENTINEL-2 bands (B1 – B12) were employed as input variables.

Table 7.8: Risk Zone Variables

SENTINEL-2A	Spectral Bands	B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11, B12
	Mean Distance to Water Bodies	Percentage of Water within 1*1km
MODIS	EVI	Mean, Var, Std, Min, Max, Q
	NDVI	Mean, Var, Std, Min, Max, Q
	LST Day	Mean, Var, Std, Min, Max, Q
	LST Night	Mean, Var, Std, Min, Max, Q
	LAI	Mean, Var, Std, Min, Max, Q
	FPAR	Mean, Var, Std, Min, Max, Q
LANDSAT-8	Indices: EVI, NDVI, SAVI, MSAVI, NDWI	Mean, Var, Std, Min, Max, Q
	Metrics	Mean, Var, Std, Min, Max, Q
ASCAT	Metrics	Mean, Var, Std, Min, Max, Q

The computation of the applied metrics and indices is described in the following two chapters. The first chapter describes the calculation of metrics from all MODIS input data including the land surface temperature, the vegetation indices, the leaf area index and the fraction of photosynthetically active radiation. The second chapter elaborates on the computation of vegetation indices from the LANDSAT-8 scenes and the calculation of metrics (LANDSAT-8). The calculation of the ASCAT metrics in not further described in detail as it was calculated in the same way as the MODIS and LANDSAT-8 metrics.

Calculation of MODIS Metrics

For this thesis, several metrics were calculated from the MODIS data including the standard deviation, variance, mean, minimum, maximum and quantiles. The following equation for calculating metrics based on pixel values was implemented in R:

The mean for all pixel in all 23 available scenes was calculated using the R base package and the function rowMeans. The rowMeans function is defined as rowMeans(x, na.rm = FALSE, dims = 1), where x is input values for the mean calculation, na.rm defines if missing values (NaN) should be ignored and dims which dimensions are regarded as rows (ETH, no date). Further, the standard deviation was calculated using the R matrixStats package. Where x are values, and N the number of data points (Science Buddies, 2017). The rowSds function is defined as rowSds(x, rows = NULL, cols = NULL, ...), where x are the input values and rows, and col can be defined if the standard deviation of a subset of the rows and columns within matrix should be processed. Where x are again the values, and N the number of data points (Science Buddies, 2017). The variance is computed by applying the following function in R: It estimates the variance for each row in the data frame.

```
RowVar <- function(x) {
  rowSums((x - rowMeans(x))^2)/(dim(x)[2] - 1)
}
```

The minimum and maximum pixel values within the 23 layers were calculated using the apply function from the base R package, which returns a vector or array by applying a function. In this case, the function was set to minimum and maximum (Becker et al., 1988).

Finally, the quantiles for each row were calculated. Quantiles can efficiently

be computed using the matrixStats R package (Bengtsson et al., 2017).

The following code example shows the implementation of the functions above on the Normalized Difference Vegetation Index (NDVI):

```
NDVI <- all.16[26:48]
NDVI.m <- as.matrix(NDVI)
NDVI_Mean <- rowMeans(NDVI)
NDVI_Var <- RowVar(NDVI)
NDVI_Min <- apply(NDVI,1,min)
NDVI_Max <- apply(NDVI,1,max)
NDVI_Std <- rowSds(NDVI.m)
NDVI_Q <- rowQuantiles(NDVI.m)
colnames(NDVI_Q)<-
c('Q.NDVI.1','Q.NDVI.2','Q.NDVI.3','Q.NDVI.4','Q.NDVI.
5')
NDVI                                                  <-
cbind(NDVI_Mean,NDVI_Var,NDVI_Max,NDVI_Min,NDVI_Std,ND
VI_Q)
```

The first line accesses the NDVI data via indices from all.16. The second line convert the format from data frame to matrix. This needs to be performed in order to work with the matrix statistics package *matrixStats*. The third line uses the *rowMeans* function to calculate NDVI mean values from all available NDVI values. The next few lines calculate all the metrices listed above. Then names of the columns are assigned according to the metrices and finally the metrices are merged to one data frame for further processing.

Thus, for each MODIS product including EVI, NDVI, LSTD Day, LSTD Night, LAI and FPAR all metrices as defined above were calculated resulting in six data frames.

Calculation of LANDSAT-8 Metrics and Vegetation Indices

The further processing of the LANDSAT-8 images was composed of the calculation of metrics and vegetation indices from the LANDSAT pixel values. The first step was comprised of the computation of vegetation indices from different bands.

The first vegetation index, which was computed, was the Normalized Difference Vegetation Index (NDVI). According to the Landsat Product Guide (2017), which describes LANDSAT surface reflectance-derived spectral indices, the NDVI ratio should be calculated from the Red (band 4) and NIR (band 5) spectral bands using the standard NDVI equation:

$$NDVI = \frac{Red - NIR}{Red + NIR} \tag{7.5}$$

The NDVI investigates the contrast in the reflectance of vegetation in the Red and NIR spectra. It is commonly used in remote sensing for examining vegetation dynamics. The NDVI has been extensively used in epidemiological studies as vegetation is influenced by variation in precipitation and temperature. Thus, a combination of remotely sensed surface reflectance values and vegetation indices derived from reflectance values and climate variables has been employed to model vector dynamics (Kalluri et al., 2007; USGS, 2017).

The second index, the Enhanced Vegetation Index (EVI) was implemented by using the NIR, Red and Blue spectral band. The usage of the Blue band in calculating vegetation index has the advantages that it reduces background reflectance and the effects of atmospheric aerosols. The EVI is applied to model vegetation dynamics as predictors for vector-borne diseases. For LANDSAT-8 images the EVI should be calculated using the equation (Fuller et al., 2009; USGS, 2017):

$$EVI = 2.5 * \frac{(Band5–Band4)}{(Band5 + 6 * Band4–7.5 * Band2 + 1)} \tag{7.6}$$

The implementation was realized by the following function based on the equation and has three input arguments, band 5, band 4, and band 2:

```
EVI <- function(b5,b4,b2){
  2.5* ((b5 - b4)/(b5 + 6*b4 - 7.5*b2 + 1))
}
```

The soil adjusted vegetation index (SAVI) is a ratio between the Red and NIR band including a soil brightness correction factor (0.5). The modified soil adjusted vegetation index (MSAVI) is calculated using the Red and NIR band and an inductive soil brightness factor function. The function is applied to reduce the soil effects on the vegetation signal. The SAVI and MSAVI indices were developed from the traditional vegetation indices EVI and NDVI to overcome the effect of soil assuming that the relationship between NIR and Red reflectance from bare soil is linear. Both indices are implemented using the formulas below (Rondeaux et al., 1996; USGS, 2017):

$$SAVI = 2.5 * \frac{(Band5 – Band4)}{(Band5 + Band4 + 0.5) * 1.5} \tag{7.7}$$

$$MSAVI = 2 * Band5 + 1 – \sqrt{\frac{((2 * Band5 + 1)2 – 8 * (Band5 – Band4))}{2}} \tag{7.8}$$

Further a Normalized Difference Moisture Index (NDMI) was calculated from the available LANDSAT scenes using the NIR and SWIR1 bands:

$$NDWI = \frac{Band5 - Band6}{Band5 + Band6} \tag{7.9}$$

The indices were stacked together in one data frame and exported for further processing.

Metrics were calculated from a stack of each LANDSAT-8 band and the previously calculated vegetation indices. Metrics from each band and all

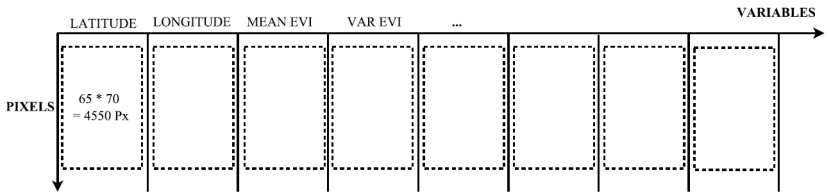
indices were calculated to assess the importance of each LANDSAT-8 band and evaluate if vegetation indices calculated from LANDSAT-8 are having a significant impact on the definition of risk zones. Thus, first bands from all images were stacked together using the cbind function, resulting in eleven (11 LANDSAT-8 bands) data frames with six columns from six different images. Second, all vegetation indices from all scenes stacked together, resulting in five (all calculated indices) data frames with six columns from six different images. These two data frames were then used to calculate metrics including the mean, the variance, the minimum, the maximum, the standard deviation and the quantiles. The following code shows the stacking and calculation of the spectral band 1.

```
LC08.1<-   cbind(landsat$l1.0211,landsat$l1.0618,land-
sat$l1.0906,
landsat$l1.1109,landsat$l1.1125,landsat$l1.1211)
LC08.1_Mean <- rowMeans(LC08.1) # mean
LC08.1_Var <- RowVar(LC08.1)# variance
LC08.1_Min <- apply(LC08.1,1,min) # min
LC08.1_Max <- apply(LC08.1,1,max) # max
Lc08.1_Std <- rowSds(LC08.1) # standard deviation
LC08.1_Q <- rowQuantiles(LC08.1) # quantiles
colnames(LC08.1_Q)                                    <-
c('Q.1.1','Q.1.2','Q.1.3','Q.1.4','Q.1.5')
L1<- cbind(LC08.1_Mean,LC08.1_Var,LC08.1_Min,
LC08.1_Max, Lc08.1_Std,LC08.1_Q)
```

The first two lines stack the Coastal Blue LANDSAT bands from six images together using the cbind function. The following lines calculate the metrics described in the previous section and the last two lines are responsible for the binding the metrics results.

7.3 Habitat Modelling using Ensemble Learning

Figure 7.10: Data Frame Design



A data frame object was created to further analyze the acquired and preprocessed data in R using Ensemble Learning. A data frame, which is a list of vectors with equal length, stores data tables in R. The first row, which is called the header, contains the variable names. Figure 7.10 visualizes the data frame design. The first two columns contain the geographic latitude and longitude, the location of each pixel within the study area. The pixel values of the variables are added as vectors to the location. The length of each vector within the data frame is 4.550, resulting from 65*70 pixels from each image, resulting in a data frame with the dimension of 4.550 rows and 245 columns.

Random Forest Ensemble Learning

The classification was based on the random forest ensemble learning algorithm (chapter 4). The algorithm was implemented using the R package *'randomForest'*. The random forest implemented in R is based on Breiman's random forest algorithm.

```
randomForest(x, y=NULL,  xtest=NULL, ytest=NULL, ntree=500,
             mtry=if (!is.null(y) && !is.factor(y))
             max(floor(ncol(x)/3), 1) else floor(sqrt(ncol(x))),
             replace=TRUE, classwt=NULL, cutoff, strata,
             sampsize = if (replace) nrow(x) else ceiling(.632*nrow(x)),
             nodesize = if (!is.null(y) && !is.factor(y)) 5 else 1,
             maxnodes = NULL,
             importance=FALSE, localImp=FALSE, nPerm=1,
             proximity, oob.prox=proximity,
             norm.votes=TRUE, do.trace=FALSE,
             keep.forest=!is.null(y) && is.null(xtest), corr.bias=FALSE,
             keep.inbag=FALSE, ...)
```

The function, which is called *randomForest*, can have several input arguments. X consists of a data frame or matrix of predictors. Y is the response vector, if it is a factor (categorical variable), classification is assumed otherwise regression will be computed. XTest and YTest may be implemented when a test set is used (equal to X and Y). Ntree is the number of trees to grow. The number of trees to grow is set to 500 by default. However, the number should not be set too small, to ensure that all input rows get predicted at least a couple of times. The Mtry argument is the number of variables randomly sampled as candidates at each split. The default value is calculated via the square root of all input variables. Replace indicates if the sampling should be done with or without replacement. The importance input argument can either be set to true or false, if true the importance of each variables within the random forest algorithm is assessed.

According to Liaw and Wiener (2015) a random forest object is calculated using the function above with a list of the following values:

Values

- Call: the original call to randomForest
- Type: one of regression, classification, or unsupervised
- Predicted: the predicted values of the input data based on out-of-bag samples

- Importance: a matrix with nclass + 2 (for classification) or two (for regression) columns. For classification, the first nclass columns are the class-specific measures computed as mean descrease in accuracy. The nclass + 1st column is the mean descrease in accuracy over all classes. The last column is the mean decrease in Gini index. For Regression, the first column is the mean decrease in accuracy and the second the mean decrease in MSE. If importance=FALSE, the last measure is still returned as a vector.
- importanceSD: The "standard errors" of the permutation-based importance measure. For classi- fication, a p by nclass + 1 matrix corresponding to the first nclass + 1 columns of the importance matrix. For regression, a length p vector.
- localImp: a p by n matrix containing the casewise importance measures, the [i,j] element of which is the importance of i-th variable on the j-th case. NULL if localImp=FALSE.
- Ntree: number of trees grown
- Mtry: number of predictors sampled for spliting at each node
- Forest: (a list that contains the entire forest; NULL if randomForest is run in unsupervised mode or if keep.forest=FALSE
- err.rate: (classification only) vector error rates of the prediction on the input data, the i-th element being the (OOB) error rate for all trees up to the i-th
- Confusion: (classification only) the confusion matrix of the prediction (based on OOB data)
- Votes: (classification only) a matrix with one row for each input data point and one column for each class, giving the fraction or number of (OOB) 'votes' from the random forest
- oob.times: number of times cases are 'out-of-bag' (and thus used in computing OOB error estimate)
- proximity: if proximity=TRUE when randomForest is called, a matrix of proximity measures among the input (based on the frequency that pairs of data points are in the same terminal nodes)
- Mse: (regression only) vector of mean square errors: sum of squared residuals divided by n
- Rsq: (regression only) "pseudo R-squared": 1 - mse / Var(y)

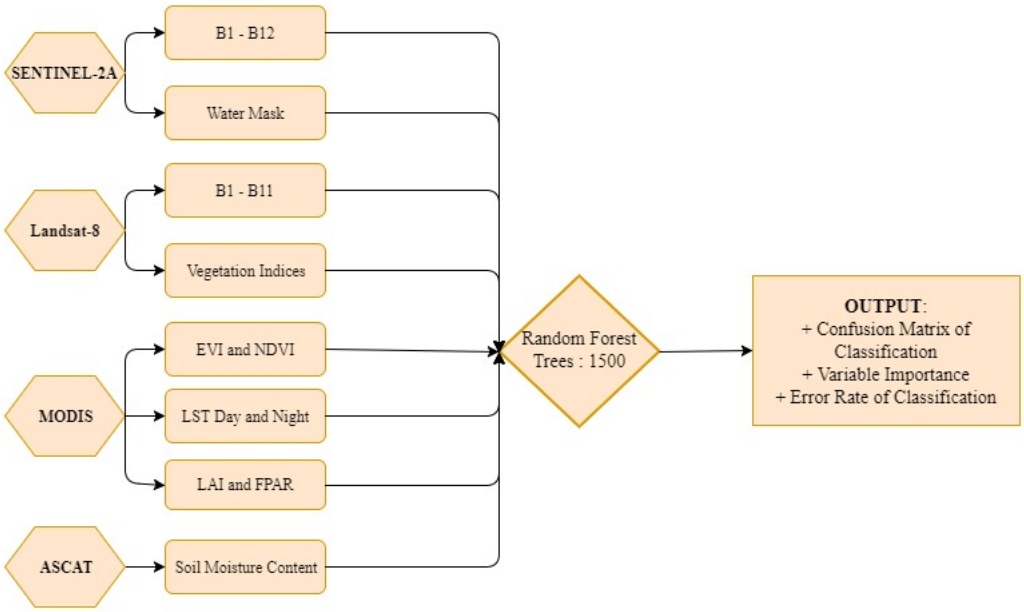
Selection of Variables

To assess the importance of the acquired indicators several habitat models were tested using different subsets of the data (Table 7.9):

Table 7.9: Random Forest Models

Model Name	Input Data	Indicators Count
M_1	All Indicators	243
M_2	MODIS Indicators	60
M_3	SENTINEL-2A Spectral Bands	13
M_4	LANDSAT-8 Time Series Metrics of Spectral Bands	109
M_5	LANDSAT-8 Indices	50
M_6	MODIS and ASCAT Indicators	70
M_7	MODIS and LANDSAT-8 Indicators	169
M_8	MODIS and SENTINEL-2A Indicators	83
M_9	MODIS, ASCAT and LANDSAT-8 Indicators	179
M_10	MODIS, ASCAT and SENTINEL-2A Indicators	83
M_11	ASCAT Indicators	10
M_12	Subset of Indicators with Mean Gini Decrease Index >3	35
M_13	ASCAT and LANDSAT-8	119
M_14	ASCAT and SENTINEL-2A	23
M_15	ASCAT and subset of Indicators with Mean Decrease Index >3	40
M_16	ASCAT and most important Indicator	11

Figure 7.11: Flowchart of Model 1



The first model was determined using all available input variable (Table 7.9) to assess the overall importance of each variable (7.11) including the SENTINEL-2 bands and the derived mean distances to water bodies, the LANDSAT-8 indices and metrics, the MODIS indices and metrics and the ASCAT soil moisture content. The CCI land cover map was not included as input variable, but applied for the plausibility check (chapter 7.4.3)

The models M_2, M_3, M_4, M_5 and M_11 tested the usage of only one remote sensing imagery source to assess the importance of each sensor for habitat modelling. Model 2 (M_2) was implemented to assess the acquired MODIS indicators EVI, NDVI, LST, LAI and FPAR. Model 3 (M_3) was based on the SENTINEL-2A scene to explore SENTINEL-2A spectral bands for modelling *Ae. Species*. Model 4 (M_4) and model 5 (M_5) were used to evaluate the usage of time series and indices derived from the two Landsat-8 sensors OLI and TIRS. Model 11 was implemented using ASCAT soil moisture content indicators.

Further, a combination of several data sources was utilized. M_6, M_7 and M_8 were used to evaluate the importance of MODIS data in combination with ASCAT indicators, LANDSAT-8 time series and indicators, and SENTINEL-2A indicators. A combination of three different remote sensing products was implemented within model 9 (M_9) and model 10 (M_10). Beyond that, two models were developed using ASCAT and LANDSAT-8 (M_13), and ASCAT and SENTINEL-2 (M_14).

Model 12 (M_12) was developed using the mean Gini Decrease Index of the first random forest model (M_1). The most important variables were explored and employed as input indicators for model 12. The basic assumption behind model 12 was that the classification might improve, if only those variables are considered, which were most important during the first classification based on model 1.

Finally, two models were tested using the soil moisture data derived from the ASCAT sensor and a subset of the most important variables (M_15) and the most import indicator beside soil moisture (M_16).

7.4 Results and Accuracy Assessment

7.4.1 Results and Accuracy

Table 7.10 shows the results of the random forest habitat models. The table consists of the model number and the employed data. The classification was based on two habitat classes: (1) non-suitable means that those habitats are not suitable for the presence of Ae. Species using the specified variables; and (2) suitable implies habitats were Ae. Species might be present. The class error for each class is indicating to what extent the random forest model was able to predict the classes.

Table 7.10: Results of the Random Forest Habitat Models

Model No.	Model Description	Class Error Non-suitable	Class Error Suitable
1	All indicators	0.0032	0.3670
2	MODIS	0.0056	0.4249
3	SENTINEL-2 band metrics	0.0107	0.6154
4	LANDSAT-8 band metrics	0.0051	0.5018
5	LANDSAT-8 indices	0.0072	0.5934
6	MODIS and ASCAT	0.0051	0.3297
7	MODIS and LANDSAT-8	0.0044	0.3640
8	MODIS and SENTINEL-2A	0.0061	0.4579
9	MODIS, ASCAT, and LANDSAT-8	0.0040	0.3736
10	MODIS, ASCAT, and SENTINEL-2	0.0047	0.3590
11	ASCAT	0.0065	0.1905
12	Most important indicators	0.0049	0.3480
13	ASCAT and LANDSAT-8	0.0047	0.4066
14	ASCAT and SENTINEL-2	0.0061	0.3626
15	ASCAT and subset of all	0.0047	0.3223
16	ASCAT and LST	0.0047	0.1538

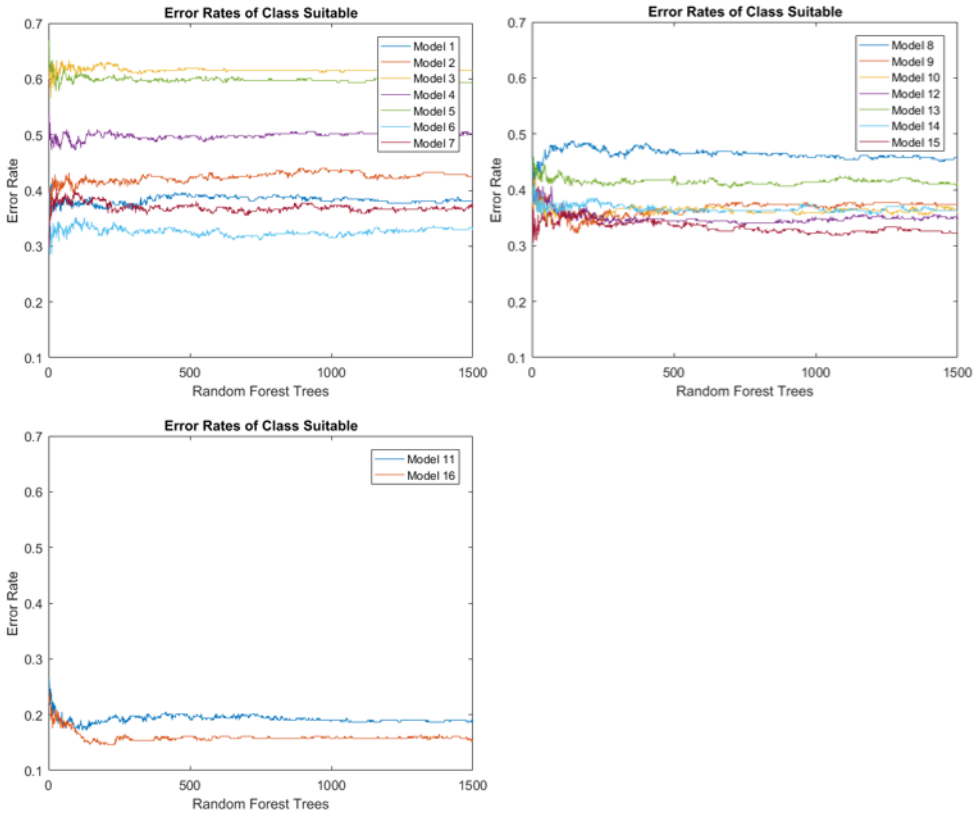
Table 7.11 gives an overview of the classification results, detailed discussions of the models will be given in chapter 7.4. The best classification results

were obtained using model 16, which was a combination of the ASCAT data and the night land surface temperature derived from MODIS. Regarding the accuracy, the model is able to predict using the prepared training data, the class which is suitable for the abundance of Ae. Species and therefore the probability of occurring YF cases with 84.62 % (Table 7.10: 1 - 0.0.1538). The accuracy of the non-suitable class is 99.5 % (Table 7.10: 1 - 0.0047). Considering model 11, the accuracy of the class suitable is 80. 95 % (Table 7.10: 1 - 0.1905) and for the non-suitable class 99.35% (Table 7.10: 1 - 0.0065). Similar classification accuracies for the class suitable were obtained using ASCAT plus a subset of all variables (model 15) and ASCAT in combination with MODIS data (model 6), which were approx. 67 %. Classification results with an accuracy over 60 % were received using a subset of all variables and all variables (model 12 and model 1), and combinations including ASCAT, MODIS, SENTINEL-2, and LANDSAT-8 metrics (Table 7.10). Finally, using one or two different data source input resulted in less than 60 % predictability of the class suitable. The predictability of class non-suitable is not discussed on further detailed but remains high with more than 99 % for all implemented models (Table 7.10).

Table 7.11: Classification Accuracies for all models

Model ID	Model Description	Classification Accuracy Class Suitable in %
16	ASCAT and LST	84.62
11	ASCAT	80.95
15	ASCAT and subset of all	67.77
6	ASCAT and MODIS	67.03
12	Subset of all	65.20
9	ASCAT, MODIS, and SENTINEL-2	64.10
14	ASCAT and SENTINEL-2	63.74
7	MODIS and LANDSAT-8	63.60
1	All indicators	63.30
10	ASCAT, MODIS, and SENTINEL-2	62.64
13	ASCAT and LANDSAT-8	59.34
2	MODIS	57.51
8	MODIS and SENTINEL-2A	54.21
4	LANDSAT-8 band metrics	49.82
5	LANDSAT-8 indices	40.66
3	SENTINEL-2 band metrics	38.46

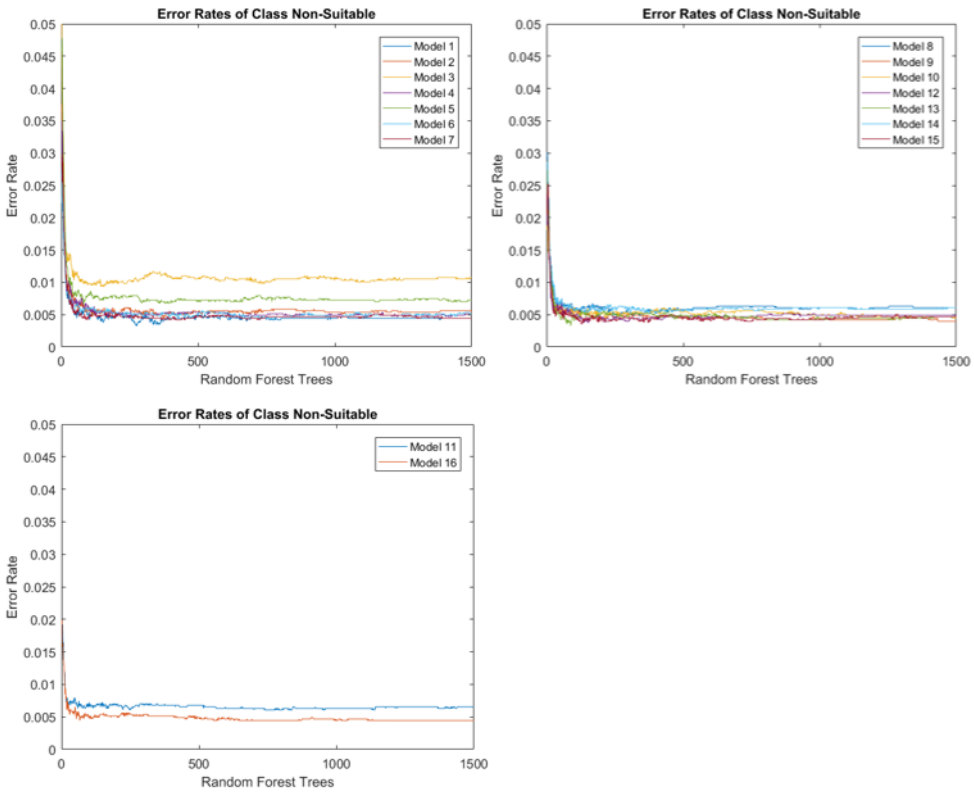
Figure 7.12: Error Rates of Class Suitable



To assess the influence of the built trees during the random forest classification process error rates of the class suitable (Figure 7.12) and error rates of the class non-suitable (Figure 7.13) are plotted against the built random forest trees. It seems that the random forest tree building process (1500 trees were built) over time considering all class errors. This might be affiliated to the fact that the random forest takes a subset of all variables with replacement at each iteration, which means if one variable is not considered significant in one iteration it is not excluded for the next iterations. Assuming that non-important variables are excluded after each iteration, the curve might be smoother. However, regarding the first plot (Figure 7.12) including model 1 – 7 all models perform moderate of the tree building process, resulting in moderate classification results (Table 7.11). All

models are characterized by significant variations in the first 50 built trees. Model 3, which uses the SENTINEL-2A spectral bands metrics as indicators, has an average error rate of 0.6161. A similar moderate performance is obtained by applying the LANDSAT-8 indices including EVI, NDVI, SAVI, MSAVI, and NDWI from model 5. Model 5 has a mean error rate of 0.5978. Slightly better performance can be achieved using the calculated LANDSAT-8 metrics. However, the mean error rate is still moderate with 0.4979. Better classification results can be obtained using model 1 and 2 (Figure 7.12), and model 6 and 7. Those models have an average error between 0.32 and 0.42. The models 8 – 10 and 12 – 15 (Figure 7.12) behave similar to the models from plot 1, and all error rates are between 0.3 and 0.5. The last plot including model 11 and model 16 perform considerably better the other models. Model 11 uses the ASCAT soil moisture metrics for the classification process. The mean error rate of model 11 is 0.1922, which is significantly smaller than the mean error rates of the other models. The last model consists of indicators including the ASCAT soil moisture metrics and one of the most crucial variables the Land Surface Temperature performs much better than the other models with an average error of 0.1597.

Figure 7.13: Error Rates of Class Non-Suitable



Considering the error rates of the class Non-suitable, all models perform similar (Figure 7.13). Within the first ten decision trees, all models are characterized by moderate error rates. When ten decisions were made, the error rates decrease to a mean error rate of about 0.005. However, model 3, 5 and 11 are characterized by higher error.

However, in lack of validation data, which would be needed to evaluate all models independently from the training data, uncertainty remains within the classification accuracy of all models. Consequently, solely the training data is employed to assess the importance of the variables within each model, but no validation is applied to evaluate the results. To reliably assess the importance of the variables, all models need to be rerun using

independent validation data instead of the training data. Therefore, considering the training data conclusions about the classification accuracies and variables are limited.

7.4.2 Variable Importance

The following section presents and discusses in detail the variable importance for selected models. A detailed description of the variables as well as the other models are attached to this thesis in appendix 1. The variable importance of the other models, which are not described in this section, can be obtained from appendix 1. The Mean Decrease in Accuracy (MDA) and the Mean Decrease in Gini Index (MDG) of each variable are visualized.

"*The index MDA, utilizes permuting OOB samples to compute the importance of the variable. The OOB sample is the set of observations which are not used for building the current tree. It is used to estimate the prediction error and then to evaluate variable importance. The OOB error importance is defined as follows: For each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for classification, mean square error (MSE henceforth) for regression). Then the same is done after permuting each predictor variable. The differences between the two are then averaged over all trees. The larger of MDA value, the more important of the variable.*" (Han et al., 2016)

Model 1

Model 1 was implemented employing all potential indicators for defining suitable Ae. Species habitats. The first model aimed to test the importance of the products from each sensor including SENTINEL-2, LANDSAT-8, MODIS, and ASCAT

The first model is characterized by a class error rate of 36.70 % for the class suitable, resulting in 174 correctly predicted test areas and 99 wrongly predicted (Table 7.12) test areas using all input variables. Regarding class non-suitable, the random forest was able to correctly predict 4254 test areas,

which are non-suitable for the abundance of Ae. Species and the potential risk of occurring YF, resulting in a classification accuracy using the training data of 99.68 % (class error 0.32%) (Table 7.12). The average classification results may be ascribed to the use of too many predictive variables.

Table 7.12: Accuracy Model 1

	Non-suitable	Suitable	Class Error
Non-suitable	4254	15	0.0032
Suitable	99	174	0.3670

Figure 7.14: Variable Importance of Model 1

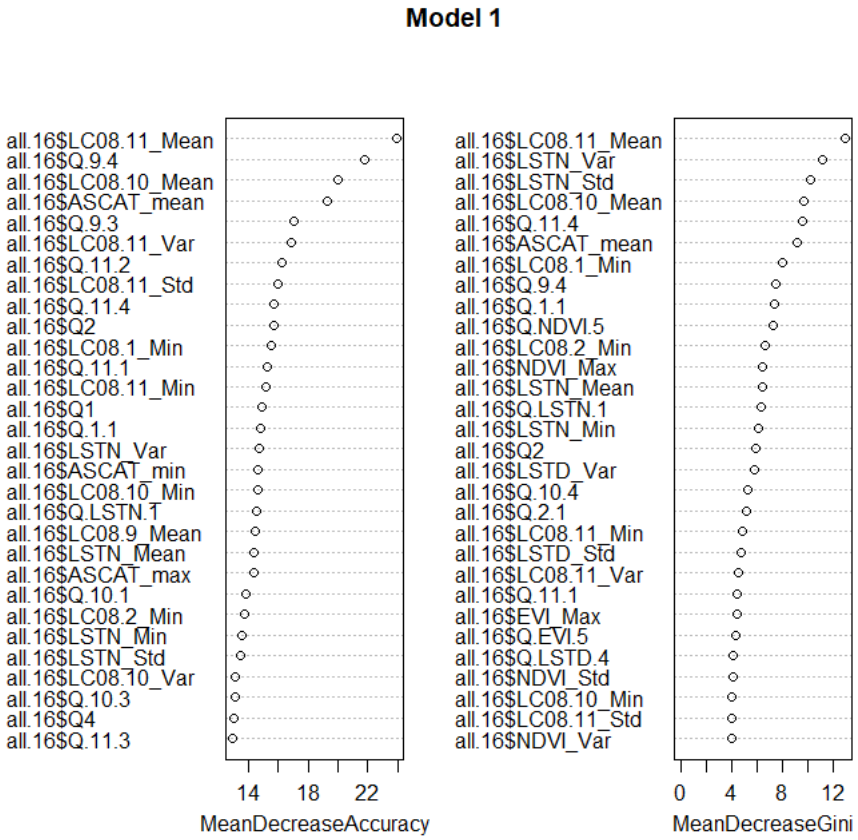


Figure 7.14 shows the 30 most important variables of model 1. According to the MDA, the most significant variables for the classification in suitable and non-suitable habitats is the mean of the at-satellite Brightness Temperature values of the LANDSAT-8 Thermal Infrared (TIRS) 2 time series stack (all.16$LC08.11_Mean). Referring to several studies (chapter 3.1., table 3.2.) temperature is known to be an essential variable for defining Ae. Species habitats. The second vital variables, which is surprising, is to forth quantile of the ninth LANDSAT-8 band (all.16$Q.9.4.), the Cirrus band. This may be traced to the fact, that several LANDSAT-8 scenes were characterized by moderate cloud coverage. If the training points are located under cloud-covered areas, the random forest might take the information of the Cirrus band as an important variable. Further, the mean at-satellite Brightness Temperature values of the LANDSAT-8 TIRS 1 time series stack (all.16$LC08.10_Mean) and the mean soil moisture content of the ASCAT sensor (all.16$ASCAT_mean) are considered essential for the classification process. Similar to the TIRS 2 band, TIRS 1 (all.16$LC08.10_Mean) includes information about temperature and is, therefore, a significant variable. The soil moisture content is also considered as important variable referring to the defined environmental indicators (chapter 3.1, table 3.2.). Further, the third quantile of the Cirrus band (all.16$Q.9.3.) is vital for the classification process (cloud coverage). The variance of LANDSAT-8 TIRS 2 (all.16$LC08.11_Var), the second quantile of the LANDSAT-8 TIRS 2 (all.16$Q.11.2), the standard deviation of the LANDSAT-8 TIRS 2 (all.16$LC08.11_Std) and the forth quantile of the LANDSAT-8 TIRS 2 (all.16$Q.11.4) are considered significant, all capturing information about temperature. Beyond that, the first (all.16$Q1), second (all.16$Q2), and forth quantile (all.16$Q4) of the ASCAT soil moisture is significant for model 1. Besides, the Aerosol band (band 1) seems to have an influence on the classification results (all.16$LC08.1_Min, all.16$Q1.1), which is also surprising as no previous publication defines the LANDSAT-8 Aerosol band as important. In addition to LANDSAT-8 and ASCAT derived metrics, the MODIS night land surface temperature (all.16$LSTN_Var, all.16$LSTN_Mean, all.16$LSTN_Min and all.15$LSTN_Std) seems to capture significant information.

Concluding that for model 1, when using all available variables, the most critical variables are the temperature from LANDSAT-8 and MODIS, the Cirrus and Aerosol band from LANDSAT-8 and the soil moisture derived

from the ASCAT sensor.

Model 2

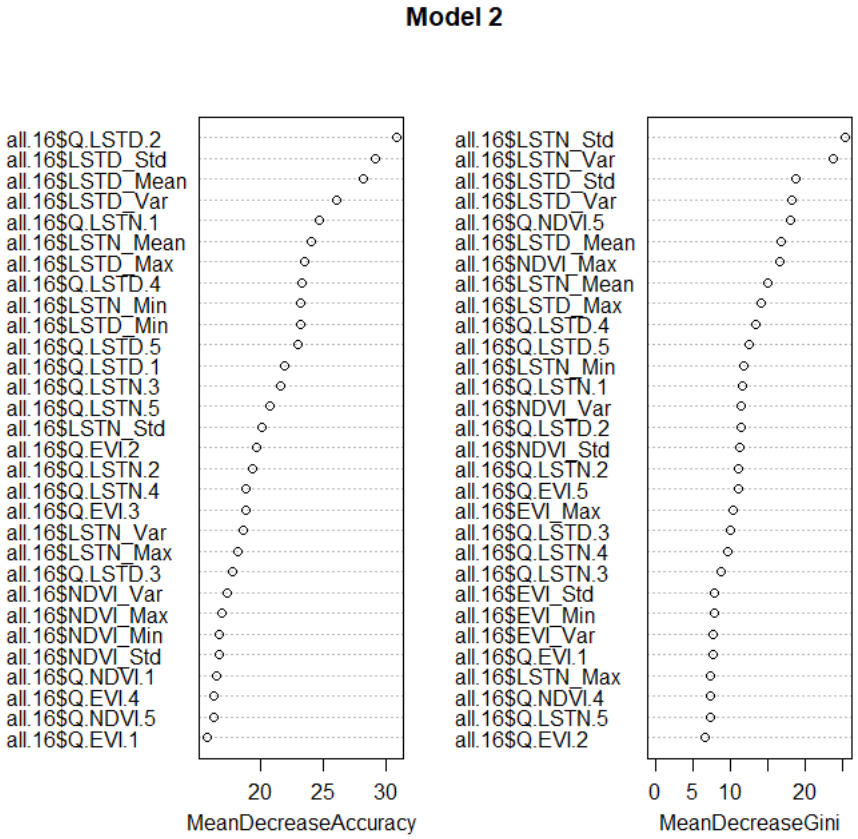
Model 2 was created to the test different products from the MODIS Terra and Aqua sensors for predicting habitats of mosquitoes.

The second model uses variables from all acquired MODIS products including land surface temperature (LST), vegetation indices (VI), leaf area index (LAI) and the fraction of photosynthetically active radiation (FPAR). The model was able to successfully predict 157 suitable test areas, resulting in an accuracy of 57.51 %. However, 116 test areas were classified wrongly. Considering the prediction of non-suitable areas 24 areas were mispredicted eventuating in a classification accuracy of 99.44% (class error 0.56%).

Table 7.13: Accuracy Model 2

	Non-suitable	Suitable	Class Error
Non-suitable	4253	24	0.0056
Suitable	116	157	0.4249

Figure 7.15: Variable Importance of Model 2



The most significant MODIS variables were metrics from the day land surface temperature (all.16$Q.LSTD.2, all.16$LSTD_Std, all.16$LSTD_mean, all.16$LSTD_Var, all.16$LSTD_Max, all.16$Q.LSTD.4, and all.16$LSTD_Min), which has as already been mentioned with model 1 and which can be traced back to several publications (chapter 3.1, table 3.2.). In general, land surface temperature (LSTN and LSTD) seems to be more informative than metrics from the two vegetation indices NDVI and EVI. However, the LAI and the FPAR are considered not crucial for model 2.

In conclusion, for the available test sites, the land surface temperature

is more informative for the prediction than the vegetation indices, the leaf area index and the fraction of photosynthetically active radiation. Thus, in this case, temperature captures more information than vegetation.

Model 3

The third model, which was employed to test the use of SENTINEL-2 spectral bands, performed moderate with a classification accuracy of 57.51 %, resulting in 157 correctly predicted test sites. Considering the non-suitable classification accuracy, model 3 is characterized by the same performance as model 2, 99.44 % (class error 0.56%) test areas were predicted correctly (Table 7.14).

Table 7.14: Accuracy Model 3

	Non-suitable	Suitable	Class Error
Non-suitable	4253	24	0.0056
Suitable	116	157	0.4249

Figure 7.16: Variable Importance of Model 3

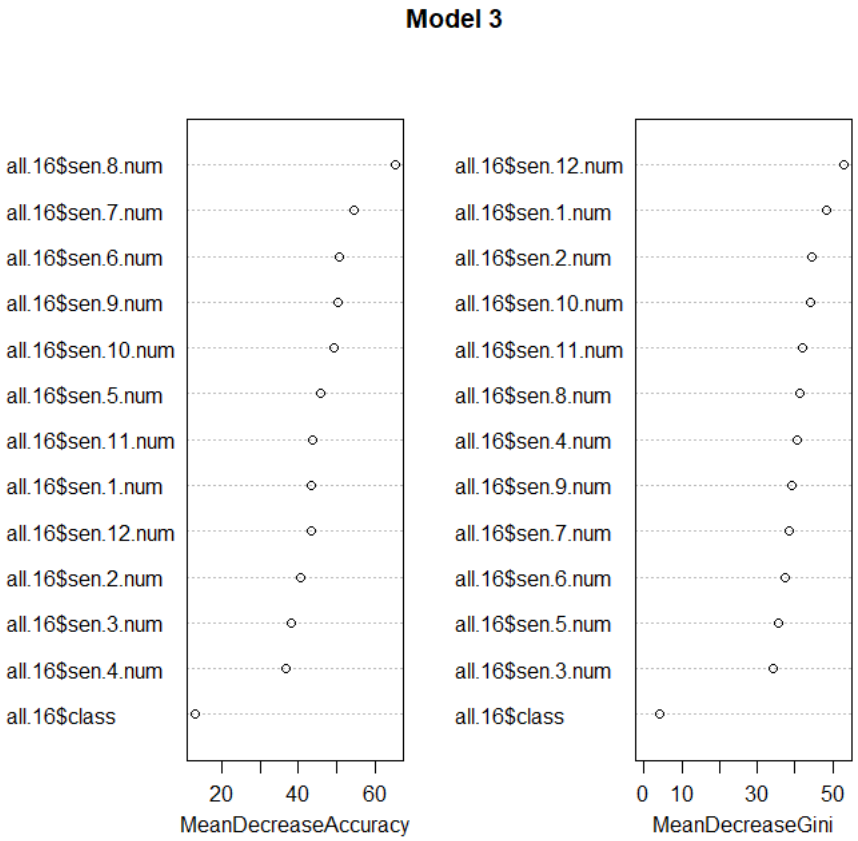


Figure 7.16 visualizes the importance of each SENTINEL-2 band for the prediction in suitable and non-suitable areas. The most significant spectral band is the near infrared band B8 (all.16$sen.8.num), which is used to investigate biomass content (Barsi et al., 2014).

Additionally, to the near-infrared spectral band, the vegetation red-edge bands B7, B6 and B5 (all.16$sen.7.num, all.16$sen.6.num, all.16$sen.5.num) are considered as essential variables for the classification process. Further, the Cirrus band B10 and both SWIR bands B11 and B12 are influencing the classification. The influence of Cirrus band may be based on the same

assumption as the influence of the LANDSAT-8 Cirrus band. Moderate remaining clouds (Figure 7.2) are probably influencing the classification. A combination of different bands or the implementation of vegetation indices might have been more informative than single bands. Further, the use of time series would probably give more insights for predicting habitats.

In conclusion, for the present research, the scene from the SENTINEL-2 sensor is limited revealing. However, testing band combination or the use of several scenes would be more revelatory.

Model 4

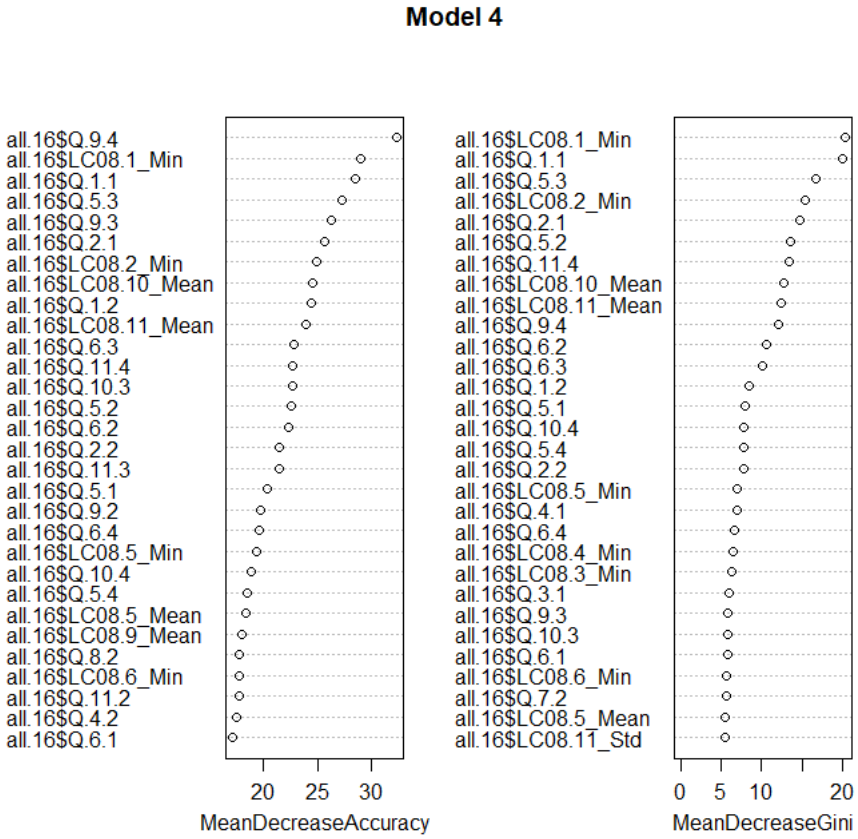
Model 4 was implemented to test the use of time series metrics of the 11 LANDSAT-8 spectral bands. Several studies investigated the use of LANDSAT data for mapping mosquito habitats (chapter 3.1.). Therefore, for present thesis, LANDSAT-8 data was chosen to examine the possibility of predicting potential habitats.

Regarding class suitable, about the half of the test areas (49.82%) were classified correctly using the random forest algorithm. 50.18% were classified wrongly, leading to a moderate performance of model 4. The class error of the non-suitable class is again low at 0.51%, resulting in a successful prediction of non-suitable habitats for the abundances of Ae. Species.

Table 7.15: Accuracy Model 4

	Non-suitable	Suitable	Class Error
Non-suitable	4255	22	0.0051
Suitable	137	136	0.5018

Figure 7.17: Variable Importance of Model 4



The 4. Quantile of the LANDSAT-8 Cirrus band (all.16%Q.9.4) is the most critical variable for model 4. As already mentioned within the discussion of model 1, it is assumed that some of the random forest test sites are located below cloud covered areas. Further, the Aerosol band B1 (all.16$LC08.1_Min, all.16$Q.1.1) is again significant for the prediction. Based on the literature review (chapter 3.1., chapter 3.2.), the LANDSAT-8 Aerosol band was never mentioned to be an essential indicator for the determination of vector habitats or risk zones. However, B1 useful for identifying water and aerosols in the blue region, which could lead to the hypothesis that B1 might capture information about the LULC class 'water', which is an indicator for the

abundance of the Ae. Mosquito (Barsi et al., 2014).

Besides, according to the variable importance, the near infrared band B5 (all.16$Q.5.3) is considered as a significant variable. The infrared band is useful for mapping biomass content, which comprises information about vegetation density, which in turn is associated with the abundance of Ae. Species (Vezanni et al. 2005; Barsi et al., 2014). The blue band, which is employed for distinguishing soil from vegetation, is another crucial variable (all.16$Q1.1). According to Rey et al. (2006) who investigated LULC for predicting the abundance of Ae. Species concluded that soil is having a negative impact on the abundance of the mosquito. Thus, information about soil captured by LANDSAT-8 blue band may be revealing for the habitat prediction (Barsi et al., 2014).

Additionally, the mean of the two SWIR bands B10 and B11 are informative variables, for estimating soil moisture, which is known to be a significant indicator for mosquito habitat mapping (Barsi et al., 2014; Patz et al., 1998).

Noticeable when examining model 4 is that the most important metrics are those from the calculated quantiles, whereas the computed variances and standard deviations are not considered as significant as in model 1 (Figure 7.14). It can be assumed, when solely using LANDSAT-8 metrics instead of a variable combination of different remote sensing products, quantiles of spectral bands are more important than other statistical parameters.

Model 11

To test the importance of soil moisture derived from the active ASCAT sensor, a model was created using calculated soil moisture metrics.

The accuracy of predicting class suitable using the test polygons was remarkably improved. 221 test sites out of 273 were successfully predicted, resulting in an accuracy of 80.95%. Model 11 was the first model with accuracy for class suitable of more than 80%. However, the class error of class non-suitable increased by more than 0.1% compared to the other built

models, resulting in a class error of 0.65% (Table 7.16).

Table 7.16: Accuracy Model 11

	Non-suitable	Suitable	Class Error
Non-suitable	4249	28	0.0065
Suitable	52	221	0.1905

Figure 7.18: Variable Importance of Model 11

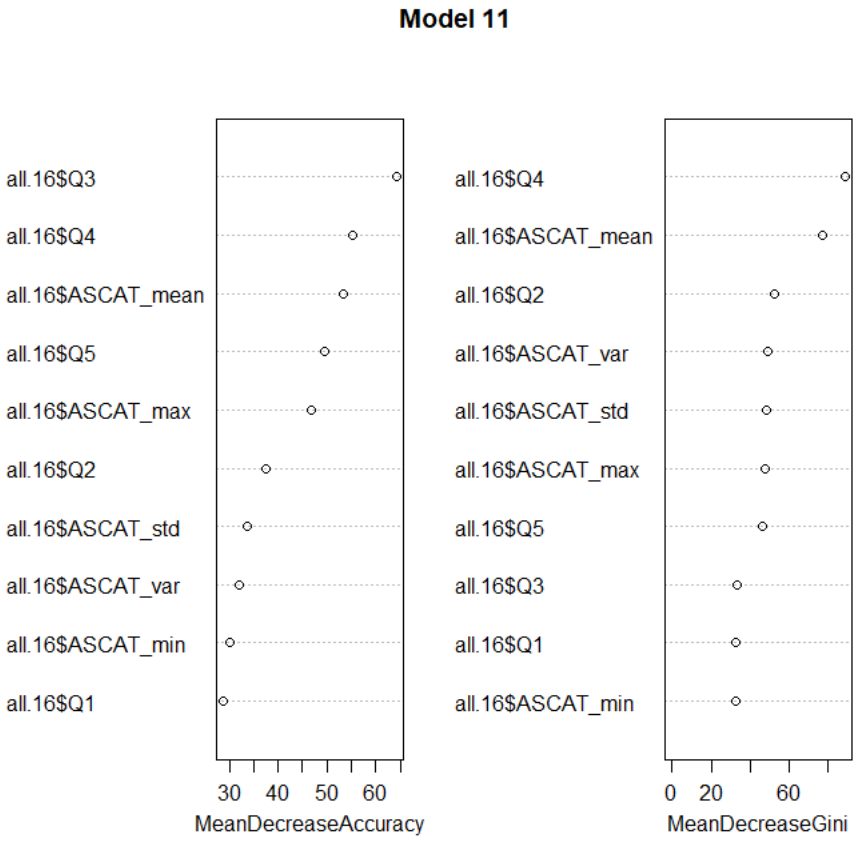


Figure 7.18 includes all calculated soil moisture metrics from the ASCAT sensor. The most critical variables are the third and fourth quantile of the

ASCAT time series stack. Further, the mean soil moisture content has a significant impact on the classification of habitats. Less significant are the standard deviation, the variance and the minimum parameters, similar to the model 4 (LANDSAT-8 time series metrics of spectral bands). Considering the study by Patz et al. (1998) soil moisture is known to be an essential indicator for predicting mosquitoes. The authors tested the use of soil moisture for predicting malaria transmission factors and concluded that soil moisture as indicators improves the prediction of biting rates compared to the usually employed precipitation. Compared to the present study, the soil moisture variables derived from ASCAT improved the classification result of suitable habitats by at least 10% (Table 7.11).

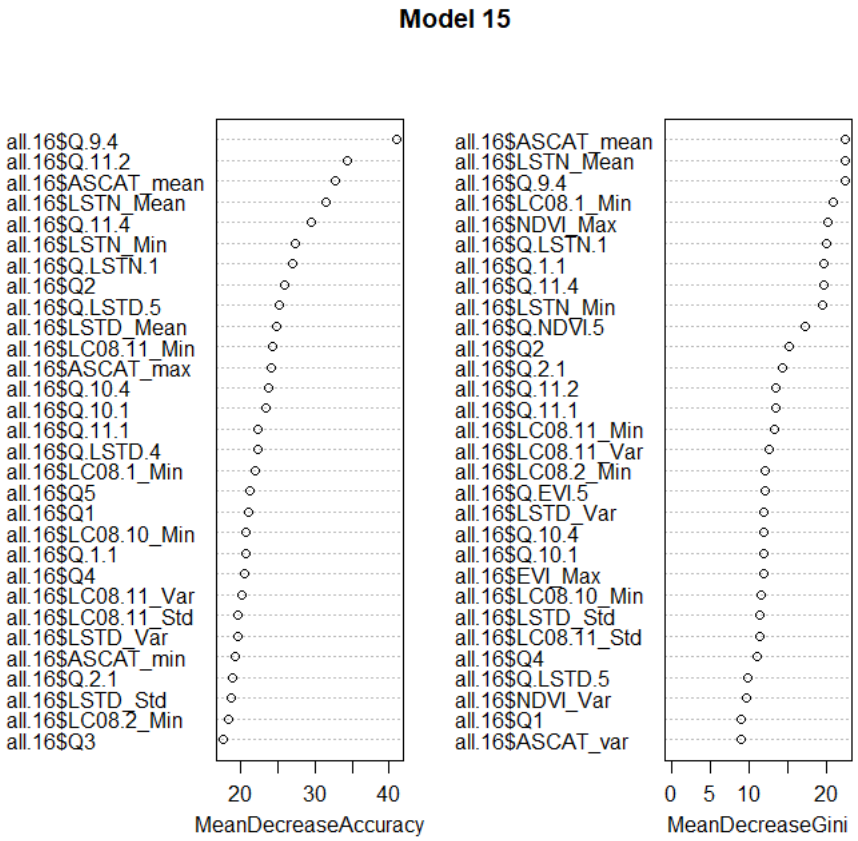
Model 15

Following the surprisingly good classification results of model 11, a combination of ASCAT and a subset of other relevant variables were tested. The subset was defined by investigating the importance of all variables from model 1. The variables with a Mean Decrease Gini more than 3, were taken as input variables, which resulted in a decreased accuracy compared to model 11 of 67.77%. Nevertheless, model 15 performs better than most of the other compiled models (Table 7.11). The classification accuracy of class non-suitable slightly improved again eventuating in a class error of 0.47% including 20 wrongly predicted test areas.

Table 7.17: Accuracy Model 15

	Non-suitable	Suitable	Class Error
Non-suitable	4257	20	0.0047
Suitable	88	185	0.3223

Figure 7.19: Variable Importance of Model 15



Considering the variable importance of model 15 (Figure 7.19) the most critical variables are those derived from the LANDSAT-8 time series, the Cirrus band (all.16$Q.9.4) and the SWIR 2 band (all.16$Q11.2). The importance of the Cirrus band might be again ascribed to the hypothesis of the cloud coverage, and the SWIR 2 captures information soil moisture, which is as already mentioned, a significant variable (Barsi et al., 2014) Beyond that the mean of the ASCAT (all.16$ASCAT_mean) time series and the variance of night land surface temperature (all.16$LSTN_Mean) are vital for the prediction of potential habitats. The land surface temperature, which was already identified as relevant variables within model 2 (Figure 7.15), has a

significant impact again, yet in this case the night land surface temperature. The ASCAT data seems to be less critical when applying a combination of ASCAT and a subset of all variables, which leads to the hypothesis that a combination of several data sources might decrease the classification results. Therefore, another model was tested (Figure 7.20) to confirm the hypothesis.

Model 16

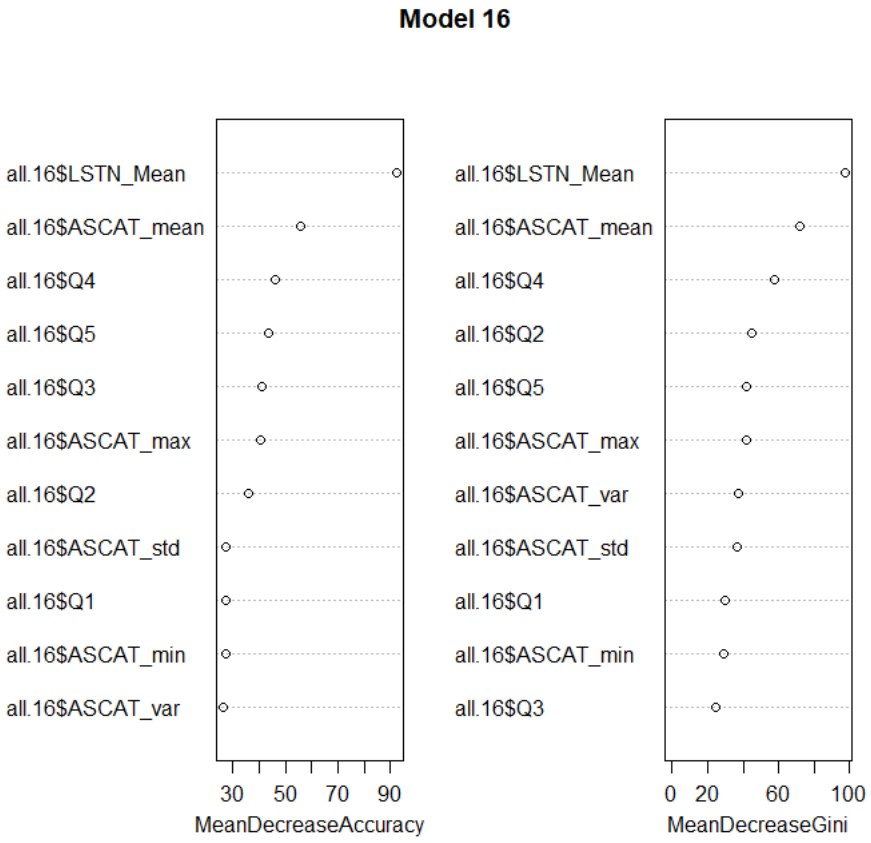
Model 16 was implemented to explore the combination of ASCAT soil moisture and MODIS temperature and to investigate the previous hypothesis from model 15. Therefore, the importance of ASCAT soil moisture data and the mean of the night land surface temperature, which was one of the most critical variable for the classification process, were used as input variables.

The last model resulted in the best classification results for class suitable. 231 test areas were classified successfully using the combination of temperature and soil moisture. 42 test sites were classified wrongly though, which may be affiliated to missing variables, resulting in a classification accuracy of 84.62%. Considering the classification result of the class non-suitable, 4257 out of 4277 test sites were classified adequately. Thus, the classification accuracy is high with 99.53% (Table 7.18).

Table 7.18: Accuracy Model 16

	Non-suitable	Suitable	Class Error
Non-suitable	4257	20	0.0047
Suitable	42	231	0.1538

Figure 7.20: Variable Importance of Model 16

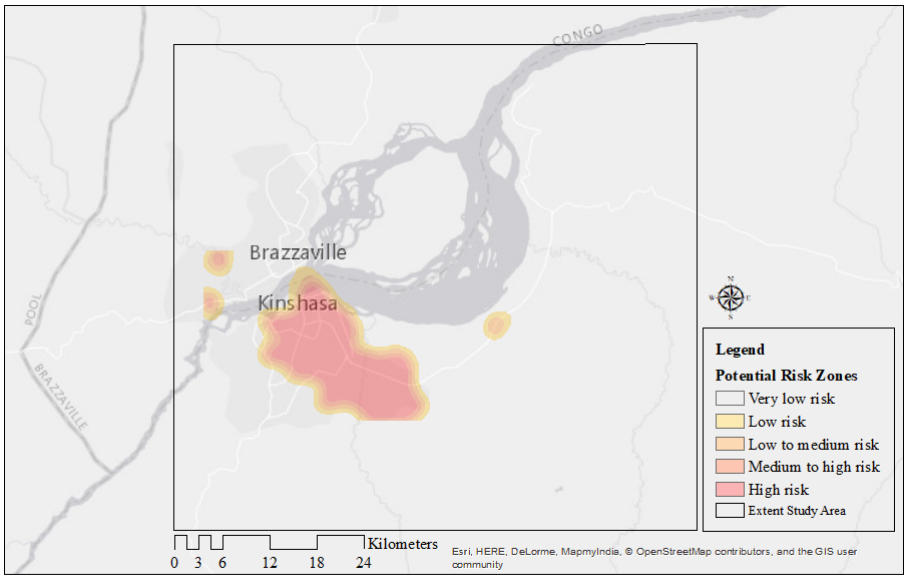


Here the most important variable is the mean of the night land surface temperature (all.16$LSTN_Mean) derived from MODIS. Further, the mean of the soil moisture (all.16$ASCAT_mean) and the third, fourth and fifth quantiles (all.16$Q3, all.16$Q4, all.16$Q5) are considered essential. Various studies examined the use of temperature (chapter 3.1.) to investigate mosquito habitats and concluded the divers' temperature metrics are meaningful variables. In this case, the mean of the night land surface temperature is the most important variable for the classification. Considering the ASCAT metrics, especially quantiles are highly ranked within the mean decrease accuracy, meaning that the quantiles are capturing significant information. This might

be affiliated to the fact the fourth and fifth quantiles are containing values of high soil moisture content.

In conclusion, a combination of soil moisture and land surface temperature is eminently suitable for predicting the available test sites (Figure 7.21).

Figure 7.21: Potential risk zones



Editor: Jasmine Planitzer
Datasource: MSF 2016, Kraemer et al., 2016, ESRI 2017;

7.4.3 Plausibility Check using CCI land cover

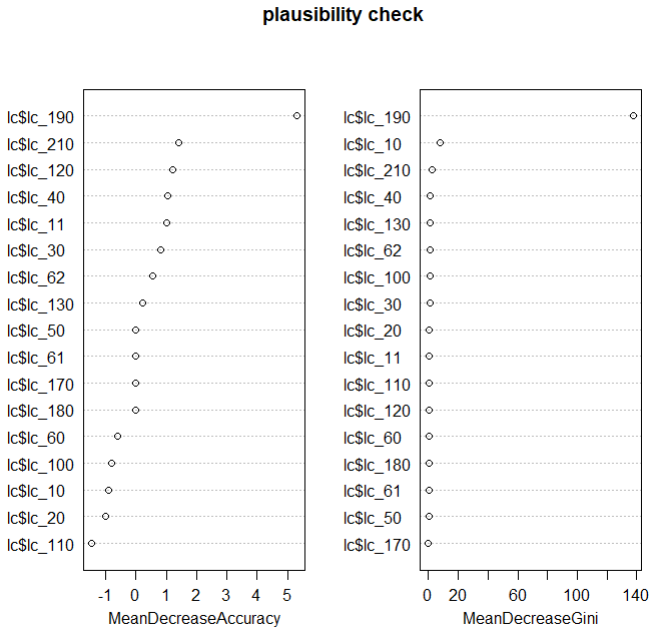
To assess the plausibility of the classification results and further investigate the use of CCI land cover data a plausibility check was applied using the training data and the LULC classes.

After the preprocessing, which included the resampling and cropping, the CCI land cover classes were split into single variables, resulting in 17 land cover classes (Figure 7.22). Each variable was binary coded (0/1) using the ArcGIS attribute query function and further a random forest model was applied. The class error is with both, suitable and non-suitable, compared to the classification results of the applied models (Table 7.10), high. Considering class suitable, 105 training areas were classified successfully, whereas 168 training areas were not classified correctly, resulting in a class error of 61.54 %, thus an accuracy of 38.46 % was reached. Regarding class non-suitable, 140 training areas were classified wrongly, which leads to a class error of 3.27 %. Consequently, the CCI land cover classes are not suitable for determining habitats and YF risk zones. The information loss (Table 7.7), which arose from the resampling during the preprocessing of the land cover data, might be responsible for the poor results of the applied random forest (Table 7.19).

Table 7.19: Accuracy Plausibility Check

	Non-suitable	Suitable	Class error
Non-suitable	4137	140	3.27
Suitable	168	105	61.54

Figure 7.22: Variable Importance Plausibility Check



However, considering the variable importance (Figure 7.22), the class urban (190) and further the class water (210) are having a significant impact for describing the habitats and risk zones. Several studies (chapter 3.1) investigated the use of LULC classes for describing habitats and concluded that both classes, urban and water are significant indicators.

Figure 7.23: Plausibility Check

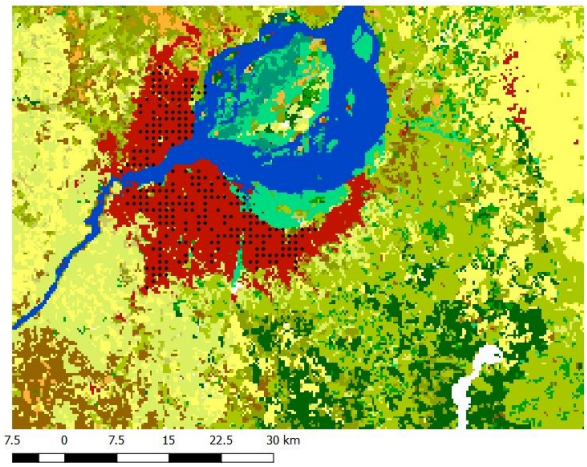
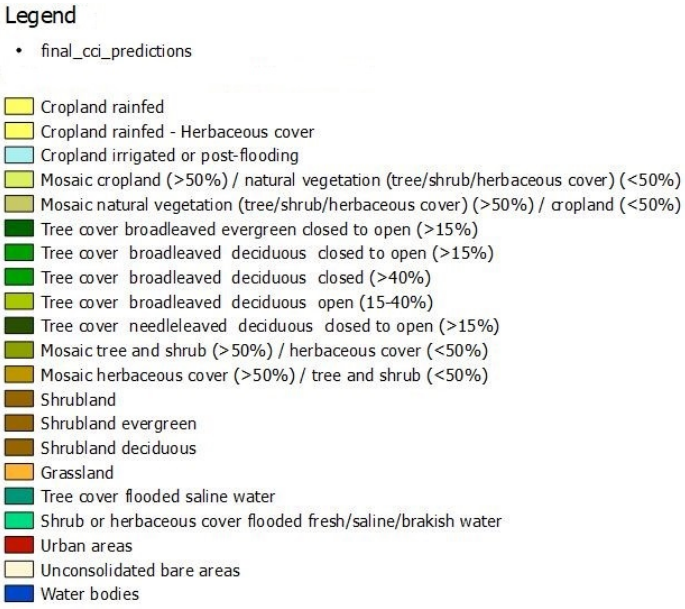


Figure 7.24: CCI Land Cover Classes Legend



Almost all predicted risk zones are located within the class urban (Figure 7.23). Considering the plausibility of the classification results soil moisture, which appeared as important variable, might be connected to both classes urban and water bodies. Regarding the class urban, impervious surfaces and artificial structures such as roads, buildings or parking lots are characterized by low soil moisture contents since asphalt and concrete do not accumulate water. Thus, low soil moisture content can be expected within those areas, which means a low soil moisture content might be a significant indicator for the abundance of the Ae. species and therefore for the risk of YF occurrence. While, water bodies and flooded areas are characterized by high soil moisture contents. Consequently, it is assumed that especially high and low soil moisture is an important variable for the habitat description. Further, land surface temperature is known to be a significant indicator for describing habitats and risk zones (chapter 3.3). Again, connecting the classes urban and water to surface temperature, impervious surfaces are characterized by relatively high land surface temperature (urban heat islands), whereas water bodies are characterized by low land surface temperature. Thus, it can be assumed that both indicators, soil moisture content and land surface temperature can be connected to the classes urban and water. Therefore, additional research should be conducted to assess the connection between LULC, soil moisture and temperature for investigating Ae. species habitats and the risk of the occurrence of YF fever within urban areas.

7.5 Discussion

The models implemented in this thesis are used to identify areas with high predictive probability for the abundance of Ae. Species using environmental indicators derived from remote sensing data. Therefore, multiple environmental indicators, which can be detected using remote sensing imagery were identified including temperature, precipitation, vegetation indices, land cover and soil moisture and compared with other studies. The results of the models are partly similar to those from previous studies (chapter 3.1.) including the importance of land surface temperature and soil moisture, whereas vegetation indices were considered less important. Although the plausibility check, which was implemented to test the plausibility of the results and further examine the use of LULC classes for predicting suitable habitats, resulted in moderate classification results, it confirmed the importance of the classes urban and water.

However, several findings provide incentives for further investigation. First, the importance of the Cirrus band from both LANDSAT-8 and SENTINEL-2 sensors, which has not been considered significant in other recent studies, should be further discussed. The fundamental hypothesis behind the significance of the Cirrus band is that partly cloud covered scenes were acquired and test sites are located below cloud covered pixels. An expedient approach would have been to conduct a LULC classification using the classes water, vegetation, clouds, cloud shadows and urban areas on both, the LANDSAT-8 and SENTINEL-2 scene. Following this approach, a cloud mask could have been created and employed to confirm the hypothesis (Huang et al., 2010).

Other possible error sources, which should be discussed in this section, is the data preprocessing. A comprehensive preprocessing was necessary due to the use of data from multiple sensors and sources. The use of nearest neighbor resampling for aggregating the LANDSAT-8 scenes and CCI land cover map might have resulted in position errors, especially along linear features (Santosh and Renuka, 2010). Referring to the resampling of the CCI land cover map (chapter 7.1.) a severe information loss regarding several land cover classes appeared, which might be a reason for the moderate classification results within the plausibility check. Further, the interpolation

of the ASCAT soil moisture data, which was conducted by TU Vienna might have caused errors, as the spatial resolution of the original data is 12.5*12.5 km.

Another aspect that should be considered is that the mean distance to water bodies is not significant. Although the classification results of the MLC of water bodies was successful, the mean distances are not vital for the prediction of potential habitats. This could be explained by the fact that the Ae. Aegypti mosquito, which is responsible for the transmission of YF in urban areas in Africa, favors artificial breeding containers such as cans or tires that collect rainfall water and not natural water habitats such as rivers or ponds (Gubler and Clark, 1995; Vainio and Cutts, 1998). To overcome this issue, VHR remote sensing data could be applied to detect smaller objects. However, VHR data is barely available and accessible, especially in developing countries.

Further, the use of the random forest ensemble learning algorithm for estimating the importance of the various environmental indicators may have introduced errors. One major disadvantage that needs to be discussed is that the random forest is a "black-box" classifier, which means relationships between predictive variables and the response may not be detected. Referring to the random forest use in this study, 1500 trees were built for each model and investigating 1500 trees for each model is virtually impossible. However, to classification process allows examining the importance of each variable (Wiesmeier et al., 2011). Moreover, the classification accuracy depends on the user-defined selection of variables. Thus, it is certain that implemented 16 models did not detect the optimal variable combination. Beyond that, the classification accuracy also depends on the training data, which gives rise to another error source, the training data. Since training data was only available to a limited extent as map provided by MSF, the preparation of test sites was problematic. The sample points are not well allocated over the whole study area. Consequently, the training data set lacks of test sites within sub-urban and rural areas, where YF outbreaks emerge (chapter 5.6.). Further, the defined possible infection zones around test sites were selected empirically. Field surveys and expert knowledge of epidemiologists would be necessary to define convincing infection zones.

However, the random forest ensemble learning also has several advantages. The random forest is one of the most accurate machine learning algorithms, not only for remote sensing applications but also for other research fields. Further, it is suitable for large data sets using numerous variables. For this research, the random forest was implemented since many variables from various data sources were tested. Besides, it estimates the importance of each variable. Therefore, it was possible to predict the significance of each environmental indicator for each classification process. The random forest can handle missing data, which was especially important when using land surface temperature from MODIS data. Some scenes were characterized by missing values, which might be caused when the scene is not acquired under clear-sky conditions (Jha, 2012).

The initial motivation behind conducting research in epidemiology using environmental indicators was to use the results determined from the implemented models to support doctors and national health authorities. However, since the results of most models give little satisfaction, the use of the approach in medical studies may be limited. Though model 11 using soil moisture and model 16 using soil moisture and night land surface temperature resulted in a classification accuracy of more than 80 %, which means assuming that error sources are limited to preprocessing and considering the lack of validation data, the majority of potential Ae. mosquito habitats were classified successfully. Nevertheless, it must be admitted that the abundance of the vector does not necessarily involve a YF outbreak since several indicators beside environmental must be considered (chapter 3.3., table 3.1.). Further, to field-test the approach during outbreaks higher degree of automatization would be necessary to save precious time.

Regarding the credibility and reliability of the results, the lack of validation data needs to be discussed. To validate the classification results independently, validation data is necessary. For the present thesis, validation data was not available, due to restriction from MSF and the health ministry of the Democratic Republic of Congo (DRC). Thus, the classification was not validated. Consequently, the results are solely based on the training data and might be therefore not suitable for health studies and in field use.

Further, the applicability of the proposed approach on other test sites and

with other VBDs should be illuminated. Using the approach on other test sites is generally possible since the employed remote sensing data is globally available. It should be considered though, that training data is needed to train the random forest ensemble learning algorithm. Besides, as already mentioned validation would be necessary to verify the classification result. Beyond that applying the approach on other test sites is only reasonable when both, the vector and host (non-human primates), are present, which applies solely in endemic or partly endemic countries (WHO, 2016a). Regarding the application of the approach on other VBDs, it may be applied to diseases which are also transmitted via the Ae. mosquito species including dengue fever, chikungunya or the zika virus (ECDC, 2016).

8 Conclusion

This thesis introduced a methodology for investigating the relation between environmental characteristics derived from satellite remote sensing imagery and the risk of yellow fever occurrence in Africa. The approach investigated the use of a combination of active and passive remote sensing products including imagery from SENTINEL-2, LANDSAT-8, MODIS Terra and Aqua, and ASCAT. Further, a plausibility check was employed using CCI land cover classes, to cross-check the classification results. The proposed method was based on data level machine learning and avoids cost-extensive in-situ surveys.

Environmental indicators that influence the presence of the yellow fever vector and thus the potential occurrence of yellow fever were assessed. Based on the assessment of possible influencing environmental indicators, further environmental indicators were defined which can be derived from remote sensing data. The results confirmed the use of environmental indicators for predicting potential vector habitats. Applying soil moisture content derived from ASCAT and land surface temperature derived from MODIS resulted in acceptable accuracies, while other indicators including vegetation indices, mean distance to water bodies, metrics from spectral bands from different sensors, the leaf area index and the fraction of photosynthetically active radiation are less significant. For the present thesis, it is not possible to infer a yellow fever virus risk hazard index, since the random forest classification does not allow statements about environmental conditions which need to be fulfilled to favor the abundance of the Ae. species and the possible outbreak of the disease. Thus, it is not possible to define for example the threshold of soil moisture content that needs to be reached to facilitate the development of habitats and therefore the abundance of the Ae. species.

The real distribution of the mosquito vector is influenced by a variety

of indicators though, not just those investigated in the present study. Beyond that the presence of the Ae. species does not necessarily introduce disease outbreaks since numerous complex parameters are involved in the emerging of an epidemic. Nevertheless, the proposed method might not only help policymakers and stakeholders but also might support health ministries. It is therefore anticipated that the approach and results presented in this thesis play a part in contributing further research on yellow fever and its vector.

A limited amount of studies has included complex and long-term data. Thus, further assessment of predicted areas should consider time series data to infer future Ae. species habitats. These results combined with human exposure and other factors such as demographic change would be essential to identify potential risk zones and future outbreaks in endemic countries.

Bibliography

Anderson B. (2016). Packages 'weathermetrics'. CRAN. from https://cran.r-project.org/web/packages/weathermetrics/weathermetrics.pdf.

Arboleda, S., Jaramillo-O, N., & Peterson, A.T. (2012). Spatial and temporal dynamics of Aedes aegypti larval sites in Bello, Colombia. Journal of vector ecology : journal of the Society for Vector Ecology, 37, 37–48.

Bailey, S.C., & Linthicum, K.J. (1989). Satellite remote sensing: the newest technology for monitoring vector populations and predicting arbovirus outbreaks. Australia.

Barrett, A.D.T., & Higgs, S. (2007). Yellow fever. A disease that has yet to be conquered. Annual review of entomology, 52, 209–229.

Barsi, J., Lee, K., Kvaran, G., Markham, B., & Pedelty, J. (2014). The Spectral Response of the Landsat-8 Operational Land Imager. Remote Sensing, 6, 10232–10251.

Beck, L.R., Lobitz, B.M., & Wood, B.L. (2000). Remote sensing and human health. New sensors and new opportunities. Emerging infectious diseases, 6, 217–227.

Becker, R.A. (1988). The New S Language. https://www.rdocumentation.org/packages/base/versions/3.4.1/topics/apply.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing. A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31.

Benediktsson, J.A. (2015). Spectral-Spatial Classification of Hyperspectral Remote Sensing Images. Norwood: Artech House.

Bleyhl, B., Baumann, M., Griffiths, P., Heidelberg, A., Manvelyan, K., Radeloff, V.C., Zazanashvili, N., & Kuemmerle, T. (2017). Assessing landscape connectivity for large mammals in the Caucasus using Landsat 8 seasonal image composites. Remote Sensing of Environment, 193, 193–203.

Breiman, L., & Cuttler, A. (no date). Random Forest. Description. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_copyright.htm

Brown, H.E., Diuk-Wasser, M.A., Guan, Y., Caskey, S., & Fish, D. (2008). Comparison of three satellite sensors at three spatial scales to predict larval mosquito presence in Connecticut wetlands. Remote Sensing of Environment, 112, 2301–2308.

Buczak, A.L., Koshute, P.T., Babin, S.M., Feighner, B.H., & Lewis, S.H. (2012). A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC medical informatics and decision making, 12, 124.

Canadian Space Agency (2014). RADARSAT-1. from http://www.asc-csa.gc.ca/eng/satellites/radarsat1/default.asp

Castillo, J.A.A., Apan, A.A., Maraseni, T.N., & Salmo, S.G. (2017). Estimation and mapping of above-ground biomass of mangrove forests and their replacement land uses in the Philippines using Sentinel imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 134, 70–85.

Cheng, G., Han, J., & Lu, X. (2017). Remote Sensing Image Scene Classification. Benchmark and State of the Art. Proceedings of the IEEE, 105, 1865–1883.

Climate-Org (2017). Climate Data for Cities Worldwide. https://en.climate-data.org/

Cortijo, F.J., & Pérez de la Blanca, N. (1996). Image Classification using

non-parametric classifiers and contextual information. http://www.isprs.org/proceedings/xxxi/congress/part3/120_XXXI-part3.pdf

Defourny, P.e.a. (2017). Land Cover CCI. Product User Guide. http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf.

Diallo, D., Talla, C., Ba, Y., Dia, I., Sall, A.A., & Diallo, M. (2011). Temporal distribution and spatial pattern of abundance of the Rift Valley fever and West Nile fever vectors in Barkedji, Senegal. Journal of vector ecology : journal of the Society for Vector Ecology, 36, 426–436.

Didan, K. MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/modis/mod13a2.006.

Duro, D.C., Girard, J., King, D.J., Fahrig, L., Mitchell, S., Lindsay, K., & Tischendorf, L. (2014). Predicting species diversity in agricultural environments using Landsat TM imagery. Remote Sensing of Environment, 144, 214–225.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2. ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment, 120, 25–36.

Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W., & Li, X. (2016). Water Bodies' Mapping from Sentinel-2 Imagery with Modified Normalized Difference Water Index at 10-m Spatial Resolution Produced by Sharpening the SWIR Band. Remote Sensing, 8, 354.

EPA (2017). Environmental Indicators Gateway. What is an Environmental Indicator. https://web.archive.org/web/20110127060903/http://www.epa.gov/igateway/whatIndicator.html

ESA (2017). Sentinel 2 - Missions. https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

Espinosa, M., Weinberg, D., Rotela, C.H., Polop, F., Abril, M., Scavuzzo, C.M. (2016). Temporal Dynamics and Spatial Patterns of Aedes aegypti Breeding Sites, in the Context of a Dengue Control Program in Tartagal (Salta Province, Argentina). PLoS neglected tropical diseases, 10, e0004621.

ESRIN (2016). The use of remote sensing data to derive environmental indicators. http://documentacion.ideam.gov.co/openbiblio/bvirtual/017931/DocumentosIndicadores/Temasvarios/Docum14.pdf

Estallo, E.L., Benitez, E.M., Lanfri, M.A., Scavuzzo, C.M., & Almiron, W.R. (2016). MODIS Environmental Data to Assess Chikungunya, Dengue, and Zika Diseases Through Aedes (Stegomia) aegypti Oviposition Activity Estimation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9, 5461–5466.

EUMETSAT (2017). Monitoring Weather and Climate from Space. https://www.eumetsat.int/website/home/Satellites/CurrentSatellites/index.html.

European Centre for Disease Prevention and Control (2016). Aedes aegypti - Factsheet for experts. https://ecdc.europa.eu/en/disease-vectors/facts/mosquito-factsheets/aedes-aegypti.

Foody M.G. et al. (1992) Derivation and Applications of Probabilistic Measures of Class Membership from the Maximum-Likelihood Classification. https://www.asprs.org/wp-content/uploads/pers/1992journal/sep/1992_sep_1335-1341.pdf.

Fuller, D.O., Troyo, A., & Beier, J.C. (2009). El Niño Southern Oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica. Environmental research letters : ERL [Web site], 4, 140111–140118.

Gatti, A. & Bertolini, A. (2015). Sentinel-2 Products Specification Document.

Gatti, A. & Naud, C. (2017). Sentinel-2 Products Specification Document.

Gardner, M.W., & Dorling, S.R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment, 32, 2627–2636.

Gislason, P.O., Benediktsson, J.A., & Sveinsson, J.R. Random forest classification of multisource remote sensing and geographic data, 1049–1052.

Gislason, P.O., Benediktsson, J.A., & Sveinsson, J.R. (2006). Random Forests for land cover classification. Pattern Recognition Letters, 27, 294–300.

Gopal, S., & Woodcock, C. (1996). Remote sensing of forest change using artificial neural networks. IEEE Transactions on Geoscience and Remote Sensing, 34, 398–404.

Gubler, D.J., & Clark, G.G. (1996). Community involvement in the control of Aedes aegypti. Acta tropica, 61, 169–179.

Hagenlocher, M. (2014). Integrated spatial indicators for modeling, exploring and visualizing vulnerability to vector-borne diseases. Paris-Lodron University Salzburg. pp. 1 - 235.

Hammond, A., Adriaanse, A., Rodenburg, E:, Byrant, D., & Woodward, R. (1995). Environmental Indicators. A systematic approach to measuring and reporting on environmental policy performance in the context of sustainable development. `http://infohouse.p2ric.org/ref/30/29288.pdf`

Hay, S.I., Packer, M.J., & Rogers, D.J. (1997). Review article The impact of remote sensing on the study and control of invertebrate intermediate hosts and vectors for disease. International Journal of Remote Sensing, 18, 2899–2930.

Hay, S.I., Snow, R.W., & Rogers, D.J. (1998). From Predicting Mosquito Habitat to Malaria Seasons Using Remotely Sensed Data. Practice, Problems and Perspectives. Parasitology Today, 14, 306–313.

Hayden, M.H., Uejio, C.K., Walker, K., Ramberg, F., Moreno, R., Rosales, C., Gameros, M., Mearns, L.O., Zielinski-Gutierrez, E., & Janes, C.R. (2010).

Microclimate and human factors in the divergent ecology of Aedes aegypti along the Arizona, U.S./Sonora, MX border. EcoHealth, 7, 64–77.

Healey, S.P., Cohen, W.B., Yang, Z., Kenneth Brewer, C., Brooks, E.B., Gorelick, N., Hernandez, A.J., Huang, C., Joseph Hughes, M., Kennedy, R.E., Loveland, T.R., Moisen, G.G., Schroeder, T.A., Stehman, S.V., Vogelmann, J.E., Woodcock, C.E., Yang, L., & Zhu, Z. (2018). Mapping forest change using stacked generalization. An ensemble approach. Remote Sensing of Environment, 204, 717–728.

Herbreteau, V., Salem, G., Souris, M., Hugot, J.-P., & Gonzalez, J.-P. (2007). Thirty years of use and improvement of remote sensing, applied to epidemiology. From early promises to lasting frustration. Health & place, 13, 400–403.

Hijmans R.J.(2017) Package raster. Cran Project. `https://cran.r-project.org/web/packages/raster/raster.pdf`.

Han, H., Guo, X., & Yu, H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest, 219–224.

Huang, C., Davis, L.S., & Townshend, J.R.G. (2002). An assessment of support vector machines for land cover classification. International Journal of Remote Sensing, 23, 725–749.

Huang, C., Thomas, N., Goward, S.N., Masek, J.G., Zhu, Z., Townshend, J.R.G., & Vogelmann, J.E. (2010). Automated masking of cloud and cloud shadow for forest change analysis using Landsat images. International Journal of Remote Sensing, 31, 5449–5464.

Ibañez-Justicia, A., & Cianci, D. (2015). Modelling the spatial distribution of the nuisance mosquito species Anopheles plumbeus (Diptera. Culicidae) in the Netherlands. Parasites & vectors, 8, 258.

Intergraph (2013). Erdas field guide. `http://www.hexagon-solutions.com.cn/Libraries/Misc_Docs/ERDAS_FieldGuide_PDF_Intergraph_brand.sflb.pdf`.

IPCC (2014). Climate Change 2014 Synthesis Report. Summary for Policymakers. https://www.ipcc.ch/pdf/assessment-report/ar5/syr/AR5_SYR_FINAL_SPM.pdf

Irons J. R. Landsat science. Landsat 7. https://landsat.gsfc.nasa.gov/landsat-7/

Jacob, B.G., Shililu, J., Muturi, E.J., Mwangangi, J.M., Muriu, S.M., Funes, J., Githure, J., Regens, J.L., & Novak, R.J. (2006). Spatially targeting Culex quinquefasciatus aquatic habitats on modified land cover for implementing an Integrated Vector Management (IVM) program in three villages within the Mwea Rice Scheme, Kenya. International journal of health geographics, 5, 18.

Jentes, E.S., Poumerol, G., Gershman, M.D., Hill, D.R., Lemarchand, J., Lewis, R.F., Staples, J.E., Tomori, O., Wilder-Smith, A., & Monath, T.P. (2011). The revised global yellow fever risk map and recommendations for vaccination, 2010. Consensus of the Informal WHO Working Group on Geographic Risk for Yellow Fever. The Lancet Infectious Diseases, 11, 622–632.

Jha, V. (2012). Random Forest algorithm. http://amateurdatascientist.blogspot.co.at/2012/01/random-forest-algorithm.html

Kalluri, S., Gilruth, P., Rogers, D., & Szczur, M. (2007). Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques. A review. PLoS pathogens, 3, 1361–1371.

Kaplan, G., & Avdan, U. (2017). Object-based water body extraction model using Sentinel-2 satellite imagery. European Journal of Remote Sensing, 50, 137–143

Kaufman, Y.J., Tanré, D., Remer, L.A., Vermote, E.F., Chu, A., & Holben, B.N. (1997). Operational remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer. Journal of Geophysical Research: Atmospheres, 102, 17051–17067.

Knudby, A., LeDrew, E., & Brenning, A. (2010). Predictive mapping of

reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. Remote Sensing of Environment, 114, 1230–1241.

Kotchi I., & Brazeau S. (2016).). Earth Observation and Indicators Pertaining to Determinants of Health – an Approach Earth Observation and Indicators Pertaining to Determinants of Health – an Approach to Support Local Scale Characterization of Environmental Determinants of Vector-Borne Diseases. Prague.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift, 15, 259–263.

Kraemer, M.U.G., Sinka, M.E., Duda, K.A., Mylne, A.Q.N., Shearer, F.M., Barker, C.M., Moore, C.G., Carvalho, R.G., Coelho, G.E., van Bortel, W., Hendrickx, G., Schaffner, F., Elyazar, I.R.F., Teng, H.-J., Brady, O.J., Messina, J.P., Pigott, D.M., Scott, T.W., Smith, D.L., Wint, G.R.W., Golding, N., & Hay, S.I. (2015). The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. eLife, 4, e08347.

Kuhn, M. (2017). Package 'caret'. `https://cran.r-project.org/`.

Landau, K.I., & van Leeuwen, W.J.D. (2012). Fine scale spatial urban land cover factors associated with adult mosquito abundance and risk in Tucson, Arizona. Journal of vector ecology : journal of the Society for Vector Ecology, 37, 407–418.

Lary, D.J., Alavi, A.H., Gandomi, A.H., & Walker, A.L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7, 3–10.

Li, W., Cao, S., Campos-Vargas, C., & Sanchez-Azofeifa, A. (2017). Identifying tropical dry forests extent and succession via the use of machine learning techniques. International Journal of Applied Earth Observation and Geoinformation, 63, 196–205.

Linthicum, K., Bailey, C., Davies, F., & Tucker, C. (1987). Detection of Rift

Valley fever viral activity in Kenya by satellite remote sensing imagery. Science, 235, 1656–1659.

Lockaby, G., Noori, N., Morse, W., Zipperer, W., Kalin, L., Governo, R., Sawant, R., & Ricker, M. (2016). Climatic, ecological, and socioeconomic factors associated with West Nile virus incidence in Atlanta, Georgia, U.S.A. Journal of vector ecology : journal of the Society for Vector Ecology, 41, 232–243.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing, 28, 823–870. Mack, B., Roscher, R., Stenzel, S., Feilhauer, H., Schmidtlein, S., & Waske, B. (2016). Mapping raised bogs with an iterative one-class classification approach. ISPRS Journal of Photogrammetry and Remote Sensing, 120, 53–64.

Martinuzzi, S., Vierling, L.A., Gould, W.A., Falkowski, M.J., Evans, J.S., Hudak, A.T., & Vierling, K.T. (2009). Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. Remote Sensing of Environment, 113, 2533–2546.

Mas, J.F., & Flores, J.J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. International Journal of Remote Sensing, 29, 617–663.

Masuoka, P.M., Claborn, D.M., Andre, R.G., Nigro, J., Gordon, S.W., Klein, T.A., & Kim, H.-C. (2003). Use of IKONOS and Landsat for malaria control in the Republic of Korea. Remote Sensing of Environment, 88, 187–194.

Messina, J.P., Kraemer, M.U., Brady, O.J., Pigott, D.M., Shearer, F.M., Weiss, D.J., Golding, N., Ruktanonchai, C.W., Gething, P.W., Cohn, E., Brownstein, J.S., Khan, K., Tatem, A.J., Jaenisch, T., Murray, C.J., Marinho, F., Scott, T.W., & Hay, S.I. (2016). Mapping global environmental suitability for Zika virus. eLife, 5.

Michaud, J.-S., Coops, N.C., Andrew, M.E., Wulder, M.A., Brown, G.S., & Rickbeil, G.J.M. (2014). Estimating moose (Alces alces) occurrence and

abundance from remotely derived environmental indicators. Remote Sensing of Environment, 152, 190–201.

Moreno-Madriñán, M., Crosson, W., Eisen, L., Estes, S., Estes Jr., M., Hayden, M., Hemmings, S., Irwin, D., Lozano-Fuentes, S., Monaghan, A., Quattrochi, D., Welsh-Rodriguez, C., & Zielinski-Gutierrez, E. (2014). Correlating Remote Sensing Data with the Abundance of Pupae of the Dengue Virus Mosquito Vector, Aedes aegypti, in Central Mexico. ISPRS International Journal of Geo-Information, 3, 732–749.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing. A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66, 247–259.

Mueller-Wilm, U. (2016). Sentinel-2 MSI – Level-2A Prototype Processor Installation and User Manual. `http://step.esa.int/thirdparties/sen2cor/2.2.1/S2PAD-VEGA-SUM-0001-2.2.pdf`.

Myueni, R. (2012). MODIS LAI/FPAR Product User's Guide. `https://lpdaac.usgs.gov/dataset_discovery/`.

Nakhapakorn, K., & Tripathi, N.K. (2005). An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. International journal of health geographics, 4, 13.

NASA (2004). Aster. Advanced Spaceborne Thermal Emission and Reflectance Radiometer. `https://asterweb.jpl.nasa.gov/instrument.asp`.

NASA (2017). MODIS – Moderate Resolution Imaging Spectroradiometer. `https://modis.gsfc.nasa.gov/index.php`.

O'Connell, J., Bradter, U., & Benton, T.G. (2015). Wide-area mapping of small-scale features in agricultural landscapes using airborne remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing, 109, 165–177.

Patz, J.A. (1998). Predicting key malaria transmission factors, biting and

entomological inoculation rates, using modelled soil moisture in Kenya. Tropical Medicine and International Health, 3, 818–827.

Paules, C.I., & Fauci, A.S. (2017). Yellow Fever - Once Again on the Radar Screen in the Americas. The New England journal of medicine, 376, 1397–1399.

Pope, K.O., Sheffner, E.J., Linthicum, K.J., Bailey, C.L., Logan, T.M., Kasischke, E.S., Birney, K., Njogu, A.R., & Roberts, C.R. (1992). Identification of central Kenyan Rift Valley Fever virus vector habitats with landsat TM and evaluation of their flooding status with airborne imaging radar. Remote Sensing of Environment, 40, 185–196.

Pope, K.O., Sheffner, E.J., Linthicum, K.J., Bailey, C.L., Logan, T.M., Kasischke, E.S., Birney, K., Njogu, A.R., & Roberts, C.R. (1992). Identification of central Kenyan Rift Valley Fever virus vector habitats with landsat TM and evaluation of their flooding status with airborne imaging radar. Remote Sensing of Environment, 40, 185–196.

Pouncey, R. (1999). ERDAS Field Guide. Fifth edition, revised and expanded. `http://web.pdx.edu/~emch/ip1/FieldGuide.pdf`.

Rebelo, A.J., Scheunders, P., Esler, K.J., & Meire, P. (2017). Detecting, mapping and classifying wetland fragments at a landscape scale. Remote Sensing Applications: Society and Environment, 8, 212–223.

Regis, L.N., Acioli, R.V., Silveira, J.C., Melo-Santos, M.A.V. de, da Cunha, M.C.S., Souza, F., Batista, C.A.V., Barbosa, R.M.R., Oliveira, C.M.F. de, Ayres, C.F.J., Monteiro, A.M.V., & Souza, W.V. (2014). Characterization of the spatial and temporal dynamics of the dengue vector population established in urban areas of Fernando de Noronha, a Brazilian oceanic island. Acta tropica, 137, 80–87.

Rejmankova, E., Rejmanek, M. Pitcairn, M.J. Washino, R.K (1998). Aquatic vegetation in rice fields as a habitat for Culex tarsalis and Anopheles freeborni.

Rejmankova, E., Polanco, J., Manguin, S., Woodruff, R., Roberts, D.R., Har-

bach, R.E., Paris, J.F., Legters, L.J., & Wullschleger, B. (1996). Predictions of Malaria Vector Distribution in Belize Based on Multispectral Satellite Data. The American Journal of Tropical Medicine and Hygiene, 54, 304–308.

Rejmankova, E., Roberts, D.R., Harbach, R.E., Pecor, J., Peyton, E.L., Manguin, S., Krieg, R., Polanco, J., & Legters, L. (1993). Environmental and Regional Determinants of Anopheles (Diptera. Culicidae) Larval Distribution in Belize, Central America. Environmental Entomology, 22, 978–992.

Rey, J.R., Nishimura, N., Wagner, B., Braks, M.A.H., O'Connell, S.M., & Lounibos, L.P. (2006). Habitat Segregation of Mosquito Arbovirus Vectors in South Florida. Journal of Medical Entomology, 43, 1134–1141.

Richards, J.A., & Jia, X. (1999). Remote sensing digital image analysis. An introduction. (3rd ed.). Berlin: Springer.

Richter, R., Louis, J., & Müller-Wilm, U. (2012). Sentinel-2 MSI - Level 2A Products Algorithm Theoretical Basis Document. ESRIN.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J.P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93–104.

Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. Remote Sensing of Environment, 55, 95–107.

Robel J., & Graumann A. (2014). NOAA KLM User's Guide with NOAA-N, N Prime, and MetOp SUPPLEMENTS, National Centres for Environmental Information.

Rotela, C., Fouque, F., Lamfri, M., Sabatier, P., Introini, V., Zaidenberg, M., & Scavuzzo, C. (2007). Space-time analysis of the dengue spreading dynamics in the 2004 Tartagal outbreak, Northern Argentina. Acta tropica, 103, 1–13.

Samson, D.M., Archer, R.S., Alimi, T.O., Arheart, K.L., Impoinvil, D.E.,

Oscar, R., Fuller, D.O., & Qualls, W.A. (2015). New baseline environmental assessment of mosquito ecology in northern Haiti during increased urbanization. Journal of vector ecology : journal of the Society for Vector Ecology, 40, 46–58.

Santhosh Baboo, S., & Renuka Devi, M. (2010). An Analysis of Different Remsampling Methods in Coimbatore, District.

Santos, J., & Meneses, B.M. (2017). An integrated approach for the assessment of the Aedes aegypti and Aedes albopictus global spatial distribution, and determination of the zones susceptible to the development of Zika virus. Acta tropica, 168, 80–90.

Satellite Imagine Corporation (2017). IKONOS Satellite Sensor. http://www.satimagingcorp.com/satellite-sensors/ikonos/.

Schuster, C., Schmidt, T., Conrad, C., Kleinschmit, B., & Förster, M. (2015). Grassland habitat mapping by intra-annual time series analysis – Comparison of RapidEye and TerraSAR-X satellite data. International Journal of Applied Earth Observation and Geoinformation, 34, 25–34.

Suhet (2013). Sentinel-2 User Handbook. ESA.

Shapiro M. R.stats - Generates area statistics for raster map. https://grass.osgeo.org/grass73/manuals/r.stats.html.

Tran, A., Ponçon, N., Toty, C., Linard, C., Guis, H., Ferré, J.-B., Lo Seen, D., Roger, F., La Rocque, S. de, Fontenille, D., & Baldet, T. (2008). Using remote sensing to map larval and adult populations of Anopheles hyrcanus (Diptera. Culicidae) a potential malaria vector in Southern France. International journal of health geographics, 7, 9.

Ugorji F.O., & Lartey A.O. Malaria surveillance with Remote Sensing and GIS in Owerri, Nigeria.

USGS (2011). Sensors – Hyperion. https://eo1.usgs.gov/sensors/hyperion.

USGS (2013). Landsat – Earth Observation Satellites. Fact Sheet 2015-3081.

USGS (2016). Landsat Collections. https://landsat.usgs.gov/.

USGS (2017). Product Guide. Land surface reflectance-derived spectral indices. https://landsat.usgs.gov/sites/default/files/.

Vainio J., & Cutts F. (1998). Yellow Fever. Division of Emerging and other Communicable Diseases Surveillance and Control. WHO. http://www.who.ch/gpv-documents/.

Vanwambeke, S.O., Bennett, S.N., & Kapan, D.D. (2011). Spatially disaggregated disease transmission risk. Land cover, land use and risk of dengue transmission on the island of Oahu. Tropical Medicine and International Health, 16, 174–185.

Vasconcelos, C.H., & Novo, E. Influence of precipitation, deforestation and Tucurui reservoir operation on malaria incidence rates in southeast Para, Brazil, 4567–4569.

Venkata, J. (2017). Random Forest Template for TIBCO Spotfire. https://community.tibco.com/wiki/

Vezzani, D., Rubio, A., Velázquez, S.M., Schweigmann, N., & Wiegand, T. (2005). Detailed assessment of microhabitat suitability for Aedes aegypti (Diptera. Culicidae) in Buenos Aires, Argentina. Acta tropica, 95, 123–131.

Vreugdenhil, M., Dorigo, W.A., Wagner, W., Jeu, R.A.M. de, Hahn, S., & van Marle, M.J.E. (2016). Analyzing the Vegetation Parameterization in the TU-Wien ASCAT Soil Moisture Retrieval. IEEE Transactions on Geoscience and Remote Sensing, 54, 3513–3531.

Wagner, V.E., Hill-Rowley R., Narlock S. A., & Newson H. D. (1979). Remote Sensing: A rapid and accurate method of data acquisition for a newly formed mosquito control distrcit. https://www.biodiversitylibrary.org/content/part/JAMCA/MN_V39_N2_P283-287.pdf.

Wan, Z.S.H. MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/modis/mod11a2.006.

Wang, C., Li, J., Liu, Q., Zhong, B., Wu, S., & Xia, C. (2017). Analysis of Differences in Phenology Extracted from the Enhanced Vegetation Index and the Leaf Area Index. Sensors (Basel, Switzerland), 17.

Weiss, M., Baret, F., Smith, G.J., Jonckheere, I., & Coppin, P. (2004). Review of methods for in situ leaf area index (LAI) determination. Agricultural and Forest Meteorology, 121, 37–53.

WHO (2014). WHO factsheet. Vector borne diseases. WHO. http://www.who.int/kobe_centre/mediacentre/vbdfactsheet.pdf.

WHO (2014). Risk Assessment on yellow fever virus circulation in enedmic countries. http://apps.who.int/iris/bitstream/10665/112751/1/WHO_HSE_PED_CED_2014.2_eng.pdf.

WHO (2016a). Yellow fever fact sheet. WHO. http://www.who.int/mediacentre/factsheets/fs100/en/.

WHO (2016b). Yellow fever situation report. WHO. http://apps.who.int/iris/bitstream/10665/250661/1/yellowfeversitrep28Oct16-eng.pdf?ua=1.

WHO (2017). The yellow fever outbreak in Angola and Democratic Republic of Congo ends.

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and Soil, 340, 7–24.

Wilke, A.B.B., Medeiros-Sousa, A.R., Ceretti-Junior, W., & Marrelli, M.T. (2017). Mosquito populations dynamics associated with climate variations. Acta tropica, 166, 343–350.

Williams, J.P., Hanavan, R.P., Rock, B.N., Minocha, S.C., & Linder, E. (2017). Low-level Adelges tsugae infestation detection in New England through partition modeling of Landsat data. Remote Sensing of Environment, 190, 13–25.

Wilsey, C.B., Lawler, J.J., & Cimprich, D.A. (2012). Performance of habitat suitability models for the endangered black-capped vireo built with remotely-sensed data. Remote Sensing of Environment, 119, 35–42.

Wood, B., & Muir, D. The Use Of Remote Sensing Data In Vector-borne Disease Control, 1221–1222.

Young, N.E., Anderson, R.S., Chignell, S.M., Vorster, A.G., Lawrence, R., & Evangelista, P.H. (2017). A survival guide to Landsat preprocessing. Ecology, 98, 920–932.

Zanter, K. Landsat 8 (OLI) Data Users Handbook.

Zeng, W., Cui, X., Liu, X., Cui, H., & Wang, P. Remote Sensing and GIS for Identifying and Monitoring the Environmental Factors Associated with Vector-borne Disease. An Overview, 1443–1446.

Zhang, C. (2015). Applying data fusion techniques for benthic habitat mapping and monitoring in a coral reef ecosystem. ISPRS Journal of Photogrammetry and Remote Sensing, 104, 213–223.

Zhang, Z., Ward, M., Gao, J., Wang, Z., Yao, B., Zhang, T., & Jiang, Q. (2013). Remote sensing and disease control in China. Past, present and future. Parasites & vectors, 6, 11.

Zhao, K., Popescu, S., Meng, X., Pang, Y., & Agca, M. (2011). Characterizing forest canopy structure with lidar composite metrics and machine learning. Remote Sensing of Environment, 115, 1978–1996.

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L., Samanta, A., Piao, S., Nemani, R., & Myneni, R. (2013). Global Data Sets of Vegetation Leaf Area Index (LAI)3g and Fraction of Photosynthetically Active Radiation (FPAR)3g Derived from Global Inventory Modeling and Mapping Studies

(GIMMS) Normalized Difference Vegetation Index (NDVI3g) for the Period 1981 to 2011. Remote Sensing, 5, 927–948.

Appendices

Appendix 1

Results and Accuracy Assessment

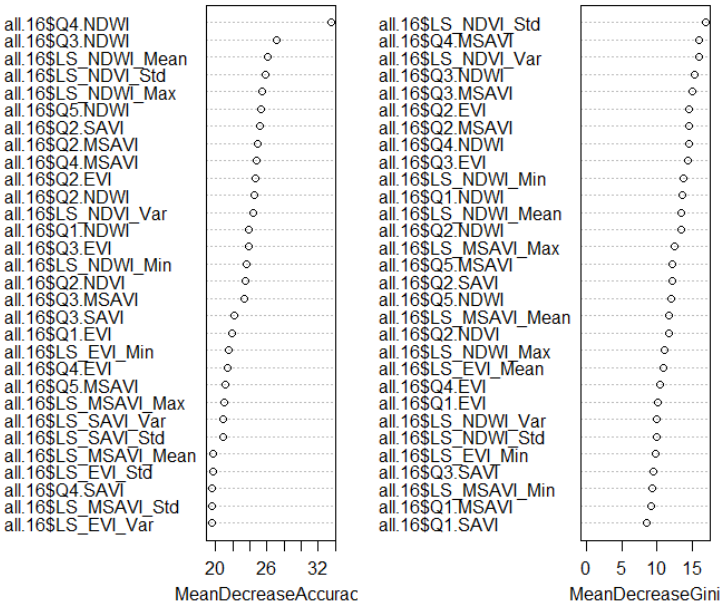
Variable	Description	Variable	Description
all.16$class	mean distance to water bodies	all.16$LC08.1_Max … all.16$LC08.11_Max	LANDSAT-8 B1 maximum … LANDSAT-8 B11 maximum
all.16$sen.1.num … all.16$sen.12.num	SENTINEL-2 B1 … SENTINEL-2 B12	all.16$Q.1.1 … all.16$Q11.1	LANDSAT-8 B1 1. quantile … LANDSAT-8 B11 1. quantile
all.16$cci_landcover	CCI land cover	all.16$Q.1.2 … all.16$Q11.2	LANDSAT-8 B1 2. quantile … LANDSAT-8 B11 2. quantile
all.16$EVI_Mean	MODIS EVI mean	all.16$Q.1.3 … all.16$Q11.3	LANDSAT-8 B1 3. quantile … LANDSAT-8 B11 3. quantile
all.16$EVI_Var	MODIS EVI variance	all.16$Q.1.4 … all.16$Q11.4	LANDSAT-8 B1 4. quantile … LANDSAT-8 B11 4. quantile
all.16$EVI_Std	MODIS EVI standard deviation	all.16$Q.1.5 … all.16$Q11.5	LANDSAT-8 B1 5.quantile … LANDSAT-8 B11 5. quantile
all.16$EVI_Min	MODIS EVI minimum	all.16$LS_EVI_Mean	LANDSAT-8 EVI mean
all.16$EVI_Max	MODIS EVI maximum	all.16$LS_EVI_Var	LANDSAT-8 EVI variance
all.16$Q.EVI.1	MODIS EVI 1. quantile	all.16$LS_EVI_Std	LANDSAT-8 EVI standard dev.
all.16$Q.EVI.2	MODIS EVI 2. quantile	all.16$LS_EVI_Min	LANDSAT-8 EVI minimum
all.16$Q.EVI.3	MODIS EVI 3. quantile	all.16$LS_EVI_Max	LANDSAT-8 EVI maximum
all.16$Q.EVI.4	MODIS EVI 4. quantile	all.16$Q1.EVI	LANDSAT-8 EVI 1. quantile
all.16$Q.EVI.5	MODIS EVI 5. quantile	all.16$Q2.EVI	LANDSAT-8 EVI 2. quantile
all.16$NDVI_Mean	MODIS NDVI mean	all.16$Q3.EVI	LANDSAT-8 EVI 3. quantile
all.16$NDVI_Var	MODIS NDVI variance	all.16$Q4.EVI	LANDSAT-8 EVI 4. quantile
all.16$NDVI_Std	MODIS NDVI standard dev.	all.16$Q5.EVI	LANDSAT-8 EVI 5. quantile
all.16$NDVI_Min	MODIS NDVI minimum	all.16$LS_NDVI_Mean	LANDSAT-8 NDVI mean
all.16$NDVI_Max	MODIS NDVI maximum	all.16$LS_NDVI_Var	LANDSAT-8 NDVI variance
all.16$Q.NDVI.1	MODIS NDVI 1. quantile	all.16$LS_NDVI_Std	LANDSAT-8 NDVI standard dev.
all.16$Q.NDVI.2	MODIS NDVI 2. quantile	all.16$LS_NDVI_Min	LANDSAT-8 NDVI minimum
all.16$Q.NDVI.3	MODIS NDVI 3. quantile	all.16$LS_NDVI_Max	LANDSAT-8 NDVI maximum
all.16$Q.NDVI.4	MODIS NDVI 4. quantile	all.16$Q1.NDVI	LANDSAT-8 NDVI 1. quantile
all.16$Q.NDVI.5	MODIS NDVI 5. quantile	all.16$Q2.NDVI	LANDSAT-8 NDVI 2. quantile
all.16$LSTD_Mean	MODIS LST day mean	all.16$Q3.NDVI	LANDSAT-8 NDVI 3. quantile

all.16$LSTD_Var	MODIS LST day variance	all.16$Q4.NDVI	LANDSAT-8 NDVI 4. quantile
all.16$LSTD_Std	MODIS LST day standard dev.	all.16$Q5.NDVI	LANDSAT-8 NDVI 5. quantile
all.16$LSTD_Min	MODIS LST day minimum	all.16$LS_SAVI_Mean	LANDSAT-8 SAVI mean
all.16$LSTD_Max	MODIS LST day maximum	all.16$LS_SAVI_Var	LANDSAT-8 SAVI variance
all.16$Q.LSTD.1	MODIS LST day 1. quantile	all.16$LS_SAVI_Std	LANDSAT-8 SAVI standard dev.
all.16$Q.LSTD.2	MODIS LST day 2. quantile	all.16$LS_SAVI_Min	LANDSAT-8 SAVI minimum
all.16$Q.LSTD.3	MODIS LST day 3. quantile	all.16$LS_SAVI_Max	LANDSAT-8 SAVI maximum
all.16$Q.LSTD.4	MODIS LST day 4. quantile	all.16$Q1.SAVI	LANDSAT-8 SAVI 1. quantile
all.16$Q.LSTD.5	MODIS LST day 5. quantile	all.16$Q2.SAVI	LANDSAT-8 SAVI 2. quantile
all.16$LSTN_Mean	MODIS LST night mean	all.16$Q3.SAVI	LANDSAT-8 SAVI 3. quantile
all.16$LSTN_Var	MODIS LST night variance	all.16$Q4.SAVI	LANDSAT-8 SAVI 4. quantile
all.16$LSTN_Std	MODIS LST night standard dev.	all.16$Q5.SAVI	LANDSAT-8 SAVI 5. quantile
all.16$LSTN_Min	MODIS LST night minimum	all.16$LS_MSAVI_Mean	LANDSAT-8 MSAVI mean
all.16$LSTN_Max	MODIS LST night maximum	all.16$LS_MSAVI_Var	LANDSAT-8 MSAVI variance
all.16$Q.LSTN.1	MODIS LST night 1. quantile	all.16$LS_MSAVI_Std	LANDSAT-8 MSAVI standard dev.
all.16$Q.LSTN.2	MODIS LST night 2. quantile	all.16$LS_MSAVI_Min	LANDSAT-8 MSAVI minimum
all.16$Q.LSTN.3	MODIS LST night 3. quantile	all.16$LS_MSAVI_Max	LANDSAT-8 MSAVI maximum
all.16$Q.LSTN.4	MODIS LST night 4. quantile	all.16$Q1.MSAVI	LANDSAT-8 MSAVI 1. quantile
all.16$Q.LSTN.5	MODIS LST night 5. quantile	all.16$Q2.MSAVI	LANDSAT-8 MSAVI 2. quantile
all.16$LAI_mean	MODIS LAI mean	all.16$Q3.MSAVI	LANDSAT-8 MSAVI 3. quantile
all.16$LAI_var	MODIS LAI variance	all.16$Q4.MSAVI	LANDSAT-8 MSAVI 4. quantile
all.16$LAI_std	MODIS LAI standard dev.	all.16$Q5.MSAVI	LANDSAT-8 MSAVI 5. quantile
all.16$LAI_min	MODIS LAI minimum	all.16$LS_NDWI_Mean	LANDSAT-8 NDWI mean
all.16$LAI_max	MODIS LAI maximum	all.16$LS_NDWI_Var	LANDSAT-8 NDWI variance
all.16$LAI.1	MODIS LAI 1. quantile	all.16$LS_NDWI_Std	LANDSAT-8 NDWI standard dev.
all.16$LAI.2	MODIS LAI 2. quantile	all.16$LS_NDWI_Min	LANDSAT-8 NDWI minimum
all.16$LAI.3	MODIS LAI 3. quantile	all.16$LS_NDWI_Max	LANDSAT-8 NDWI maximum
all.16$LAI.4	MODIS LAI 4. quantile	all.16$Q1.NDWI	LANDSAT-8 NDWI 1. quantile
all.16$LAI.5	MODIS LAI 5. quantile	all.16$Q2.NDWI	LANDSAT-8 NDWI 2. quantile

all.16$FPAR_mean	MODIS FPAR mean	all.16$Q3.NDWI	LANDSAT-8 NDWI 3. quantile
all.16$FPAR_var	MODIS FPAR variance	all.16$Q4.NDWI	LANDSAT-8 NDWI 4. quantile
all.16$FPAR_std	MODIS FPAR standard dev.	all.16$Q5.NDWI	LANDSAT-8 NDWI 5. quantile
all.16$FPAR_min	MODIS FPAR minimum	all.16$ASCAT_mean	ASCAT mean
all.16$FPAR_max	MODIS FPAR maximum	all.16$ASCAT_var	ASCAT variance
all.16$FPAR.1	MODIS FPAR 1. quantile	all.16$ASCAT_std	ASCAT standard dev.
all.16$FPAR.2	MODIS FPAR 2. quantile	all.16$ASCAT_min	ASCAT minimum
all.16$FPAR.3	MODIS FPAR 3. quantile	all.16$ASCAT_max	ASCAT maximum
all.16$FPAR.4	MODIS FPAR 4. quantile	all.16$Q1	ASCAT 1. quantile
all.16$FPAR.5	MODIS FPAR 5. quantile	all.16$Q2	ASCAT 2. quantile
all.16$LC08.1_Mean … all.16$LC08.11_Mean	LANDSAT-8 B1 mean … LANDSAT-8 B11 mean	all.16$Q3	ASCAT 3. quantile
all.16$LC08.1_Var … all.16$LC08.11_Var	LANDSAT-8 B1 variance … LANDSAT-8 B11 variance	all.16$Q4	ASCAT 4. quantile
all.16$Lc08.1_Std ... all.16$LC08.11_Std	LANDSAT-8 B1 standard dev. … LANDSAT-8 B11 standard dev.	all.16$Q5	ASCAT 5. quantile
all.16$LC08.1_Min … all.16$LC08.11_Min	LANDSAT-8 B1 minimum … LANDSAT-8 B11 minimum		

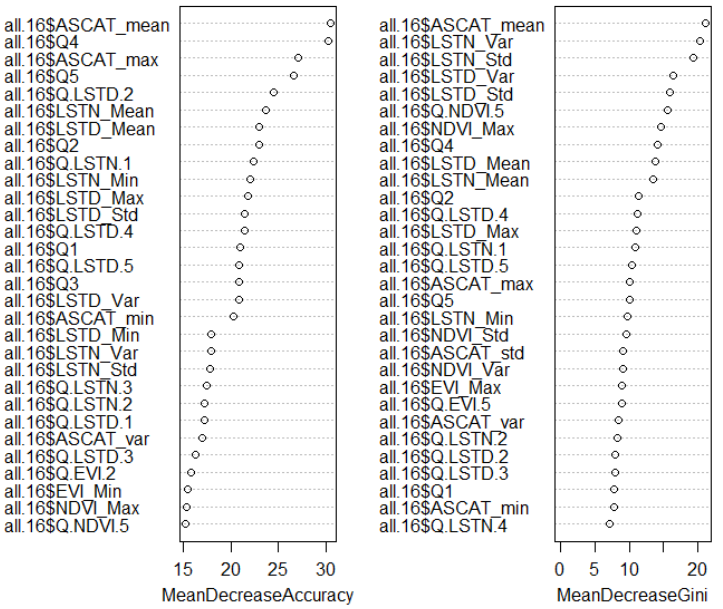
Model 5: LANDSAT-8 Indices

	Non-suitable	Suitable	Class Error
Non-suitable	4246	31	0.0072
Suitable	162	111	0.5934



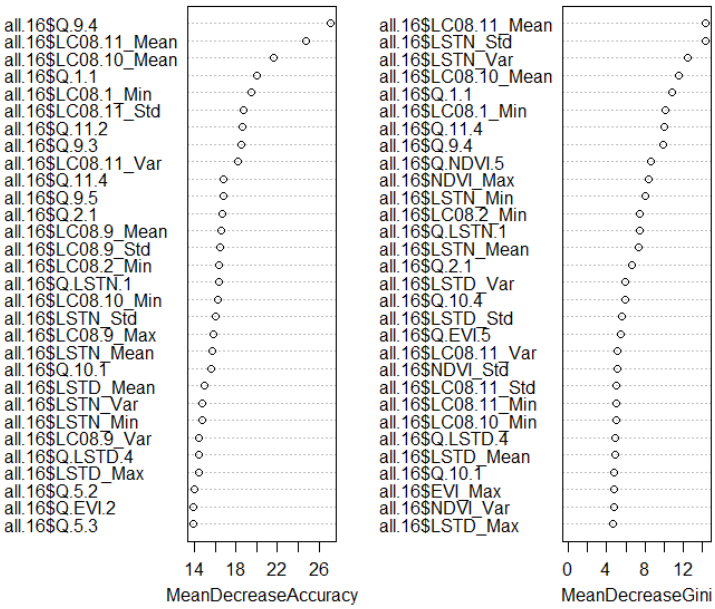
Model 6: MODIS and ASCAT Indicators

	Non-suitable	Suitable	Class Error
Non-suitable	4255	22	0.0051
Suitable	90	183	0.3297



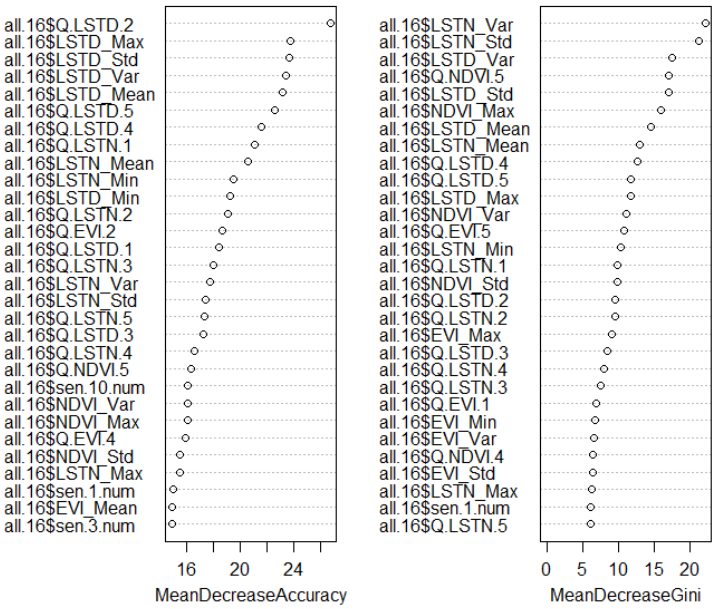
Model 7: MODIS and LANDSAT-8 Indicators

	Non-suitable	Suitable	Class Error
Non-suitable	4258	19	0.0044
Suitable	101	172	0.3640



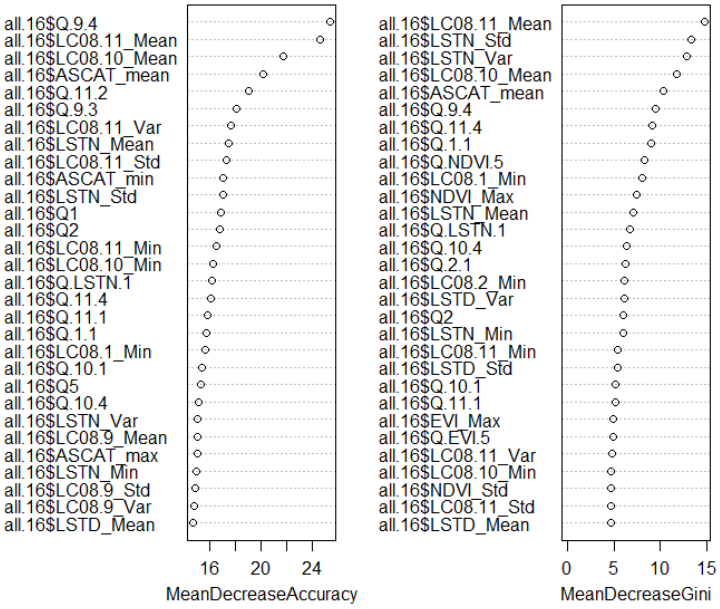
Model 8: MODIS and SENTINEL-2A Indicators

	Non-suitable	Suitable	Class Error
Non-suitable	4251	26	0.0061
Suitable	125	148	0.4579



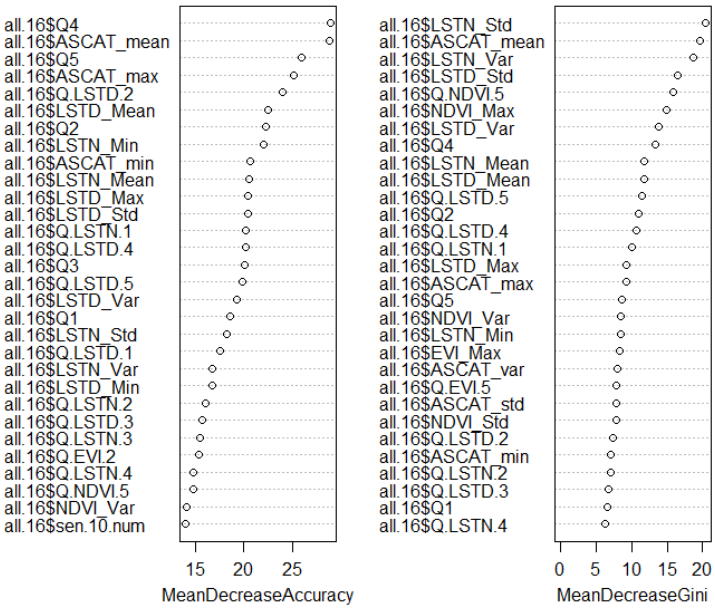
Model 9: MODIS, ASCAT and LANDSAT-8 Indicators

	Non-suitable	Suitable	Class Error
Non-suitable	4260	17	0.0040
Suitable	102	171	0.3736



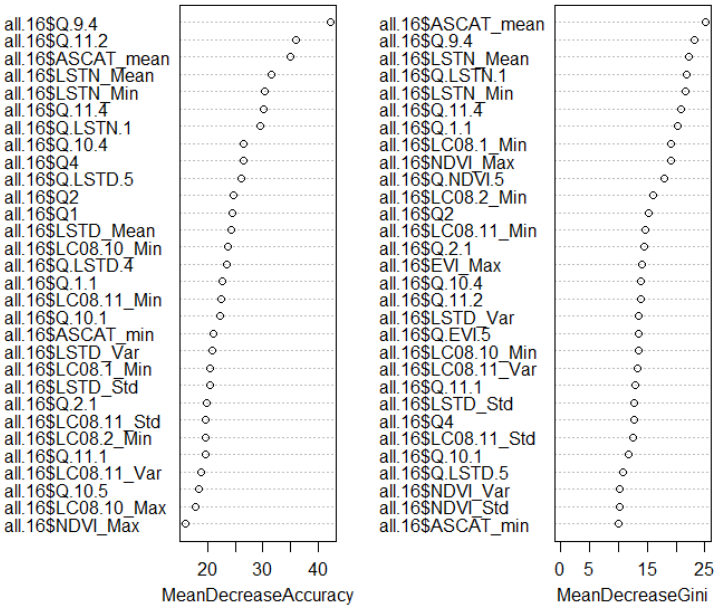
Model 10: MODIS, ASCAT and SENTINEL-2 Indicators

	Non-suitable	Suitable	Class Error
Non-suitable	4257	20	0.0047
Suitable	98	175	0.3590



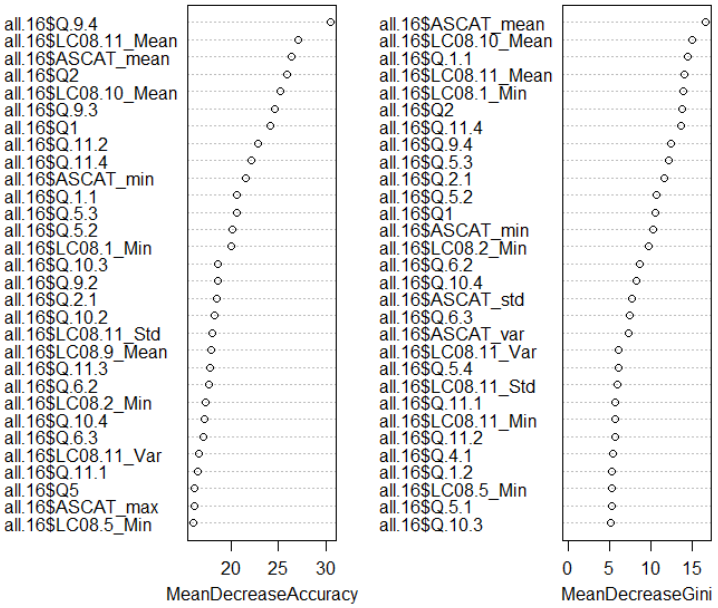
Model 12: Subset of all indicators

	Non-suitable	Suitable	Class Error
Non-suitable	4256	21	0.0049
Suitable	95	178	0.3480



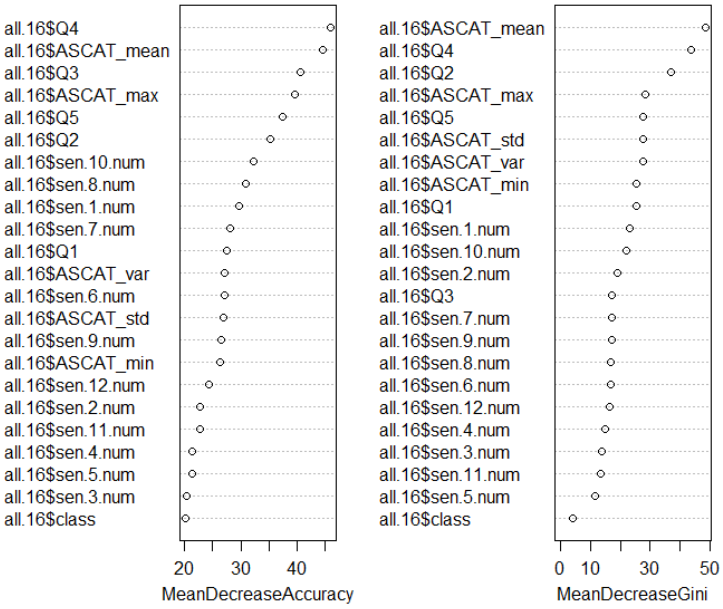
Model 13: ASCAT and LANDSAT-8

	Non-suitable	Suitable	Class Error
Non-suitable	4257	20	0.0047
Suitable	111	162	0.4066



Model 14: ASCAT and SENTINEL-2

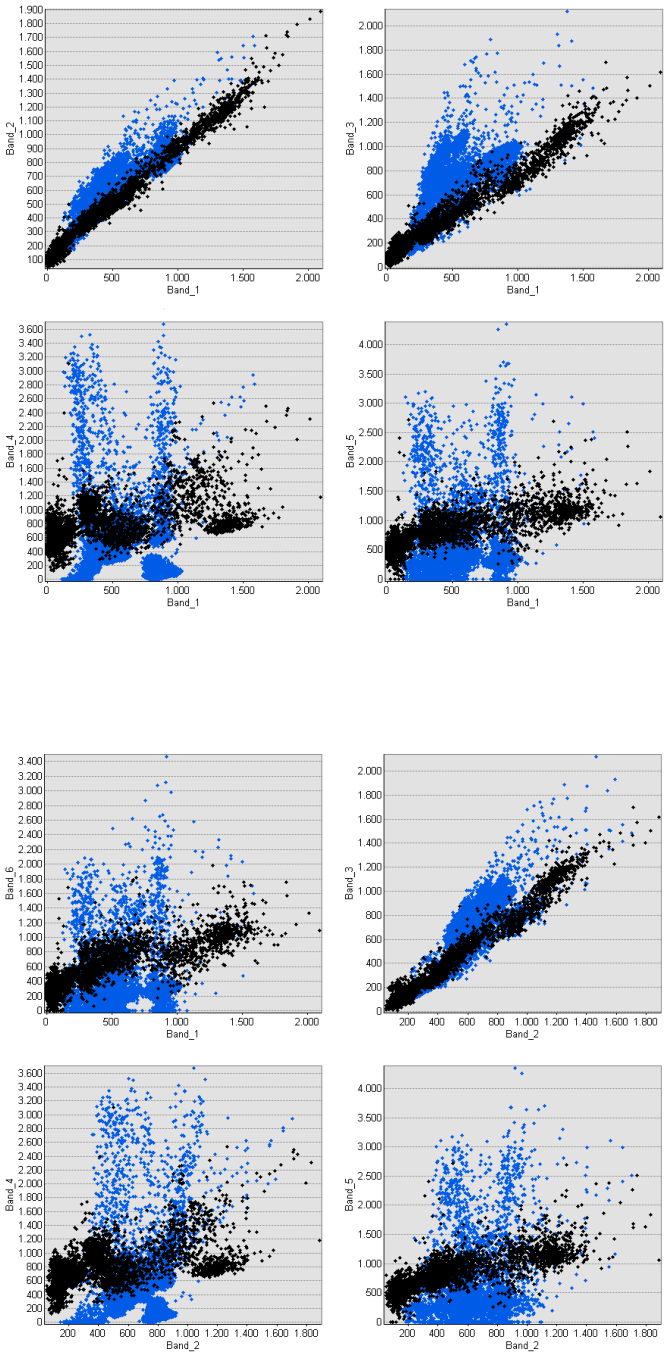
	Non-suitable	Suitable	Class Error
Non-suitable	4251	26	0.0061
Suitable	99	174	0.3626

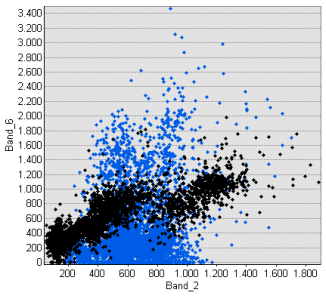


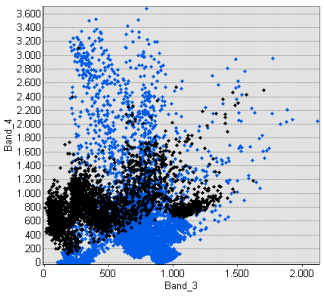
Appendix 2

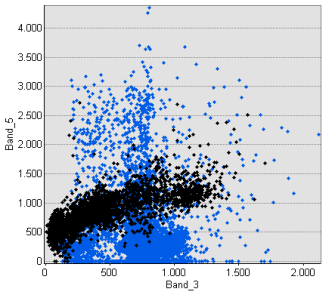
Signature Analysis of Class Water (blue) and Cloud Shadows (black)

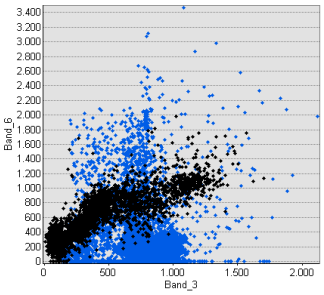
Band_1 = Blue, Band_2 = Green, Band_3 = Red, Band_4 = NIR, Band_5 = SWIR 1, Band_6 = SWIR 2

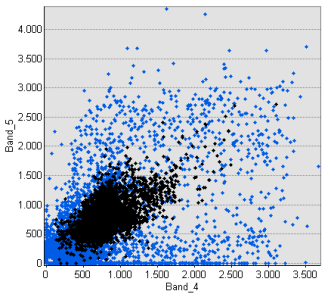


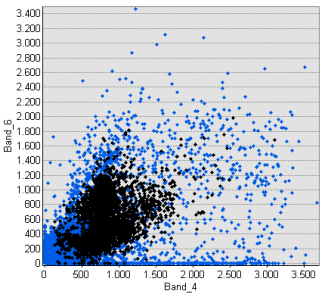

Band_6
Band_2

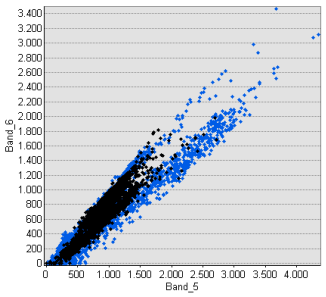

Band_4
Band_3


Band_5
Band_3


Band_6
Band_3


Band_5
Band_4


Band_6
Band_4


Band_6
Band_5